UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

RODRIGO SCHRAMM

SISTEMA AUDIOVISUAL PARA ANÁLISE DE SOLFEJO

Tese apresentada como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação

Orientador: Prof. Dr. Cláudio Rosito Jung

Porto Alegre
2015

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Schramm, Rodrigo

SISTEMA AUDIOVISUAL PARA ANÁLISE DE SOLFEJO / Rodrigo Schramm. – Porto Alegre: PPGC da UFRGS, 2015.

117 f.: il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2015. Orientador: Cláudio Rosito Jung.

1. Reconhecimento de padrões. 2. Processamento de sinais. 3. Classificador Bayesiano. 4. Dynamic Time Warping. 5. Transcrição melódica. 6. Solfejo. 7. Marcação de Compasso. 8. Avaliação automática. 9. Educação musical. I. Jung, Cláudio Rosito. II. Título.

"Defying the laws of gravity" (Don't Stop Me Now)

— Freddie Mercury

AGRADECIMENTOS

À minha esposa, Helena, que me incentivou durante todo o período em que estive envolvido com o doutorado, dando-me carinho e todo o suporte de que precisei. Às queridas meninas Biju, Lili e Carol, que sempre torceram por mim. Também aos meus pais, Alexandre e Liane, que sempre estiveram ao meu lado, me apoiando e se alegrando com minhas conquistas. A todos os meus professores, pelos ensinamentos e pelos desafios. Ao Prof. Dr. Eduardo Reck Miranda, juntamente com sua equipe do ICCMR – Interdisciplinary Centre for Computer Music Research, em Plymouth / UK, os quais fizeram parte de grandes momentos da minha vida acadêmica na Inglaterra. Aos colegas do PPGC, pelas trocas de experiências. Aos professores do Departamento de Música, que me deram suporte durante o meu afastamento do país. À CAPES, que através de Bolsa Doutorado Sanduíche do Programa Ciência Sem Fronteiras, viabilizou meus estudos no exterior. À UFRGS, instituição que garantiu toda a infra-estrutura necessária para a realização das pesquisas envolvidas no meu trabalho. Ao meu orientador, Prof. Dr. Cláudio Rosito Jung, que esteve sempre presente não apenas ensinando conteúdos acadêmicos, mas principalmente apontando a direção na qual eu deveria olhar.

RESUMO

O solfejo, em seu conceito mais amplo, é uma técnica usual no processo de ensino-aprendizagem musical, o qual envolve a realização vocal de melodias, considerando as alturas e as durações dos sons musicais registrados em partitura, devidamente associado à marcação de compassos por intermédio de gestos que definem a estrutura métrica e o andamento da peça musical. Este trabalho apresenta uma abordagem audiovisual para avaliação automática dessa prática de estudo pertinente à leitura e à estruturação musicais. O sistema proposto é dividido em três partes. A primeira efetua o reconhecimento visual dos gestos de marcação de compassos realizados pela mão, por intermédio de um classificador probabilístico. Um processo de alinhamento temporal garante o reconhecimento dos padrões de movimento mesmo em casos com variação de andamento, permitindo também a avaliação da precisão rítmica do aluno, quando comparado com um referencial metronômico. A segunda parte deste sistema obtém a transcrição melódica do canto a partir da análise do respectivo sinal de áudio. Os fragmentos melódicos detectados são agrupados e mapeados em relação às notas da partitura do exercício de solfejo, permitindo uma avaliação direta nota-a-nota da performance do canto. Por fim, a terceira parte do sistema proposto faz a integração entre o gesto de marcação de compassos e a transcrição melódica. Nesse caso, o gesto atua como um metrônomo, controlando o fluxo temporal. Assim, a avaliação nota-a-nota do solfejo pode ser empregada também em casos onde exista grande variação no andamento da peça. Tanto o processo de avaliação do gesto de marcação de compassos quanto a avaliação do canto são obtidos por intermédio de um classificador Bayesiano gerado a partir de avaliações reais, feitas por especialistas em música. Dessa forma, o sistema desenvolvido efetua o mapeamento advindo da opinião de especialistas humanos em um sistema de avaliação automática de solfejo executado por máquina, que é capaz de identificar as notas musicais cantadas pelo aluno em cada instante métrico determinado, devidamente conduzido pelo gesto, sem a necessidade de sincronização por um metrônomo ou manutenção de um andamento fixo.

Palavras-chave: Reconhecimento de padrões. Processamento de sinais. Classificador Bayesiano. Dynamic Time Warping. Transcrição melódica. Solfejo. Marcação de Compasso. Avaliação automática. Educação musical.

AUDIOVISUAL SYSTEM FOR SOLFÈGE ANALYSIS

ABSTRACT

Solfége is a general technique used in the music learning process, which involves the vocal performance of melodies, regarding the time and duration of musical sounds as specified in the music score, properly associated the meter-mimicking performed by the hand movement. This thesis presents an audiovisual approach for automatic assessment of this relevant musical study practice. The proposed system is divided into three parts. First, a probabilistic classifier recognizes the musical metric patterns drawn by the hand movement. A time alignment process assures the proper recognition of the movement patterns even in cases where there are changes in the musical tempo. Also, this process allows to estimate the accuracy of the rhythmic performance. In the second part of this work, audio analysis is applied to achieve the melodic transcription of the sung notes. The detected melodic fragments are then grouped and mapped into single notes, which are connected to their related notes on the music score of the solfège exercise. This mapping procedure allows the direct assessment (note by note) of the singing performance, even if there are slight discrepancies between the transcribed notes and the music score. Finally, the last part of the proposed system combines the gesture of meter-mimicking (video information) with the melodic transcription (audio information), where the hand movement works as a metronome, controlling the time flow (tempo) of the musical piece. Thus, the meter-mimicking is used to align the music score (ground truth) with the sung melody, allowing the assessment even in time dynamic scenarios. Both meter-mimicking and sung notes are evaluated by a set of Bayesian classifiers that were generated from real evaluations done by experts listeners. In this way, the developed model performs the mapping of the point of view from human experts into an automatic system which is able to make the solfege assessment, regarding the pitch, onset and duration of the music notes, without the need for external synchronization with a metronome or the maintenance of a fixed tempo.

Keywords: Pattern recognition, signal processing, melodic transcription, solfège, automatic assessment, music education.

LISTA DE ABREVIATURAS E SIGLAS

HMM	Hidden Markov Model
DTW	Dynamic Time Warping
PDF	Probability Density Function
RGB-D	Red-Green-Blue-Depth
DTree	Decision Tree
SNR	Signal-to-Noise Ratio
ANN	Artificial Neural Network
BPM	Batidas Por Minuto
EAD	Ensino a Distância

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Alfabetização musical e musicalização, em seu sentido mais amplo, requerem acompanhamento próximo do aprendiz por parte de um professor da área, músico mais experiente (MACHADO, 2012). Principalmente ao longo dos passos iniciais do processo de ensino-aprendizagem, é importante que todo eventual erro possa ser identificado no momento em que ocorre, evitando aquisições futuras sobre bases falsas. Nesse contexto, o solfejo é uma técnica usual no processo de ensino-aprendizagem musical, o qual envolve o canto de notas musicais, considerando o ritmo e a altura, bem como a noção de métrica de compasso e estruturação musical.

Por ocupar-se principalmente com melodia e ritmo devidamente organizados em frases, que são unidades com sentido delimitadas pela respiração, o solfejo é a base de tudo o que, num nível mais complexo, vai implicar aspectos de estruturação, forma, análise, compreensão e expressividade em música (SWANWICK, 1994). A utilização do solfejo e sua necessidade de rápida avaliação por um especialista motivou a elaboração dessa tese, a qual busca desenvolver um sistema para avaliação automática dessa prática de estudo musical.

1.1 Contextualização

A prática do solfejo inicia por estruturas simples e vai avançando por outras mais complexas, consistindo em um modo particular de lidar com o caos perceptivo, que segue as vivências informais com música e antecedem a musicalização formal. Isso porque o solfejo se constitui e tem utilidade, precisamente, na interface entre a vivência sensório-motora e a abstração, dentro do conhecimento musical. A organização cinestésica, advinda da leitura da partitura pareada com a marcação de compassos, e a organização auditiva, advinda da leitura da partitura pareada com a emissão vocal, são igualmente necessárias ao solfejo.

Resumindo, solfejo, em conceitos atuais, implica cantar alturas de notas, afinadas a partir de um "diapasão", devidamente associadas a durações de valores, determinados por uma métrica. A métrica é verificável a partir do agrupamento de unidades de tempo dentro de um compasso, os quais, por sua vez, seguem padrões básicos de marcação representada pelo gesto da mão. E todo esse processo precisa ser acompanhado e avaliado com precisão, garantindo um aprendizado correto, desde o primeiro contato de um aluno

com uma partitura.

Ao se constituir na interface entre vivência sensório-motora e abstração, no processo de aquisição do conhecimento musical, o estudo do solfejo depara-se com a expressividade em música. Dentre os muitos fatores relacionados à expressividade se destacam as variações de andamento da peça em seu conjunto e de cada uma de suas partes. Quando a execução musical acontece em uma velocidade única e constante, sustentada unicamente por batidas de um metrônomo, ela soa de modo mecânico, frio e sem sentido (LEONIDO, 2007). Assim, se a peça é rápida, num andamento *Allegro*, permanecerá assim em todos os seus momentos; e o mesmo acontecerá, se a peça for lenta, num andamento *Largo*. Tal execução monótona, entretanto, é desinteressante sob o ponto de vista artístico, pois leva a obra a soar como se estivesse sendo executada por uma máquina, perdendo em força comunicativa.

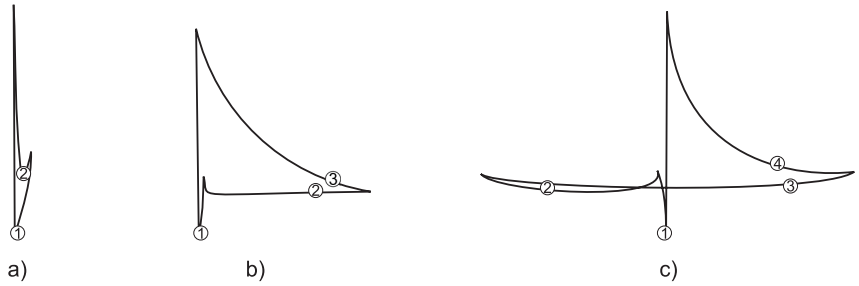
O conjunto de técnicas propostas nesta tese permite que a mão, desenhando compassos no ar, possa exercer a função das batidas do metrônomo e, por consequência, controlar a velocidade de avanço de cada uma das unidades de tempo, posto que a cada uma corresponde a uma posição pré-definida no dito desenho. Concomitantemente, o módulo do sistema proposto para a avaliação automática do solfejo extrai as notas cantadas através de um processo de transcrição melódica, e compara a performance vocal e gestual com um modelo probabilístico gerado a partir de avaliações realizadas previamente por músicos especialistas.

O problema em foco integra dois aspectos diretamente relacionados à musicalização: i) a internalização das alturas e a internalização das durações dos sons, ambas classificadas em relação a parâmetros específicos, respectivamente, o Lá Internacional (440Hz) na escala temperada; e ii) o conceito de pulso (materializado em unidades de tempo e de compasso). Por isso, este trabalho aborda o solfejo em sua prática contemporânea, isso é, o ato de cantar as alturas e durações das notas musicais, dizendo seus nomes ou não, e marcar simultaneamente os compassos, garantindo um pulso métrico baseado nos padrões de regência (GORDON, 2011; OTTMAN; ROGERS, 2011).

A marcação dos compassos é evidenciada por intermédio de fluxos organizados de movimentos corporais, em particular das mãos, ao desenharem padrões específicos no ar, exteriorizando visualmente a noção de métrica em música. Em conjuntos sequenciais, coesos e coerentes de agrupamentos binários, ternários ou quaternários de pulsos, estabelecem-se as unidades de tempo e as de compasso. A cada nova unidade de tempo, verifica-se uma nova posição da mão dentro de um mesmo desenho; a cada novo com-

passo, seja ele qual for, um mesmo desenho é reiniciado. Para ilustrar esse procedimento, a Figura 1.1 mostra os três padrões básicos de regência utilizados para a marcação de compassos, e na Figura 1.2 são apresentados alguns quadros consecutivos ilustrando o movimento de marcação de compassos na métrica quaternária, com a trajetória da mão representada por pontos vermelhos.

Figura 1.1: Padrões de movimento para marcação de compasso. (a) Padrão Binário. (b) Padrão Ternário. (c) Padrão Quaternário.



Fonte: O Autor

O exercício de internalização das alturas sonoras está intimamente ligado a um processo conhecido como aculturação tonal (TILLMANN, 2008), por intermédio do qual a imagem sonora é registrada sob a forma de um mapeamento de faixas de frequências contidas no som para uma nota musical de altura específica definida pela frequência fundamental do sinal. As notas musicais são discretizadas em sequências temporais de acordo com as alturas definidas pela escala cromática igualmente temperada[1]. Juntamente com a internalização das durações, esse processo de internalização de alturas é visualizado, e, por consequência, evidenciado por intermédio da transcrição do canto na partitura. A Figura 1.3a ilustra a partitura de um exemplo de exercício de solfejo e a transcrição melódica de uma performance do respectivo exemplo. Na Figura 1.3b é possível identificar a frequência fundamental contida no sinal de áudio (linha azul) e a sequência de notas musicais cantadas, as quais são delimitadas por retângulos em vermelho.

1.2 Motivação

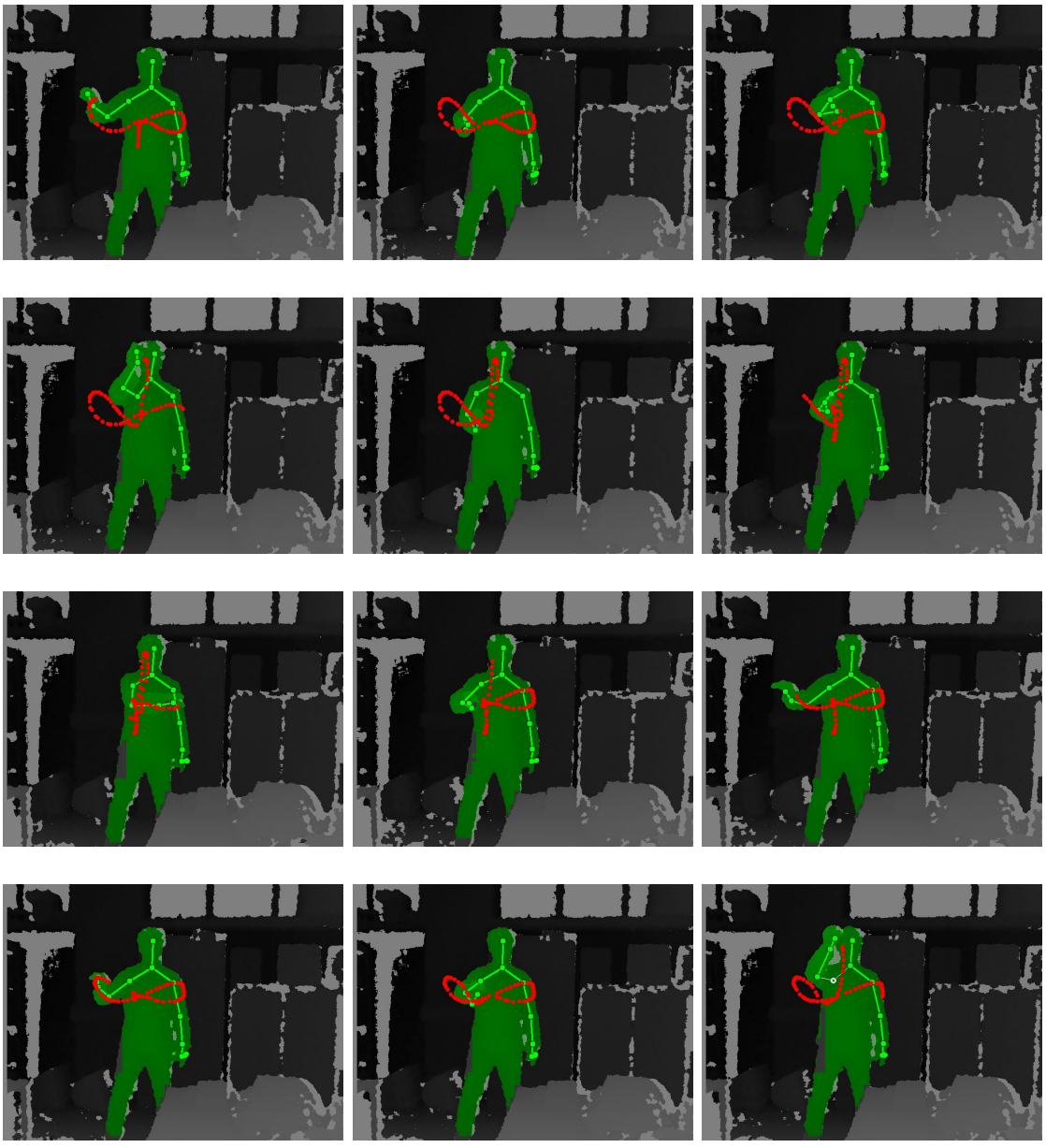
O ensino de música, até muito recentemente ministrado apenas em aulas individuais, raras e caras (GOMES, 2003), democratizou-se com o emprego de sempre novas e renovadas tecnologias da informação e comunicação. A UFRGS é pioneira, no Brasil,

[1]Existem diversas outras formas de temperamento para as escalas musicais. Este trabalho foca apenas na escala temperada cromática, a qual divide a oitava em 12 semitons iguais.

no ensino formal de cursos de graduação Licenciatura em Música por meio da Internet, e este pesquisador integra essas iniciativas desde seu primeiro momento (Edital SEIF/MEC 01/2003 e Resolução SEB/MEC 034/2005).

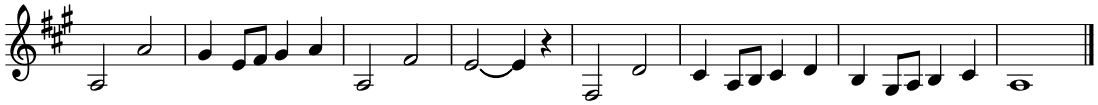
Nesse contexto, o ensino-aprendizagem do solfejo requer acompanhamento permanente e atento, por parte de um especialista músico. Tal tarefa é extenuante e impossibilita a avaliação em larga escala pela mesma pessoa, motivando o desenvolvimento de uma ferramenta computacional.

Figura 1.2: Quadros consecutivos com captura do movimento, através de uma câmera RGB-D, durante a marcação de compassos na métrica quaternária.

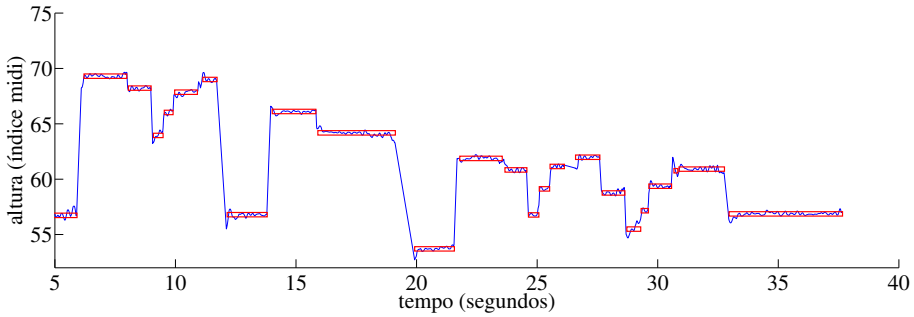


Fonte: O Autor

Figura 1.3: Visualização do processo de transcrição melódica: (a) partitura do exercício de solfejo. (b) linha azul representa a frequência fundamental do sinal obtido a partir de uma gravação do canto. As notas musicais detectadas são representadas por retângulos em vermelho.



(a)



(b)

Fonte: O Autor

A revisão bibliográfica (que será apresentada no Capítulo 2) mostrou que não existem técnicas propostas para atender a avaliação automática de solfejos com o controle concomitante de melodia, ritmo e expressividade, no que se refere às variações de andamento. Sendo assim, a originalidade deste trabalho está no reconhecimento de padrões gestuais utilizados para a marcação de compassos associado à identificação concomitante de notas musicais a partir da análise de áudio, principalmente quando há variação no fluxo metronômico do texto musical.

1.3 Objetivos

Analisando a prática de solfejo, é possível identificar duas etapas distintas: uma delas deve ser capaz de avaliar o movimento da mão ao efetuar a marcação do compasso e, assim como ocorre na realidade, manter-se coerente à agógica e aos momentos expressivos da peça sem perder-se no fluxo natural das notas associadas aos diferentes pontos desse movimento; a outra deve avaliar a precisão rítmico-melódica da voz durante o canto das notas musicais, independente do momento em que elas ocorram, mas sempre de acordo com aquelas escritas na partitura. Assim, associa-se gesto e som.

1.3.1 Objetivo Geral

O objetivo final desse trabalho foi desenvolver um sistema automático de avaliação de solfejo, considerando também a marcação da métrica de compassos, que associa o movimento da mão à emissão vocal. Seu foco está voltado tanto para o reconhecimento de padrões gestuais, no caso, os utilizados para a marcação de compassos, quanto para a identificação concomitante de notas musicais cantadas a partir da análise de áudio.

1.3.2 Objetivos Específicos

Ao longo deste estudo, desenvolveram-se técnicas específicas de processamento de imagens/sinais e de reconhecimento de padrões para a classificação de gestos realizados pela mão e para a identificação de notas musicais. O trabalho foi dividido nas seguintes etapas, com seus correspondentes objetivos específicos:

- **Detecção do gesto**: Detectar o gesto de marcação de compassos, tomando por referência padrões básicos de regência, em métricas binária, ternária e quaternária.
- **Detecção e reconhecimento de notas musicais**: Efetuar o reconhecimento de alturas e durações sonoras individuais, extraídas de um *continuum* rítmico-melódico emitido pela voz.
- **Elaboração do sistema de avaliação audiovisual**: Integrar dados obtidos pela detecção visual de movimentos da mão com dados relativos à emissão vocal em um sistema de avaliação parametrizado por uma partitura.

1.4 Cenário de Uso

Esta seção da tese tem como objetivo descrever o cenário de uso do sistema audiovisual para análise de solfejo, explicitando a forma de utilização do sistema a partir do ponto de vista do usuário, através de uma sequência de etapas que ocorre na interação do ator (aluno de música iniciante) com o sistema. A Figura 1.4 descreve esse cenário de uso através de texto estruturado.

O uso do sistema requer uma simples configuração do espaço físico, conforme ilustrado na Figura 1.5. O hardware utilizado é composto por um sensor RGB-D e um

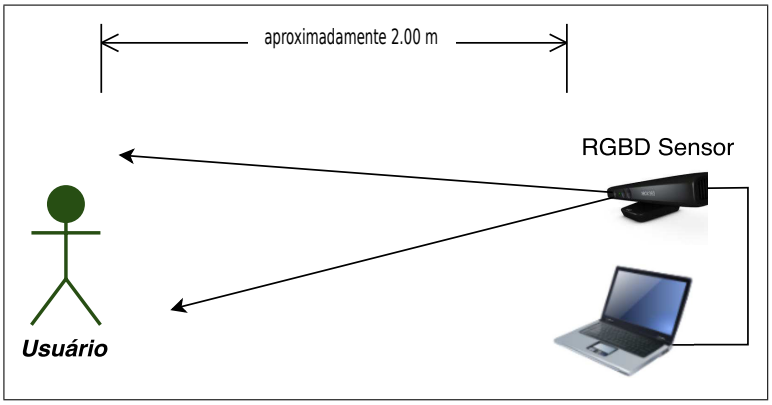
computador com conexão USB. Para a execução adequada do sistema, o ambiente não deve sofrer de interferência por ruído ou quantidade excessiva de reverberação. O aluno, durante a execução do exercício de solfejo, deve se posicionar em frente do sensor a uma distância aproximada de dois metros. Ao final da performance do exercício, o sistema exibe o resultado avaliativo na tela do computador, conforme ilustrado na Figura 1.6. Nessa tela, a parte inferior apresenta em azul as notas musicais definidas na partitura do exercício e as marcas em vermelho apresentam as notas cantadas pelo aluno que foram detectadas pelo sistema. No canto superior direito é apresentada, em versão ampliada da imagem, a classificação nota-a-nota do solfejo, considerando a precisão do ataque, duração e afinação (acurácia da altura da nota).

Figura 1.4: Cenário de uso do sistema de avaliação automática de solfejo

Cenário de Uso:
Gravação e avaliação da execução de um exercício de solfejo pelo aluno.
Ator:
Usuário (aluno de música iniciante)
Fluxo Normal:

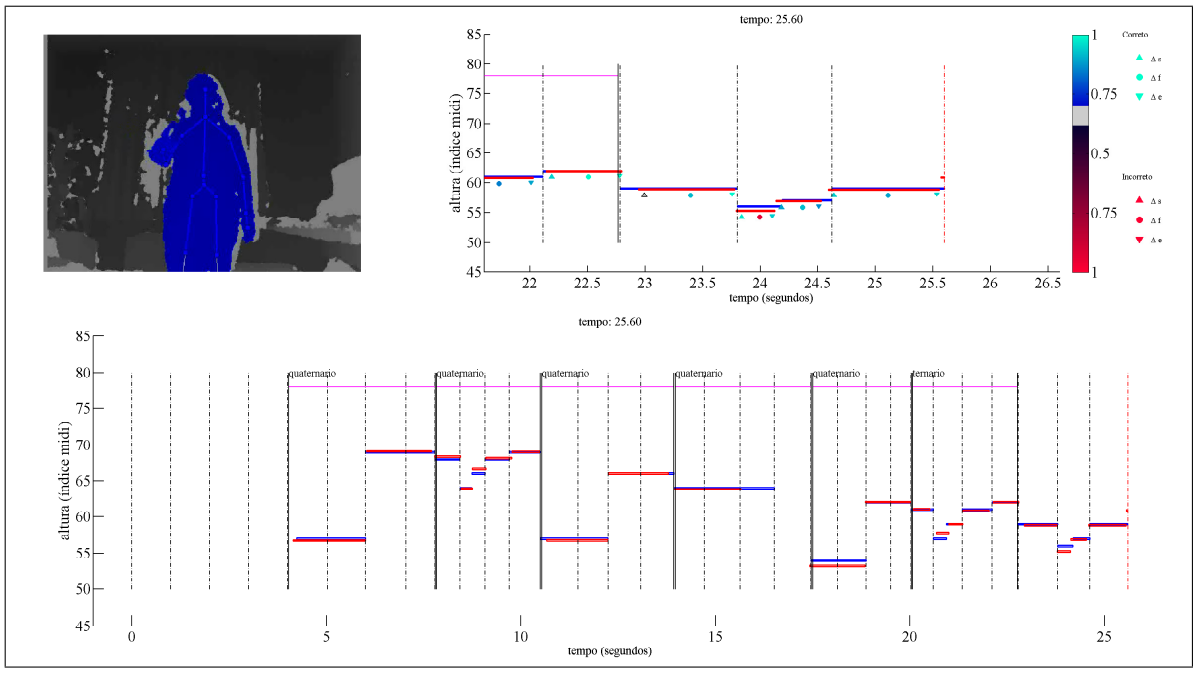
1. O *usuário* escolhe uma sala adequada para a prática do solfejo, com boa iluminação, e baixa presença de ruído e reverberação sonora.
2. O *usuário* conecta o sensor RGB-D (Microsoft Kinect) no computador com o software instalado.
3. O *usuário* clica no ícone do software para iniciar o sistema de avaliação automática de solfejo.
4. O *usuário* escolhe o exercício de solfejo e um andamento de execução (BPM) dentre as opções disponíveis listadas na tela.
5. O *usuário* se posiciona de pé em frente ao sensor a aproximadamente dois metros de distância desse (vide Figura 1.5).
   O *sistema* exibe uma janela contendo a captura da imagem e o rastreamento do esqueleto do usuário (vide Figura 1.2).
   O *sistema* exibe contagem de um compasso (*beat count*) para indicar o início da gravação do exercício.
6. O *usuário* realiza o exercício de solfejo, cantando as notas contidas na partitura, a qual é exibida na tela do computador, e regendo concomitantemente os respectivos compassos no andamento desejado.
7. O *usuário* pressiona o botão "Concluído" para finalizar o exercício.
   O *sistema* realiza a avaliação automática do solfejo e apresenta o resultado na tela (vide Figura 1.6).

Figura 1.5: Configuração do espaço físico para a realização do exercício de solfejo e respectiva captura de áudio e vídeo.



Fonte: O Autor

Figura 1.6: Tela com avaliação do solfejo (*feedback* visual) gerada pelo sistema proposto ao final do processo de avaliação.



Fonte: O Autor

1.5 Representação Esquemática

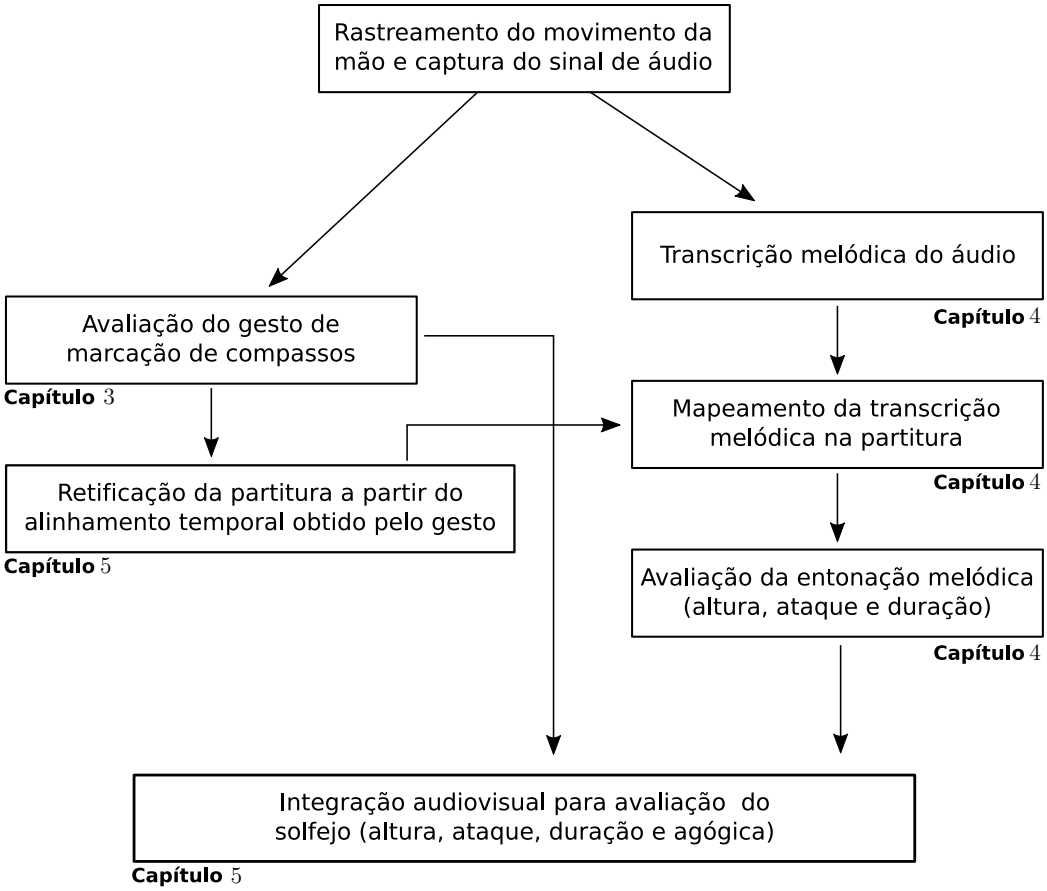
O sistema de avaliação final é construído com bases em avaliações audiovisuais realizadas por músicos especialistas, gerando-se assim um modelo probabilístico que efetua um mapeamento entre a percepção humana e a avaliação automática realizada pela máquina. O conjunto de técnicas computacionais desenvolvido para atingir o objetivo final

desta tese pode ser resumida em sete etapas principais:

1. Rastreamento do movimento da mão e captura do sinal de áudio.
2. Avaliação do gesto de marcação de compassos.
3. Retificação da partitura a partir do alinhamento temporal obtido pelo gesto.
4. Transcrição melódica do áudio.
5. Mapeamento da transcrição melódica na partitura.
6. Avaliação da entonação melódica (altura, ataque e duração).
7. Integração audiovisual para avaliação do solfejo (altura, ataque e duração, agógica).

A interação entre essas etapas pode ser visualizada de forma sintetizada no fluxograma da Figura 1.7.

Figura 1.7: Representação esquemática do sistema proposto para avaliação audiovisual de solfejo.



Fonte: O Autor

1.6 Contribuições desta Tese

No desenvolvimento desse trabalho foram combinados técnicas e algoritmos computacionais amplamente utilizados em problemas de reconhecimento de padrões e análise de séries temporais. Para alcançar os objetivos propostos nesta tese, foram feitas modificações e proposições matemáticas que geraram contribuições científicas para o campo de pesquisa relacionado a este trabalho. A principais, dentre elas, foram:

1. Avaliação da acurácia do movimento de marcação de compassos por meio de um modelo Bayesiano gerado a partir de avaliação de especialistas em música. Detalhes da técnica na Seção 3.3.
2. Proposição de uma medida de custo local para o algoritmo DTW, a qual é mais eficiente que medidas tradicionais (por exemplo, baseadas em norma L^p) para efetuar a comparação entre a trajetória do movimento efetuado pela mão e os modelos de marcação de compassos utilizados neste trabalho. Maiores detalhes na Seção 3.3.1.
3. Avaliação da precisão rítmica do movimento da mão em relação aos tempos do compasso. A técnica proposta utiliza o caminho de alinhamento temporal gerado pelo algoritmo DTW para estimar avanços ou atrasos na condução do movimento de marcação de compasso. Essa abordagem supera propostas anteriores, pois não necessita da informação de máximos e mínimos locais contidas no padrão de movimento para identificar as unidades de tempo do compasso (*beat time*). Descrição detalhada na Seção 3.5.
4. Mapeamento da transcrição melódica do sinal de áudio para as notas da partitura por meio do agrupamento coerente dos segmentos melódicos detectados. Esse processo de mapeamento permite a comparação individual das notas, mesmo que as mesmas apresentem pequeno desalinhamento com relação ao fluxo regular das unidades de tempo. Detalhamento na Seção 4.3.
5. Avaliação da acurácia da precisão rítmica (ataque e duração) e da afinação (altura) do canto por meio de um classificador Bayesiano gerado a partir de avaliações humanas realizadas por um comitê de especialistas em música. Descrição na seção 4.4.
6. Integração audiovisual do sistema de avaliação de solfejo, a qual permite ao usuário a execução do solfejo com variação de andamento, sendo a mesma regida pelo movimento da mão. Detalhes no Capítulo 5.

Além dessas contribuições científicas, destacam-se também as potenciais contribuições sociais que o sistema proposto pode gerar no contexto da educação musical, em especial, em cursos em modalidade a distância e com número elevado de alunos. Algumas dessas contribuições são:

1. Automatização do processo de avaliação de exercícios de solfejo: em cursos tradicionais de música com elevado número de alunos, seja na modalidade presencial ou a distância, o processo de avaliação da prática de solfejo pode ser muito trabalhoso e demorado. Nesses casos, um processo avaliativo criterioso, de todo o conjunto de alunos, a partir da análise visual e auditiva de gravações, pode levar dias ou semanas. O sistema proposto neste trabalho permite a automatização desse processo, podendo ser executado inclusive em paralelo, sendo independente da quantidade de professores disponíveis no respectivo curso.
2. Feedback imediato ao aluno: o sistema permite avaliar imediatamente a performance do aluno, não precisando esperar pela disponibilidade de um professor.
3. Uniformização do resultado avaliativo: A avaliação do solfejo por especialistas humanos pode ser influenciada por fatores como fadiga e subjetividade (LARROUY-MAESTRI et al., 2013). O sistema proposto não possui desgaste e utiliza critérios objetivos e idênticos para todos os alunos avaliados.
4. Eliminação de constrangimento emocional: A prática do solfejo pode ser comprometida devido ao constrangimento do aluno em executar o exercício na presença do professor e/ou colegas, principalmente se com fins avaliativos (COELHO, 1994). O sistema permite ao aluno a sua autoavaliação, de forma individual e independente, garantindo assim sua privacidade durante a prática do exercício de solfejo.

1.7 Estruturação desta Tese

O texto desta tese foi organizado numa lógica que segue os objetivos específicos acima citados. Assim, o estudo e desenvolvimento do trabalho também foi organizado em três momentos distintos, a saber: detecção de gestos, captação e reconhecimento de notas musicais, e elaboração de um sistema de avaliação audiovisual, que integra o controle manual do andamento com a avaliação da altura e da duração dos sons emitidos pela voz, de acordo com uma partitura dada.

O Capítulo 2 apresenta a revisão bibliográfica dos trabalhos relacionados ao tema

desta tese. Esse Capítulo é divido em duas partes, citando, primeiramente, os trabalhos que têm relação direta ou indireta com a análise do movimento de marcação de compassos; e, num segundo momento, as técnicas desenvolvidas e relacionadas à transcrição melódica e avaliação do canto.

O conteúdo do Capítulo 3 aborda a detecção de gestos para marcação de compassos. Neste trabalho, tomou-se por referência padrões básicos de regência em métricas binária, ternária e quaternária. A principal contribuição aqui é o desenvolvimento de uma nova técnica capaz de reconhecer o gesto de marcação de um compasso dado, bem como a velocidade de avanço dos tempos do mesmo, conforme materializados pela posição da mão em cada novo ponto do desenho correspondente a esse compasso.

O Capítulo 4 trata da detecção e do reconhecimento de notas musicais emitidas pela voz, as quais são comparadas com aquelas que deveriam estar sendo emitidas, tomando por parâmetro uma partitura dada. Para tanto, efetuou-se o reconhecimento de alturas e durações sonoras individuais, extraídas de um *continuum* rítmico-melódico de áudio e pareadas em sua relação entre evidências sonora e musicografada.

No Capítulo 5 chega-se à elaboração do sistema de avaliação audiovisual que integra dados obtidos pela detecção visual de movimentos controladores da mão e capazes de funcionarem como metrônomo para o dados relativos à emissão vocal em um sistema de avaliação parametrizado por uma partitura.

Por fim, o Capítulo 6 apresenta as conclusões deste estudo e delinea trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Para o desenvolvimento deste trabalho, foi realizada uma revisão bibliográfica em busca de técnicas com objetivo semelhante, qual seja, o de analisar o gestual de condução musical com vistas ao aprendizado dos padrões métricos de compasso incluindo variações internas de andamento e avaliar, concomitantemente, a prática de entonação melódica associada a tais unidades de tempo não fixas. Não foram encontrados trabalhos estritamente relacionados ao foco desta tese; contudo, seguindo a subdivisisão natural do sistema proposto, foi possível organizar os trabalhos similares encontrados em duas partes principais. Na primeira, relacionam-se sistemas desenvolvidos para a análise do gestual de regência, incluindo técnicas, sensores e características de rastreamento e classificação do movimento. Na segunda parte, apresentam-se trabalhos desenvolvidos que estão relacionados à detecção de notas musicais (detecção de alturas, ataques e durações), a partir de sequências de áudio, os quais servirão de base para o desenvolvimento da proposta de avaliação do canto.

2.1 Registro e Análise da Marcação de Compasso

Além das referências estritamente conectadas à aplicação musical e educacional da regência, também serão abordados outros trabalhos que focam em análise e interpretação do gesto, incluindo algoritmos de classificação e técnicas para extração e rastreamento de características. O uso de sistemas computacionais aplicados à regência ou utilizados como instrumentos musicais alternativos iniciou com o desenvolvimento do Radio Baton (MATHEWS, 1990). O Radio Baton utilizava duas antenas em forma de batutas que emitiam sinais de rádio, os quais eram capturados por um painel que efetuava o rastreamento da posição das antenas por triangulação. Max Mathews mapeou a variação do movimento das batutas à variação do andamento musical e volume do som.

Ao longos dos últimos anos, muitos outros trabalhos seguiram as ideias contidas no Radio Baton, estendendo suas funcionalidades iniciais. No trabalho intitulado *Personal Orchestra* (BORCHERS; SAMMINGER; MUHLHAUSER, 2002), os autores desenvolveram um sistema interativo que utiliza audio e vídeo gravados da Orquestra Filarmônica de Viena. Nesse sistema, o usuário pode controlar o volume e o andamento da música através de gestuais simples, executando um movimento periódico para cima e para baixo com as mãos, parecido com o movimento de um pêndulo. Outro trabalho semelhante,

desenvolvido por pesquisadores do Bell Labs (SEGEN; KUMAR; GLUCKMAN, 2000), foi o *Visual Conducting Interface (VCI)*, o qual utiliza um par de câmeras de vídeo estéreo para identificar a posição 3D da mão e/ou batuta do maestro. Nessa interface, os autores controlam o volume da música por intermédio da posição da mão, enquanto que o andamento é definido pelo movimento periódico vertical da batuta. A posição horizontal da mão, apontada para uma região da orquestra, define o naipe (grupo de instrumentos) que sofrerá a influência dos comandos gerados pelo maestro. Ilmonen e Takala (1999) usam sensores magnéticos para extrair o movimento 3D das mãos do regente, e redes neurais artificiais para interpretá-los. O resultado do processamento também é aplicado ao andamento e dinâmica da música.

Uma abordagem utilizando *Hidden Markov Models* (HMMs) foi desenvolvida por Kolesnik and Wanderley (2004). Nessa técnica, os movimentos do maestro são capturados por duas câmeras, uma frontal e outra lateral. As coordenadas 3D da mão são obtidas após a segmentação da mão baseada em cor (o algoritmo necessita de luvas/marcadores coloridos) e, por fim, o movimento é interpretado por uma HMM discreta. Lee et al. (2006) desenvolveram o CONGA, um *framework* adaptativo para análise de gestuais de condução musical. Este *framework*, implementado numa linguagem baseada em blocos semelhante a linguagens de programação para música, como PureData (PUCKETTE, 1996), utiliza as coordenadas 2D da mão, para avaliar o movimento de regência. Esses blocos, representados por nodos visuais, podem ser conectados para gerar regras de avaliação do movimento de condução musical. Por fim, os blocos são organizados para formar uma máquina de estados finitos, a qual é responsável por classificar o movimento gestual numa das possíveis métricas de compasso, ilustradas na Figura 1.1. A ideia de uma máquina de estados serviu como base para o desenvolvimento do trabalho proposto por Schramm e Jung (2007). Nessa proposta, os autores utilizam uma câmera RGB para rastrear a posição da mão e extrair suas coordenadas 2D, a quais são avaliadas por uma máquina de estados finitos, que identifica os padrões gestuais, incluindo os tempos do compasso. Uma proposta mais recente de Argueta, Ko e Chen (2009) segue a mesma ideia, baseada em sensores de posição e orientação (*data glove*) que são utilizados para capturar os movimentos do regente. A técnica também utiliza a informação de inversão do sentido do movimento para segmentar o padrão métrico do compasso a partir de um conjunto de regras pré-estabelecidas.

Mandanici e Sapir (2012), bem como Toh, Chao e Chen (2013) usaram câmeras RGB-D, as quais permitem rastrear movimentos da mão no espaço 3D. Esses dispositivos

são mais robustos em ambientes não controlados e são facilmente encontrados no mercado com preços acessíveis. Ambos os trabalhos utilizam a posição da mão para simular a condução musical ou para gerar sons com o auxílio de um instrumento virtual. O uso de sensores, capazes de rastrear o movimento da mão com maior precisão, permitiu o desenvolvimento de técnicas que focam no rastreamento contínuo do gestual de regência. Nesse caso, os algoritmos buscam fazer o mapeamento da evolução temporal do gesto com a evolução temporal da música. Um exemplo desse tipo de aplicação, o *Gesture Follower*, foi desenvolvido por Bevilacqua et al. (2010), o qual usa a sequência de estados de um HMM para seguir modelos de gestuais previamente treinados.

Nenhuma das técnicas citadas anteriormente se preocupa em identificar subdivisões do compasso, nem de extrair medidas objetivas para avaliar a performance dos gestuais realizados por estudantes de música. Em sua maioria, as técnicas citadas até o presente momento utilizam a inversão do sentido do movimento na direção vertical para definir o *beat time* (andamento) e, em paralelo, um algoritmo para identificar o padrão de movimento, como HMM, Redes Neurais e/ou Máquina de Estados Finitos. Poucas são as técnicas que têm o foco na análise para a aprendizagem do padrão de gestual e não na sua aplicação para performance musical, como proposto por Schramm e Jung (2007). Uma exceção é a técnica descrita em Maes et al. (2013), que foca na análise e avaliação do movimento da mão com o objetivo de assistir o processo de aprendizagem dos padrões de gestuais. Os autores dessa técnica extraem *templates* a partir de amostras de gestuais realizados por maestros especialistas, os quais são utilizados para avaliação do movimento. A técnica possibilita a detecção dos padrões de compasso e do andamento, usando um algoritmo baseado em *template matching*, a partir de correlação cruzada entre os sinal analisado e os modelos previamente treinados. A técnica é robusta mesmo em situações onde há variação de andamento (*time lag*), porém não é claro como tal estratégia pode explicitamente avaliar a variação individual dos tempos dentro do compasso (*beat time*), uma vez que a técnica não utiliza alinhamento temporal.

A partir desse breve panorama, é possível perceber que as técnicas para análise do movimento de regência utilizam como dados de entrada a informação da posição das mãos ao longo do tempo. Percebe-se, também, que o processo de classificação do gestual de regência depende do movimento periódico para cima e para baixo com as mãos, semelhante ao movimento de um pêndulo. Apesar desse movimento ser uma regra dentre os padrões ocidentais para a marcação de compasso musical, verificou-se a partir de amostras grava-

das[1] que nem sempre a inversão no sentido do movimento, o qual caracteriza o *beat time*, é efetuado ao longo de todos os tempos do compasso. Por isso, para uma representação mais genérica do padrão de movimento com vistas à análise da aprendizagem da execução cinestésica de padrões de marcação de compasso, é necessária uma técnica de reconhecimento gestual mais robusta, a qual deverá ser capaz de identificar também as subdivisões do tempo do compasso. Com o objetivo de propor um método mais eficaz em relação aos problemas acima apresentados, foram estudados algoritmos para reconhecimento de gestos que são utilizados em aplicações diversas. Além disso buscou-se catalogar os tipos de características capturadas mais utilizadas por esses algoritmos, com vistas a determinar a melhor abordagem para a condução dessa proposta. Um resumo desses trabalhos investigados está descrito a seguir.

2.1.1 Tipos de Sensores e de Características

Um aspecto observado durante o estudo dos trabalhos relacionados foi o tipo de informação utilizada para efetuar o reconhecimento ou classificação de gestos. Obviamente, o tipo de informação está condicionado ao tipo de sensor utilizado no desenvolvimento da técnica. No caso de câmeras de vídeo, existem técnicas que utilizam informações locais ou globais da imagem.

A partir de sequências de vídeos, alguns autores efetuam o reconhecimento de movimentos ou ações, utilizando-se de características de movimento (*motion features*), ou de alguma combinação dessas com padrões de forma (*templates*) gerados a partir da silhueta do corpo humano, conforme detectada pelas câmeras. Ahmad e Lee (2006) obtêm essas características das imagens a partir de múltiplas vistas (múltiplas câmeras) e aplicam Análise de Componentes Principais – PCA para reduzir a dimensionalidade dos dados. Ankerst et al. (1999) introduzem uma característica baseada no histograma da forma (*shape histogram*), o qual é uma extensão da ideia de padrões de forma 2D para o 3D, empregada posteriormente na técnica de Tran e Trivedi (2008), criando uma representação volumétrica do objeto. Outras técnicas globais utilizam características semelhantes, como padrões de forma temporais (BOBICK et al., 2001; WANG; SUTER, 2006), que codificam o histórico do movimento através de máscaras binárias, quase sempre geradas a partir da silhueta do corpo humano ao longo do tempo. Segundo Gorelick et al. (2005), as silhuetas

[1]Experimentos foram gravados com alunos do curso de Licenciatura em Música da UFRGS. A captura da posição da mão foi feita com uma câmera RGB-D.

são excelentes descritores, pois codificam a forma do objeto, incluindo detalhes. Outro tipo de característica global bastante utilizada são pontos de interesse, obtidos através de *corners* ou textura da imagem. Abdolahi, Ghasemi e Gheissari (2012) propuseram uma técnica baseada em descritores dinâmicos de textura, a qual reconhece o movimento do corpo humano através de um dicionário visual, enquanto Laptev (2005) utiliza pontos de interesse (*corners*) como características para representar gestos ao longo do tempo. Apesar dessas técnicas exibirem resultados interessantes, elas dependem de grandes bases de treinamento e são sensíveis a variações de iluminação. Por exemplo, características baseadas em textura, pontos de interesse, ou mesmo técnicas que utilizam fluxo óptico (*optical flow*) (ZUFFI et al., 2013) encontram problemas quando as imagens possuem regiões homogêneas. Em relação à extração dos padrões de forma (*templates*), é comum a utilização de subtração de fundo da imagem, para identificar a silhueta dos objetos. Nestes casos, tais técnicas geralmente necessitam de câmeras e fundo (*background*) estáticos, o que limita a aplicação em situações reais.

Nas técnicas acima citadas, as características (*features*) não têm relação direta com o padrão de movimento, sendo necessários classificadores para que seja possível um subconjunto de gestos do universo de todas possibilidades gestuais que poderiam pertencer a uma deteminada sequência de imagens. Em outras palavras, a existência de um conjunto de características na imagem e/ou vídeo é associada a uma classe de movimento (ou padrão gestual) durante uma etapa de treinamento e, posteriormente, sempre que o algoritmo encontrar essas características (ou um subconjunto mínimo delas) na imagem e/ou vídeo, ele deverá identificar sua classe, baseando-se numa medida de proximidade dos dados. Como os gestuais são executados no ambiente 3D e câmeras de vídeo RGB capturam tais cenas em projeções 2D, o processamento dessa informação é dependente do posicionamento da câmera, o que restringe muitas vezes a funcionalidade dos algoritmos a ambientes controlados. Nessas aborgagens, a precisão de extração de detalhes do movimento depende diretamente da distância entre objetos e câmera, da resolução da imagem e também do ângulo de visão. Ou seja, o modelo treinado é dependente da posição e configuração da câmera. Por esses motivos, técnicas que possuem como dados de entrada as posições 3D dessas articulações simplificam o processo de classificação. Por outro lado, a obtenção dessas posições 3D (ou características) passam a ser também um problema que precisa de solução. Considerando o objetivo desse trabalho, onde pretende-se reconhecer um gesto realizado com as mãos, é natural esperar que a posição das mesmas seja relevante ao processo. Mais do que isso, os pontos que descrevem a trajetória do movi-

mento em coordenadas de mundo ao longo do tempo podem ser utilizados para derivar outros tipos de informações, como velocidade, aceleração, variação angular, etc. A seguir, serão apresentadas brevemente técnicas que utilizam como dados iniciais as posições 2D ou 3D da mão para a classificação do movimento gestual. Sendo assim, as técnicas muitas vezes diferem entre si apenas pelo tipo de sensor utilizado. Em outros, as técnicas utilizam o mesmo tipo de informação de entrada, mas desenvolvem um tipo específico de classificador.

Algumas técnicas utilizam acelerômetros para capturar a variação do movimento das mãos (WANG; CHUANG, 2012; AKL; FENG; VALAEE, 2011). Outras técnicas utilizam a informação 3D capturada por *Motion Capture* (ZHOU; FRADE, 2012; DAHL, 2014), ou por *Data Glove* e sensores magnéticos (ARGUETA; KO; CHEN, 2009).

A vantagem desses tipos de sensores está na alta precisão e também na captura dos dados em 3D; porém, esse tipo de abordagem é muito invasiva, pois o usuário necessita vestir aparatos como luvas, marcadores, ou mesmo sensores complexos. Alguns dos trabalhos encontrados nessa revisão bibliográfica utilizavam o sensor *Wiimote* para extrair a trajetória 3D da mão (BRADSHAW; NG, 2008; HAN; KIM; KIM, 2012), enquanto outros (BEHRINGER, 2005; BORCHERS; SAMMINGER; MUHLHAUSER, 2002) combinam o uso de câmeras de vídeo com marcadores com luz infravermelha. Apesar de serem menos invasivos do que o caso dos sistemas baseados em *Motion Capture*, o usuário ainda precisa segurar um aparato para garantir a eficiência do sistema. Alternativamente, existem trabalhos que buscam obter a posição da mão a partir de rastreamento baseado na informação de cor de pele em imagens obtidas por câmeras monoculares RGB. Essa abordagem é bastante explorada (KAKUMANU; MAKROGIANNIS; BOURBAKIS, 2007; SCHRAMM; JUNG, 2007; DARDAS; GEORGANAS, 2011; GUO et al., 2012). Entretanto, como já mencionado, esse tipo de abordagem com câmeras monoculares sofre com inúmeros problemas, como inicialização, variação de iluminação, perda do objeto rastreado e reinicialização, necessidade de marcadores, entre outros.

O rastreamento da mão pode ser melhorado com a utilização de câmeras estéreo, as quais permitem extrair a profundidade da cena (LI et al., 2011; HUANG et al., 2012; JE; KIM; KIM, 2007; SEGEN; KUMAR; GLUCKMAN, 2000). A informação da profundidade da cena permite facilmente segmentar os objetos em relação ao fundo, simplificando o rastreamento da mão. O cálculo para gerar o mapa de profundidade, também conhecido como mapa de disparidades, é baseado no casamento de pixels das imagens provenientes do arranjo de câmeras (SCHARSTEIN; SZELISKI, 2002). Todavia, na maioria das vezes

não é possível encontrar um mapeamento para todos pixels e, por causa disso, é comum que esse mapa seja esparso, contendo muitas falhas. Alguns autores (ZITNICK et al., 2004; SCHRAMM; JUNG, 2014) propuseram técnicas que buscam melhorar a estimativa do mapa de profundidade; mas como consequência acabam aumentando o custo computacional e na maioria das vezes não permitem a execução em tempo real. Já outros autores como Zhao e Taubin (2011) propõem alternativas nas quais a prioridade está no desempenho em detrimento da qualidade do mapa de profundidade.

A utilização de sensores ativos, especialmente câmeras RGB-D, graças à sua popularização com o *Microsoft Kinect*, vem sendo uma alternativa bastante eficaz. Por exemplo, o algoritmo desenvolvido por Shotton et al. (2011) permite a extração e o rastreamento de 31 articulações do esqueleto do corpo humano. Com base na facilidade de acesso ao equipamento, custo e disponibilidade de técnicas de rastreamento das mãos, esse sensor foi escolhido para gerar os dados de entrada que serão utilizados nessa tese.

Com relação à classificação de gestos, pode-se verificar que as técnicas que utilizam o rastreamento da posição da mão, ou mesmo das articulações do esqueleto humano, implementam classificadores temporais que utilizam as coordenadas e suas derivações, obtidas durante o rastreamento. Nestes casos, as características mais comuns encontradas para cada ponto rastreado foram sua posição 2D ou 3D, sua velocidade, sua aceleração e variação angular (AKL; FENG; VALAEE, 2011; UCHIDA et al., 2012; HUSSAIN; RASHID, 2012). Também foram encontrados trabalhos que utilizam representações no domínio de frequência (HARDING; ELLIS, 2004; SIGAL et al., 2010), e combinações das opções anteriores com medidas de densidade de ocupação geográfica, como proposto por Wang et al (WANG et al., 2012). Existem também estudos que usam descritores globais, como a quantidade de movimento e o índice de contração do corpo (CAMURRI et al., 2005; FENZA et al., 2005). Essas características, incluindo as coordenadas das articulações e os descritores globais, foram usadas em diferentes análises (LUCK; TOIVIAINEN, 2006; SARASÚA; GUAUS, 2014) e mostram relevância em suas capacidades para descrever os gestuais de regência.

2.1.2 Classificadores Temporais

Neste trabalho, um dos objetivos propostos é avaliar se um determinado gesto realizado pela mão de um usuário está sendo realizado corretamente, considerando o formato do desenho resultante do movimento e também sua precisão rítmica. Ou seja,

pretende-se verificar se o trajeto realizado pela mão segue um padrão pré-estabelecido, e se o mesmo está em coerência temporal com a métrica de compasso e com o andamento, fixo ou variado, da música. Para realizar essa tarefa é necessário um classificador capaz de identificar qual padrão de gestual está sendo executado pelo aluno, considerando um conjunto finito de modelos previamente definidos. Esse classificador deverá ser capaz de identificar os padrões executados, mesmo que em diferentes andamentos, e não apenas os decorrentes da velocidade do conjunto de compassos marcados em um fluxo contínuo. Além disso, outro objetivo dessa pesquisa é identificar e avaliar a precisão rítmica do aluno. Assim, além de robusto às variações de andamento, o algoritmo classificador deverá ser capaz de gerar alguma métrica relativa à precisão temporal da execução real.

Como esperado, o movimento apresentado pelo gesto de marcação de compassos é periódico. É possível observar certos padrões em cada um dos tipos de movimentos relativos aos compassos de métrica binária, ternária e quaternária (vide Figura 3.1). Por exemplo, todos os modelos têm pouca ou nenhuma variação na posição horizontal imediatamente antes do primeiro tempo do compasso. Durante a evolução do segundo tempo do compasso, contudo, há variação negativa – positiva no eixo horizontal (movimento da mão para a esquerda e só depois para a direita) do padrão ternário, enquanto no padrão quaternário essa variação horizontal é positiva – negativa (movimento da mão primeiro para a direita e só depois para a esquerda), e no padrão binário, somente positiva. Esses e outros padrões, como a presença de mínimos e máximo locais, foram utilizados para gerar um conjunto de regras e montar um máquina de estados finito capaz de reconhecer tais padrões de movimento na proposta de Schramm e Jung (2007) e posteriormente em (ARGUETA; KO; CHEN, 2009). Porém, essa abordagem não é suficiente para estimar a precisão rítmica em pequena granularidade. Além disso, a presença de ruído, ou mesmo filtragem excessiva, pode adicionar ou suprimir esse elementos de máximo e mínimo, impossibilitando o funcionamento da máquina de estados.

Considerando que o usuário pode executar o movimento em diferentes andamentos, ou seja, variar a velocidade, a trajetória descrita por um mesmo gesto pode ter comprimentos diferentes, o que é um dos maiores desafios no desenvolvimento de sistemas de reconhecimento de gestos (AKL; FENG; VALAEE, 2011). Nestes casos, onde há variação da velocidade do movimento, uma solução é definir atributos que sejam invariantes no tempo, para efetuar o treinamento dos classificadores. Exemplos desse tipo de atributos são a média e variância dos dados amostrados, pois elas são medidas de tendência central e dispersão, respectivamente. O inconveniente dessas medidas é que elas reduzem a

resolução temporal dos dados.

Hidden Markov Models – HMM e *Dynamic Time Warping* – DTW tem sido largamente empregados, efetuando sincronização temporal e sendo robustos em relação aos problemas acima mencionados. Classificadores que implementam técnicas baseadas em DTW (UCHIDA et al., 2012; ADISTAMBHA; RITZ; BURNETT, 2008; HUSSAIN; RASHID, 2012; AKL; FENG; VALAEE, 2011) usam um procedimento determinístico para comparar exemplos. O algoritmo DTW busca a melhor deformação temporal (sincronização) entre duas séries temporais (trajetória do gestual), usando uma métrica de distância para cada par de pontos provenientes de ambas sequências. Por outro lado, classificadores baseados em HMM (MIN et al., 1997; BEVILACQUA et al., 2010; RAJKO et al., 2007; KOLESNIK, 2004) trabalham com uma abordagem probabilística. Em um simples caso ilustrativo, o gestual poderia ser representado, em um HMM, por uma sequência de estados escondidos e características observáveis. Essas características, obtidas pelo sistema de rastreamento do gestual, geram transições entre estados, o que representa a evolução do gestual ao longo do tempo.

Bianne-Bernard et al. (2011) assinalam que classificadores baseados em HMM e que consideram informação contextual e dinâmica sofrem com a alta dimensionalidade e grande variância nas densidades de probabilidade dos estados observáveis resultantes. Como consequência, estas condições implicam grandes bases de dados de treinamento para garantir uma modelagem correta do problema. Em outras palavras, métodos baseados em HMM necessitam bases de dados de treinamento suficientemente grandes para que se possa modelar o conjunto de transições dos estados e as probabilidades dos respectivos observáveis, para que sejam atingidos bons resultados de classificação. Por outro lado, em casos onde não é possível obter grande bases de treinamento, algoritmos determinísticos podem ser uma melhor alternativa aos algoritmos probabilísticos, permitindo a construção, a partir de poucos exemplos, de modelos eficientes e com significado semântico coerente. Esta tese leva isso em consideração e apresenta uma nova abordagem baseada na DTW para avaliar os gestuais de marcação de compasso. De fato, a técnica descrita no Capítulo 3 permite a inclusão de novos modelos de gestos com apenas um único exemplo, o que não seria uma tarefa trivial para a maioria das abordagens não determinísticas.

O algoritmo DTW permite calcular uma medida de distância entre duas sequências de tamanhos distintos. Desta forma, o DTW pode ser utilizado para classificar movimentos representados por sequências que não possuem alinhamento perfeito, problema que pode ter acontecido por causa de taxas de amostragem distintas ou deslocamento tempo-

ral. Por esse motivo, esse algoritmo é apropriado para resolver o problema de classificação de gestos e tem sido utilizado em diferentes trabalhos.

Por exemplo, Uchida et al. (2012) utilizam DTW para reconhecer dígitos capturados por uma caneta digitalizadora. Semelhante a ele, podem ser citados trabalhos de Adistambha, Ritz e Burnett (2008) e Hussain e Rashid (2012), os quais reconhecem gestos, utilizando as coordenadas da posição da mão ao longo do tempo. Akl, Feng e Valaee (2011), ao invés de utilizar DTW para efetuar a classificação, a utilizam para alinhar as amostras capturadas por um acelerômetro. A partir dessas amostras alinhadas, eles geram uma base de dados, que será utilizada por um classificador não temporal. Após gerado o classificador, os movimentos gestuais capturados são alinhados com DTW e classificados de acordo com o menor custo obtido em relação aos modelos previamente armazenados. Zhou e Frade (2012) efetuam o alinhamento entre sequências temporais, contendo informação de diferentes sensores, como *motion capture* e acelerômetros e vídeo. Isso é possível porque o DTW pode ser implementado para suportar o alinhamento de dados multidimensionais (ZHOU; FRADE, 2009). Por exemplo, Wang, Cheng and Wang (2011) e Junejo et al. (2011) propõem o uso de DTW para alinhar um videoclipe, contendo algum tipo de ação/gesto humano com sequências de vídeo previamente armazenadas em uma base de dados.

2.1.3 Algoritmo DTW

O DTW é um algoritmo que mede a similaridade entre duas sequências temporais $X = (x_1, x_2, ..., x_N)$ de tamanho $N \in \mathbb{N}$ e $Y = (y_1, y_2, ..., y_M)$ de tamanho $M \in \mathbb{N}$. Para isso, ele busca o melhor casamento entre as duas sequências, fazendo uma busca exaustiva que considera todas as combinações entre os elementos de ambas sequências, respeitando certas restrições. Seguindo a definição e notação de Müller (2007), considere um espaço de características definido por $\mathcal{F}$, onde $x_n,\ y_m \in \mathcal{F}$ para $n \in [1 : N]$ e $m \in [1 : M]$. Para comparar dois atributos $x, y \in \mathcal{F}$ é preciso definir uma medida local de custo c (ou distância), de tal forma que:

$$c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0} \tag{2.1}$$

Na prática, a medida de custo entre dois pontos $c(x, y)$ deve ser pequena quando x e y são similares. Ao avaliar a medida local de custo para cada possível par de elementos

entre as sequencias X e Y, obtém-se uma matriz de custo $C \in \mathbb{R}^{N \times M}$, definida como $C(n, m) = c(x_n, y_m)$. A partir dessa matriz de custo, o objetivo passa a ser a obtenção do alinhamento entre X e Y que gerar o menor custo global.

Um *warping path* é a sequência de pontos $p = (p_1, ..., p_L)$ tal que $p_\ell = (n_\ell, m_\ell) \in [1 : N] \times [1 : M]$ e satisfaz três condições:

1. **fronteira:** $p_1 = (1, 1)$ e $p_L = (N, M)$.
2. **monotonicidade:** $n_1 \leq n_2 \leq ...n_L$ e $m_1 \leq m_2 \leq ...m_L$.
3. **largura do passo:** $p_{\ell+1} - p_\ell \in \{(1, 0), (0, 1), (1, 1)\}$ para $\ell \in [1 : L - 1]$.

A condição de fronteira força o alinhamento tanto do primeiro, quanto do último elemento da sequência X com o primeiro e último elemento da sequência Y, respectivamente. Sendo assim, o alinhamento é feito, considerando todo o comprimento de cada sequência. A condição de monotonicidade implica a condição natural da evolução temporal, ou seja o movimento em ambas sequências não pode voltar no tempo. A condição de largura do passo expressa a ideia de continuidade do movimento, e sendo assim ela deve incluir todos os elementos de ambas sequências X e Y e proibir elementos duplicados no *warping path*.

O custo total $C_p(X, Y)$ de um *warping path* p entre X e Y é definido como:

$$c_p(X, Y) = \sum_{\ell=1}^{L} c(x_{n\ell}, y_{m\ell}), \tag{2.2}$$

e o *warping path* ótimo entre X e Y é aquele que obtiver o custo mínimo na matriz de custos C entre todos os possíveis caminhos p iniciados em $p_1 = (1, 1)$ e terminados em $p_L = (N, M)$. A DTW é definida, então, como:

$$DTW(X, Y) = c_p^*(X, Y) = min\{C_p(X, Y)\}. \tag{2.3}$$

O algoritmo DTW efetua uma busca por força bruta para encontrar $c_p^*(X, Y)$, a qual possui complexidade exponencial. Entretanto, ele pode ser implementado através de programação dinâmica, a qual reduz sua complexidade para $O(NM)$, onde N e M representam o comprimento das duas sequências comparadas. Maiores detalhes de implementação podem ser verificados em Müller (2007).

2.2 Captação e Classificação da Melodia Cantada

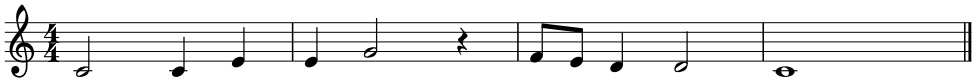
No contexto desta proposta, o solfejo é o ato de cantar uma melodia e ao mesmo tempo marcar o andamento musical com o gestual da mão. A primeira seção desse capítulo tratou da revisão dos trabalhos relacionados à marcação de compasso e regência musical. Nesta seção, serão tratados os trabalhos relacionados ao outro aspecto desta tese, ou seja, será feita uma revisão do trabalhos que buscam analisar o sinal de áudio e também definir técnicas automatizadas para a avaliação da entonação melódica.

O estudo de solfejo por músicos iniciantes é um processo que necessita a verificação de um avaliador externo, a fim de garantir sua correta execução e, ao mesmo tempo, propiciar um retorno avaliativo sobre o processo de aprendizagem. Usualmente, esse processo é realizado numa sala de aula com a presença de um especialista (professor de música). Um exemplo de exercício de solfejo, considerando apenas o aspecto melódico, é ilustrado na Figura 2.1. No item (a) dessa figura, é ilustrada a partitura desse exercício. Nos itens (b) e (c) são exibidas a amplitude e o espectrograma (espetro de frequências ao longo do tempo) do sinal de áudio capturado durante a execução do exercício.

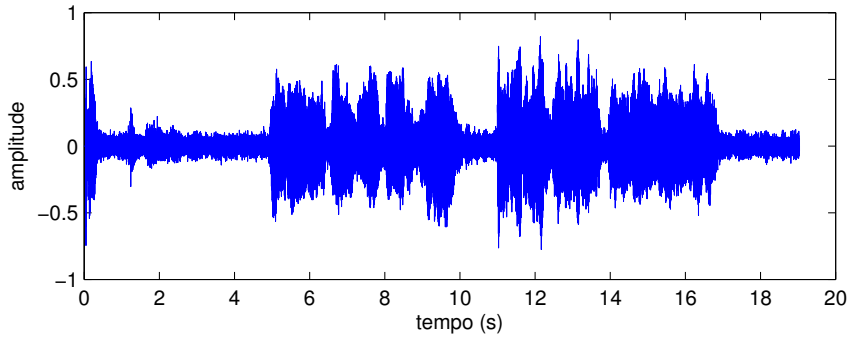
Para que se possa avaliar o solfejo melódico é preciso identificar três aspectos principais: instante de início da nota musical, instante final da mesma (duração) e altura (frequência fundamental). Outros aspectos como dinâmica (volume) e timbre também são importantes para uma avaliação criteriosa; mas aqui serão considerados secundários, pois podem ser analisados como propriedades do som numa etapa posterior à identificação das notas musicais.

Poucos dos trabalhos encontrados focam explicitamente na análise de áudio para avaliação do canto (MOLINA et al., 2013; MAKA, 2013; MAUCH; FRIELER; DIXON, 2014; ABESSER et al., 2013; MAUCH et al., 2015). Em todos esses trabalhos os autores efetuam a transcrição melódica (identificação das notas a partir do sinal de áudio capturado), mas apenas a proposta de Molina et al. (2013) implementa o alinhamento temporal da melodia detectada com a melodia de referência (partitura). Molina et al. (2014) fizeram uma revisão dos métodos para transcrição melódica da voz, incluindo um estudo comparativo dos diferentes critérios utilizados para avaliar a performance dessas técnicas. Apesar da grande variedade de métricas comparadas, a maioria delas focam em verificar se o algoritmo foi capaz de detectar o instante inicial, a duração, e a frequência fundamental das notas cantadas. A partir desses trabalhos, pode-se perceber que a tarefa de avaliação do canto apresenta componentes muito semelhantes aos procedimentos de segmentação

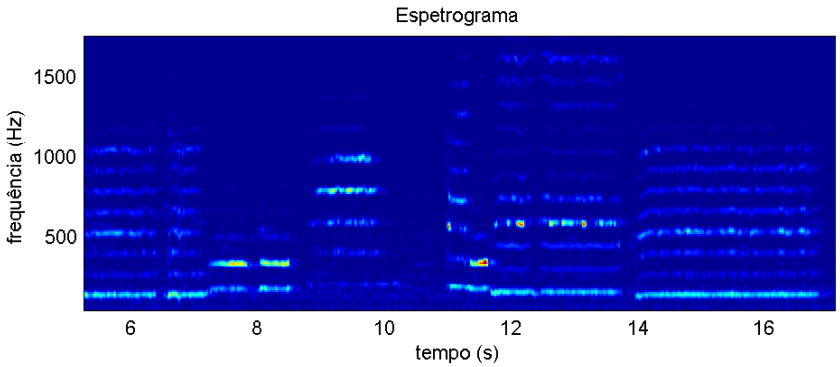
Figura 2.1: Ilustração de um exemplo de exercício de solfejo. (a) Exemplo de uma partitura com um simples exercício de solfejo para uma voz, em compasso de métrica quaternária simples e tonalidade Dó Maior. (b) Sinal de áudio de voz capturado durante a execução do exercício do solfejo. (c) Espectrograma obtido a partir do sinal de áudio.



(a)



(b)



(c)

Fonte: O Autor

de áudio baseado na identificação de notas musicais (BELLO et al., 2005), transcrição automática (BENETOS et al., 2012; BAY et al., 2012) ou mesmo sincronização de audio e partitura (EWERT; MULLER; GROSCHE, 2009; YU et al., 2010) .

Embora existam inúmeros métodos para localizar eventos musicais ao longo do sinal de áudio, essa tarefa ainda é um problema em aberto, dependendo do seu conteúdo sonoro. A classe de algoritmos que visa identificar o início de eventos musicais é conhecida como *onset detection*. O objetivo principal dos métodos de *onset detection* é identificar o início de um evento musical, que acontece normalmente quando há uma transição acentuada na amplitude de energia do sinal de áudio ou em alguns de seus componentes de frequência. Bello et al. (2005) apresentam uma revisão de técnicas para detecção de even-

tos musicais e caracterizam esse processo em três etapas principais: pré-processamento, redução e *peak-picking*. A etapa de pré-processamento é responsável por transformar o sinal original, enfatizando ou atenuando alguns aspectos do áudio, tornando as etapas subsequentes mais simples. A etapa de redução busca uma nova representação do sinal, de forma que os eventos musicais sejam facilmente identificados por um crescente aumento de amplitude. O último estágio, *peak-picking* avalia o sinal resultante das etapas anteriores e detecta as características que são agora bem definidas. Geralmente, essas características são pontos de máximo local gerados a partir da avaliação da primeira derivada do sinal resultante. Existem técnicas mais robustas que utilizam o domínio de frequências (TAN et al., 2010), decompondo o sinal em múltiplas bandas (EWERT; MULLER; GROSCHE, 2009; GAINZA; COYLE, 2011), ou construindo modelos probabilísticos (DEGARA et al., 2011) .

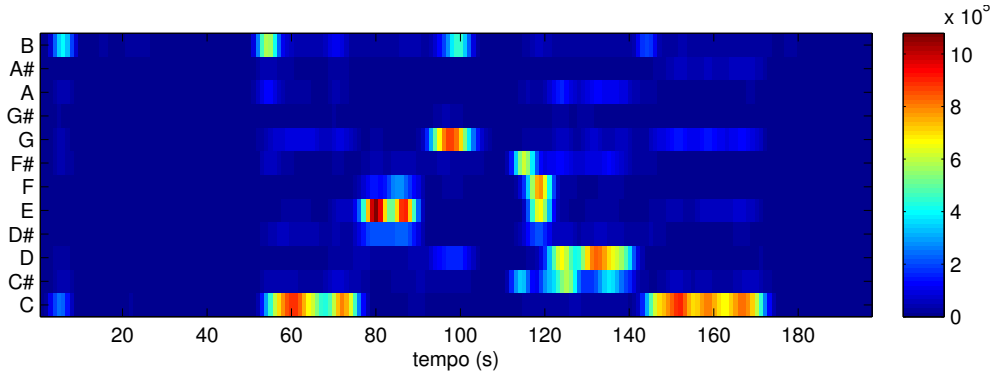
O sinal pode ser decomposto em múltiplas bandas de frequências, onde cada banda corresponde a um *pitch* da escala temperada Ocidental. Essa decomposição pode ser feita com banco de filtros (*filter banks*), em múltipla escala, ou diretamente no domínio de frequências, com a Transformada de Fourier. Essa nova representação permite a detecção de *onsets* localizados em frequências específicas, ou seja, permite uma detecção mais robusta das notas musicais já determinadas pela sua altura (frequência fundamental).

A representação do sinal no domínio de frequências ao longo do tempo por meio de uma análise em quadros (janelas) pode ser feita usando o espectrograma, o qual permite analisar os parâmetros tempo, frequência e amplitude. No exemplo da Figura 2.1c, é possível visualizar, no eixo horizontal, a evolução temporal da execução vocal de um trecho de solfejo. No eixo vertical estão representadas as frequências do espectro, enquanto as amplitudes de cada componente são representadas em cores.

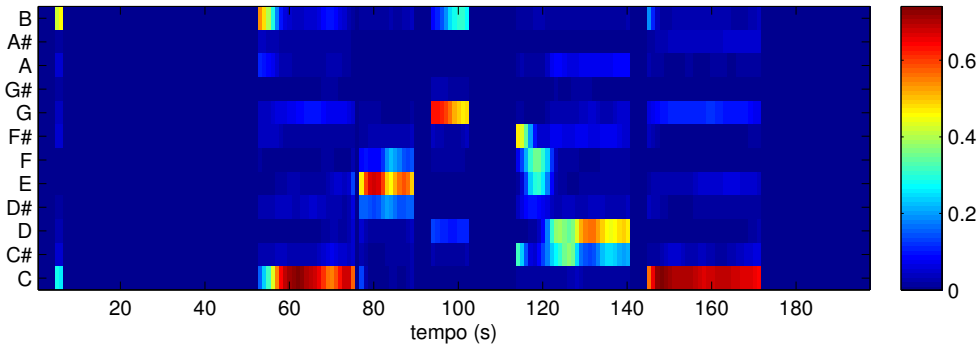
A partir do espectrograma é possível construir novos tipos de características mais eficazes para representar as notas musicais. Ao agrupar os harmônicos sobre as frequências fundamentais da Escala Temperada (temperamento igual, onde a oitava é dividida em doze semitons iguais), é gerado um atributo cromático (*chroma feature*) (MÜLLER; EWERT; KREUZER, 2009), o qual é mais indicado para representar as características musicais do sinal de áudio. Desta forma, todo o espectro de frequências é projetado sobre doze *bins* representando todos os semitons distintos da Escala Temperada em uma única oitava, de onde deriva o nome *chroma*. Essa ideia vem do fato que, no sistema auditivo humano, frequências espaçadas com intervalos exatos de oitava (frequências múltiplas de potência de 2, como $f_0, 2f_0, 4f_0, 8f_0, ...$) são percebidas como notas muito semelhantes,

fenômeno chamado de "*chroma perception*". Seguindo a analogia do espectrograma, um chromagrama pode ser considerado um histograma das notas musicais (doze semitons) ao longo do tempo. Na Figura 2.2a, é ilustrado o chromagrama gerado a partir do espectrograma.

Figura 2.2: Ilustração do chromagrama obtido a partir da trecho de áudio gravado e exibido na Figura 2.1b. (a) Chromagrama sem normalização. (b) Chromagrama com normalização seguida de filtragem Gaussiana (veja MÜller (2007)).



(a)



(b)

Fonte: O Autor

A análise temporal dos atributos cromáticos permite gerar uma matriz de similaridade, com a qual pode-se agrupar características comuns entre cada nota individual e, por consequência, segmentar a faixa de áudio. Foote (2000) propôs uma medida de novidade no conteúdo do sinal de áudio, o qual é conhecida como "*Novelty*". Para identificar esses pontos de mudanças significativas na música, é efetuada a convolução de uma máscara (*checkerboard kernel*) ao longo da diagonal da matriz de similaridade. Essa máscara realça os instantes onde há mudança no conteúdo do sinal de áudio. Diferentes tipos de máscaras podem ser utilizados, visando o realce de estruturas pré-definidas (KAISER; PEETERS, 2013). O uso da abordagem sugerida por Foote permite a identificação do início e fim de

cada nota musical contida no sinal de áudio.

Uma vez determinado o início e fim de uma nota musical, é possível identificar a altura da nota, a qual é predominantemente definida pela frequência fundamental do sinal contido nesse intervalo de tempo. A análise de atributos cromáticos é normalmente utilizada para a identificação da altura da nota musical no Sistema Temperado, sendo comum a normalização dos atributos cromáticos para eliminar a sensibilidade a variações na intensidade do som. Esses atributos em versão normalizada são conhecidos como CENS (*chroma energy normalized statistics*). Segundo MÜller (2007), os CENS, que são correlacionados à estrutura harmônica do conteúdo do sinal, são menos sensíveis a perturbações que envolvem dinâmica, timbre, articulação e pequenas variações temporais.

A grande maioria das técnicas de transcrição melódica citadas neste capítulo utiliza como informação de entrada a sequência de frequências fundamentais extraídas do sinal de áudio. Babacan et al. (2013) realizaram um estudo comparativo entre diferentes técnicas para a extração da frequência fundamental. Nesse estudo, o algoritmo YIN (CHEVEIGNÉ; KAWAHARA, 2002), que é uma técnica baseada na autocorrelação do sinal para detectar periodicidades, apresentou a melhor acurácia dentre seis abordagens comparadas. Recentemente, Mauch e Dixon (2014) desenvolveram uma variação do YIN, incluindo informação probabilística e aplicando consistência temporal nas estimativas. O novo algoritmo é chamado de pYIN e possui acurácia e precisão superiores ao algoritmo original.

Entre as técnicas já desenvolvidas para a transcrição melódica, pode-se identificar uma metodologia comum, a qual divide o processo em extração de características de baixo nível (*low-level*), segmentação e rotulação de notas musicais, e, finalmente, pós-processamento para refinar o resultado (MOLINA et al., 2015; GÓMEZ; BONADA, 2013). Por exemplo, Vitaniemi, Klapuri e Eronen (2003) implementaram um algoritmo de transcrição melódica através da detecção de sequências de frequências fundamentais extraídas a partir de uma análise quadro-a-quadro do sinal de áudio. Essas sequências são posteriormente convertidas em probabilidades (observáveis), os quais são utilizados num algoritmo baseado em Hidden Markov Model (HMM). Esse procedimento é estendido por Ryynanen e Klapuri (2004), que implementaram um conjunto maior de características de áudio de baixo nível. Assim, além das estimativas da frequência fundamental, eles também mapearam em distribuições de probabilidades as características que indicam a presença de voz, acentuação rítmica (ataque) e estimativa de andamento.

Frequentemente, modelos musicológicos são incluídos nos algoritmos de transcri-

ção melódica, visando melhorar a acurácia do sistema. Geralmente, eles atuam como probabilidades *a priori*. Seguindo essa linha, os autores de (VIITANIEMI; KLAPURI; ERONEN, 2003) incorporaram um modelo de duração das notas musicais que mapeia funções de densidade de probabilidade em inteiros, subdivisões e múltiplos da unidade de tempo. Dessa forma, ele funciona como um "quantizador" rítmico das notas detectadas. Modelos musicológicos podem ser usados para detectar a tonalidade da peça, bem como a estrutura rítmica e a forma da performance musical, restringindo as opções e consequentemente melhorando a acurácia (KLAPURI; DAVY, 2006). Infelizmente, modelos musicológicos não podem ser aplicados diretamente como informação *a priori* em ferramentas de avaliação, uma vez que não é possível ter qualquer tipo de expectativa sobre a performance do canto realizada pelo estudante.

O trabalho de Molina et al. (2013) tem um objetivo similar ao proposto nesta tese. Ele explora a avaliação do canto através da análise individual de medidas de similaridade das notas musicais transcritas, bem como através da informação de alinhamento temporal entre a performance do estudante e a melodia alvo (partitura). Apesar do uso de medidas de similaridade individuais para nota musical detectada, a avaliação final em (MOLINA et al., 2013) é construída usando-se uma pontuação global estimada a partir de testes com especialistas humanos, os quais, durante uma fase de avaliação manual, definiram pontuações numa escala entre 1 e 10 para cada gravação de trecho cantado numa fase de treinamento. Os autores estimaram correlações entre essas pontuações e as medidas de similaridade, obtendo assim uma estimativa de qualidade apropriada para um contexto global. Porém, um ponto negativo dessa abordagem é que ela descarta a informação local, no nível individual de cada nota musical, ou seja, não é possível localizar e quantificar a(s) nota(s) responsáveis pelo mau ou bom desempenho da performance do canto. Uma pequena extensão dessa abordagem é apresentada em (LIN et al., 2014), a qual inclui outras características de áudio para estimar a correlação da pontuação obtida pelos especialistas.

Um trabalho recente (MOLINA et al., 2014) apresenta uma taxonomia de medidas de avaliação utilizadas em diversos algoritmos de transcrição automática de canto. A maioria das abordagens tabuladas pelos autores utilizam medidas para avaliação dos algoritmos baseada no erro estimado quadro-a-quadro, incluíndo o *pitch*, *onset* e *offset*. Há também algumas estratégias que usam a informação do alinhamento temporal (DTW ou HMM) entre o *ground truth* (partitura) e a melodia transcrita. Apesar da variedade e do esforço em construir medidas robustas e compreensivas, essas ideias não se prestam diretamente ao contexto de avaliação de solfejo. De fato, a definição de "correto", quando

analisadas a propriedades *pitch/onset/offset*, é geralmente aplicada quando os respectivos valores caem dentro de faixas de tolerância com limiares fixos (MOLINA et al., 2014). Essa abordagem pode ser adequada para realizar a comparação entre algoritmos de transcrição melódica, porém, ela pode não ser coerente com a percepção humana durante a avaliação do exercício de solfejo. Mesmo que existam algumas tentativas (MOLINA et al., 2013; LIN et al., 2014) para resolver esse problema através da conexão da análise feita por especialistas com as métricas de avaliação, a avaliação final nestes casos carrega uma interpretação global, faltando com o *feedback* de detalhes individuais das notas cantadas.

No desenvolvimento desse trabalho, um subconjunto das técnicas acima descritas foi combinado para identificar as notas musicais obtidas a partir de gravações de emissões vocais durante a execução do exercício de solfejo, e uma nova técnica para a avaliação individual de cada nota com base na opinião de especialistas foi desenvolvida. Nos próximos capítulos serão apresentados as técnicas desenvolvidas para identificar os padrões gestuais de marcação da métrica de compasso, bem como a abordagem para a avaliação do solfejo por meio da análise do sinal de áudio, incluindo a detecção das notas musicais e respectivas alturas.

3 PADRÕES GESTUAIS DA MÉTRICA EM MÚSICA

Neste capítulo, será apresentada a descrição e o correspondente detalhamento de aspectos pertinentes aos padrões gestuais da métrica em Música, considerando-se neles eventuais variações internas de andamento, em um mesmo trecho musical. Para tanto, foram desenvolvidas técnicas específicas para detecção de gestos, possibilitando o reconhecimento de padrões de movimentos da mão, ao conduzirem a marcação de compassos.

Neste trabalho foi utilizada uma câmera RGB-D (Microsoft Kinect) para rastrear a posição 3D da mão ao longo do tempo, devido a seu baixo custo e aplicabilidade em ambientes pouco controlados. A Figura 3.1 ilustra cada uma das dimensões (x, y, z) relativas à posição da mão ao longo do tempo em exemplos de movimentos de marcação de compasso.

O sistema proposto segmenta o movimento em seções periódicas e as classifica em padrões (a) binários, (b) ternários e (c) quaternários (ou rejeita o movimento). O esquema de classificação proposto explora o algoritmo Dynamic Time Warping (DTW) com uma nova métrica de custo para fazer o casamento entre os elementos das sequências de características que é posteriormente combinado com um classificador Bayesiano, superando limitações de rejeição de *outliers* existentes na proposta de Bevilacqua et al. (2010), sendo capaz de rejeitar *outliers* mesmo em situações onde o treinamento é realizado com um conjunto pequeno de amostras. Além da classificação do tipo de movimento, o sistema proposto introduz uma nova metodologia, a qual supera abordagens anteriores (SCHRAMM; JUNG, 2007; ILMONEN; TAKALA, 1999; NAKRA et al., 2009) e permite estimar a precisão rítmica da marcação do compasso sem utilizar explicitamente os pontos de mínimos locais (eixo vertical) da trajetória descrita pela mão.

Seguindo a ideia proposta em (LICHTENAUER; HENDRIKS; REINDERS, 2008), a sincronização temporal (*time warping*) e a classificação foram separadas em dois passos distintos. O classificador Bayesiano utiliza probabilidades condicionais estimadas a partir da saída da DTW (custo total), quando avaliada para cada uma das classes de gestual. Além disso, a DTW utiliza uma medida de custo que é mais adequada para efetuar o alinhamento dos sinais (trajetórias da mão) neste tipo de aplicação. Para descartar os *outliers* (ou gestos que não pertencem à nenhum dos três tipos de compasso), a técnica aplica uma medida de confiança estimada a partir das respecitvas funções de densidade acumulada. Dessa forma, esta abordagem não necessita exemplos de gestuais não-classe na base de treinamento, o que é altamente desejável nesse tipo de aplicação. Finalmente,

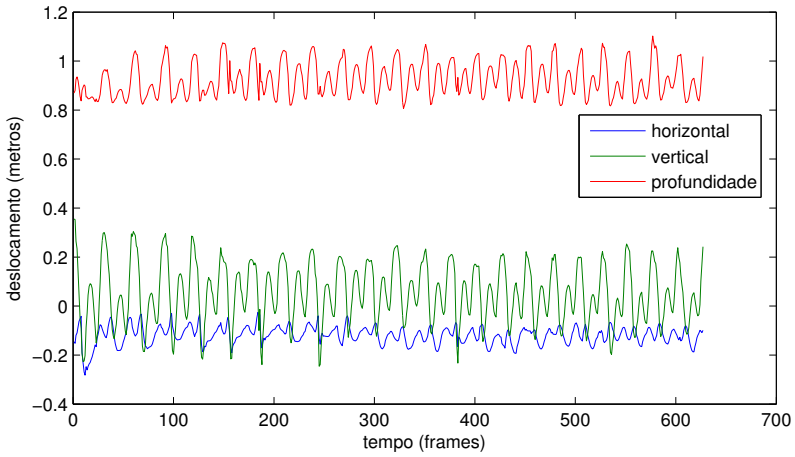
a precisão do tempo/andamento dos gestuais de marcação de compasso é extraída diretamente do alinhamento das sequências (*time warping*) obtido pelo algoritmo de caminho ótimo da DTW. Os detalhes da abordagem proposta são mostrados a seguir.

3.1 Extração da Trajetória

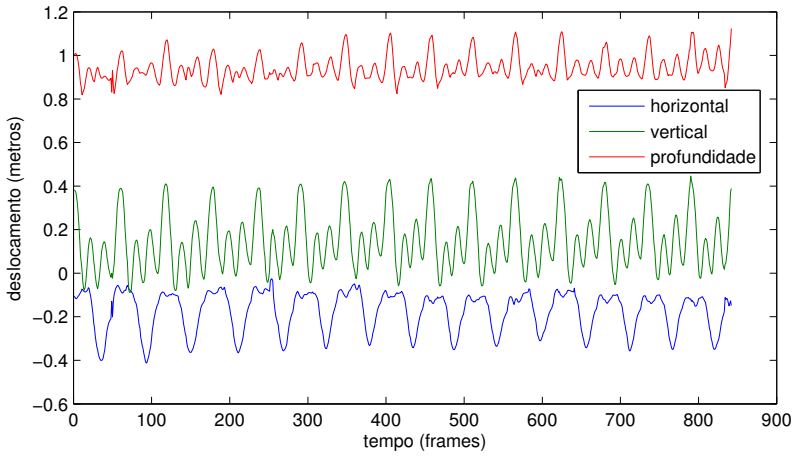
O dispositivo de aquisição de dados utilizados nessa pesquisa é uma câmera RGB-D montada em posição paralela ao plano do solo e de frente para o regente. Ao invés de usar o rastreamento 3D da mão diretamente, foi explorado o algoritmo rastreamento de esqueleto (skeleton tracking) (SHOTTON et al., 2011), o qual mostrou melhores resultados uma vez que ele valida as coordenadas da mão com o restante do corpo humano. A cada quadro t, são extraídas apenas as posições 3D $x(t)$, $y(t)$ and $z(t)$ (horizontal, vertical e profundidade, respectivamente) da mão. As demais coordenadas relativas às outras juntas do esqueleto são descartadas. Uma vez que os modelos de marcação de compasso no escopo deste trabalho são praticamente planares (i.e. z é praticamente constante), a componente de profundidade também é descartada. Para evitar eventual desfiguração do padrão de movimento de marcação de compassos, gestos com variação de profundidade superior a 25 centímetros são invalidados. A trajetória é então representada como um conjunto de pontos 2D $\boldsymbol{P}(t) = (x(t), y(t))$. Exemplos de trajetórias horizontais e verticais para os padrões binário, ternário e quaternário são ilustradas na Figura 3.1.

Para reduzir a possibilidade de ruído no processo de captura do movimento, um pré-processamento usando um filtro causal passa-baixa é aplicado ao sinal. Mais precisamente, um filtro Butterworth, digital e causal, de segunda ordem com frequência de corte $w_c = 0.4\pi$, a qual corresponde a 6 Hz se a sequência de vídeo é adquirida a 30 quadros por segundo. Comparações com o filtro Savitzky-Golay (SAVITZKY; GOLAY, 1964), o qual usualmente preserva melhor as informações de mínimo e máximos, apresentaram resultados semelhantes. O filtro passa baixa Butterworth mantém, nos gestuais de marcação de compasso, as unidades de tempo (*beats*) representadas por mínimos locais em intervalos de pelo menos 167 ms, mantendo as características dos movimentos da mão em velocidades de até 360 BPM.

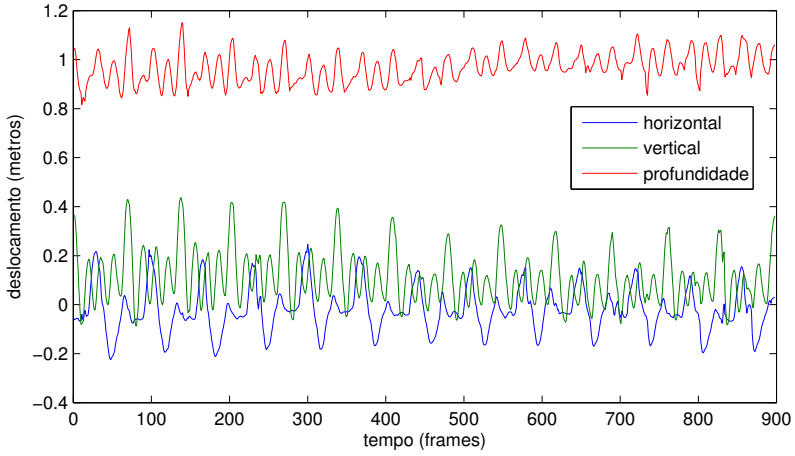
Figura 3.1: Coordenadas (x-horizontal, y-vertical, z-profundidade) dos modelos de marcação de compasso.



(a) Movimento Binário



(b) Movimento Ternário



(c) Movimento Quaternário

Fonte: O Autor

3.2 Segmentação da Trajetória

O sistema captura os dados do gestual utilizando a câmera RGB-D e sincroniza a sequência de pontos com um metrônomo externo controlado por um sistema MIDI. Um conjunto de N amostras é gravado para cada compasso, onde N depende de três parâmetros: fórmula de compasso ($binário = 2$, $ternário = 3$, $quaternário = 4$), taxa de aquisição do sensor (FPS), e metrônomo BPM (andamento). Por exemplo, a performance de um gestual ternário ao longo de um compasso, com $FPS = 30$ e $BPM = 70$ gera $3 \times 30 \times (60/70) \approx 77$ pontos amostrados. Uma vez que o gestual de marcação de compassos é sincronizado com o metrônomo, a sequência de compassos é trivialmente segmentada a cada primeiro tempo de um novo compasso. Dependendo da configuração dos parâmetros citados acima, as trajetórias podem ter diferentes comprimentos. Por isso, todas as sequências gravadas durante a fase de treinamento foram reamostradas para um tamanho constante $N = 60$.

O passo de reamostragem é apenas necessário na fase de treinamento para a geração dos modelos de gestuais, conforme explicado na próxima seção. Porém, não há necessidade de reamostrar os dados de entrada durante a etapa de avaliação. Dado que o foco deste trabalho está na avaliação, ele demanda intrinsicamente um conjunto verdade (*ground truth*), o qual pode ser definido como uma sequência de vários compassos contendo distintas fórmulas de compasso e diferentes andamentos. Assim, a etapa de avaliação (classificação) da marcação de compasso não exige um andamento fixo: a trajetória é capturada e segmentada de acordo com o andamento definido no conjunto verdade, e comparada contra o tipo de gestual também definido no conjunto verdade.

Considerando esses aspectos, a variável $M_i^r = \{\boldsymbol{P}_i^r(1), \boldsymbol{P}_i^r(2), ..., \boldsymbol{P}_i^r(N)\}$ denota a i-ésima amostra da trajetória relacionada à classe r, para $i \in \{1, ..., N_s^r\}$, onde N_s^r é o número de amostras relacionados à classe r, e $r \in \{d, t, q\}$[1]. Lembrando que todas as trajetórias usadas no treinamento possuem o mesmo número de amostras (N).

3.3 Classificador

Dada uma sequência de exemplo, extraída da etapa de segmentação e candidata à uma classe de gestual, o objetivo do classificador é atribuir um dos possíveis rótulos: biná-

[1] d, t, q são relacionados aos gestuais de marcação de compasso binário, ternário e quaternário, respectivamente.

rio, ternário, quaternário ou inválido (no caso do gestual não ser realizado corretamente). Para esse propósito, foi proposto um classificador estatístico baseado numa modificação da DTW, como explicado a seguir. O primeiro passo do classificador é construir um protótipo de trajetória $\hat{M}^r$ para cada classe r. Como assinalado por Wöllner e colegas, a escolha de um modelo gerado a partir da média de um conjunto de sequências é mais adequado do que utilizar apenas um único exemplo, uma vez que estudos (WÖLLNER et al., 2012) mostraram ser mais fácil para as pessoas reconhecerem protótipos de gestuais gerados a partir da média dos exemplos de treinamento, em particular, os pontos de ataque de cada tempo (*beat time*) dentro do compasso. Além disso, ao usar um protótipo de gestual a partir da média de diferentes exemplos, pode-se extrair medidas estatísticas comparando o conjunto de treinamento e o próprio protótipo de cada classe, obtendo-se uma estimativa da variância intra-classe.

Neste trabalho foi adotada a *metrically trimmed mean* (KIM, 1992) das amostras ao londo do tempo. Essa técnica remove exemplos que estão distantes da mediana (possivelmente relacionados à *outliers*) e calcula a média com os exemplos restantes. Para uma dada classe r e tempo t, $\hat{\boldsymbol{P}}^r(t)$ denota a mediana (calculados independentemente nas coordenadas x e y) considerando todos as amostras $\boldsymbol{P}_1^r(t)$, $\boldsymbol{P}_2^r(t)$, ..., $\boldsymbol{P}_{N_s^r}^r(t)$. Além disso, K_n denota a distribuição empírica representando a distância Euclidiana $\|\boldsymbol{P}_i^r(t) - \hat{\boldsymbol{P}}^r(t)\|$ para cada ponto da mediana. Usando a mesma notação de (KIM, 1992), $1\{.\}$ denota função de indicação, $\lceil . \rceil$ denota arredondamento para o maior inteiro, e a *metrically trimmed mean* a partir da media é definida como a média aritimética de $N_s^r - \lceil \alpha N_s^r \rceil$ observações:

$$\begin{aligned} \hat{\boldsymbol{P}}_\alpha^r(t) &= \tfrac{1}{N_s^r - \lceil \alpha N_s^r \rceil} \sum_{i=1}^{N_s^r} \boldsymbol{P}_i^r(t) 1\{.\} \\ 1\{.\} &= \|\boldsymbol{P}_i^r(t) - \hat{\boldsymbol{P}}^r(t)\| \leq K_n^{-1}(1-\alpha), \end{aligned} \tag{3.1}$$

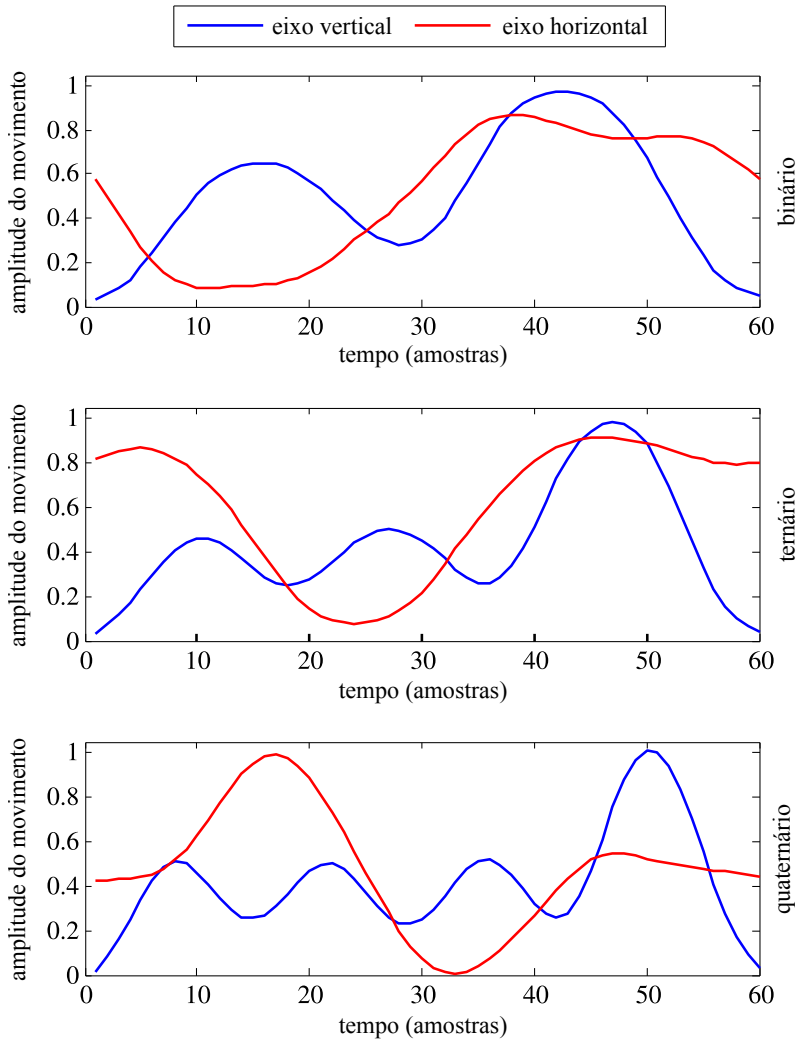
onde $\alpha = 0.5$ é a fração de amostras removidas do cálculo da média. Cada modelo de gestual é então denotado por

$$\hat{M}^r = \{\hat{\boldsymbol{P}}_\alpha^r(1), ..., \hat{\boldsymbol{P}}_\alpha^r(N)\}. \tag{3.2}$$

Análise de Procrustes (W.; MOON, 1992) é aplicada entre o conjunto de treinamento e os respectivos modelos, visando maximizar o alinhamento dos dados com possíveis transformações (translação, rotação ou escala) ocorridas ao longo da captura de diferentes exemplos. Após, os dados são normalizados entre 0 e 1 e os modelos são atualizados por

meio de uma nova aplicação da *metrically trimmed mean* ao longo da variável temporal (usando desta vez $\alpha = 0.2$) sobre a base de dados previamente transformada e normalizada pela análise de Procrustes. Finalmente, esses modelos são filtrados novamente com o mesmo filtro Butterworth utilizado previamente nos exemplos individuais para reduzir distorções residuais que foram involuntariamente adicionadas ao sinal após o cálculo da média. A Figura 3.2 ilustra o modelo final para cada padrão (classe) de gestual de marcação de compasso, obtidos a partir da base de dados utilizada nos experimentos desse trabalho (detalhes sobre a construção da base de dados são discutidos na seção 3.6.1). As linhas azuis e vermelhas mostram a variação da posição da mão nos eixos vertical e horizontal, respectivamente.

Figura 3.2: Modelos de gestos gerados na etapa de treinamento e usados na classificação dos movimentos: (a) binário, (b) ternário e (c) quaternário. A linha azul representa a posição da mão no eixo vertical e a linha vermelha representa a posição da mão no eixo horizontal.



Fonte: O Autor

O reconhecimento de gestuais é efetuado pela comparação entre um exemplo capturado e os três modelos $\hat{M}^r$ obtidos na fase de treinamento. Mais do que isso, essa proposta usa um *pipeline* que combina a medida de erro de saída da DTW com uma avaliação probabilística, permitindo estimar um valor de confidência como feedback ao usuário. Desta forma, o sistema é capaz de retornar um número que presenta a confidência de que um gestual foi realizado corretamente. Esse valor de confidência permite rejeitar *outliers*, os quais são relacionados à performance incorreta do gestual.

3.3.1 Função de Custo Local

O algoritmo DTW permite calcular a distância entre duas sequências de diferentes tamanhos e/ou com relativo desalinhamento. Essas características são comuns no contexto desse trabalho, uma vez que podem acontecer variações na taxa de amostragem, deslocamento temporais ou execução de gestuais com velocidades irregulares. Para comparar duas sequências, a DTW busca o caminho de melhor alinhamento, o qual minimiza uma função de custo global através da busca exaustiva entre todas as possibilidades de combinações de caminhos ao longo de todos os elementos de ambas sequências. Para alcançar esse objetivo, a DTW utiliza também uma função de custo local c que é avaliada a cada possibilidade de par de valores obtidos a partir das duas sequências que estão sendo comparadas. Tradicionalmente, essa função de custo é baseada em normas L^p, como por exemplo a distância Euclidiana. Porém, os gestos de marcação de compasso podem ser executados com diferentes larguras de amplitudes ao longo dos eixos vertical e horizontal, de tal forma que normas L^p tendam a produzir erros grandes mesmo nos casos em que os gestos são considerados similares por especialistas.

Outros autores (ZHANG; EDGAR, 2008) propuseram medidas de custo incluindo informação da derivada do sinal, evitando pontos de singularidade e reduzindo alinhamentos tendenciosos. Keogh e Pazzani (KEOGH; PAZZANI, 2001) também combinaram o filtro Savitzky-Golay para estimar a derivada do sinal e ao mesmo tempo evitar problemas com ruído. Seguindo essas ideias, foi proposto nesse trabalho uma nova função de custo que consegue estimar o alinhamento de sequências gestuais com grandes variações de amplitude, obtendo resultados melhores (vide Seção 3.6.2 para uma avaliação comparativa dos resultados) do que as medidas que usam apenas a derivada. Essa nova função reduz o valor do custo quando o sinal das derivadas de ambas as sequências comparadas são iguais. De fato, tal formulação prioriza o casamento de porções do sinal que são coerentemente

monotônicas em ambas as trajetórias (i.e. comportamentos ascendentes e descendentes, os quais são caracterizados pelos sinal da derivada), o que também implicitamente auxilia no alinhamento de extremos locais, mesmo quando a variação de amplitude é grande. Essa é uma propriedade importante no contexto da marcação de compassos, uma vez que a amplitude do movimento é altamente dependente do usuário e pode variar significativamente entre diferentes pessoas (ex. crianças são menores em tamanho e consequentemente seus movimentos tem amplitudes menores do que os movimentos realizados por adultos.).

Considere a comparação entre duas trajetórias 2D: $A = \{\boldsymbol{a}(1), ..., \boldsymbol{a}(N)\}$ e $B = \{\boldsymbol{b}(1), ..., \boldsymbol{b}(N)\}$, com $\boldsymbol{a} = (a_x, a_y)$ e $\boldsymbol{b} = (b_x, b_y)$. Também definem-se $\boldsymbol{a'} = (a'_x, a'_y)$ e $\boldsymbol{b'} = (b'_x, b'_y)$ como sendo as derivadas de primeira ordem das trajetórias, calculadas usando diferenças finitas. Inicia-se a comparação de similaridade no eixo x, definida com base na amplitude e derivada:

$$u_x = |a_x - b_x| + \beta_x |a'_x - b'_x|, \tag{3.3}$$

onde β_x controla o peso do termo da derivada (atribuído 0.5 com base em nossos experimentos). Também foi introduzido um termo de "encolhimento" baseado na similaridade das derivadas, obtido por

$$w_x = \frac{(a'_x - b'_x)^2}{(a'_x - b'_x)^2 + \epsilon_x^2}, \tag{3.4}$$

onde ϵ_x controla o decaimento de w_x e é foi definido com o valor de 0.07, baseado na média do valor de $|a_x|$, considerando todos os exemplos da base de dados de treinamento. Em particular, w_x tende a apresentar valores muito baixos quando perto do extremo local, uma vez que ambos a'_x e b'_x serão pequenos. A função de custo proposta (apenas em x) é então dada por

$$c_x(a_x, b_x, a'_x, b'_x) = \begin{cases} w_x u_x & \text{se sign}(a'_x) = \text{sign}(b'_x) \\ u_x & \text{caso contrário} \end{cases}. \tag{3.5}$$

É importante notar na Equação (3.5) que a trajetória de pontos com mesmo sinal na derivada tende a apresentar custos pequenos. Assim, os pontos relacionados com movimentos ascendentes na primeira trajetória tendem a serem casados com os pontos relacionados com movimentos ascendentes da segunda trajetória. Como consequência, extremos locais também tendem a serem casados entre si.

Uma função de custo c_y análoga é definida para efetuar a comparação dos com-

ponentes verticais das trajetórias ($\beta_y = 2.5,\ \epsilon_y = 0.07$), e a função de custo final, que compara as trajetórias 2D é dada por

$$c(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{a}', \boldsymbol{b}') = \alpha_x c_x(a_x, b_x, a'_x, b'_x) + \alpha_y c_y(a_y, b_y, a'_y, b'_y), \tag{3.6}$$

onde α_x e α_y são pesos que dão prioridades distinta para os componentes verticais e horizontais. Nos testes realizados, foram usados $\alpha_x = 1.0$ e $\alpha_y = 1.2$, com base em experimentos.

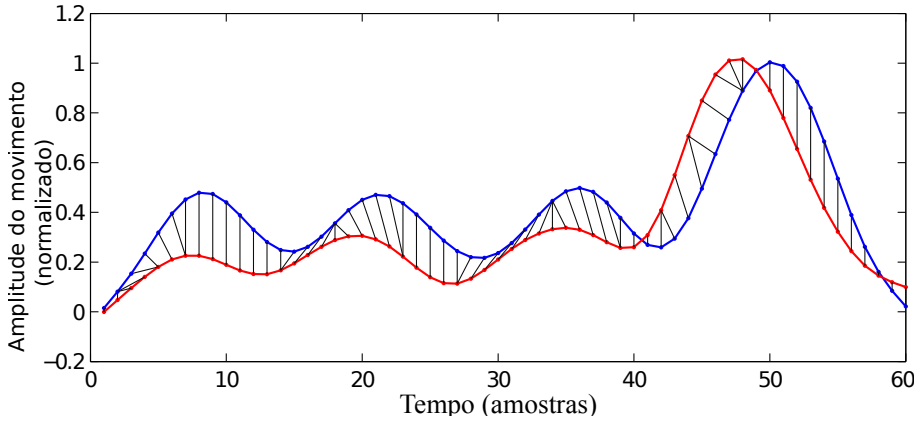
Como exemplo, a Figura 3.3 ilustra a comparação entre o casamento obtido usando a DTW com a função de custo tradicional (parte superior), baseada na distância Euclidiana e, o casamento obtido com a DTW usando a métrica proposta por este trabalho (parte inferior). Como pode ser visto, a técnica proposta obtém um resultado melhor ao conectar os pontos relacionados ao mesmo tempo (*beat time*), e também ao manter boa correspondência nos extremos locais ao longo do componente vertical do gestual.

O reconhecimento de gestuais poderia ser realizado através da comparação dos resultados da DTW entre o sinal de entrada e os três modelos previamente treinados. Neste caso, o gestual com a menor distância DTW representa a classe vencedora. Uma vantagem dessa abordagem é que o modelo pode ser implementado usando apenas um único exemplo para cada classe (ex. o professor grava um exemplo de cada movimento), ao contrário de classificadores tradicionais (ex. Máquina de Vetores de Suporte (SVM), Rede Neurais Artificiais (ANN), Modelo Oculto de Markov (HMM)), os quais requerem uma quantidade de dados de treinamento mínima e suficiente. Um ponto negativo dos classificadores de distância mínima é que se um movimento aleatório é executado, então provavelmente todas as distâncias DTW serão grandes, e mesmo o menor valor, obtido quando comparado com os diferentes modelos, não representa um gestual válido. Um procedimento simples para rejeitar *outliers* neste cenário seria a aplicação de um limiar na saída da distância DTW. Porém, a seleção de tal limiar não é uma tarefa fácil em casos onde a base de treinamento não é grande suficiente. Apesar disso, de acordo com os resultados da seção 3.6.2, o classificador proposto neste trabalho apresenta bons resultados, particularmente nos casos onde a base de dados de treinamento é pequena (menos de 10 exemplos por classe).

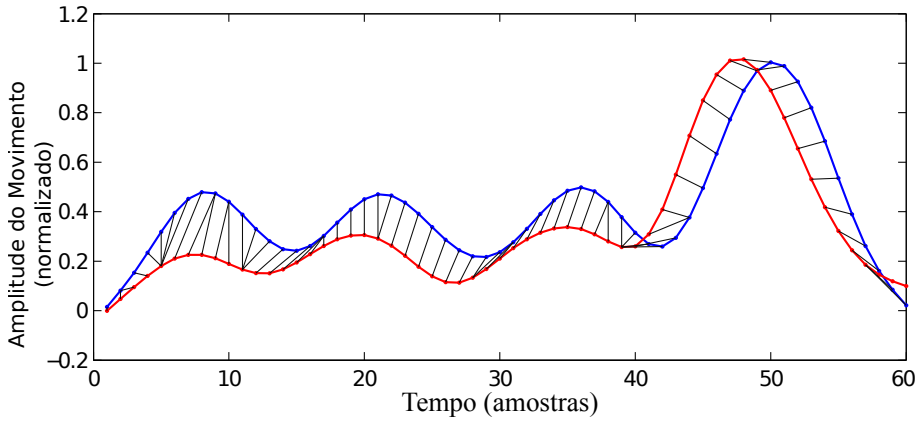
Quando o conjunto de treinamento é grande, é possível estimar a distribuição intra-classe das distâncias DTW entre cada exemplo de treinamento e o modelo de gestual, de tal forma que um classificador Bayesiano pode ser idealizado. Uma vantagem deste tipo de classificador é que ele produz uma saída probabilística, a qual pode ser interpretada

como uma medida de confiança, que por sua vez pode ser usada para avaliar o grau de corretude dos movimentos de marcação de compasso realizados por estudantes de regência musical. Além disso, o conhecimento da distribuição intra-classe permite formular um método formal para rejeitar *outliers*. Algoritmos de Regressão Logística para múltiplas classes poderiam ser aplicados diretamente nesse tipo de problema, porque a saída desse tipo de algoritmo são as probabilidades *a posteriori* de cada classe do modelo (BISHOP, 2006). Porém, esse tipo de algoritmo precisa incorporar exemplos de não-classe na fase de treinamento para permitir futura rejeição de *outliers*. O treinamento de exemplos não-classe é uma tarefa complicada e, neste tipo de aplicação, tais exemplos podem não estar disponíveis.

Figura 3.3: Alinhamento DTW entre duas sequências temporais. (a) função de custo com distância Euclidiana. (b) função de custo proposta (Equação 3.6). As linhas pretas mostram o alinhamento (*warping*) gerado pela DTW. A função de custos proposta reduz a influência de movimentos mais expressivos, comprimindo as diferenças de amplitude, como pode ser visto em (b).



(a)



(b)

Fonte: O Autor

Para essa proposta, foi usada a função densidade de probabilidade condicional acumulada para rejeitar os *outliers*, evitando a necessidade de exemplos não-classe para gerar o modelo. Na fase de treinamento, é possível estimar a função densidade de probabilidade condicional acumulada $p(\delta|r)$ que modela a distribuição dos erros DTW δ para cada classe $r \in \{d, t, q\}$. Neste trabalho, foi utilizada a função densidade de probabilidade *Gamma* para modelar os custos DTW finais. Entre diferentes escolhas existentes para variáveis aleatórias positivas, a distribuição *Gamma* foi escolhida porque tem sido utilizada com sucesso para modelar problemas de "life-testing" (JOHNSON; KOTZ; BALAKRISHNAN, 1995; BALAKRISHNAN; MITRA, 2013), os quais têm características similares com os nossos dados, como, por exemplo, a distribuição assimétrica positiva (cauda mais longa à direita). A função densidade de probabilidade *Gamma*, parametrizada por dois parâmetros positivos, forma α_r e escala θ_r, é dada por:

$$p(\delta|r) \sim Ga(\delta; \alpha_r, \theta_r) = \frac{\delta^{\alpha_r - 1} e^{\frac{-\delta_r}{\theta_r}}}{\Gamma(\alpha_r)\theta_r^{\alpha_r}}, \tag{3.7}$$

onde Γ é função *Gamma*. Os parâmetros α_r e θ_r podem ser estimados usando máxima verossimilhança (*maximum likelihood estimation*). Maiores detalhes deste procedimento podem ser obtido em (RAMACHANDRAN; TSOKOS, 2015; GILES; FENG, 2009).

Na fase de classificação, cada trajetória candidata M é comparada em relação aos modelos que correspondem às três classes, gerando um vetor de distâncias DTW $\boldsymbol{\delta}(M) = (\delta_d, \delta_t, \delta_q)^T$. Assumindo que essas distâncias δ_r seguem a *Gamma* PDF correspondente, conforme Equação (3.7), é possível estimar a probabilidade *a posteriori* de cada exemplo de gestual M que pertencente a classe r, usando a regra de Bayes (DUDA; HART; STORK, 2001):

$$p(r|\boldsymbol{\delta}(M)) = \frac{p(\delta_r|r)P(r)}{\sum\limits_{w \in \{d,t,q\}} p(\delta_w|w)P(w)}, \tag{3.8}$$

onde as probabilidades *a priori* $P(r)$ são definidas como equiprováveis.

3.4 Rejeição de Não-Classe

Na seção anterior, a Equação (3.8) define a probabilidade *a posteriori* de um dado exemplo pertencente a uma classe r, e o maior valor de probabilidade quando comparado

com todas os modelos define a classe vencedora. Porém, exemplos contendo *outliers* (i.e exemplos que não pertencem a nenhum dos três modelos de gestual) serão também associados a uma classe válida. Para identificar tais amostras é necessário definir um esquema de rejeição de *outliers*. Neste contexto, trajetórias discrepantes são aquelas em que a distância DTW δ_r da classe vencedora recai sobre a cauda da *Gamma PDF* correspondente. Mais precisamente, é possível definir um limiar probabilístico $0 \leq T_{dtw} \leq 1$, baseado no resultado da função de distribuição de probabilidade acumulada *Gamma* $F(\delta; \alpha_r, \theta_r)$ (JOHNSON; KOTZ; BALAKRISHNAN, 1995). Definido o limiar T_{dtw}, exemplos com distância DTW δ_r tal que

$$F(\delta_r; \alpha_r, \theta_r) > T_{dtw} \tag{3.9}$$

têm probabilidade menor ou igual a $1-T_{dtw}$ de serem gerados pela distribuição $Ga(\delta; \alpha_r, \theta_r)$. Valores pequenos para T_{dtw} tendem a introduzir falsos negativos (rejeição de exemplos bons), enquanto valores altos tendem a introduzir falsos positivos (*outliers* não são descartados). Para definir esse limiar, um conjunto de experimentos foi realizado usando diferentes valores para T_{dtw}. A partir da análise da acurácia de classificação, definiu-se $T_{dtw} = 0.99$ como o melhor *tradeoff*, o que equivale a uma confiança de 99%.

3.5 Estimativa de Precisão Rítmica

Um segundo objetivo deste trabalho é definir uma métrica para avaliar a consistência temporal do movimento de marcação de compassos ao longo das unidades de tempo (*beat time*). Abordagens anteriores usaram a informação de mínimo local para representar a posição do *beat* (SCHRAMM; JUNG, 2007; ILMONEN; TAKALA, 1999; NAKRA et al., 2009), mas tais técnicas são particularmente vulneráveis quando gestuais com baixa amplitude são realizados. Nestes casos, o ruído na aquisição dos dados pode gerar mínimos/máximos extremos espúrios. Além disso, o processo de filtragem passa-baixas pode eliminar alguns desses pontos, como ilustrado na Figura 3.4a. Neste exemplo, o gestual de marcação de um compasso ternário foi capturado (sem variação de andamento) e comparado com o modelo correspondente, usando a distância DTW proposta. Apesar da trajetória do gestual não seguir exatamente o modelo ternário, ela foi validada pelo comitê de especialistas. Neste exemplo, a unidade de tempo (*beat time*) deveria ser detectada próximo da amostra 20, porém não há nenhum mínimo local no exemplo de trajetória (linha vermelha). Inevitavelmente, nestas situações, as técnicas baseadas em mínimos

locais irão falhar.

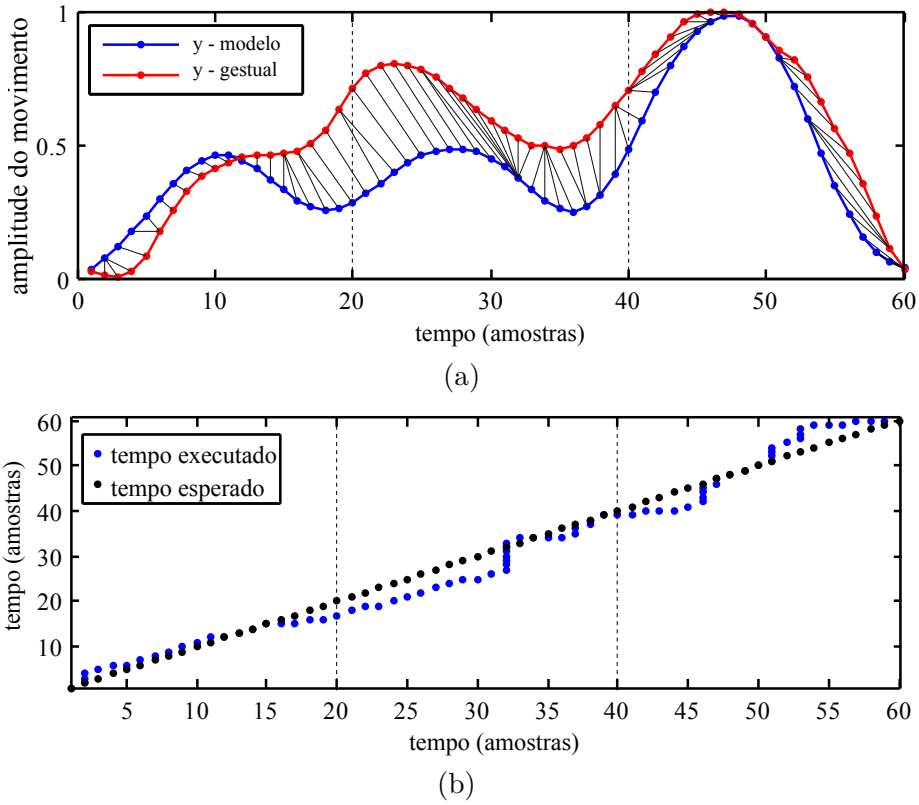
Outra observação importante, obtida a partir do conjunto de treinamento, é que os mínimos locais não acontecem exatamente na posição (*timestamp*) de cada tempo sincronizado pelo metrônomo. Desta forma, é possível que a posição temporal de cada mínimo local, visualmente classificado e validado pelos músicos especialistas, possa ter algum deslocamento no tempo. Por isso, sua posição absoluta pode não ser uma boa estimativa. Este fato corrobora com análises prévias (LUCK; TOIVIAINEN, 2006; DAHL, 2014), as quais demonstram que as unidades de tempo (*beats*) realmente não coincidem com as posições de extrema amplitude. Como alternativa mais eficiente, este trabalho usa o caminho de alinhamento (*warping path*) da DTW para comparar a posição de cada unidade de tempo, a partir da classificação prévia do gestual, com as posições das unidades de tempo do modelo $\hat{M}^r$, evitando o uso explícito do mínimo local. Isso é possível pois cada gestual capturado é previamente segmentado em intervalos (janelas temporais) que contém um compasso inteiro, conforme descrito na seção 3.2. Portanto, as posições ideiais das unidades de tempo são estimadas como frações do tamanho da janela (previamente conhecido), como ilustrado pelas linhas verticais pontilhadas na Figura 3.4a.

Observou-se também, a partir do processo de validação da base de dados de treinamento, que os músicos especialistas permitem certa flexibilidade em relação à precisão rítmica. Isso mostra que a percepção humana relacionada ao limiar de erro pode ser diferente entre avaliadores distintos. Portanto, no caso de apenas duas classes, onde a precisão da unidade de tempo precisa ser classificada como correta (classe φ_1) ou incorreta (classe φ_2), a fronteira que divide esses dois conjuntos é não-determinística. Considerando que até mesmo para os especialistas a fronteira entre φ_1 e φ_2 não é um consenso, o desafio nesta etapa do trabalho é definir uma medida de confiança que pode ser utilizada como feedback para o usuário, reportando quão acurado no tempo um gestual de marcação de compassos pode ser.

No momento em que é calculado o caminho ótimo do DTW, obtém-se também uma estimativa do deslocamento temporal entre cada ponto no modelo e o relativo ponto no gestual candidato. Por exemplo, a Figura 3.4b ilustra o casamento dos pontos do modelo (pontos em preto) com os pontos obtidos a partir de uma gestual da classe ternária (pontos em azul). As linhas verticais pontilhadas representam os tempos dos compasso (*beat timestamp*). A distancia vertical entre os pontos azuis e pretos representa o deslocamento temporal (desalinhamento) entre o gestual capturado e o modelo, assumindo que o movimento tenha sido previamente classificado corretamente na etapa anterior. Neste

exemplo, o andamento utilizado foi $BPM = 70$ e o movimento da mão foi capturado pelo sensor numa taxa de aquisição de 30 FPS, de tal forma que cada deslocamento de tempo (quadro amostrado) representa aproximadamente 28.57 ms.

Figura 3.4: (a) As linhas pretas conectando os pontos vermelhos e azuis representam o alinhamento temporal. (b) A distância vertical entre o ponto azul e o ponto preto representa o deslocamento temporal (atraso ou avanço) do movimento na comparação entre o gesto capturado e o modelo gestual identificado.



(a)

(b)

Fonte: O Autor

Assim, a acurácia do movimento em relação ao tempo é estimada relacionando as distâncias DTW com uma medida de probabilidade (ou medida de confiança), i.e., quanto mais perto um tempo (*beat*) marcado pela mão estiver de um tempo do compasso, maior será a probabilidade de o mesmo pertencer à classe φ_1 e menor será a probabilidade de pertencer à classe φ_2.

Considere-se que o valor discreto λ representa o deslocamento temporal (desalinhamento) medido a cada unidade de tempo do compasso, estimado através da comparação do *ground-truth* (metrônomo) com as posições obtidas pelo caminho ótimo da DTW, computados usando todos os exemplos da base de treinamento previamente validada por especialistas. As distribuições de probabilidade $p(\lambda|\varphi_1)$ e $p(\lambda|\varphi_2)$, relacionadas com as unidades de tempo corretas e incorretas, foram modeladas usando uma função densidade de

probabilidade discreta Poisson, uma vez que essa distribuição é adequada para expressar a probabilidade de eventos que ocorrem no domínio das variáveis aleatórias discretas (KIM et al., 2006). A Figura 3.5a ilustra a distribuição das classes $p(\lambda|\varphi_1)$ e $p(\lambda|\varphi_2)$. Similar ao procedimento de classificação gestual, apresentado na Seção 3.3, optou-se por adotar um *framework* probabilístico para estimar a acurácia das unidades de tempo, o qual é baseado nas probabilidades *a posteriori* $p(\varphi_i|\lambda)$, usando *a priori* igualmente prováveis para as classes φ_1 and φ_2, i.e. $P(\varphi_1) = P(\varphi_2) = 1/2$.

Como pode ser observado na Figura 3.5b, as duas classes apresentam considerável sobreposição, corroborando com as discrepâncias obtidas nas avaliações realizadas pelos músicos especialistas. Erros de classificação são esperados, especialmente perto da fronteira de decisão, definida por $p(\varphi_i|\lambda) > 1/2$ e ilustrada com uma linha vermelha pontilhada na Figura 3.5c. Uma vez que um dos objetivos desse trabalho é prover o feedback para o estudante, uma possível opção é rejeitar exemplos que caem dentro dessa região nebulosa. Os erros de classificação são convertidos em rejeições usando a regra de rejeição de Bayes para o erro mínimo (WEBB, 2011), a qual divide o espaço de amostras em duas regiões. A região de aceitação A e a região de rejeição R são obtidas por

$$R(T_\lambda) = \{\lambda|1 - \max_i p(\varphi_i|\lambda) > T_\lambda\}, \tag{3.10}$$

$$A(T_\lambda) = \{\lambda|1 - \max_i p(\varphi_i|\lambda) \leq T_\lambda\}, \tag{3.11}$$

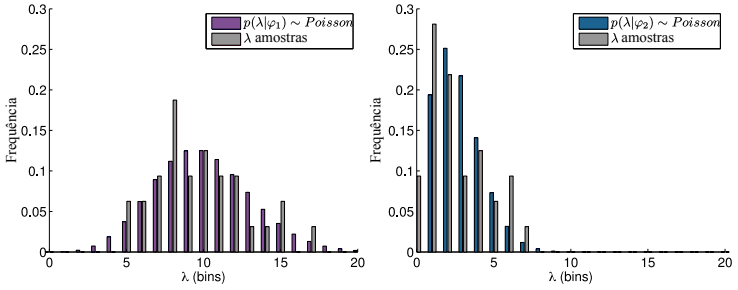
onde o limiar T_λ regula o equilíbrio entre o número de exemplos rejeitados e a taxa de erro de classificação, a qual é definida como

$$e(T_\lambda) = \sum_{\lambda \in A(T_\lambda)} \Big(1 - \max_i p(\varphi_i|\lambda)\Big)\, p(\lambda), \tag{3.12}$$

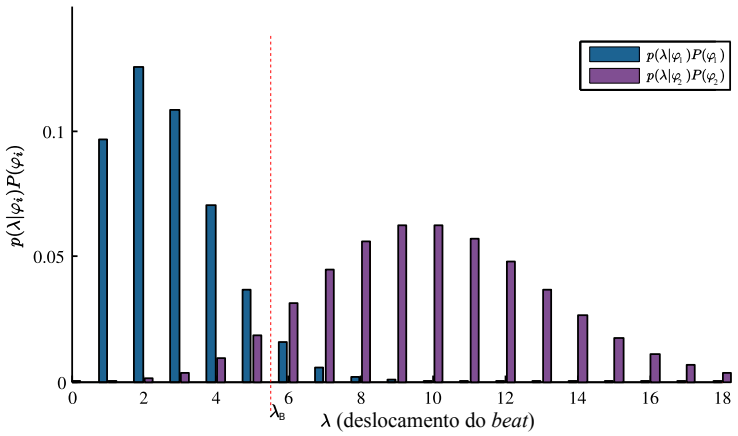
onde $p(\lambda) = P(\varphi_1)p(\lambda|\varphi_1) + P(\varphi_2)p(\lambda|\varphi_2)$ é a distribuição total de λ.

A escolha do limiar $T_\lambda = 0.25$ foi determinada a partir do conjunto de treinamento, onde a acurácia de classificação e o número de rejeições foram levados em conta. As fronteiras entre as regiões A e R, geradas por esse limiar são indicadas pelas linhas pretas verticais e pontilhadas na Figura 3.5c.

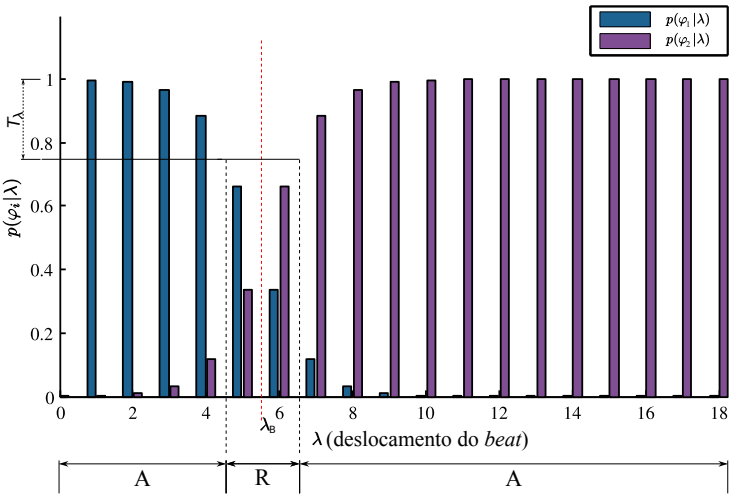
Figura 3.5: (a) Histograma das distâncias λ obtidas a partir do *warping path* da DTW em relação às classes φ_1 e φ_2, e respectiva estimativa (*fitting*) da função densidade de probabilidade discreta Poisson. (b) PDFs $p(\lambda|\varphi_i)$ reescaladas pelas respectivas probabilidades *a priori*, incluindo a fronteira de decisão λ_e. (c) Probabilidades *a posteriori*, com respectivas regiões de aceitação e rejeição.



(a)



(b)



(c)

Fonte: O Autor

3.6 Experimentos e Resultados Relativos à Marcação de Compasso

Nesta seção serão apresentados os resultados obtidos com a aplicação do algoritmo para a classificação de gestos de marcação de compassos e avaliação da precisão rítmica.

3.6.1 Construção da Base de Dados Relativa ao Gesto

O processo de construção e validação da base de dados utilizou exemplos de cada um dos três tipos de movimentos para marcação de compasso (binário, ternário e quaternário). Esses exemplos foram capturados a partir de seis usuários, dos quais, quatro eram músicos profissionais com mais de dez anos de experiência e os outros dois eram músicos amadores. Cada sessão de gravação capturou doze compassos, dos quais os primeiros quatro serviram para aquecimento e por isso foram descartados. Cada sessão de gravação foi sincronizada com um metrônomo controlado através de uma interface MIDI. Alguns exemplos foram executados propositadamente com movimentos errados, incluindo imprecisão rítmica nas unidades de tempo e/ou trajetória incorreta (o movimento da mão não corresponde a nenhum dos modelos de marcação de compasso). As sessões de gravação foram repetidas diversas vezes para gerar um conjunto suficientemente grande de exemplos na base de dados, permitindo a análise estatística.

Numa etapa posterior, cada exemplo gravado foi avaliado por um conjunto externo de avaliadores especialistas. Ao total, foram utilizados dez avaliadores, que rotularam a corretude dos gestuais, a cada compasso, em três possibilidades: correto, incorreto, incerto (no caso em que o avaliador não tinha certeza em sua opinião). Todos os compassos, rotulados como incerto ou que tiveram um rótulo em desacordo entre os avaliadores, foram removidos da base de treinamento relacionada à classificação dos gestuais. O conjunto de exemplos restantes (validados) formaram a base de dados que foi utilizada para treinar todos os classificadores de gestuais avaliados nessa seção.

Um processo semelhante foi adotado para construir a base de dados relativa à precisão rítmica. Porém, apenas os exemplos que foram rotulados como incerto foram removidos. Ambiguidades na avaliação foram mantidas para garantir que o classificador final contivesse os mesmos aspectos não determinísticos, inerentes do processo de avaliação manual por humanos.

3.6.2 Testes Comparativos com o Classificador de Gestos

Para efetuar uma avaliação objetiva e comparativa da técnica proposta para classificar os gestos de marcação de compasso, este trabalho realizou testes comparando os resultados de diferentes classificadores, tradicionalmente empregados em problemas multi-classe. Dentre os classificadores utilizados, estavam aqueles baseados em redes neurais artificiais, árvores de decisão, hidden Markov models e dynamic time warping. Além disso, a técnica proposta foi implementada em duas versões. A primeira segue o método descrito na Seção 3.3 e utiliza a DTW. A segunda versão faz uma adaptação à técnica e utiliza a HMM no lugar da DTW. Ambas implementações foram construídas com duas opções para a medida de custo local. Assim, tanto a versão DTW quanto a versão HMM foram avaliadas usando a distância Euclidiana e a distância proposta neste trabalho (Eq. 3.6).

Todos os classificadores multi-classe usados nos testes de avaliação foram implementados no MATLAB, usando algumas rotinas do *machine learning toolbox*. O algoritmo de árvore de decisão (DTree) foi implementado usando *Gini's Diversity Index* (RAILEANU; STOFFEL, 2004) (medida de ganho de informação) para fazer a divisão dos nodos. A rede neural artificial foi implementada usando um topologia com uma camada escondida contendo 60 neurônios, conectados em *feedforward* e treinados usando o algoritmo *backpropagation*. O hidden Markov model foi implementado com duas configurações. A primeira configuração, HMM-E, segue a técnica descrita em (BEVILACQUA et al., 2010), onde a emissão de probabilidade dos observáveis $\mathbf{\Omega}$ em cada estado k é modelada por uma distribuição Normal $\Omega_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, com média e variância extraídos a partir dos exemplos de treinamento. Quando o número de exemplos de treinamento era menor que 5, foi aplicado o valor padrão $\sigma^2 = 0.2$, conforme proposto no artigo original. A outra configuração, HMM-P usa a mesma função de custo (definida na Equação 3.6) aplicada no algoritmo DTW para gerar a emissão de probabilidade dos observáveis $\Omega \sim e^{-\frac{c(\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}')}{\sigma^2}}$ (omitimos o índice do estado para facilitar a leitura). Ambas configurações usam a arquitetura esquerda-direita com 60 estados para cada tipo de gestual. As probabilidades de transição $\mathbf{\Phi}$ são os valores empíricos $\Phi_0 = \Phi_1 = \Phi_2 = 1/3$, como definido em (BEVILACQUA et al., 2010). Essas probabilidades de transição significam que há igual probabilidade de a cadeia de Markov, após uma transição (leitura de um símbolo de entrada), permanecer no mesmo estado, avançar um estado, ou avançar dois estados.

Antes de testar todos os classificadores, os exemplos foram segmentados em subsequências contendo um único compasso cada, conforme descrito na Seção 3.2. O conjunto

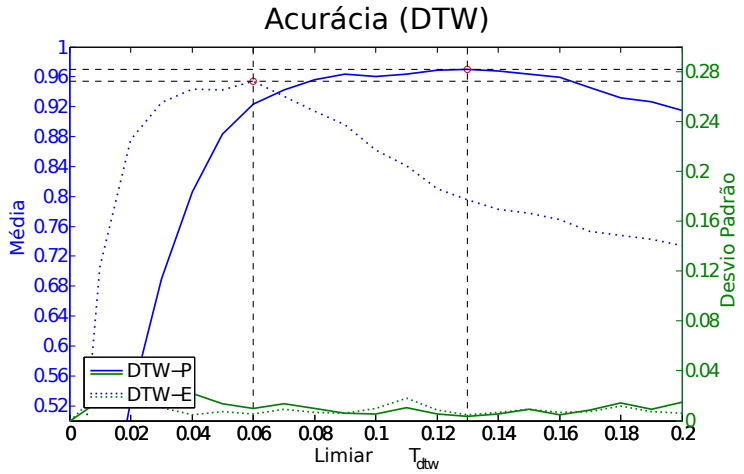
final contendo todos os exemplos validados foi avaliado usando validação cruzada, com diferentes combinações de subconjuntos (n dobras, com n variando entre 1 e 10). A acurácia final de cada configuração, foi calculada usando a média sobre 30 tentativas.

O primeiro conjunto de testes compara as duas abordagens mais utilizadas (conforme identificado na revisão bibliográfica) para classificar gestuais: HMM e DTW. Ambos foram avaliados utilizando a distância Euclidiana e a medida de custo proposta nesse trabalho. Em especial, os testes foram focados em cenários onde a base de dados de treinamento contém poucos exemplos (ex. o professor repete um exemplo de movimento poucas vezes para treinar o modelo). Neste caso, como mencionado na Seção 3.3, um classificador de mínima distância baseado no custo global da DTW foi utilizado. E um limiar fixo δ_{dtw} foi empregado para descartar *outliers*. Este limiar foi estimado a partir de testes onde os *templates* (modelos de gestuais) foram gerados com apenas 10 exemplos. A Figura 3.6 ilustra a acurácia de classificação variando os valores para o limiar δ_{dtw} (e respectivo limiar δ_{hmm}). A partir da análise desses resultados, definiu-se $\delta_{dtw-p} = 0.13$ e $\delta_{dtw-e} = 0.06$ como opções adequadas.

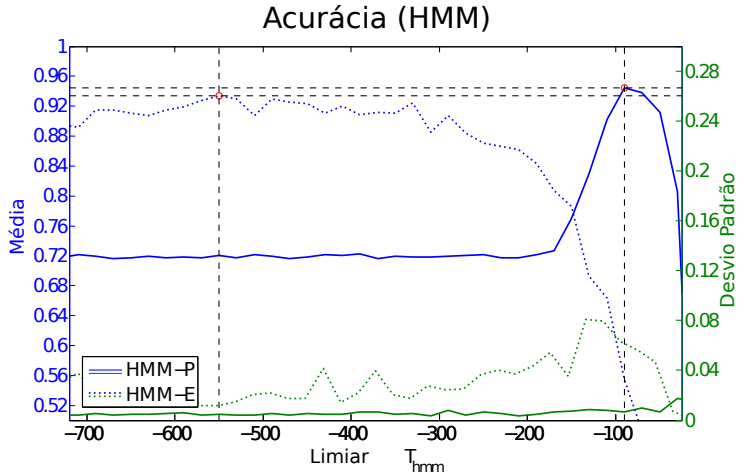
Para a versão em HMM, cada modelo λ_r foi treinado com a mesma base de dados utilizada na versão DTW. A classe com maior *log-probability* $\log P(\mathcal{O}; \lambda_r)$ foi selecionada como classe vencedora, onde $\mathcal{O}$ é a trajetória (sequência de observações). Para a rejeição dos *outliers*, os respectivos limiares $\delta_{hmm-p} = -90$ e $\delta_{hmm-e} = -550$ (também obtidos com os mesmos modelos treinados com 10 exemplos na versão DTW) foram usados sobre $\log P(\mathcal{O}; \lambda_r)$. Fixados esses limiares, foram realizados testes variando a quantidade de exemplos de treinamento entre 1 e 15, onde a acurácia de ambos os classificadores (DTW e HMM) foram comparados. A Figura 3.7 ilustra esse processo comparativo. Como pode ser observado, a abordagem DTW-P obteve os melhores resultados, considerando bases de treinamento com poucos exemplos. A versão DTW-P atingiu uma acurácia média acima de 94%, mesmo no caso onde apenas 6 exemplos foram utilizados no conjunto de treinamento. No segundo teste, foi utilizada a base de dados completa para o treinamento e avaliação dos classificadores. Neste teste, a versão probabilística (Equação (3.8)) da abordagem DTW foi utilizada, incluindo o esquema de rejeição de *outliers* baseado na função distribuição de probabilidade acumulada, conforme descrito na Seção 3.4. Para garantir uma comparação justa com a versão HMM, também foi modelado o $|\log P(\mathcal{O}; \lambda_r)|$ como PDFs *Gamma* , e usado um esquema de rejeição de *outliers* análogo ao utilizado na versão DTW. Os classificadores DTW e HMM, usando seus respectivos procedimentos de rejeição de *outliers*, bem como os outros tradicionais classificadores multi classe, foram

comparados usando validação cruzada sobre uma base de dados contendo 164 exemplos validados para cada classe (Binário, Ternário, Quaternário, e Não Classe). A Tabela 3.2 apresenta a média e o desvio padrão da acurácia, obtidos a partir de 30 tentativas usando diferentes combinações de n dobras. Os resultados indicaram que a abordagem proposta com a DTW apresentou o melhor resultado de classificação, alcançando mais de 95%, mesmo quando foram utilizados 2 dobras. Em particular, é interessante notar que a função de custo proposta (DTW-P) obteve resultados superiores do que a tradicional distância Euclidiana (DTW-E). Além disso, também mostrou superioridade quando comparado com a versão HMM.

Figura 3.6: Acurácia do algoritmo de classificação usando os modelos treinados com apenas 10 exemplos. A linha azul representa a média da acurácia, calculada sobre 30 tentativas, e a linha verde representa o respectivo desvio padrão. (a) Acurácia usando DTW-P (linha sólida) e DTW-E (linha pontilhada), em função de T_{dtw}. (b) Acurácia usando HMM-P (linha sólida) e HMM-E (linha pontilhada), em função de T_{hmm}.



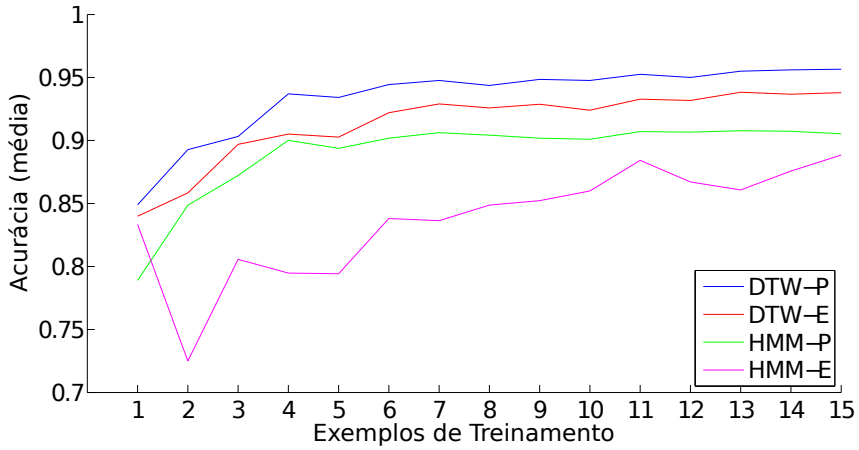
(a)



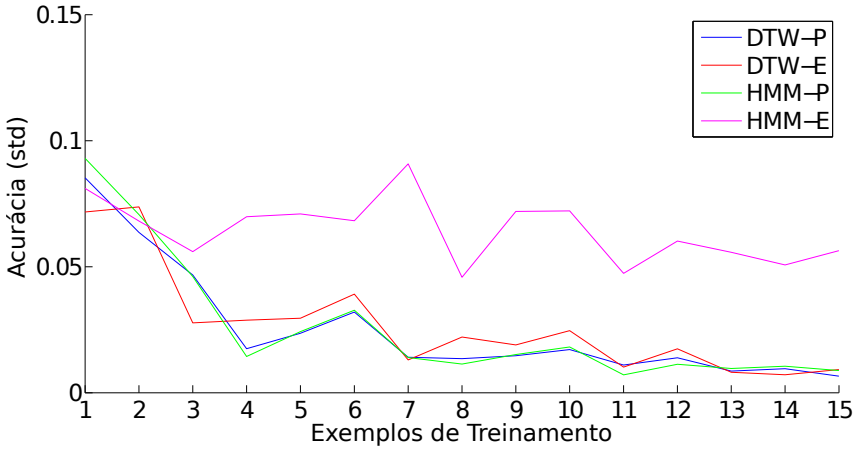
(b)

Fonte: O Autor

Figura 3.7: Média da acurácia (a) e respectivo desvio padrão (b), obtidos a partir da utilização do algoritmo de classificação de marcação de compassos. O classificador foi comparado usando as versões DTW-P, DTW-E, HMM-P e HMM-E, enquanto o número de exemplos utilizados para treinamento de cada modelo foi variado entre 1 e 15. As estatísticas foram obtidas sobre 30 tentativas, utilizando os respectivos limiares fixos de rejeição: mbox$\delta_{dtw-p} = 0.13$, $\delta_{dtw-e} = 0.06$, $\delta_{hmm-p} = -52$ and $\delta_{hmm-e} = -320$.



(a)



(b)

Fonte: O Autor

A Tabela 3.1 apresenta a matriz de confusão obtida na avaliação do classificador proposto com DTW-P, usando validação cruzada com 2 dobras (média sobre todos os resultados). A partir desses resultados é possível verificar que há uma concentração de erros de classificação entre as classes binário e não-classe. A razão principal para isso é que o movimento binário não possui variação significativa da posição da mão quando comparados os eixos horizontal e vertical. Por isso, movimentos aleatórios podem ser confundidos como movimentos binário mais frequentemente. Apesar disso, a acurácia geral do sistema continua elevada.

Visando detectar possível *overfitting*, além dos testes com validação cruzada, foram também realizados novos testes com adição de ruído Gaussiano ao exemplos da base de dados. A Tabela 3.3 mostra a acurácia obtida usando validação cruzada com 2 dobras, e variando a relação sinal/ruído (SNR) entre 29dB e 25dB. O valor mínimo de 25dB foi selecionado a partir da aceitação visual por um especialista durante análise dos exemplos corrompidos com ruído. Apesar da redução já esperada da acurácia do classificador, à medida que o SNR diminui, a performance do DTW-P continua superior aos demais classificadores e também superior a 94%.

Tabela 3.1: Matriz de confusão contendo resultados da avaliação quantitativa (validação cruzada com 2 dobras e 30 tentativas) em relação à corretude de execução do movimento de marcação de compasso (B)inário, (T)ernário, (Q)uaternário, (N)ão-classe.

		Classe Detectada				
		B	T	Q	N	
Classe Alvo	B	**92.28**	0.24	0	7.48	
	T	0.41	**96.26**	0	3.33	
	Q	0	0	**98.01**	1.99	
	N	3.67	0.69	0	**95.64**	
						95.54%

Tabela 3.2: Validação cruzada n dobras.

Técnica	Acurácia *versus* n dobras									
	$n = 2$		$n = 3$		$n = 4$		$n = 5$		$n = 10$	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
DTW-P	**95.54 %**	**0.85 %**	**95.23 %**	**1.20 %**	**95.78 %**	**1.22 %**	**96.06 %**	**1.54 %**	**95.83 %**	**2.30 %**
DTW-E	91.43 %	1.71 %	91.28 %	1.62 %	91.84 %	1.74 %	92.34 %	2.16 %	91.95 %	2.95 %
HMM-P	92.37 %	1.99 %	92.16 %	1.37 %	92.70 %	1.58 %	93.33 %	1.82 %	92.58 %	2.59 %
HMM-E	92.30 %	2.04 %	92.14 %	1.37 %	92.67 %	1.62 %	93.31 %	1.84 %	92.57 %	2.59 %
DTree	81.20 %	3.36 %	80.47 %	2.92 %	80.55 %	3.01 %	80.07 %	2.88 %	79.45 %	4.01 %
ANN	71.45 %	3.55 %	71.83 %	2.75 %	72.52 %	2.31 %	72.36 %	2.36 %	72.15 %	3.38 %

Tabela 3.3: Adição de ruído nos exemplos de treinamento para verificar possível *overfitting*.

Técnica	Acurácia - Adição de ruído Gaussiano									
	SNR = 29 dB		SNR = 28 dB		SNR = 27 dB		SNR = 26 dB		SNR = 25 dB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
DTW-P	**95.00 %**	**1.00 %**	**95.01 %**	**1.04 %**	**94.79 %**	**1.23 %**	**94.66 %**	**1.33 %**	**94.68 %**	**1.06 %**
DTW-E	91.74 %	1.76 %	91.52 %	1.54 %	90.90 %	2.02 %	91.21 %	1.64 %	90.83 %	2.05 %
HMM-P	92.64 %	1.75 %	92.23 %	1.98 %	91.98 %	1.88 %	92.17 %	1.97 %	91.84 %	2.03 %
HMM-E	92.64 %	1.76 %	92.32 %	1.87 %	92.01 %	1.89 %	92.24 %	1.90 %	91.86 %	2.04 %

3.6.3 Acurácia do Classificador para Avaliação da Precisão Rítmica

Foram realizados testes para verificar a acurácia do sistema em relação à capacidade de avaliação da precisão rítmica. Neste caso, a verificação da acurácia relativa à precisão rítmica foi feita utilizando o classificador Bayesiano descrito na Seção 3.5.

Também foi necessário construir uma base de dados com exemplos validados por especialistas para testar esse classificador. O processo de validação dessa base de dados é bastante custoso, pois é necessário rotular todas as unidades de tempo dos compassos gravados. Esse processo foi feito de forma totalmente manual, e necessitou inúmeras revisões. Ao final do processo de validação, obteve-se um total de 32 exemplos válidos para cada classe φ_1 e φ_2. Por isso, utilizou-se o método de validação cruzada *leave-one-out*, garantindo assim 31 amostras para estimar a distribuição de probabilidade Poisson, e a restante para avaliação do modelo.

A implementação do classificador sem a região de rejeição alcançou uma acurácia média de 90.63%, em trinta tentativas. Para definir as regiões de aceitação e rejeição, foi avaliado exaustivamente o limiar T_λ, determinado com base no compromisso entre a acurácia de classificação e o número total de exemplos restantes (não descartados) no processo de classificação. O gráfico da Figura 3.8 ilustra a acurária de classificação versus a porcentagem de exemplos classificados enquanto a região de aceitação é variada.

Uma vez que o objetivo dessa proposta é criar uma ferramenta para dar suporte ao processo de aprendizagem de marcação de compassos, entende-se que a acurácia deve ser priorizada. Por exemplo, um limiar de $T_\lambda = 0.25$ corresponde a uma acurácia de 98.18%, porém causa a incapacidade do sistema dar *feedback* em 14.06% dos casos. Apesar disso, ainda é mais desejável omitir o *feedback* do que gerar uma resposta incorreta. A matrizes de confusão obtidas a partir do processo de validação cruzada, usando $T_\lambda > 0.5$ e $T_\lambda = 0.25$, são apresentadas nas Tabelas 3.4 e 3.5, respectivamente.

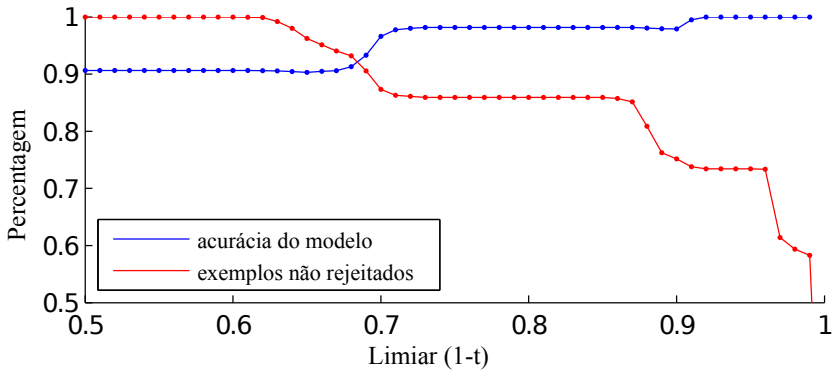
Tabela 3.4: Matriz de confusão obtida a partir de avaliação quantitativa dos resultados usando validação cruzada *leave-one-out* sobre 30 tentativas. A tabela apresenta a média da acurácia do classificador de precisão rítmica sem a função de rejeição.

		Classe Detectada	
		φ	$\overline{\varphi}$
Classe Alvo	φ	**87.50%**	12.50%
	$\overline{\varphi}$	6.25%	**93.75%**
			90.63%

Tabela 3.5: Matriz de confusão obtida a partir de avaliação quantitativa dos resultados usando validação cruzada *leave-one-out* sobre 30 tentativas. A tabela apresenta a média da acurácia do classificador de precisão rítmica com a função de rejeição, onde $T_\lambda = 0.25$.

		Classe Detectada	
		φ	$\overline{\varphi}$
Classe Alvo	φ	**96.30%**	3.70%
	$\overline{\varphi}$	0.00%	**100%**
			98.18%

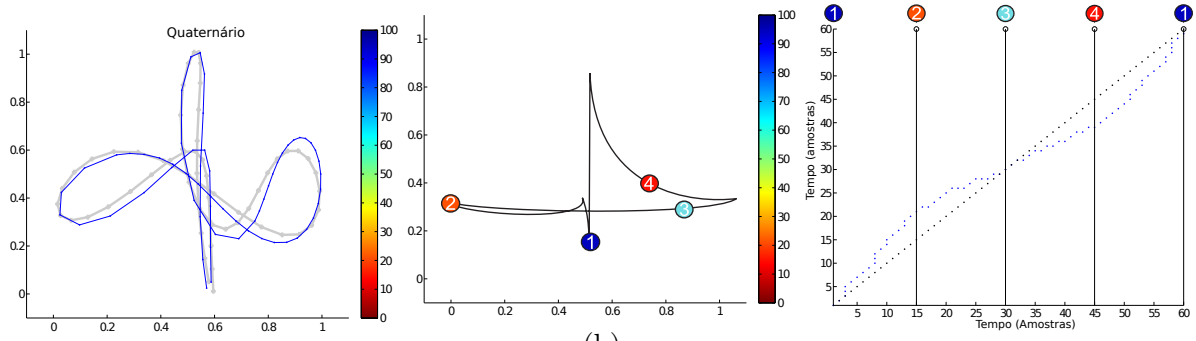
Figura 3.8: Acurácia teórica, baseada na Eq. (3.12). Acurácia obtida (azul) versus fração de exemplos classificados (vermelho), em função da limiar de rejeição T_λ.



Fonte: O Autor

A Figura 3.9 apresenta a interface visual do sistema proposto, onde o aluno pode visualizar o *feedback* gerado pelo modelo. A primeira tela, Figura 3.9a, apresenta as diferenças espaciais entre o movimento executado e os modelos previamente treinados (cinza). Nesta tela, a cor do caminho realizado pelo movimento representa a medida de confiança que o mesmo está sendo executado corretamente. A Figure 3.9b apresenta o *feedback* em relação à precisão rítmica. As cores em cada círculo indicam o grau de confiança sobre a precisão rítmica sobre cada unidade de tempo do compasso. Por fim, a Figure 3.9c ilustra outra possível visualização da precisão rítmica, medida ao longo de todo o compasso. Acelerações e retardos podem ser visualizados através das diferenças (distância vertical, ponto a ponto) entre o template (preto) e o movimento corrente (azul).

Figura 3.9: Interface visual do sistema. a) Diferença especial entre o movimento de marcação de compasso e o template previamente treinado (cinza). Cores representam o grau de confidência na corretude do movimento. b) Cores em cada círculo representam o grau de confiança na precisão rítmica da respectiva unidade de tempo. c) Desalinhamento temporal entre o template e o movimento corrente, ao longo da execução do gestual de marcação de compasso.



(a) (b) (c)

Fonte: O Autor

4 NOTAS MUSICAIS CANTADAS

Entre as técnicas encontradas na revisão bibliográfica que buscam identificar a frequência fundamental ao longo do tempo em um sinal de áudio, duas se destacam por apresentarem os melhores resultados em testes comparativos (BABACAN et al., 2013). A primeira, conhecida como YIN (CHEVEIGNÉ; KAWAHARA, 2002), foi desenvolvida em 2002 e, atualmente, continua sendo largamente empregada, visto que sua implementação é simples, seu custo computacional é baixo e sua acurácia é alta em comparação com outras abordagens (BABACAN et al., 2013). A segunda técnica, chamada de pYIN (MAUCH; DIXON, 2014), é uma extensão da primeira, pois inclui também um modelo probabilístico para melhorar a acurácia de detecção da frequência fundamental. Devido à reconhecida acurácia dessa técnica, ela foi utilizada nesta tese para extrair a altura do som, bem como para segmentar as notas musicais ao longo do tempo.

Contudo, tais qualidades não atendem todas as necessidades do modelo buscado por esta pesquisa. Assim, emprega-se também a técnica de Molina et al. (MOLINA et al., 2013), que propõe novas métricas para a avaliação do canto, considerando especialmente a possibilidade de execução em diferentes andamentos. A maior limitação dessa última, contudo, é que o andamento precisa ser constante ao longo da performance do exercício; ou seja, o exercício de leitura musical pode ser executado em diferentes andamentos, porém o mesmo precisa ser constante do início ao fim da execução. Esse fato impossibilita a execução expressiva do canto, onde o aluno pode variar o andamento durante a prática do exercício.

O presente estudo propõe um sistema que faz uma combinação dessas três técnicas, integrando-as de modo complementar. A técnica proposta permite identificar as notas (altura) e os valores (duração) dos sons musicais, ao mesmo tempo que extrai o respectivo alinhamento temporal em relação à partitura do exercício (*ground truth*), permitindo a plasticidade temporal própria às execuções musicais expressivas.

Para tanto, duas contribuições importantes foram desenvolvidas. A primeira é um processo de agrupamento dos segmentos de notas musicais detectadas pelo algoritmo pYIN, descrito na Seção 4.3, que transforma os grupos gerados em notas musicais (unidades atômicas), as quais são relacionadas às notas musicais da partitura. A segunda contribuição é o desenvolvimento de uma técnica para avaliação individual de cada nota, com base na opinião de especialistas, conforme descrito na Seção 4.4.

Posicionado o foco correspondente à análise de áudio, este capítulo segue com

uma breve descrição do algoritmo pYIN e do processo de transcrição melódica por histerese (MOLINA et al., 2015). Na sequência, é apresentado o algoritmo proposto de agrupamento de segmentos melódicos e o mapeamento dos mesmos em relação à partitura, com vistas a obtenção futura da conformação temporal considerando o gesto de marcação de compassos. Por fim, são aplicadas as métricas avaliativas inspiradas em (MOLINA et al., 2013), as quais são utilizadas posteriormente para gerar um classificador probabilístico da execução de exercícios de solfejo.

4.1 Extração da Frequência Fundamental

A técnica pYIN (MAUCH; DIXON, 2014) é divida em dois estágios. O primeiro estágio utiliza passos semelhantes ao YIN, extraindo a informação de periodicidade através da autocorrelação do sinal. Basicamente, num sinal periódico, a cada instante t, espera-se que a função

$$d_t(\tau) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2, \tag{4.1}$$

seja pequena quando o sinal x_t apresenta periodicidade com período fundamental $\tau = 1/f_0$, onde W é o tamanho da janela de integração. Os autores da técnica original sugerem um tamanho para W equivalente a 25ms. O algoritmo pYIN também usa uma identidade explorada em (CHEVEIGNÉ; KAWAHARA, 2002), calculando a função de diferenças a partir de dois passos. Primeiro calcula a função de autocorrelação:

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau}, \tag{4.2}$$

e depois, calcula a função de diferenças usando a identidade:

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau). \tag{4.3}$$

Para encontrar a frequência fundamental, basta localizar o mínimo local através de uma busca exaustiva variando os valores de τ. Como há inúmeras soluções para esse processo, onde todas são múltiplas do período, a solução desejada geralmente é o primeiro mínimo local (menor que certo limiar), onde $\tau \neq 0$. Entretanto, em muitas vezes os mínimos locais de frequências harmônicas são inferiores ao valor mínimo encontrado para a frequência fundamental, dificultando a seleção imediata do primeiro mínimo local.

Visando minimizar o problema causado por esse fenômeno, Cheveigné e Kawahara (CHEVEIGNÉ; KAWAHARA, 2002) ainda propuseram a média acumulada e normalizada da função de diferenças, tal que a Equação (4.1) é atualizada por

$$d'_t(\tau) = \begin{cases} 1 & \text{se } \tau = 0 \\ \frac{d_t(\tau)}{(1/\tau)\sum_{j=1}^{\tau} d_t(j)} & \text{caso contrário.} \end{cases} \tag{4.4}$$

Essa nova versão evita o problema do limite inferior, pois inicia em 1. Além disso, a normalização evita problemas com diferenças de amplitude do sinal.

No algoritmo YIN, os autores sugerem a aplicação de um limiar fixo $s = 0.1$ para identificar o mínimo local relativo ao período da frequência fundamental do sinal, denotado por $Y(x_t, s)$, tal que $d'(\tau) < s$. Nos casos, onde não há um mínimo local abaixo do limiar, a técnica original atribui o período fundamental ao mínimo global da janela de integração: $argmin_\tau d'(\tau)$. Referindo-se a isso, ao invés de usar um único limiar fixo, Mauch e Dixon (MAUCH; DIXON, 2014) fazem um mapeamento de possíveis valores para uma faixa de limiares $\{s_i| \in \{0, .01, ..., 1\}\}$, através da distribuição de probabilidade Beta com média igual a 0.15. Nesse caso, a probabilidade de que um período τ seja o período fundamental τ_0 é definida por

$$P(\tau = \tau_0|S, x_t) = \sum_{i=1}^{N} a(s_i, \tau)P(s_i)[Y(x_t, s_i) = \tau], \tag{4.5}$$

onde $[\cdot]$ são os colchetes de Iverson, os quais denotam 1 se a expressão dentro dos colchetes for verdadeira, e 0 caso contrário. Basicamente, a Equação 4.5 estima a função de densidade de probabilidade acumulada, onde $[Y(x_i, s_i) = \tau]$ define a região de integração, incluindo os períodos τ que são capturados pelo limiar s_i. A função a é dada por

$$a(s_i, \tau) = \begin{cases} 1 & \text{se } d'(\tau) < s_i \\ p_a & \text{caso contrário,} \end{cases} \tag{4.6}$$

onde $p_a = 0.01$, é a probabilidade *a priori* do modelo, definida empiricamente pelos autores da técnica original. Esse primeiro estágio da técnica pYIN extrai, para cada um dos possíveis candidatos a período fundamental, uma medida de probabilidade associada, que é usada para identificar regiões do sinal que possuem presença de voz. O candidato a período fundamental com a máxima probabilidade é equivalente ao período de mínima diferença, obtido pela Equação 4.1 na técnica YIN. O pYIN mantém os múlti-

plos candidatos para uso num segundo estágio, onde um HMM é usado para verificar a consistência temporal das estimativas das frequências fundamentais f_0. Para isso, primeiramente os espectro de frequências é convertido para uma escala linear [1] usando a relação $F0 = 12 \times \log_2 \frac{f_0}{440Hz} + 69$, e depois discretizado em 480 *bins*, equivalente a quatro oitavas, entre 55Hz (A1) a 880Hz (A5), onde $F0$ é a frequência fundamental central de cada *bin*.

Os pares de frequências fundamentais e probabilidades associadas estimados no primeiro estágio são usados diretamente como os observáveis em cada estado do HMM. Esse segundo estágio funciona como um filtro, o qual suaviza o rastreamento das alturas das notas, especialmente em regiões de descontinuidades.

A partir desse ponto do texto, a expressão "frequência fundamental" será utilizada para indicar a altura da nota musical (*pitch*) usando a escala linear acima referida ($F0$). A Figura 4.1c ilustra a extração das frequências fundamentais utilizando o algoritmo pYIN, onde a linha vermelha apresenta a evolução temporal em quadros da frequência fundamental e a linha azul apresenta as respectivas probabilidades associadas. No contexto de análise de áudio, abordada neste capítulo, cada quadro da sequência corresponde a uma janela de avaliação do algoritmo pYIN, contendo 2048 amostras do sinal de áudio e um passo de 32 amostras.

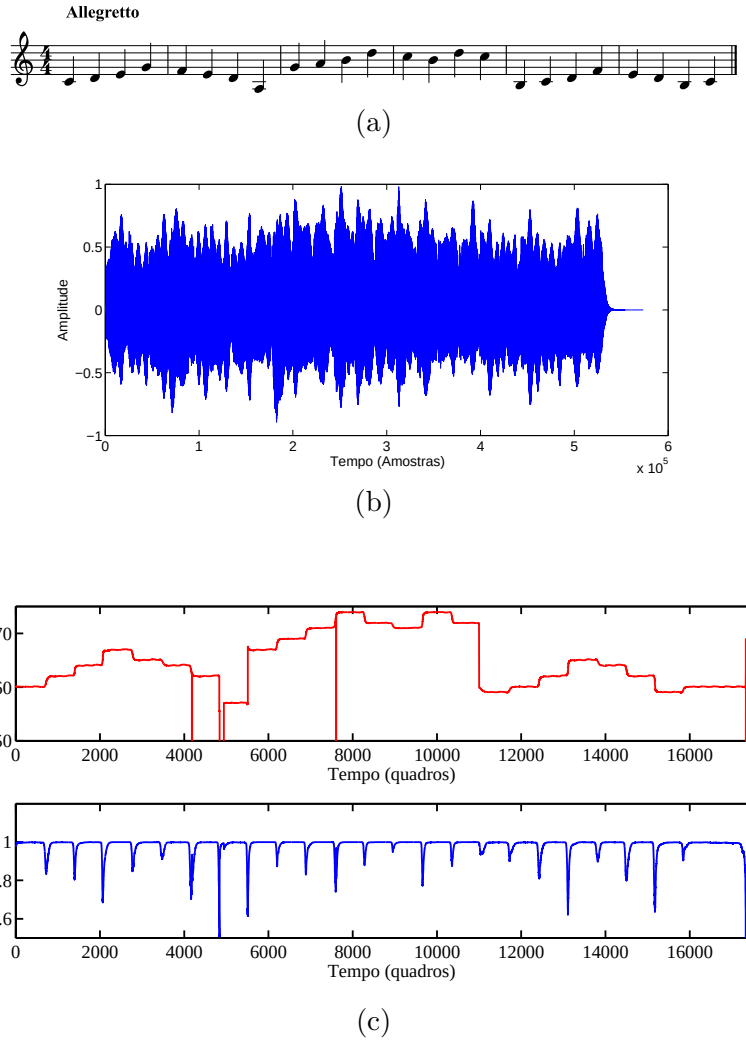
4.2 Transcrição Melódica

Para que se possa extrair métricas de avaliação de precisão rítmica e melódica da execução das notas cantadas, é preciso primeiramente obter uma representação simbólica (sequência de notas musicais com altura e duração, conforme representados na partitura) a partir da sequência de frequências fundamentais estimadas pelo processo descrito na seção anterior. Ou seja, é preciso realizar a transcrição melódica, a qual rotula os agrupamentos de frequências fundamentais em unidades atômicas, chamadas de notas musicais. Para isso, utiliza-se um método de transcrição baseado em histerese (MOLINA et al., 2015).

O algoritmo inicia o processo de rotulação das notas musicais a partir da segmentação das regiões classificadas como *voiced* pelo algoritmo pYIN. Mais precisamente, exige-se que cada segmento deve ter a duração mínima de 100ms e a probabilidade associada a cada quadro de análise contido no segmento deve ser superior a 80%.

[1] A altura do som é representada numa escala linear onde cada inteiro representa um semitom da escala temperada cromática. Essa é a mesma representação utilizada pelo protocolo MIDI (MÜLLER, 2007), onde a nota A4 (relativa a frequência 440Hz) tem seu valor igual a 69.

Figura 4.1: Extração das frequências fundamentais do sinal de áudio. (a) partitura com exemplo de exercício de solfejo. (b) sinal de áudio capturado pelo microfone. (c) frequências fundamentais (vermelho) e respectivas probabilidades (azul) para cada quadro de análise do sinal.



Fonte: O Autor

O uso da informação de probabilidade $P(\tau|x_t)$ obtida com o algoritmo pYIN evita erros de classificação em regiões que contém sinal aperiódico ou harmônicos. Além disso, a filtragem seletiva das frequências obtidas pela HMM do segundo estágio do algoritmo pYIN evita saltos indesejados de oitava provocados pela presença de harmônicos. Esse problema é evidente no algoritmo original YIN (CHEVEIGNÉ; KAWAHARA, 2002).

Ainda é preciso definir um critério para segmentar regiões continuas de voz, as quais acontecem quando há uma transição do tipo *legato* entre duas notas musicais. Nesse caso, a transição não é definida por um ponto especifico, mas sim por uma faixa de transição. O uso de um limiar fixo pode ser aplicado para definir o ponto de transição entre duas

notas adjacentes, porém perturbações de frequência durante a execução da nota podem gerar um número excessivo de segmentações.

Neste trabalho foi implementado o método proposto em (MOLINA et al., 2013), o qual usa um processo de histerese para definir o ponto de segmentação ideal, evitando a segmentação excessiva. O primeiro passo do algoritmo calcula a média dinâmica (*dynamic average*) da frequência fundamental dentro de um segmento (sequência de quadros adjacentes com probabilidade superior a 80%):

$$F0_A(l) = \frac{\sum_{k=l_0}^{l} F0(k)}{l - l_0 + 1} \tag{4.7}$$

onde l_0, com $l_0 \leq l$, é o índice do primeiro quadro do segmento. $F0_A(l)$ é a média dinâmica relativa ao quadro l, e $F0(k)$ é o pitch estimado no quadro k. À medida que o tamanho do segmento cresce, $F0(l)$ tende a ficar mais estável.

Para estimar uma possível transição entre notas musicais, o desvio acumulado do pitch, representado por $\Gamma(l)$, é estimado entre a média dinâmica $F0_A$ e a frequência fundamental $F0$ a cada quadro l. O desvio acumulado do pitch é calculado de forma recursiva, para $l > l_0$:

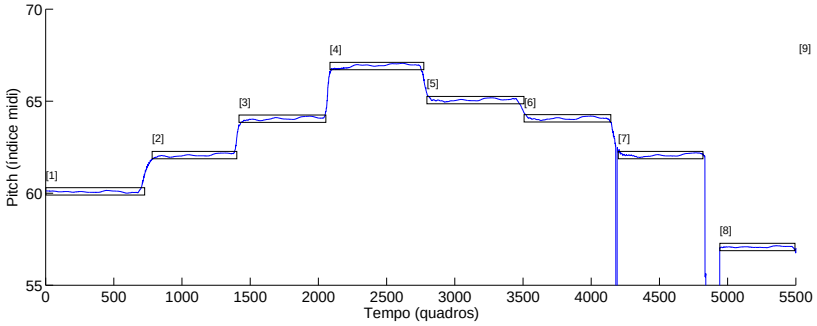
$$\Gamma(l) = \begin{cases} \Gamma(l-1) & \text{, se } \ |\delta_{F0}(l)| < \delta_{th} \\ \Gamma(l-1) + \delta_{F0}.h_s & \text{, caso contrário} \end{cases} \tag{4.8}$$

onde $\delta_{F0}(l) = F0(l) - F0_A(l)$ é o desvio instantâneo do *pitch*, o quadro inicial l_0 é definido como $\Gamma(l_0) = 0$, e h_s é o incremento (passo) temporal. δ_{th} é um limiar que pondera (maior peso) os quadros com desvios que ultrapassam seu valor. Uma nova nota musical é detectada quando $|\Gamma(l^*)| \geq \Gamma_{th}$, onde Γ_{th} é um limiar de tolerância do desvio do pitch para a nota corrente. Em outras palavras, sempre que o módulo do desvio acumulado Γ ultrapassar o limiar Γ_{th}, inicia-se uma nova nota, denotada pelo frame l^*. Nesta proposta, seguindo (MOLINA et al., 2013), os valores dos limiares foram definidos como: $\delta_{th} = 0.5$ semitons e $\Gamma_{th} = 0.1$ semitons $\times$ segundos.

A posição exata do início da nova nota é obtida como o ponto central entre os quadros l^* e l', onde o quadro l' é definido como o primeiro quadro da última região com desvio de pitch significativo (onde $|\delta_{F0}(l) \geq \delta_{th}|$) da nota que precede l^*. Sempre que uma nota é detectada, o algoritmo é reiniciado com $l_0 = l^*$, e o processo finaliza quando todos os quadros da sequência de entrada ($F0$) forem analisados. A Figura 4.2a

apresenta um exemplo da aplicação do algoritmo de segmentação por histerese, onde as notas identificadas e segmentadas são demarcadas por retângulos.

Figura 4.2: Transcrição melódica: segmentação da sequência de frequências fundamentais em notas musicais usando o processo de histerese proposto por (MOLINA et al., 2013). Os retângulos demarcam as notas identificadas (altura, ataque e duração).



(a)

Fonte: O Autor

4.3 Mapeamento da Melodia na Partitura

Para avaliar a qualidade de execução do solfejo, é necessário realizar a comparação da performance do canto com a partitura alvo, definida pelo próprio exercício. Assim, após obter a transcrição automática conforme o procedimento descrito nas seções 4.1 e 4.2, faz-se ainda necessário conectar os segmentos de notas musicais com a nota correspondente na partitura (*ground truth*).

O primeiro desafio é o fato de que a transcrição melódica gera, frequentemente, grupos de fragmentos de notas (segmentos), os quais devem ser mapeados para apenas um elemento do *ground truth* (partitura). Cada fragmento melódico é representado por f_{il}, onde i é o indice do segmento e l é o índice relativo do quadro contido no segmento.

Da mesma forma que em (MOLINA et al., 2013), não há nenhuma premissa de sincronização por metrônomo nessa abordagem. Assim, eventuais pequenos atrasos ou acelerações rítmicas do canto geram desalinhamento entre as notas transcritas e a partitura. Em (MOLINA et al., 2013), um procedimento integrado através da DTW foi empregado para realizar o alinhamento temporal dos quadros de análise. Porém, em alguns casos, a condição de fronteira do algoritmo DTW pode propagar o custo local acumulado, causando erros indesejáveis no alinhamento entre a sequência transcrita e a partitura.

Nesta seção, é proposto um novo processo para alinhamento e agrupamento dos segmentos transcritos. Além disso, neste procedimento, os grupos de segmentos resultantes são mapeados às notas correspondentes da partitura. Apesar de similar à abordagem com DTW, esse novo processo não propaga o erro uma vez que ele não precisa obedecer a condição de fronteira da DTW.

O processo conjunto de agrupamento e alinhamento foi projetado através de um algoritmo de força bruta, o qual é implementado por meio de uma matriz de custos C. Para cada nota k da partitura, o algoritmo calcula uma medida de distância cumulativa considerando todas as possibilidades de agrupamentos de segmentos adjacentes, iniciando no segmento de índice i e terminando no segmento de índice j.

Esse algoritmo é construído eficientemente usando-se do suporte de uma estrutura dados 3D, conforme ilustrado na Figura 4.3a. Assim, para cada possível combinação (k, i, j), uma medida de dissimilaridade é calculada por:

$$\begin{aligned} C(k,i,j) = \alpha_1 \Delta f(k,i,j) + \alpha_2 \Delta d(k,i,j) + \\ \alpha_3 \Delta s(k,i,j) + \alpha_4 \Delta e(k,i,j), \end{aligned} \tag{4.9}$$

onde

$$\Delta f = |f_k^{gt} - mediana(f_{i,1} ... f_{j,l_{max}})| \tag{4.10}$$

é a distância do *pitch* entre a nota k da partitura e a mediana dos valores de $F0$ pertencentes à faixa que inicia no primeiro quadro do segmento i e que termina no último quadro l_{max} do segmento j,

$$\Delta d = |D_k^{gt} - \sum_{m=i}^{j} D_m| \tag{4.11}$$

mede a diferença de duração (em segundos) entre a nota k da partitura (D_k^{gt}) e o grupo formado pelo segmento de i a j na melodia transcrita (D_i é a duração do segmento i),

$$\Delta s = |S_k^{gt} - S_i| \tag{4.12}$$

leva em consideração o atraso ou avanço (em segundos) do *onset* entre o primeiro segmento do grupo selecionado e a nota k. De forma análoga,

$$\Delta e = |E_k^{gt} - E_j| \tag{4.13}$$

leva em consideração o atraso ou avanço (em segundos) do *offset*. Os coeficientes α_i são

pesos para balancear a contribuição individual de cada medida, e nossos experimentos mostraram que $\alpha_1 = 1.0$, $\alpha_2 = 2.0$, $\alpha_3 = 2.0$, $\alpha_4 = 2.0$ são uma boa combinação.

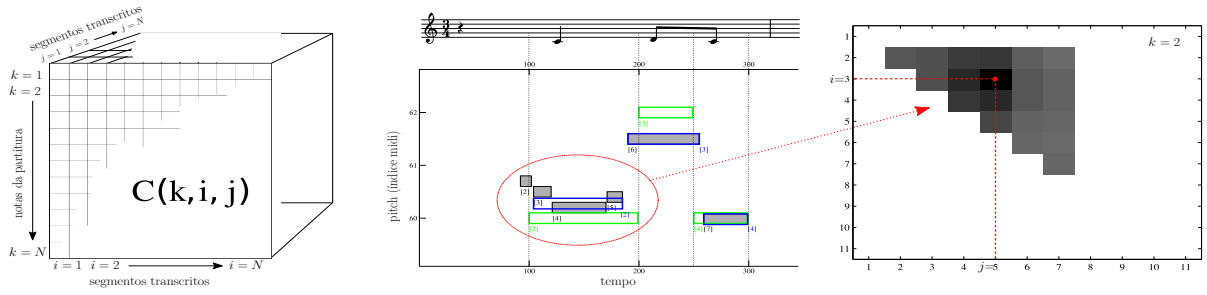
O processo de agrupamento e o respectivo mapeamento com a partitura é finalmente obtido pela função:

$$\upsilon(k) = (i_k, j_k) = \underset{i,j}{\operatorname{argmin}}\, C(k, i, j), \tag{4.14}$$

de tal forma que cada nota k na partitura é conectada a um grupo de segmentos que inicia em i (primeiro segmento do grupo) e termina em j (último segmento do grupo), obtendo-se assim a transcrição final e consolidada da nota musical cantada.

O custo computacional do processo de alinhamento no pior caso é $\mathcal{O}(MN^2)$, onde M é o número de notas musicais da partitura e N é o número de segmentos melódicos obtidos no processo da seção 4.2. Embora o custo possa ser alto se M e N forem grandes, a inclusão dos componentes Δs e Δe na Equação (4.9) causa um rápido crescimento na magnitude da medida de dissimilaridade quando segmentos estão longe da posição temporal esperada. Como consequência, é possível interromper o laço de força bruta em poucas iterações através da limitação do valor de $C(k, i, j)$. Além disso, a janela de avaliação que contém os segmentos melódicos pode ser restringida para iniciar numa posição próxima da nota alvo (partitura). Esse processo, além de reduzir o custo computacional, auxilia na redução de problemas com mínimos locais na Equação (4.14). A Figura 4.3 ilustra um exemplo do processo de agrupamento e alinhamento, onde seis segmentos são mapeados em três notas.

Figura 4.3: (a) Estrutura 3D utilizada para calcular a similaridade entre os grupos de segmentos melódicos e as notas da partitura. (b) Processo de agrupamento de diversos segmentos melódicos (cinza) em uma única nota musical (azul). (c) O melhor agrupamento para a nota k do *ground-truth* é encontrado a partir dos índices i (primeiro elemento) e j (último elemento), os quais minimizam a função $C(k, i, j)$.



Fonte: O Autor

Neste exemplo, após a construção do cubo de custos considerando todos os possíveis agrupamentos de segmentos melódicos adjacentes (Figura 4.3a), é possível visualizar um grupo formado pelos segmentos 3, 4 e 5 (em cinza), os quais estão próximos da nota 2 (em verde) da partitura, mostrados na Figura 4.3b. Esse grupo possui o menor custo associado, o qual é indicado pelo mínimo global na matriz exibida na Figura 4.3c, relativa à nota 2 da partitura. Os índices i e j (linha e coluna), indicam a posição desse mínimo local e, por consequência, definem na Figura 4.3b o segmento inicial e final do grupo utilizado para gerar a nota transcrita final (retângulo azul de índice 2) que será associada à nota musical na posição 2 da partitura. Vale a pena citar que o segmento 2 (cinza), gerado por um ataque impreciso da nota cantada, é descartado automaticamente pelo processo.

4.4 Avaliação nota-a-nota

A avaliação da execução melódica segue a mesma lógica utilizada na avaliação do movimento. Um processo de classificação é aplicado para definir se a execução de cada nota está correta, seguindo um modelo gerado a partir da avaliação previa por um comitê de especialistas (professores de música). Neste trabalho, diferentemente do procedimento adotado por Molina et al. (2013), onde os avaliadores pontuaram trechos completos de execução vocal, um processo de avaliação nota por nota foi empregado para avaliar a execução do solfejo. Sendo assim, foram extraídas, para cada nota musical transcrita, medidas de erro (distância) que contemplam tanto a informação de altura, quanto a informação temporal.

Para cada par de notas conectadas (uma proveniente da melodia cantada, e a outra da partitura), são extraídas três medidas de distância:

$$\begin{aligned} \Delta s_i &= |s_i - \hat{s}_i|, \\ \Delta e_i &= |e_i - \hat{e}_i|, \\ \Delta f_i &= |f_i - \hat{f}_i|, \end{aligned} \tag{4.15}$$

onde s_i e $\hat{s}_i$ representam o instante inicial (*onset*) da nota cantada e da nota relacionada na partitura, respectivamente. Analogamente, e_i e $\hat{e}_i$ representam os instantes finais (*offset*) e, f_i e $\hat{f}_i$, representam as alturas (*pitch*) das notas comparadas.

No caso ideal, todas essas distâncias devem ser nulas. Entretanto, os especialistas

toleram pequenas variações em cada um desses atributos, que são modeladas e consideradas no sistema de avaliação proposto.

Para realizar o treinamento do modelo, uma base de dados com diversos exemplos de gravações de solfejo foi gerada. Maiores detalhes sobre a aquisição desses dados, bem como o processo de anotação por especialistas são apresentados na Seção 4.5.1. Utilizando essa base de dados numa fase inicial de treinamento, distintas distribuições de densidade de probabilidade são modeladas para representar as notas cantadas corretas, bem como para as notas cantadas de forma incorreta, considerando individualmente os desvios *pitch* (Δf em semitons), *onset* (Δs em segundos) e *offset* (Δe em segundos). Durante a fase de classificação, para cada nota cantada, um classificador Bayesiano atribui o label correto φ ou incorreto $\overline{\varphi}$ para cada um dos parametros *pitch*, *onset* and *offset*. A seguir, o processo de classificação Bayesiana será explicado, focando apenas no parâmetro Δf. Contudo, ressalta-se aqui que o mesmo processo de classificação é também aplicado, individualmente, aos parâmetros Δs e Δe.

A Figura 4.4a apresenta os histogramas dos desvios Δf obtidos após análise do parâmetro pitch, considerando a avaliação dos especialistas para as categorias correta e incorreta, denotados por $\varphi_{\Delta f}$ e $\overline{\varphi}_{\Delta f}$, respectivamente. Como pode ser observado, o histograma de $\varphi_{\Delta f}$ apresenta um pico proeminente perto da origem (relacionado a erros pequenos no pitch), como esperado. Apesar disso, as duas classes apresentam considerável sobreposição, corroborando para as discrepâncias obtidas na avaliação da acurácia pelos especialistas, especialmente para erros intermediários do *pitch*. De fato, uma vez que foram utilizadas pontuações individuais de cada avaliador para cada nota para construir os histogramas, o desvio do *pitch* Δf, relacionado à nota que recebeu rótulos conflitantes dos avaliadores, contribui para ambos histogramas de $\varphi_{\Delta f}$ e $\overline{\varphi}_{\Delta f}$.

Uma função de densidade de probabilidade é então estimada a partir das distribuições de Δf para cada classe $r \in \{\varphi_{\Delta f}, \overline{\varphi}_{\Delta f}\}$, tal que a probabilidade a posteriori (a qual pode ser considerada uma medida de confiança) pode ser facilmente obtida. Entre diferentes funções de densidade de probabilidade paramétricas (PDFs) para modelagem de variáveis aleatórias positivas, a distribuição Gamma foi escolhida, pois o problema é similar à modelagem de dados utilizados na avaliação do gesto (vide seção 3.3), onde a distribuição dos dados possui características como unimodalidade e formato assimétrico.

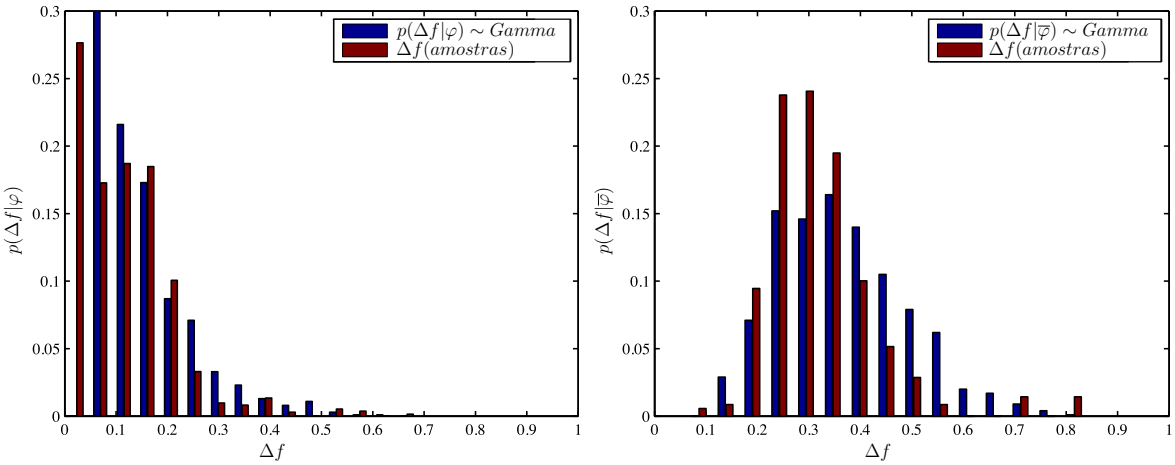
A partir das PDFs $p(\Delta f|\varphi)$ e $p(\Delta f|\overline{\varphi})$, é possível estimar a probabilidade *a posteriori* do *pitch* da nota cantada pertencer a cada uma das possíveis classes: correto/incorreto.

Pela regra de Bayes (DUDA; HART; STORK, 2001), tem-se:

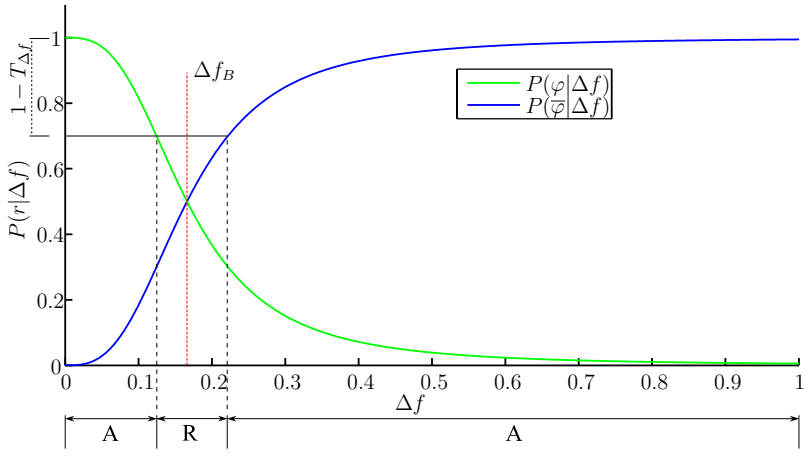
$$p(r|\Delta f) = \frac{p(\Delta f|r)P(r)}{p(\Delta f)}, \tag{4.16}$$

onde $p(\Delta f) = p(\Delta f|\varphi)P(\varphi) + p(\Delta f|\overline{\varphi})P(\overline{\varphi})$ é a distribuição total de Δf, e as probabilidades *a priori* $P(\varphi)$ e $P(\overline{\varphi})$ são definidas como equiprováveis.

Figura 4.4: (a) Histograma de Δf para as classes φ e $\overline{\varphi}$ com as respectivas Gamma PDFs ajustadas. (b) Probabilidade posterior e as respectivas regiões de aceitação e rejeição.



(a)



(b)

Fonte: O Autor

A Figura 4.4b ilustra a fronteira de decisão para $\varphi_{\Delta f}$ e $\overline{\varphi}_{\Delta f}$ como uma linha vertical pontilhada e vermelha. Em torno da fronteira de decisão há uma região nebulosa, contendo uma considerável sobreposição entre $p(\Delta f|\varphi)$ e $p(\Delta f|\varphi)$, onde a probabilidade

a posteriori da classe vencedora é apenas pouco maior que 0.5. Uma vez que essa região de sobreposição é causada em parte pelos rótulos conflitantes entre os avaliadores especialistas, um opção apropriada é a rejeição das amostras que que caem dentro dessa região nebulosa (ou seja, essas amostras não são classificadas nem como corretas nem como incorretas). Semelhante à abordagem adotada na seção 3.5, os erros (ou enganos de classificação) são convertidos em rejeições utilizando-se a regra de Bayes para erro mínimo (WEBB, 2011). A regra de rejeição divide o espaço dos exemplos em região de aceitação A e região de rejeição R, que são definidas por:

$$R(T_{\Delta f}) = \{\Delta f | 1 - \max_r p(r|\Delta f) > T_{\Delta f}\}, \tag{4.17}$$

$$A(T_{\Delta f}) = \{\Delta f | 1 - \max_r p(r|\Delta f) \leq T_{\Delta f}\}, \tag{4.18}$$

onde o limiar $T_{\Delta f}$ faz o balanço (*tradeoff*) entre o número de amostras rejeitadas e a taxa de erro $e(T_{\Delta f})$, estimada por:

$$e(T_{\Delta f}) = \sum_{\Delta f \in A(T_{\Delta f})} \Big(1 - \max_r p(r|\Delta f)\Big)\, p(\Delta f). \tag{4.19}$$

A escolha do limiar $T_{\Delta f} = 0.33$ foi determinada a partir de um conjunto de experimentos onde a acurácia de classificação e o número de rejeições foram levados em conta (mais detalhes sobre a escolha do limiar são apresentados na seção 4.5.2). As probabilidades *a posteriori* e as fronteiras de decisão entre as regiões A e R geradas por este limiar são mostradas na Figura 4.4b, abaixo do eixo horizontal.

Assim, considerando a acurácia do pitch, e usando o classificador Bayesian definido na equação (4.16) combinado com o procedimento de rejeição dado por (4.17) e (4.18), cada nota cantada é classificada em três possíveis classes: correto, incorreto, ou indeterminado (rejeitado). Quando a classificação é realizada (correto ou incorreto), a probabilidade correspondente também pode ser utilizada para prover uma medida de confiança sobre a performance do usuário (aluno).

Todo esse processo de classificação é também executado, independentemente, para avaliar a acurácia do *onset* e *offset* da nota cantada. Assim, para cada nota cantada, o sistema gera uma resposta que provê uma medida de confiança e um rótulo de classe, individualmente, para o *pitch*, *onset* e *offset*.

4.5 Experimentos e Resultados Relativos às Notas Musicais Cantadas

Para avaliar a abordagem proposta, gravações de exercícios de solfejo foram realizadas em diversas tentativas para gerar um conjunto representativo de dados, os quais permitissem a obtenção de estatísticas de performance e estudo da viabilidade da técnica proposta. A validação dessas gravações pelos especialistas é um processo bastante custoso, necessitando diversas repetições dos exemplos e muita atenção na avaliação individual de cada nota musical.

4.5.1 Construção da Base de Dados Relativa à Entonação (Somente Áudio)

Essa base de dados consiste de sequências de intervalos da escala musical cromática. As gravações dessa base de dados foram feitas com sete adultos, incluindo músicos treinados (três) e músicos amadores ou iniciantes (quatro), numa faixa etária variando entre 17 e 61 anos. Essas sequências melódicas foram gravadas durante quatro meses, em formato monofônico, taxa de amostragem de $44100Hz$, 16 bits de quantização e diferentes tipos de microfones.

Foi decidido dar suporte à performance do canto através de um faixa (*track*) auxiliar contendo os intervalos em timbre de piano, uma vez que parte dos cantores não sabiam ler partitura. Nesta faixa auxiliar (referência), os intervalos foram tocados em sequência, contendo pausas (silêncios) entre eles. Cada cantor preencheu esses espaços repetindo (cantando) o intervalo melódico previamente tocado, gravando-o no instante temporal exato do primeiro tempo do compasso subsequente. Todas as gravações foram sincronizadas por metrônomo.

Foi sugerido aos cantores usarem diferentes tipos de fonemas. Também foi solicitado a eles para cantarem livremente, porém, respeitando o *pitch*, o ataque (*onset*) e a duração (*offset*) dos sons previamente indicados. O objetivo desse procedimento foi capturar exemplos espontâneos do dia-a-dia de um cantor. Intencionalmente, buscando capturar uma maior variedade de situações naturais, as gravações foram conduzidas em dois ambientes distintos: uma parte das gravações foram realizadas em estúdio, onde o resultado do sinal de áudio é limpo; outra parte das gravações foram feitas em condições informais, apresentando ruído de fundo e reverberação.

Um total de 21 seções foram gravadas, contendo doze intervalos ascendentes e doze intervalos descendentes da escala cromática. Cada cantor realizou os intervalos

melódicos em três andamentos distintos: *Adagio*, 60 BPM; *Andante Moderato*, 90 BPM; e *Allegro*, 120 BPM. Além das gravações, um comitê de especialistas conduziu um processo de anotação para rotular cada nota gravada. O comitê foi composto por cinco músicos graduados, com mais de dez anos de experiência em audições de exercícios de teoria e percepção. Cada nota musical cantada e gravada foi rotulada em correto ou incorreto, considerando individualmente a acurácia do *pitch*, *onset* e *offset*.

Antes de cada seção de anotação, o comitê foi orientado a ouvir alguns exemplos de origem aleatória da base de dados gravada. Esse procedimento de "aquecimento" foi importante, pois ele ajudou a criar uma noção de concordância entre os especialistas, os quais compartilhavam opiniões sobre as características e aspectos das melodias gravadas. A base de dados foi dividida em partes e o processo de avaliação durou diversos dias, até que toda o conjunto de gravações fora avaliado (de fato, o processo completo para construir a base de dados anotada levou seis meses).

Para cada nota cantada na base de dados, todos os cinco avaliadores atribuíram um voto (correto ou incorreto), considerando cada parâmetro analisado (*pitch, onset and offset*). Discordâncias entre avaliadores foram mantidas e usadas para modelar o classificador probabilístico. Além disso, cada nota teve um rótulo global associado (correto ou incorreto) para cada parâmetro, baseando-se na maioria dos votos atribuídos pelos especialistas (isto é, pelo menos três votos para o mesmo rótulo). Assim, alguns rótulos (classificações pelos especialistas) podem ser considerados mais confiáveis, pois tiveram um número maior de votos em concordância. Por exemplo, considerando o *pitch*, 15.38% dos exemplos receberam três votos em concordância, o que significa um expressivo grau de dúvida entre os especialistas. A mesma análise foi feita para os parâmetros *onset* e *offset*, e a percentagem de notas com três votos (alto grau de dúvida) foi 10.71% e 12.09%, respectivamente. A base de dados final contém 3276 exemplos (notas) rotulados. A partir desse momento, essa base de dados será referenciada como DATASET_1.

4.5.2 Testes Quantitativos com o Sistema de Avaliação Nota-a-Nota.

Visando realizar uma análise objetiva do sistema proposto para avaliação automática de solfejo, um conjunto de experimentos foram conduzidos usando a base de dados DATASET_1. Para cada um dos exemplos gravados nessa base de dados, foi realizada a transcrição e o mapeamento dos segmentos melódicos com as notas da partitura do respectivo exercício, conforme descrito nas seções 4.2 e 4.3.

As distâncias do *pitch*, *onset* e *offset* (Δf_i, Δs_i e Δe_i) de cada nota musical i foi computada a partir da comparação entre a partitura e a melodia alinhada (Eq. 4.15). As distâncias calculadas foram então utilizadas num processo de validação cruzada, onde parte dos dados foi utilizada para estimar os parâmetros das distribuições de densidade de probabilidade Gamma, e a parte restante foi utilizada para testar o modelo. Mais precisamente, foi utilizado um esquema de validação cruzada com 10 dobras. Isto significa que a base de dados foi dividida aleatoriamente em dez partes iguais, e para cada rodada da validação cruzada, 9 dobras foram usadas para treinar o modelo probabilístico, e a dobra restante foi utilizada para validar o classificador Bayesiano descrito na Seção 4.4. Nestes experimentos, foi testado o classificador Bayesiano com e sem a regra de rejeição. Em ambas as situações, o sistema efetuou a classificação de cada parâmetro (pitch, onset, offset) de cada nota cantada em duas possíveis categorias: correto ou incorreto. Quando a opção da regra de rejeição foi aplicada, algumas notas foram mantidas como "não classificadas".

A Tabela 4.1 apresenta as matrizes de confusão geradas pelos classificadores Bayesianos para o pitch, onset e offset sem a regra de rejeição. Neste caso, a acurácia é maior do que 90% para os três parâmetros analisados. É possível perceber também que o sistema tende a produzir mais falso negativos (isto é, marca como incorreta uma nota cantada corretamente) do que falso positivos, particularmente para o parâmetro offset, sendo assim um avaliador "rígido". Os erros de classificação são causados por duas razões principais: primeiro, um possível erro na transcrição ou no mapeamento dos segmentos melódicos pode introduzir erros nas medidas de similaridades; segundo, desacordo entre os avaliadores humanos geram uma região nebulosa próximo da fronteira de decisão. De fato, como notado na seção 4.5.1, 10 a 15% das notas anotadas apresentam forte desacordo entre os avaliadores, tal que o conjunto verdade (rótulos) podem não ser confiáveis.

Tabela 4.1: Avaliação da abordagem proposta usando validação cruzada com 10 dobras e sem a regra de rejeição de Bayes.

Target Class \ Output Class	$\varphi_{\Delta f}$	$\overline{\varphi}_{\Delta f}$	
$\varphi_{\Delta f}$	**88.99%**	11.01%	
$\overline{\varphi}_{\Delta f}$	7.27%	**92.73%**	
			90.86%

(a) Pitch evaluation

Target Class \ Output Class	$\varphi_{\Delta s}$	$\overline{\varphi}_{\Delta s}$	
$\varphi_{\Delta s}$	**89.17%**	10.83%	
$\overline{\varphi}_{\Delta s}$	8.74%	**91.26%**	
			90.22%

(b) Onset evaluation

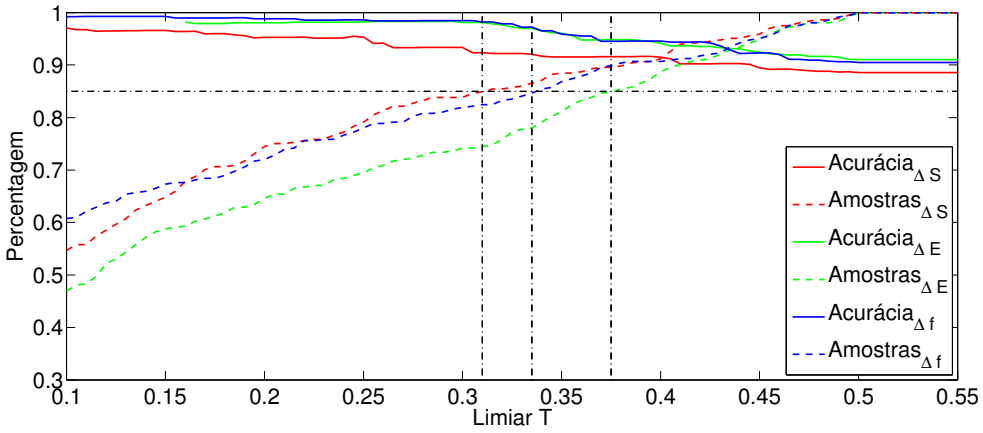
Target Class \ Output Class	$\varphi_{\Delta e}$	$\overline{\varphi}_{\Delta e}$	
$\varphi_{\Delta e}$	**84.71%**	15.29%	
$\overline{\varphi}_{\Delta e}$	2.54%	**97.46%**	
			91.08%

(c) Offset evaluation

A regra de rejeição definida na Equação (4.17) evita a classificação das amostras que potencialmente caem dentro da região nebulosa. A Figura 4.5 ilustra o efeito causado

ao variar os limiares de rejeição, tanto na percentagem de amostras aceitas quanto na acurácia do pitch, onset e offset. Como esperado, limiares baixos reduzem o número de amostras aceitas e aumentam a acurácia.

Figura 4.5: Comparação da acurácia versus o número de amostras não rejeitadas. Linhas sólidas mostram a evolução da acurácia, as quais são afetadas pelos limiares $T_{\Delta f}$ (pitch), $T_{\Delta s}$ (onset), e $T_{\Delta e}$ (offset).



Fonte: O Autor

Há uma grande dificuldade em definir um valor ótimo para o limiar de rejeição, pois é preciso escolhê-lo de forma a obter a máxima acurácia possível enquanto o número de de amostras rejeitadas precisa ser mínima. Como o foco deste trabalho está na educação musical, acredita-se que é preferível não ter uma resposta do que dar um feedback incorreto. Baseado nessa premissa, e também considerando que a percentagem de amostras com desacordo entre os avaliadores é acima de 10%, foi decidido configurar os limiares para rejeitar, na média, 15% das amostras.

A Tabela 4.2 apresenta a avaliação da acurácia para experimento com 10 dobras, usando o classificador Bayesiano com a regra de rejeição. Como pode ser observado, de forma geral, a acurácia para todos os parâmetros analisados aumentou de 3 a 5% quando comparado com a opção sem regra de rejeição, atingindo uma acurácia média de aproximadamente 94.6%.

Além disso, o número de falsos negativos foi consideravelmente reduzido, particularmente na avaliação do parâmetro offset. Esse fato indica que, quando em dúvida, os avaliadores tendem a marcar uma nota como correta ao invés de incorreta. Mais ainda, 32–35% das amostras rejeitadas receberam dos especialistas 3 votos em concordância, o que significa que o sistema proposto remove mais do que o dobro das amostras relacionadas à “dúvida” dos especialistas, quando comparado toda a base de dados.

Tabela 4.2: Avaliação da abordagem proposta usando validação cruzada com 10 dobras e com a regra de rejeição de Bayes. O sistema é capaz de responder em 90% das vezes, aumentando a acurácia final em aproximadamente 3%.

Target Class \ Output Class	$\varphi_{\Delta f}$	$\overline{\varphi}_{\Delta f}$	
$\varphi_{\Delta f}$	**94.45%**	5.55%	
$\overline{\varphi}_{\Delta f}$	2.54%	**97.46%**	
			95.96%

(a) Pitch evaluation: $T_{\Delta f} = 0.33$

Target Class \ Output Class	$\varphi_{\Delta s}$	$\overline{\varphi}_{\Delta s}$	
$\varphi_{\Delta s}$	**94.17%**	5.83%	
$\overline{\varphi}_{\Delta s}$	7.34%	**92.66%**	
			93.42%

(b) Onset evaluation: $T_{\Delta s} = 0.31$

Target Class \ Output Class	$\varphi_{\Delta e}$	$\overline{\varphi}_{\Delta e}$	
$\varphi_{\Delta e}$	**91.64%**	8.36%	
$\overline{\varphi}_{\Delta e}$	2.54%	**97.46%**	
			94.55%

(c) Offset evaluation: $T_{\Delta e} = 0.38$

5 INTEGRAÇÃO AUDIOVISUAL PARA AVALIAÇÃO DE SOLFEJO

O exercício de leitura e canto do que está escrito na partitura pode ser executado com o auxílio do metrônomo. Nesse caso, a captura do sinal está sincronizada com as unidades de tempo do metrônomo, não sendo necessário nenhum tipo de alinhamento entre as sequências. Todavia, a comparação individual das notas musicais não é possível quando o exercício é realizado com oscilações no fluxo regular das unidades de tempo, isso é, tanto desencontrado das batidas do metrônomo, como em andamentos diferentes do proposto pela partitura. Essa situação é a real e, de certo modo, desejável quando o músico interpreta o que está escrito. Todavia, em termos absolutos, resulta em incongruências temporais. Tais "erros" ocorrem porque os instantes iniciais e finais dos sons entoados não coincidem com os instantes iniciais e finais de cada batida do metrônomo, e nem com a representação desses sons musicais conforme definidos na partitura. Isso significa que, mesmo que o aluno cante as alturas e durações proporcionais corretamente caso as execute com variações de andamento (em termos musicais, deformações expressivas dos pulsos individuais), o sistema não será capaz de comparar nota contra nota.

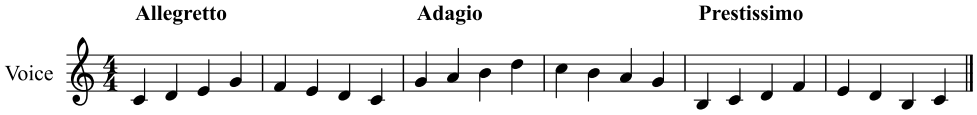
Na execução do solfejo completo, isto é, canto com marcação de compassos, os instantes iniciais e finais das notas musicais cantadas devem estar coerentes com a evolução temporal da partitura, que é sincronizada pelo movimento da mão. As próximas seções deste capítulo descrevem esse processo de sincronização do fluxo temporal entre canto, partitura e marcação de compassos, o qual reduz o problema ao mesmo caso do capítulo anterior, permitindo a aplicação do sistema de avaliação nota-a-nota descrito na Seção 4.4.

5.1 Retificação Temporal da Melodia

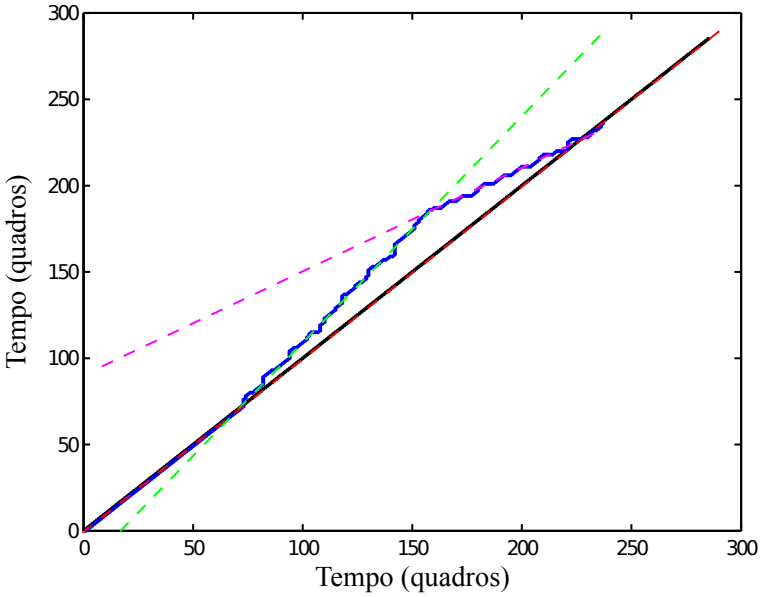
Molina et al. (MOLINA et al., 2013) utilizam o caminho de alinhamento da DTW para estimar o andamento da execução do solfejo. Porém, as medidas de similaridade utilizadas para a avaliação do solfejo nessa técnica não permitem variações intencionais no andamento. Isso ocorre pois tais variações de andamento ao longo da execução do exercício, as quais podem ser reflexo da performance expressiva do solfejo, geram um caminho de alinhamento não linear na DTW. Assim, o ajuste de reta sobre o caminho de alinhamento da DTW proposto em (MOLINA et al., 2013) não funciona adequadamente. A Figura 5.1b ilustra a vulnerabilidade do ajuste da reta em casos onde há variação do andamento ao longo da execução da melodia. As linhas tracejadas ilustram três possíveis

andamentos distintos, executados ao longo do solfejo.

Figura 5.1: (a) Partitura com exemplo de exercício de solfejo. (b) Linha azul ilustra a variação do andamento ao longo da execução da melodia cantada. As linhas tracejadas identificam os três possíveis andamentos.



(a)



(b)

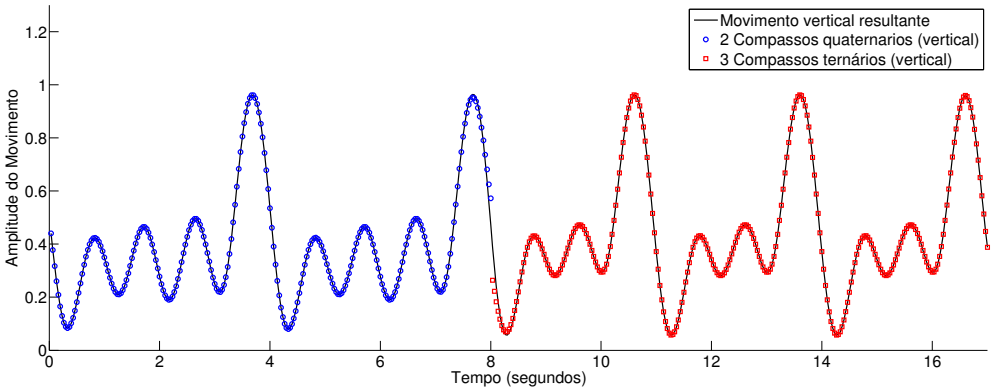
Fonte: O Autor

Para atender essas a variações circunstanciais de andamento, fenômeno comum quando o músico lê a partitura e a executa em condições reais, sem que isso implique erro de execução, foi proposto usar o alinhamento do gestual de marcação de compassos para retificar a posição temporal e a duração das notas musicais extraídas a partir da melodia entoada. Dessa forma, o conceito de solfejo, como é entendido nessa proposta, pode ser utilizado integralmente para avaliar iniciantes durante a execução de exercícios de leitura e canto da partitura. Ao invés de estimar o andamento a partir da transcrição melódica pareada a um metrônomo, a abordagem audiovisual proposta neste trabalho utiliza a marcação de compasso dada pelo movimento da mão como referência temporal, ou seja, que o metrônomo é representado pelo movimento da mão. Assim, para avaliar o exercício de solfejo, propõe-se uma fusão entre as técnicas descritas nos capítulos 3 e 4.

A partir da partitura do exercício de solfejo é possível geral uma sequência de

modelos do gesto de marcação de compasso. Usando os modelos (*templates*) de compasso obtidos na fase de treinamento do algoritmo de classificação de gesto (vide Figura 3.2), gera-se uma sequência que será comparada com a execução do exercício de solfejo. Por exemplo, se o exercício contém cinco compassos musicais, sendo os dois primeiros do tipo quaternário e os três últimos do tipo ternário, então espera-se uma sequência de movimento da mão conforme ilustrada na Figura 5.2.

Figura 5.2: Sequência de gestos usada como referência (ground truth). O gráfico apresenta uma concatenação de dois compassos quaternários seguidos de três compassos ternários.



Fonte: O Autor

O alinhamento obtido pela DTW nos permite extrair a variação de andamento do exercício, bem como estimar se a execução do movimento foi feita corretamente. É importante salientar que, diferente da abordagem utilizada na seção 3.5, onde o modelo de gestual era sincronizado com um metrônomo externo visando estimar a precisão temporal da performance do aluno, nesta seção presume-se que o aluno tem o poder de controlar o tempo. Assim, não há uma avaliação da precisão rítmica do movimento da mão.

Ao aplicar-se os atrasos e acelerações obtidos do caminho de alinhamento da DTW, pode-se "retificar" o andamento da partitura. Esse processo atualiza as posições temporais das notas musicais (*onset* e *offset*), permitindo uma comparação direta entre cada nota obtida pela transcrição melódica conectada a seu respectivo par na partitura pelo algoritmo descrito na seção 4.3.

Neste ponto, é importante mencionar que a utilização do algoritmo DTW tradicional (clássico) não é adequada para avaliar a sequência de compassos com variação de andamento. Como a sequência de compassos não é sincronizada por um metrônomo externo fixo (de fato, o próprio movimento da mão atua como ação reguladora de um metrônomo não-fixo), é preciso descartar eventuais quadros do início e do final da execução,

os quais geralmente contêm informação espúria (ex. movimentos aleatórios no início e no final do exercício). Mais ainda, como o andamento do exercício pode variar, não é possível fazer uma predição sobre a duração de gravação, fato que inevitavelmente introduzirá quadros com informação aleatória no final da sequência.

Para resolver esse impasse, aplicou-se uma modificação da DTW, conhecida como *Subsequence DTW* (MÜLLER, 2007). O algoritmo é construído de forma similar ao DTW clássico, porém com algumas pequenas modificações que efetuam a poda dos quadros no início e no final da sequência a ser comparada, buscando sempre o alinhamento ótimo. A seção a seguir descreve o algoritmo *Subsequence DTW* no contexto deste trabalho.

5.1.1 Subsequence Dynamic Time Warping

A sequência de pontos obtidos pelo rastreamento da mão é representada pela variável $V = \{\boldsymbol{P}(1), ..., \boldsymbol{P}(M)\}$, onde $\boldsymbol{P}(t) = (x(t), y(t))$ é cada ponto 2D do movimento da mão no instante temporal t. Os modelos de gestos (Eq 3.2), denotados por $\hat{M}^r = \{\hat{\boldsymbol{P}}^r_\alpha(1), ..., \hat{\boldsymbol{P}}^r_\alpha(N)\}$, são concatenados conforme a sequência de compassos da partitura, gerando um novo modelo: $\hat{Z} = \{\hat{M}^{r_i}_1, ..., \hat{M}^{r_i}_K\}$, onde r_i indica o tipo de compasso e K é o número total de compassos da partitura. A sequência $\hat{Z}$ é então redimensionada (reamostrada com interpolação cúbica) de acordo com o andamento especificado na partitura, procedimento que garante uma relação "um pra um" entre os elementos das duas sequências quando o andamento (evolução temporal) musical for igual para ambas. Em outras palavras, no caso trivial onde ambas sequências tem mesmo comprimento e não há variação temporal, cada elemento da sequência V corresponde a um único elemento na sequência $\hat{Z}$.

A sequência V (movimento rastreado) é alinhada globalmente com a sequência $\hat{Z}$ (modelos concatenados conforme os compassos da partitura). Para isso, é utilizado o algoritmo *Subsequence* DTW (MÜLLER, 2007), que busca o melhor alinhamento entre as sequências $V = \{v_1, v_2, ..., v_N\}$ e $\hat{Z} = \{z_1, z_2, ...z_M\}$, assumindo-se que $N > M$. O objetivo é encontrar uma subsequência em $V(a^* : b^*) = \{v_{a^*}, v_{a^*+1}, ..., v_{b^*}\}$, onde $1 \leq a^* \leq b^* \leq N$, e que minimiza a distância DTW em relação a $\hat{Z}$, considerando todas as possíveis subsequências de V:

$$(a^*, b^*) = \arg \min_{(a,b)|1 \leq a \leq b \leq N} DTW(\hat{Z}, V(a : b)). \tag{5.1}$$

A *Subsequence DTW* foi implementada utilizando a mesma medida de custo local definida na equação (3.6), e o caminho de alinhamento obtido pela *Subsequence* DTW é então aplicado para retificar as notas musicais da partitura. É possível atualizar os instantes temporais das notas musicais da partitura $\hat{s}_i$ (onset) e $\hat{e}_i$ (offset), usando a relação:

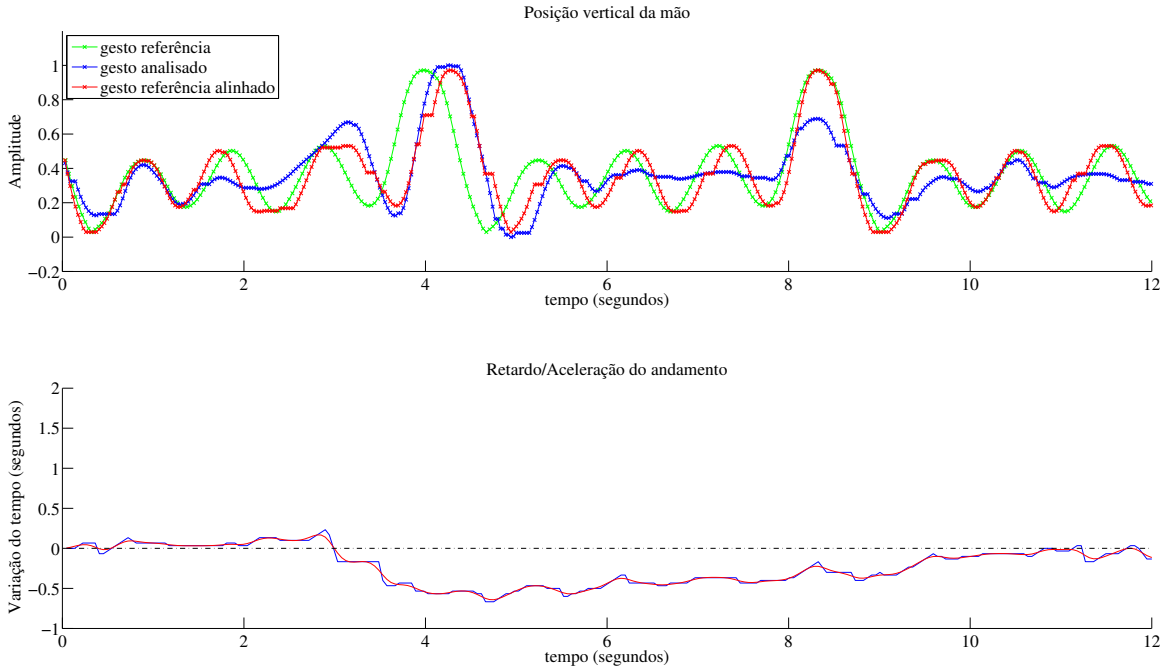
$$\begin{aligned} \hat{s}_i^* &= \hat{s}_i + Q(\hat{s}_i), \\ \hat{e}_i^* &= \hat{e}_i + Q(\hat{e}_i), \end{aligned} \tag{5.2}$$

onde

$$Q(\ell) = p_{\ell,1} - p_{\ell,2} \tag{5.3}$$

é uma função que retorna a variação temporal (em segundos) obtida na sequência de alinhamento p (*warping path*) gerada pela DTW. O gráfico da Figura 5.3 ilustra a variação do andamento ao longo do tempo em um exemplo gravado de uma performance de solfejo.

Figura 5.3: Alinhamento temporal da marcação de compassos obtido pelo algoritmo *Subsequence DTW*.



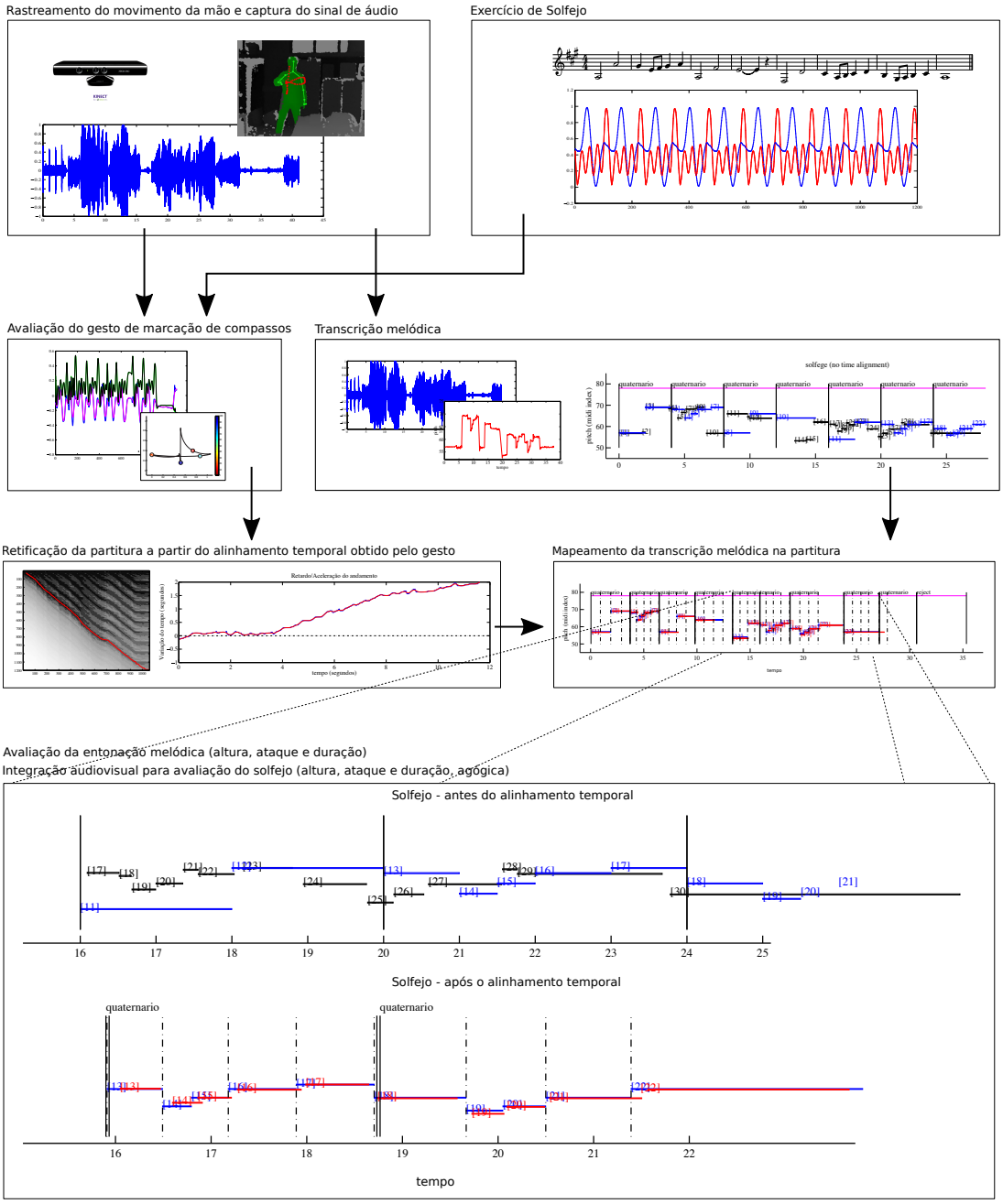
Fonte: O Autor

Após a retificação temporal da partitura, aplica-se o procedimento padrão descrito na Seção 4.3 para realizar o mapeamento dos segmentos melódicos nas notas da partitura. Finalmente, as notas cantadas, transcritas e alinhadas com a partitura são avaliadas pelo

classificador descrito na Seção 4.4.

A Figura 5.4 apresenta uma síntese, em forma de diagrama de blocos, do algoritmo audiovisual de avaliação de solfejo, o qual segue as mesmas sete etapas já apresentadas na introdução deste trabalho (Seção 1.5). Como resultado dos sete passos deste processo, obtém-se a classificação tanto do movimento de marcação de compasso, quanto a classificação das notas musicais cantadas.

Figura 5.4: Diagrama de blocos do algoritmo para avaliação audiovisual de solfejo.



Fonte: O Autor

5.2 Implementação do Sistema Computacional Proposto

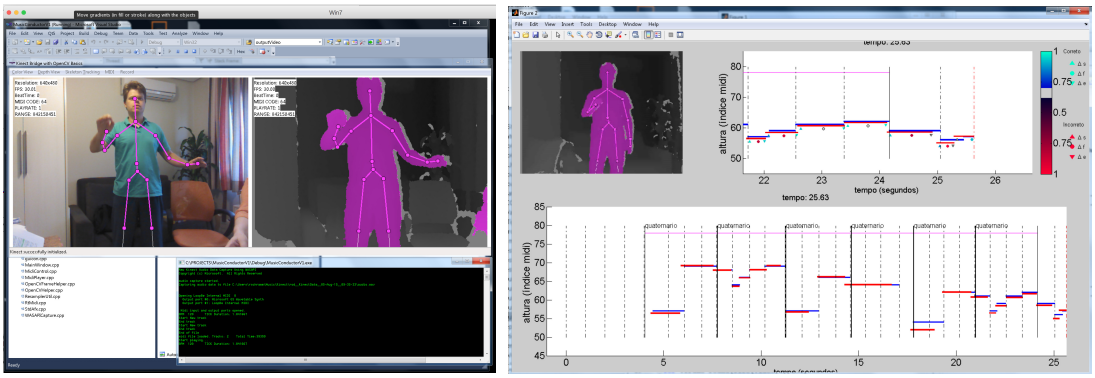
O protótipo do sistema proposto foi desenvolvido em dois módulos, e todos os experimentos foram executados em uma máquina MacBook Pro, Processador Intel Core i7 e memória DDR3 de 1600Mhz. O primeiro módulo é responsável pela captura dos sinais de áudio e vídeo e sua implementação foi feita na linguagem C++, utilizando bibliotecas padrão do sistema operacional Microsoft Windows 7 e a API de desenvolvimento do sensor Microsoft Kinect. A captura e rastreamento do esqueleto, da imagem RGB e do sinal de áudio rodam em tempo real, a 30 quadros por segundo. A Figura 5.5a apresenta a captura de tela com esse módulo em funcionamento. O segundo módulo foi desenvolvido na linguagem MATLAB, e é responsável pela classificação do movimento, transcrição melódica, retificação temporal e classificação final do solfejo. A comunicação entre essas duas partes do sistema (C++ e MATLAB) é feita através da troca de mensagens MIDI e sistema de arquivos compartilhado. O uso de memória pelo processo completo, incluindo o uso de memória necessário pela própria instância do MATLAB é, em média, aproximadamente 1GB.

Após o processamento do exercício gravado, uma rotina em MATLAB exibe na tela o resultado da avaliação do solfejo. A Figura 5.5b ilustra um exemplo de *feedback* gerado pelo sistema audiovisual integrado. A Figura 5.5c apresenta uma versão ampliada do quadro superior direito mostrado na Figura 5.5b. A figura mostra as notas especificadas na partitura (exibidas em barras horizontais azuis) e as notas cantadas detectadas e alinhadas a partir do sinal de áudio (exibidas em barras horizontais vermelhas). Além dessas barras, uma legenda também é utilizada como *feedback* ao aluno. Triângulo apontado para cima indica o ataque da nota (*onset*) e triângulo apontado para baixo indica o final da nota (*offset*), enquanto que círculo indica a afinação. As cores dessa legenda segue a especificação dada pela barra vertical ao lado direito da figura: as regiões em azul e em vermelho apresentam o grau de confiança do sistema na avaliação de cada propriedade da nota como correto e incorreto, respectivamente. Marcações em cinza definem regiões nebulosas de classificação, onde o sistema rejeita a amostra, indicando imprecisão da execução do solfejo ou indecisão do sistema na avaliação.

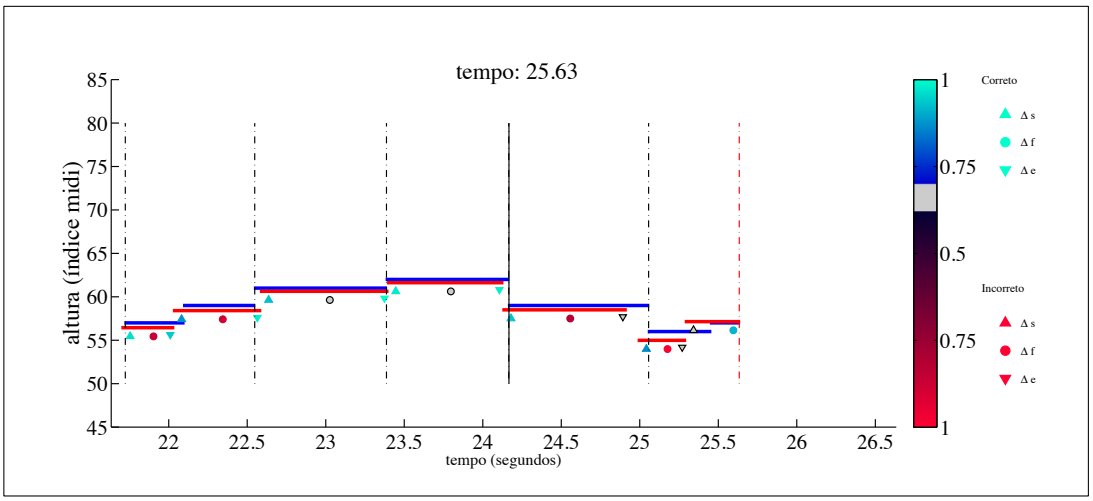
Na versão atual do sistema, o processo de avaliação é iniciado somente após o término da gravação do exercício, ou seja, o processo de avaliação é *offline*. A partir dos exemplos utilizados nos experimentos deste trabalho, para cada segundo de exercício gravado foi necessário, em média, um tempo de processamento equivalente a 0.659 segundos

para a execução do algoritmo no MATLAB. Apesar do protótipo rodar em modo *offline*, o tempo de processamento indica que é possível realizar uma implementação integral do sistema em tempo real, especialmente se todo o sistema for codificado na linguagem C++.

Figura 5.5: Imagens ilustrativas da implementação do protótipo do sistema. (a) Módulo implementado em C++, responsável pelo rastreamento do movimento da mão e da captura do sinal de áudio. (b) Módulo implementado em MATLAB, responsável pelo processamento dos sinais e classificação do movimento e do canto. (c) Imagem ampliada com exemplo de *Feedback* visual, nota-a-nota, considerando os parâmetros ataque, duração e afinação.



(a) (b)



(c)

Fonte: O Autor

5.3 Experimentos e Resultados Finais do Sistema Audiovisual Integrado

O foco dos experimentos descritos nesta Seção está no estudo do impacto da variação do andamento, a qual é modificada por intermédio da marcação de compassos definida pelos gestos do usuário, no processo de avaliação automática de solfejo.

5.3.1 Construção da Base de Dados Audiovisual

Diferentemente da base de dados DATASET_1, onde o canto das notas foi sincronizado com metrônomo, o conjunto de dados utilizado para avaliar este trabalho foi executado de forma livre, com sua atenção voltada para a execução de pequenos exercícios de solfejo onde há variação do andamento, como *accelerando*, *ritardando* e *fermatas*.

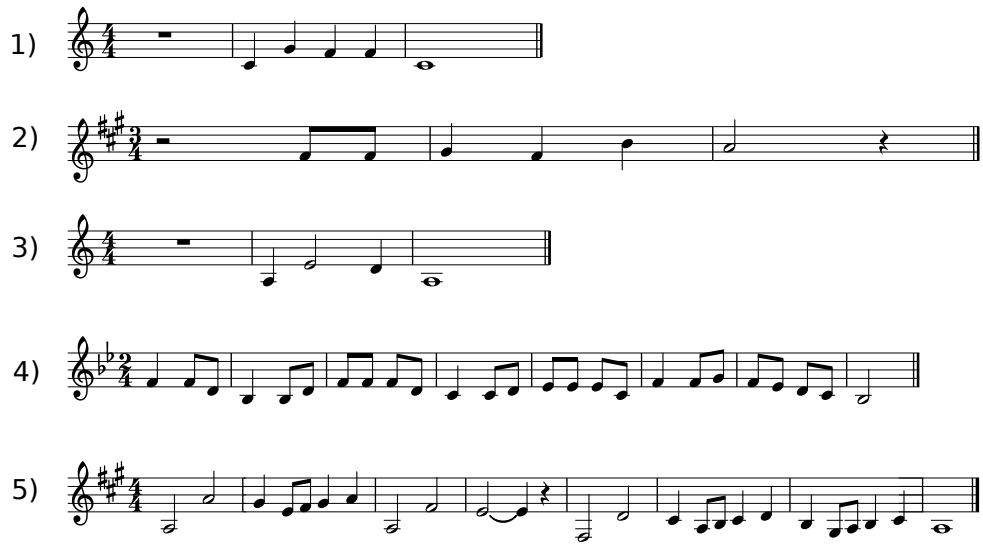
Durante as seções de gravação, a execução do exercício de solfejo foi conduzida com a marcação de compassos pela mão. Tanto o rastreamento da posição da mão quanto o sinal de áudio foram capturados pelo sensor *Kinect*. A taxa de aquisição do rastreamento da mão é de 30 quadros por segundo, e a taxa de amostragem do sinal de áudio é de 16000 amostras por segundo. Ao todo, um total de cinco pessoas foram gravadas, sendo dois professores de música com mais de dez anos de experiência e três estudantes de música em nível de graduação. A faixa etária dos participantes varia entre 19 e 58 anos. O processo de anotação seguiu o mesmo procedimento adotado na primeira base de dados (Seção 4.5.1), porém os especialistas tiveram que observar a coerência temporal das notas cantadas com a posição da mão. Novamente, 5 especialistas avaliaram toda a base de dados, atribuindo o rótulo de correto ou incorreto para cada nota cantada. As discordâncias entre avaliadores foram mantidas, as quais foram usadas posteriormente para modelar o classificador probabilístico.

Nessa nova base de dados contendo informação audiovisual (chamada de DATASET_2), a validação das gravações pelos especialistas foi um processo ainda mais custoso do que o processo utilizado para gerar a base de dados DATASET_1, uma vez que os avaliadores precisavam identificar as variações temporais originadas do movimento da mão (observação visual) concomitantemente com as variações temporais originadas do canto (observação auditiva). Como a avaliação pelos especialistas é feita nota-a-nota, cada seção de avaliação demandou um tempo bastante significativo e muita concentração por parte os músicos especialistas. Por esse motivo, a quantidade de amostras resultantes nesse processo de construção da base de dados DATASET_2 foi menor do que a utilizada na DATASET_1.

Para as gravações, foram utilizados cinco exemplos de exercícios de solfejo, e a Figura 5.6 ilustra as partituras dos exemplos utilizados. Ao todo, 1430 notas musicais foram avaliadas (286 para cada avaliador), gerando uma base de dados final com 286 exemplos rotulados pela maioria de votos. Entre o número total de exemplos, 18% são gerados a partir da partitura 1, 23% da partitura 2, 28% da partitura 3, 18% da partitura

4, e 13% da partitura 5. A maior concentração dos exemplos está nos três primeiros exercícios, pois os mesmos são mais factíveis para avaliação pelos especialistas. Apesar da quantidade inferior de notas musicais, os exemplos 4 e 5 são importantes pois permitem verificar a sensibilidade do algoritmo em relação ao alinhamento temporal da partitura, uma vez que longas sequências de movimento poderiam ser causas de erro na retificação e também no mapeamento dos segmentos melódicos à partitura.

Figura 5.6: Partituras dos exercícios de solfejo utilizados para a geração da base de dados DATASET_2.

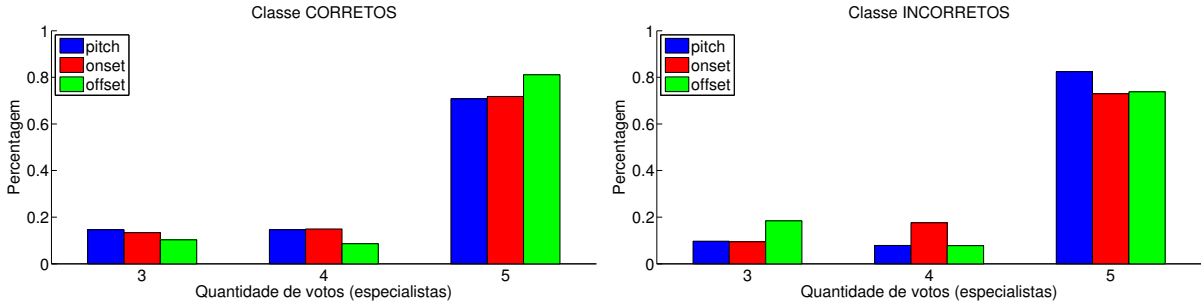


Fonte: O Autor

Uma análise a partir do número de votos dado pelos especialistas a cada nota cantada foi também realizada nos experimentos com variação de andamento. Essa análise revelou que, entre as notas com maioria de votos para a classe correto, $10 - 15\%$ das atribuições receberam apenas 3 (entre 5) votos em concordância, significando um certo grau de dúvida entre os especialistas. Percentuais semelhantes foram medidos para a classe incorreto. A Figura 5.7 apresenta as distribuições de votos em detalhes, considerando cada um dos parâmetros observados.

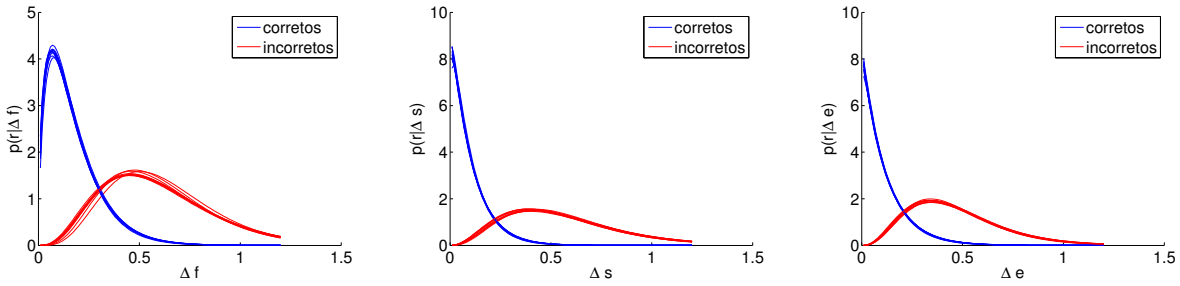
As funções de densidade de probabilidade Gamma, utilizadas pelo classificador Bayesiano para modelar as classes correto e incorreto para cada um dos parâmetros (*pitch*, *onset* e *offset*), foram estimadas a partir das distribuições de Δf, Δs e Δe. Os ajustes das funções para cada uma das 10 dobras são ilustrados na Figura 6.5. Como pode ser observado, dobras diferentes geram funções de densidade de probabilidade semelhantes, o que indica pouca influência na escolha das bases de treinamento e teste.

Figura 5.7: Percentuais das quantidades de votos atribuídos à classe vencedora (≥ 3) pelo comitê de especialistas.



Fonte: O Autor

Figura 5.8: Funções de densidade de probabilidade Gamma estimadas a partir dos dados de treinamento (10 dobras).



Fonte: O Autor

5.3.2 Testes Quantitativos com o Sistema de Avaliação Multimodal

No caso do exercício de solfejo com variação de andamento, os experimentos realizados utilizaram a base de dados DATASET_2. A metodologia adotada nestes experimentos segue exatamente o mesmo procedimento utilizado para a avaliação do solfejo sem variação de andamento. Porém, uma etapa prévia é inserida no *pipeline*, a qual avalia a capacidade do uso do movimento de marcação de compassos para realizar o alinhamento temporal da partitura.

Para cada exemplo gravado, o algoritmo primeiramente calcula o alinhamento temporal do gesto de marcação de compasso por intermédio do algoritmo *Subsequence DTW* (vide seção 5.1.1). Depois, os atrasos e/ou acelerações relativos a cada nota musical da partitura são calculados a partir do caminho de alinhamento, obtido pela comparação

entre a sequência de compassos definidos na partitura, e a sequência do movimento da mão, rastreada pelo sensor *Kinect*. Os atrasos e acelerações estimados são então aplicados a cada nota da partitura, e a partitura resultante, que está sincronizada com o movimento de marcação de compassos, é avaliada com um procedimento análogo ao da Seção 4.5.2.

As discrepâncias do *pitch*, *onset* e *offset* (Δf_i, Δs_i e Δe_i) de cada nota musical i foram computadas a partir da comparação individual de cada nota da partitura sincronizada e da melodia cantada (Eq. 4.15). O objetivo principal dos experimentos descritos nessa seção é avaliar a performance do sistema no caso onde há a execução do solfejo com variação de andamento. Para tanto, um processo de validação cruzada (10-*Dobras*) foi utilizado novamente para avaliar a acurácia do sistema, considerando a utilização do classificador Bayesiano com e sem a regra de rejeição. Cada parâmetro (*pitch*, *onset*, *offset*) de cada nota musical foi classificado como correto, incorreto, ou indefinido (este último apenas quando aplicada a regra de rejeição).

A Tabela 5.1 apresenta as matrizes de confusão geradas pelos classificadores Bayesianos sem a implementação da regra de rejeição. Neste caso, a acurácia de cada parâmetro analisado é 87.12%, 87.24%, e 82.80%, respectivamente. Os três parâmetros tiveram redução na acurácia quando comparados com os testes sem variação de andamento, feitos com a base de dados DATASET_1.

Tabela 5.1: Avaliação do sistema proposto usando validação cruzada (10 dobras), sem o uso da regra de rejeição de Bayes.

Target Class \ Output Class	$\varphi_{\Delta f}$	$\overline{\varphi}_{\Delta f}$	
$\varphi_{\Delta f}$	**82.56%**	17.91%	
$\overline{\varphi}_{\Delta f}$	7.85%	**92.15%**	
			87.12%

(a) Pitch evaluation

Target Class \ Output Class	$\varphi_{\Delta s}$	$\overline{\varphi}_{\Delta s}$	
$\varphi_{\Delta e}$	**92.86%**	7.14%	
$\overline{\varphi}_{\Delta e}$	18.38%	**81.62%**	
			87.24%

(b) Onset evaluation

Target Class \ Output Class	$\varphi_{\Delta e}$	$\overline{\varphi}_{\Delta e}$	
$\varphi_{\Delta e}$	**83.84%**	16.16%	
$\overline{\varphi}_{\Delta e}$	18.24%	**81.76%**	
			82.80%

(c) Offset evaluation

Nos experimentos com a aplicação da regra de rejeição de Bayes, houve uma melhora significativa na classificação de todos os parâmetros, como pode ser visto na comparação entre as matrizes de confusão das Tabelas 5.1 e 5.2. O procedimento para a escolha dos limiares de rejeição $T_{\Delta f}$ (*pitch*), $T_{\Delta s}$ (*onset*) e $T_{\Delta e}$ (*offset*) foi o mesmo utilizado na seção 4.5.2. Assim, buscou-se uma configuração que mantivesse 85% das amostras, obtendo-se novamente uma percentagem de rejeição semelhante à proporção de rótulos atribuídos com alto grau de desacordo (3 votos contra 2). A Figura 5.9 ilustra a evolução da acurácia *versus* a quantidade de amostras não rejeitadas a medida que os limiares são

modificados.

Tabela 5.2: Avaliação do sistema proposto usando validação cruzada (10 dobras), com o uso da regra de rejeição de Bayes. O sistema é capaz de responder em 85% das vezes e obtém uma melhora na acurácia de aproximadamente 3%.

Target Class \ Output Class	$\varphi_{\Delta f}$	$\overline{\varphi}_{\Delta f}$
$\varphi_{\Delta f}$	**89.68%**	5.68%
$\overline{\varphi}_{\Delta f}$	10.32%	**89.68%**
		90.75%

(a) Pitch evaluation: $T_{\Delta f} = 0.31$

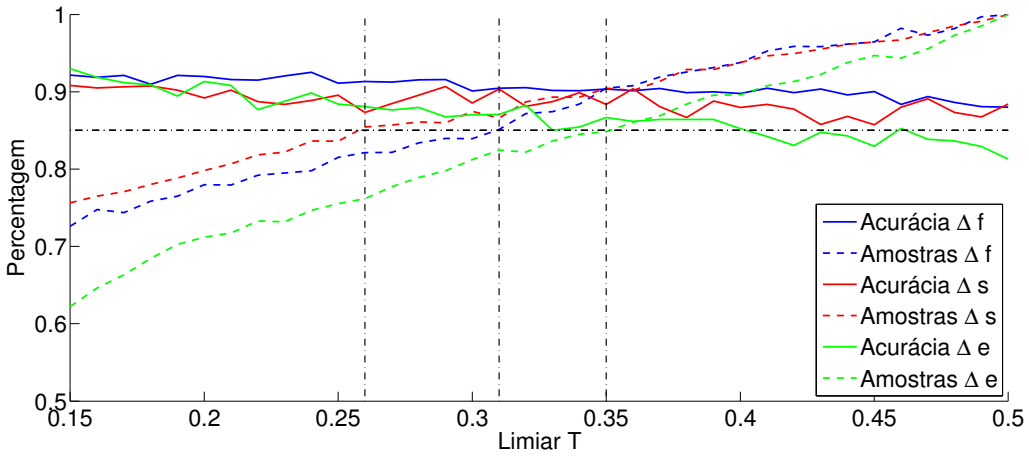
Target Class \ Output Class	$\varphi_{\Delta s}$	$\overline{\varphi}_{\Delta s}$
$\varphi_{\Delta s}$	**95.44%**	18.56%
$\overline{\varphi}_{\Delta s}$	4.56%	**81.44%**
		88.51%

(b) Onset evaluation: $T_{\Delta s} = 0.26$

Target Class \ Output Class	$\varphi_{\Delta e}$	$\overline{\varphi}_{\Delta e}$
$\varphi_{\Delta e}$	**86.93%**	14.05%
$\overline{\varphi}_{\Delta e}$	13.07%	**85.95%**
		86.23%

(c) Offset evaluation: $T_{\Delta e} = 0.35$

Figura 5.9: Comparação da acurácia versus o número de amostras não rejeitadas. Linhas sólidas mostram a evolução da acurácia, as quais são afetadas pelos limiares $T_{\Delta f}$ (pitch), $T_{\Delta s}$ (onset), e $T_{\Delta e}$ (offset).



Fonte: O Autor

Visando extrair maiores informações sobre o sistema, bem como identificar causas e origens dos erros de classificação, diferentes análises foram feitas a partir do número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. A Figura 5.10 ilustra os histogramas com o comparativo entre a percentagem de exemplos classificados (corretamente e incorretamente) e as durações das notas musicais transcritas. Essa análise busca verificar se notas de curta ou longa duração têm alguma predominância na quantidade de erro ou acerto de classificação.

A partir dos testes realizados, com excessão do parâmetro *offset*, verifica-se certa uniformidade no erro para diferentes durações de notas, não apresentando concentrações localizadas. O histograma referente ao parâmetro *offset* apresenta uma maior proporção de amostras mal classificadas. Apesar desse fato indicar algum problema com o alinhamento temporal, tal hipótese não é válida visto que as amostras relacionadas ao parâmetro

onset continuam com uma taxa de acerto em proporção quase constante à medida que o tempo e/ou a duração das notas aumenta. De fato, ao avaliar os coeficientes de correlação Spearman obtidos na comparação da proporção de erro relativo à duração das notas considerando os parâmetros Δs (*onset*) e Δe (*offset*), obteve-se 0.12 e 0.27, respectivamente. Além disso, os valores de confiança p obtidos no cálculo da correlação foram 0.62 e 0.16, respectivamente. Considerando um nível de significância de $\alpha = 0.05$ (95%), mantém-se a hipótese de que os dados não são correlacionados, indicando que não há relação de monotonicidade entre os erros e a duração das notas.

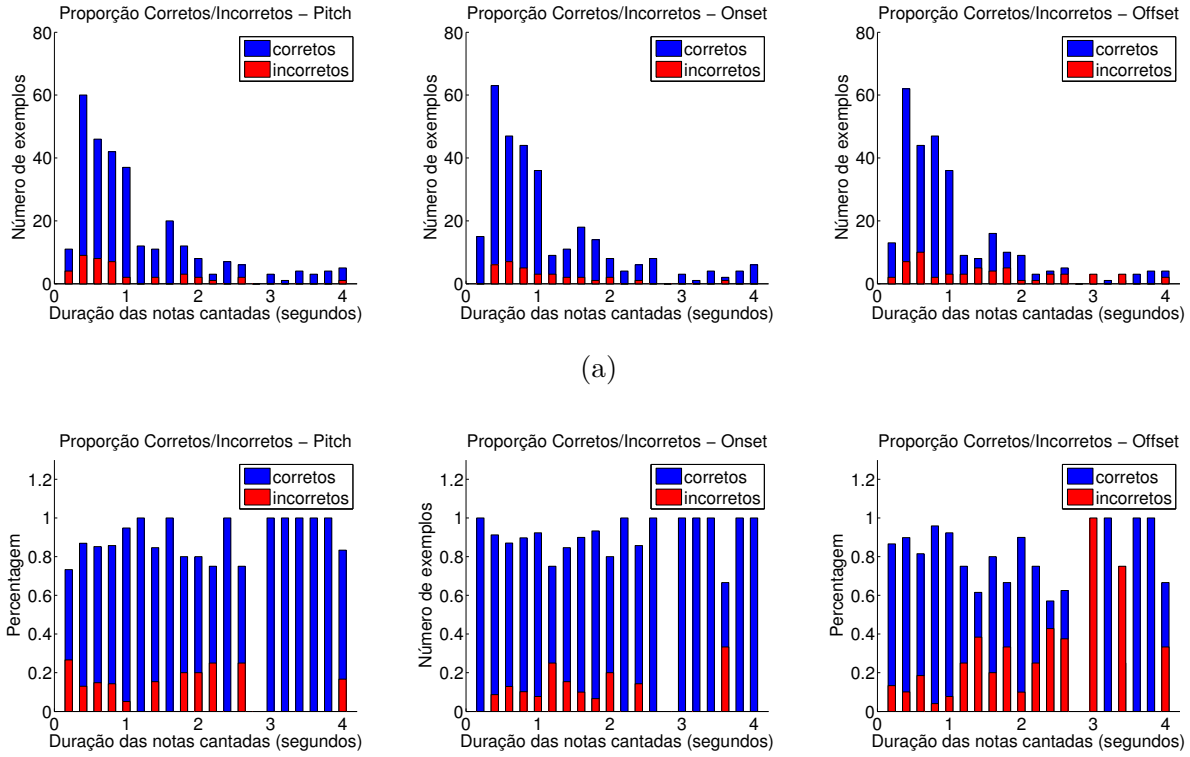
Esses valores indicam fraca correlação tanto da posição da nota quanto da duração da mesma em relação à proporção de erros obtidos nos experimentos de validação cruzada com 10 dobras. O problema do alto número de amostras mal classificadas no parâmetro *offset* é melhor explicado pelo fato de haver uma percentagem muito grande de votos (aproximadamente 20%) em desacordo entre os avaliadores, como pode ser evidenciado no histograma da Figura 5.7.

De forma análoga, a Figura 5.11 apresenta os histogramas considerando a posição temporal das notas musicais transcritas (*onset*). As mesmas conclusões obtidas para as durações das notas podem ser aplicadas para a avaliação do momento de ataque das notas musicais transcritas. Os parâmetros *pitch* e *onset* permanecem estáveis ao longo da evolução temporal, enquanto o parâmetro *offset* possui uma taxa maior de erro de classificação, independente da posição temporal de cada nota. A correlação de Spearman entre o erro e os parâmetros observados Δs (*onset*) e Δe (*offset*) obteve coeficientes iguais a 0.24 e -0.33, respectivamente. Considerando um nível de significância de $\alpha = 0.05$ (95%), os valores de confiança p obtidos no cálculo da correlação (0.20 e 0.16) também indicam que os dados não são correlacionados.

A Figura 5.12 apresenta o histograma do erro e acerto considerando a variação da posição do tempo de ataque, ajustada pela marcação do tempo com o movimento da mão. A partir do histograma da Figura 5.12a, percebe-se maior concentração de variações de *onset* entre -2 e 2 segundos, causados principalmente por *fermatas* e *rubatos*. Apesar dessa concentração, há expressiva variação com magnitude superior a 2 segundos, indicando alteração de andamento da peça. A classificação mostrou-se pouco sensível a largas variações de andamento, apresentando predominância de acertos.

Também foi analisada a influência da detecção e reconhecimento do gesto de marcação de compasso na acurácia final do sistema. A Figura 5.13 apresenta a distribuição dos erros e acertos para cada uma das classes de gestos: movimento binário, ternário,

Figura 5.10: Proporção do erro e acerto de classificação em relação à duração das notas cantadas. (a) quantidade absoluta de amostras. (b) quantidade percentual por *bin* do histograma.



(a)

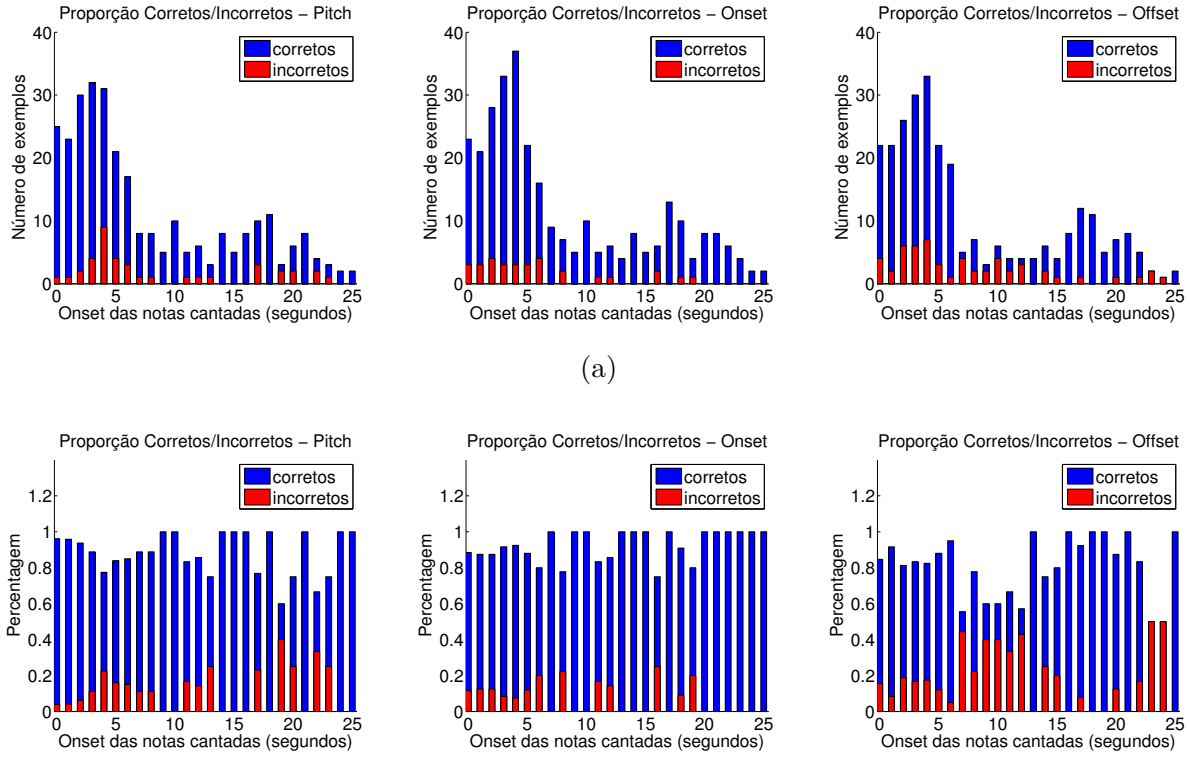
(b)

Fonte: O Autor

quaternário e indefinido (quando nenhum gesto é reconhecido). É possível concluir que não há uma predominância de erro sobre um determinado tipo de gesto, pois as proporções dos erros não apresentam um padrão de crescimento ou decrescimento. Além disso, percebe-se que mesmo quando o sistema não é capaz de reconhecer um gesto, seja por causa de um exemplo mal executado pelo intérprete ou uma falha do classificador, o alinhamento gerado pela DTW ainda permite a classificação das notas em 85% dos casos.

Por fim, a análise da relação entre os erros e acertos de classificação considerando a distribuição de votos dados pelos especialistas é ilustrada na Figura 5.14a. É evidente a concentração proporcional de erro nas amostras em que há discordância de voto entre os avaliadores especialistas, em particular, aquelas que foram rotuladas contendo apenas 3 votos em comum. A regra de rejeição de Bayes elimina parte das amostras que estão dentro dessa região "nebulosa" de classificação. A Figura 5.14b ilustra as proporções de amostras rejeitadas, as quais, ao mesmo tempo, são associadas às amostras mal classificadas. Neste caso, $18 - 30\%$ das amostras rejeitadas continham 3 votos, proporção superior

Figura 5.11: Proporção do erro e acerto de classificação em relação à posição temporal das notas cantadas (*Onset*). (a) quantidade absoluta de amostras. (b) quantidade percentual por *bin* do histograma.



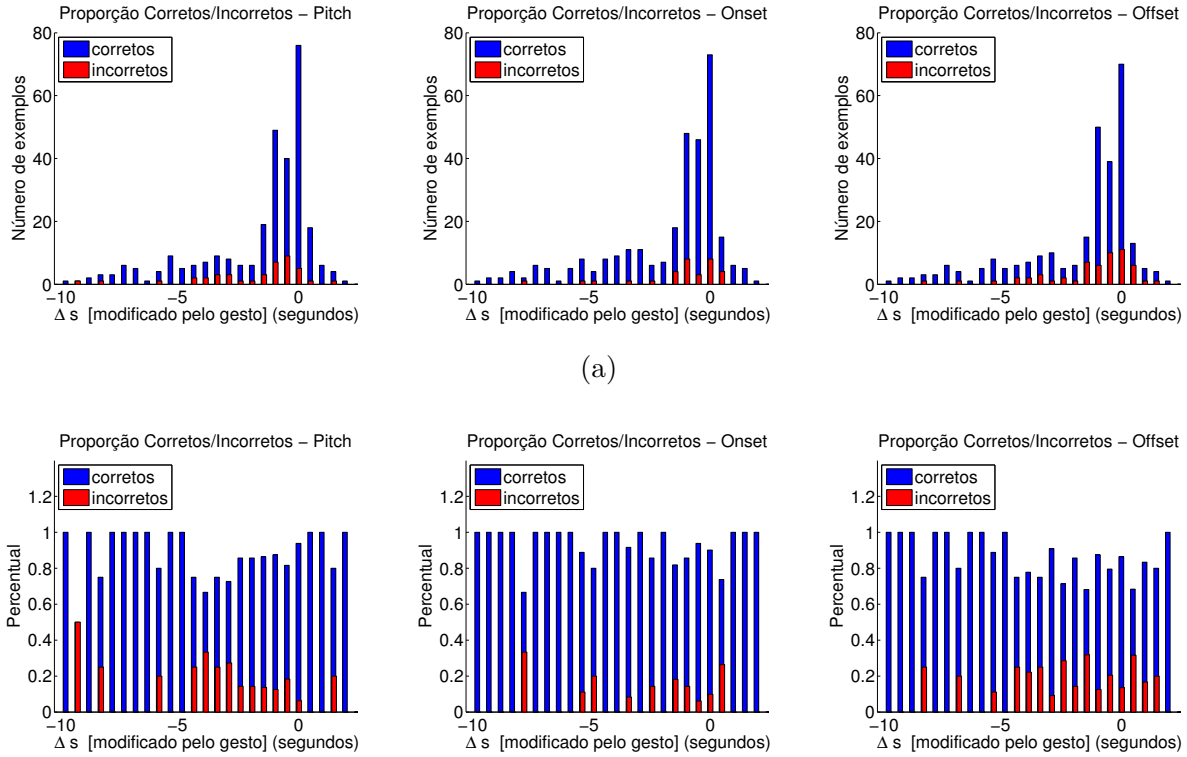
(a)

(b)

Fonte: O Autor

(aproximadamente o dobro) à percentagem correspondente de amostras, considerando todos os exemplos da base ($10 - 15\%$). Isso indica que a regra de rejeição no classificador Bayesiano atua seletivamente, removendo preferencialmente amostras que estão relacionadas aos casos de dúvida dos especialistas.

Figura 5.12: Proporção do erro e acerto de classificação em relação à variação temporal do *Onset* (Δs), causado pela modificação de andamento através do gesto de marcação de compassos. (a) quantidade absoluta de amostras. (b) quantidade percentual por *bin* do histograma.

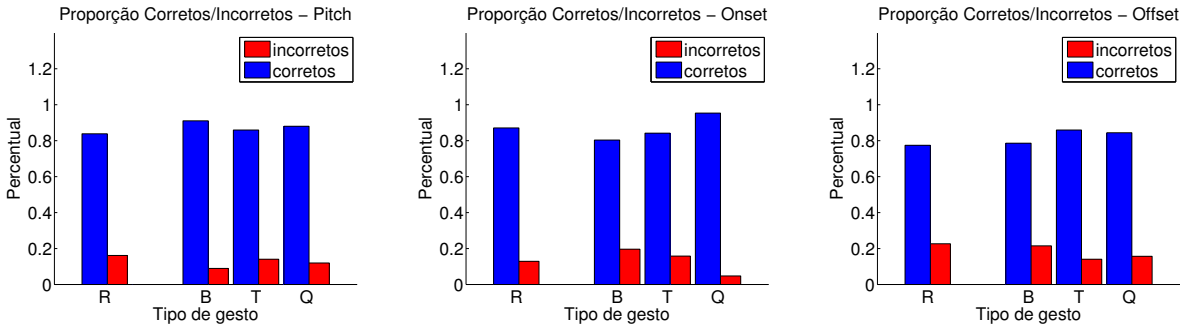


(a)

(b)

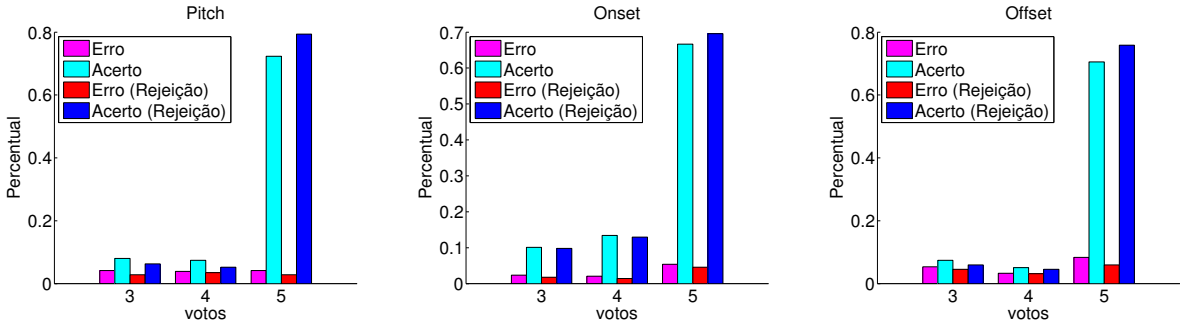
Fonte: O Autor

Figura 5.13: Proporção do erro e acerto de classificação em relação ao tipo de gesto de marcação de compassos detectado.

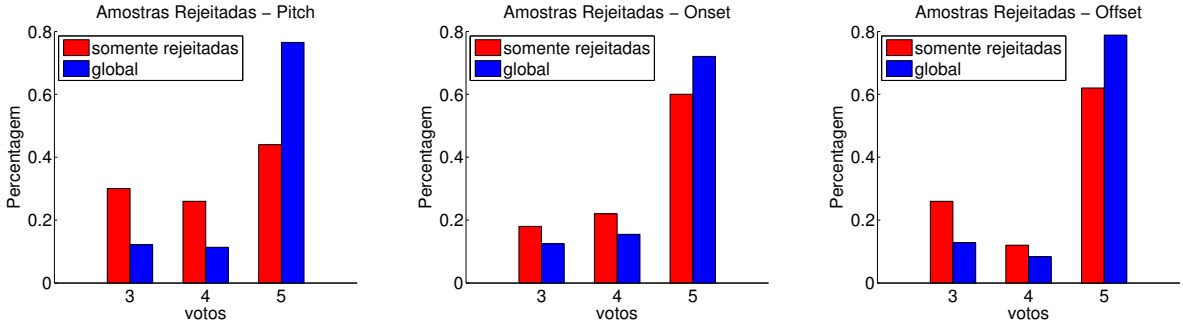


Fonte: O Autor

Figura 5.14: Erros proporcionais à quantidade de votos.



(a) Erro proporcional a cada voto dado pelos especialistas considerando a classificação com e sem regra de rejeição.



(b) Proporção de amostras rejeitadas pela regra de Bayes com relação ao número de votos dados pelos especialistas.

Fonte: O Autor

6 CONCLUSÃO

Este trabalho apresentou um sistema de avaliação automático de solfejo, o qual permite a realização do exercício musical com plasticidade expressiva das unidades de tempo. O sistema final faz a análise audiovisual conjunta da marcação de compassos e do canto da melodia. Para realizar essa tarefa, o presente estudo desenvolveu um conjunto de técnicas que busca superar limitações em abordagens pré-existentes, como a incapacidade de avaliar o exercício de solfejo contendo variações do andamento musical, e imprecisão ao gerar *feedback* individual, nota-a-nota, relativos ao ataque, duração e altura (afinação).

A fadiga e, por consequência, o risco de alteração na percepção dos especialistas humanos e nos critérios de avaliação motivaram a busca por um sistema automatizado de avaliação do solfejo, posto que tal prática é diária e imperiosa, principalmente para o aluno em fase de musicalização. Tal busca levou em consideração aspectos como viabilidade de implementação em ambientes pouco controlados e fácil acesso à infra-estrutura tecnológica. Esses critérios motivaram o uso do sensor RGB-D para capturar a informação de áudio e vídeo utilizadas nesta abordagem. A abordagem desenvolvida também permite que, futuramente, esse sensor seja substituído por outro tipo de equipamento que possua maior precisão na aquisição de dados.

Assim, o sistema desenvolvido é inovador, à medida que integra o movimento da mão, o qual controla o andamento da peça ao substituir as batidas do metrônomo, à emissão vocal, a qual é acompanhada nota-a-nota dentro de uma faixa de frequências tolerada por especialistas. A principal contribuição da pesquisa está, portanto, no reconhecimento de padrões gestuais da marcação de compassos, os quais apresentam aderência à identificação concomitante de notas musicais convencionalmente unitarizadas a partir da análise de áudio, quando há variação no fluxo metronômico do texto musical.

Primeiramente, o capítulo 3 abordou análise do gesto para marcação de compassos. A técnica desenvolvida permite a classificação dos padrões binário, ternário e quaternário, incluindo a possibilidade de verificar a precisão rítmica a cada unidade de tempo. O classificador utilizado foi modelado utilizando avaliações audiovisuais por músicos especialistas e os testes demonstraram que o sistema possui uma acurácia média de aproximadamente 95%. Esta parte do trabalho proposto pode ser utilizada, individualmente, para avaliar o estudo da marcação de compassos e manutenção do andamento, junto a alunos iniciantes em música, considerando apenas o gesto realizado pela mão. Como *feedback* ao aluno, o sistema indica se o movimento de cada compasso está correto e também atribui uma

medida de confiança, que revela o grau de precisão do movimento.

O capítulo 4 trata o problema da avaliação do solfejo, considerando apenas o aspecto do canto. Nesta parte do trabalho, um classificador Bayesiano foi construído para efetuar a avaliação nota-a-nota. Isso se constitui no diferencial que esta proposta traz em relação às técnicas já existentes, visto que essas são baseadas em informação obtida a partir de avaliações globais do trecho musical, por especialistas em música. Para tanto, neste estudo, um comitê de especialistas também realizou avaliações; porém, tais avaliações foram focadas em cada nota, individualmente. Seus pareceres foram registrados nota-a-nota, em sequências melódicas, considerando os seguintes aspectos: precisão no ataque, duração e altura. Nesta parte do trabalho, foi também desenvolvido uma técnica para mapeamento dos segmentos melódicos transcritos nas notas da partitura, o qual permite a comparação individual das notas, mesmo que as mesmas apresentem pequeno desalinhamento com relação ao fluxo regular das unidades de tempo.

Por fim, o capítulo 5, traz a integração entre as duas primeiras partes do trabalho, o que permite a avaliação da execução do solfejo com a expressividade advinda da variação da agógica. Dessa forma, o sistema de avaliação desenvolvido pode atender a condições musicais reais. O uso do movimento da mão como fator de marcação de compassos e, consequentemente, de condução temporal do fluxo da execução do canto, possibilita que notas emitidas sob plasticidade temporal possam ser devidamente acompanhadas e avaliadas, sem a necessidade de sincronização por um metrônomo. Com tal integração, além do *feedback* relativo à classificação do movimento, o sistema também retorna ao usuário a indicação de correspondência entre cada nota escrita na partitura e cada nota emitida por seu canto. Em outras palavras, o sistema indica se o movimento de marcação de compasso está correto e se as condições de emissão de cada nota, em termos de ataque, duração e afinação, encontram-se numa margem de aceitação, por sua vez obtida no treinamento do classificador a partir da base de dados rotulada pelo comitê de especialistas.

Entre a decisão correto *versus* errado sobre cada parâmetro da nota musical existe uma região nebulosa, que reflete o limiar perceptivo dos especialistas, o qual constatou-se apresentar alguma variação a cada nova circunstância. Para mapear essa região, foi desenvolvido um método de rejeição, que descarta exemplos e simula a dúvida detectada na avaliação humana. Constatou-se que entre $10 - 15\%$ dos casos avaliados por humanos tiveram alto grau de incerteza, visto que os avaliadores discordavam entre si. Isso indica que nesses casos, o rótulo atribuído pelo comitê de especialistas pode não ser confiável ou que a nota emitida se encontra numa margem de aceitação limítrofe. Conclui-se daí que a

existência de uma margem de erro é inevitável. Além disso, a regra de rejeição aplicada no classificador Bayesiano proposto exclui prioritariamente exemplos com ambiguidade; isso é, a regra exclui seletivamente as amostras com concentração de votos em dúvida, o que é desejável nessa abordagem. Usando a combinação de todas as técnicas desenvolvidas, incluindo a regra de rejeição mencionada, o sistema resultante permite classificar as notas musicais cantadas em 85% dos casos, com uma acurácia média aproximada de 88.5%.

Como trabalhos futuros, no que diz respeito à técnica relacionada à classificação de movimento e estimativa do andamento, pretende-se avaliar a influência de novos sistemas de captura de movimento, especialmente em relação à taxa de amostragem. Acredita-se que o rastreamento do movimento da mão com maior precisão melhorará a acurácia da classificação, e uma maior taxa de amostragem permitirá uma maior precisão da estimativa da variação temporal. Em relação às técnicas que abordam o processamento e análise de áudio, pretende-se incluir sensores com arranjos de microfones, os quais poderão ser utilizados para melhorar a qualidade do sinal e obter informações sobre a localização da fonte sonora. Por esse intermédio, será possível permitir a avaliação de solfejos em situações onde mais de uma pessoa estiver cantando, ao mesmo tempo.

Em uma perspectiva voltada à aplicação prática do sistema proposto, pretende-se introduzir seu uso nos cursos de música desenvolvidos pela UFRGS e Universidades Parceiras do Programa PROLICENMUS. Em especial, espera-se que o sistema auxilie as disciplinas e demais atividades que envolvem teoria e percepção musical. Uma possível forma de utilização da ferramenta nesse contexto seria a classificação da habilidade dos alunos na prática do solfejo. Por exemplo, alunos que atingem uma pontuação mínima são autorizados a progredir na sequência dos estudos ou lhes é permitido o agendamento para atendimento especial com o professor. Dessa forma, essa estratégia pode otimizar a função do professor em sala de aula, o qual terá mais tempo para focar em outros aspectos também importantes do processo de aprendizagem musical.

Em um futuro próximo, planeja-se desenvolver uma versão multiplataforma do sistema, para que seja distribuído aos professores de música. Dessa forma, o sistema pode ser introduzido em diversas escolas do país, ampliando a sua abrangência de utilização. A disseminação da ferramenta permitirá também a coleta de dados que fomentará pesquisas futuras. Acredita-se que a avaliação sistemática dos alunos em larga escala possibilite a geração de um banco de dados com características de procedimentos e comportamento relacionados aos acertos e erros recorrentes, os quais permitirão uma análise futura detalhada e a proposição de novas técnicas didáticas para resolução de problemas no processo

de ensino aprendizagem musical, em especial da prática de solfejo.

Todas essas possíveis aplicações e melhorias se inspiram na motivação inicial dessa tese e darão continuidade a busca de soluções para as lacunas detectadas no contexto da educação musical brasileira, em particular na modalidade a distância mediada por tecnologias da informação e comunicação.

REFERÊNCIAS

ABDOLAHI, B.; GHASEMI, S.; GHEISSARI, N. Human motion analysis using dynamic textures. In: **16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)**. Shiraz, Fars: IEEE, 2012. p. 151–156.

ABESSER, J. et al. Automatic Quality Assessment of Vocal and Instrumental Performances of Ninth-grade and Tenth-grade Pupils. In: **Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)**. Marseille, France: LMA, 2013. p. 975–988.

ADISTAMBHA, K.; RITZ, C.; BURNETT, I. Motion classification using dynamic time warping. In: **IEEE 10th Workshop on Multimedia Signal Processing, 2008**. Cairns, Qld: IEEE, 2008. p. 622–627.

AHMAD, M.; LEE, S.-W. HMM-based human action recognition using multiview image sequences. In: **18th International Conference on Pattern Recognition, 2006 (ICPR)**. Hong Kong: IEEE, 2006. v. 1, p. 263–266.

AKL, A.; FENG, C.; VALAEE, S. A novel accelerometer-based gesture recognition system. **IEEE Transactions on Signal Processing**, IEEE Press, Piscataway, NJ, USA, v. 59, n. 12, p. 6197–6205, dec. 2011. ISSN 1053-587X.

ANKERST, M. et al. 3D shape histograms for similarity search and classification in spatial databases. In: **Proceedings of the 6th International Symposium on Advances in Spatial Databases**. London, UK: Springer-Verlag, 1999. (SSD '99), p. 207–226. ISBN 3-540-66247-2.

ARGUETA, C. R.; KO, C.-J.; CHEN, Y.-S. Interacting with a music conducting system. In: JACKO, J. A. (Ed.). **Human-Computer Interaction. Novel Interaction Methods and Techniques**. Berlin Heidelberg: Springer, 2009, (Lecture Notes in Computer Science, v. 5611). p. 654–663. ISBN 978-3-642-02576-1.

BABACAN, O. et al. A comparative study of pitch extraction algorithms on a large variety of singing sounds. In: **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Vancouver: IEEE, 2013. p. 7815–7819. ISSN 1520-6149.

BALAKRISHNAN, N.; MITRA, D. Likelihood inference based on left truncated and right censored data from a gamma distribution. **IEEE Transactions on Reliability.**, v. 62, n. 3, p. 679–688, Sept 2013.

BAY, M. et al. Second fiddle is important too: Pitch tracking individual voices in polyphonic music. In: **Proceedings of the 13th International Society for Music Information Retrieval Conference**. Porto, Portugal: FEUP Edições, 2012. p. 319–324.

BEHRINGER, R. Conducting digitally stored music by computer vision tracking. In: **Proceedings of the First International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution**. Washington, DC, USA: IEEE Computer Society, 2005. p. 271–274. ISBN 0-7695-2348-X.

BELLO, J. P. et al. A Tutorial on Onset Detection in Music Signals. **Speech and Audio Processing, IEEE Transactions on**, IEEE, v. 13, n. 5, p. 1035–1047, sep. 2005. ISSN 1063-6676.

BENETOS, E. et al. Automatic music transcription: Breaking the glass ceiling. In: **Proceedings of the 13th International Society for Music Information Retrieval Conference**. Porto, Portugal: FEUP Edições, 2012. p. 379–384.

BEVILACQUA, F. et al. Continuous realtime gesture following and recognition. In: **Proceedings of the 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction**. Berlin, Heidelberg: Springer-Verlag, 2010. (GW'09), p. 73–84.

BIANNE-BERNARD, A.-L. et al. Dynamic and contextual information in hmm modeling for handwritten word recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 33, n. 10, p. 2066–2080, Oct 2011. ISSN 0162-8828.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. ISBN 978-0387-31073-2.

BOBICK, A. F. et al. The recognition of human movement using temporal templates. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 23, p. 257–267, March 2001. ISSN 0162-8828.

BORCHERS, J.; SAMMINGER, W.; MUHLHAUSER, M. Personal orchestra: conducting audio/video music recordings. In: **Proceedings of Second International Conference onWeb Delivering of Music, 2002.** Darmstadt, Germany: IEEE, 2002. p. 93–100.

BRADSHAW, D.; NG, K. Tracking conductors hand movements using multiple wiimotes. In: **Proceedings of the 2008 International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution**. Florence, Italy: IEEE Computer Society, 2008. (AXMEDIS '08), p. 93–99. ISBN 978-0-7695-3406-0.

CAMURRI, A. et al. Communicating expressiveness and affect in multimodal interactive systems. **IEEE MultiMedia**, v. 12, n. 1, p. 43–53, Jan 2005.

CHEVEIGNÉ, A. de; KAWAHARA, H. YIN, a fundamental frequency estimator for speech and music. **The Journal of the Acoustical Society of America**, ASA, v. 111, n. 4, p. 1917–1930, April 2002.

COELHO, H. W. **Técnica Vocal para Coros**. 10. ed. São Leopoldo: Editora Sinodal, 1994.

DAHL, L. Triggering sounds from discrete air gestures: What movement feature has the best timing? In: CARAMIAUX, B. et al. (Ed.). **Proceedings of the International Conference on New Interfaces for Musical Expression**. London, United Kingdom: Goldsmiths, University of London, 2014. p. 201–206.

DARDAS, N.; GEORGANAS, N. D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. **IEEE Transactions on Instrumentation and Measurement**, v. 60, n. 11, p. 3592–3607, November 2011. ISSN 0018-9456.

DEGARA, N. et al. Onset event decoding exploiting the rhythmic structure of polyphonic music. **IEEE Journal of Selected Topics in Signal Processing**, v. 5, n. 6, p. 1228–1239, October 2011. ISSN 1932-4553.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification (2nd Edition)**. New York: Wiley-Interscience, 2001. ISBN 0471056693.

EWERT, S.; MULLER, M.; GROSCHE, P. High resolution audio synchronization using chroma onset features. In: **IEEE International Conference on Acoustics, Speech and Signal Processing, 2009 (ICASSP)**. Taipei, Taiwan: IEEE, 2009. p. 1869–1872. ISSN 1520-6149.

FENZA, D. et al. Physical movement and musical gestures: a multilevel mapping strategy. In: **Proceedings of Sound and Music Computing Conference**. Salerno: [s.n.], 2005.

FOOTE, J. Automatic audio segmentation using a measure of audio novelty. In: **IEEE International Conference on Multimedia and Expo (ICME)**. New York: IEEE, 2000. v. 1, p. 452–455.

GAINZA, M.; COYLE, E. Tempo detection using a hybrid multiband approach. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 19, n. 1, p. 57–68, 2011. ISSN 1558-7916.

GILES, D. E. A.; FENG, H. Bias of the maximum likelihood estimators of the two-parameter gamma distribution revisited. **Econometrics Working Papers**, n. 908, p. 1–19, September 2009. ISSN 1485-6441.

GOMES, C. H. S. Formação e atuação de músicos de rua: possibilidades de atuação e de caminhos formativos. **Revista da ABEM**, n. 8, p. 25–28, March 2003.

GÓMEZ, E.; BONADA, J. Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing. **Computer Music Journal**, MIT Press, v. 37, p. 73–90, 2013. ISSN 0148-9267.

GORDON, E. **Learning Sequences in Music: A Contemporary Music Learning Theory**. 2012. ed. Chicago: GIA Publications, 2011.

GORELICK, L. et al. Actions as space-time shapes. In: **Tenth IEEE International Conference on Computer Vision (ICCV)**. Beijing: IEEE, 2005. v. 2, p. 1395–1402. ISSN 1550-5499.

GUO, J. M. et al. Improved hand tracking system. **Circuits and Systems for Video Technology, IEEE Transactions on**, v. 22, n. 5, p. 693–701, 2012. ISSN 1051-8215.

HAN, S.; KIM, J.-B.; KIM, J. D. Follow-me!: conducting a virtual concert. In: **Adjunct proceedings of the 25th annual ACM symposium on User interface software and technology**. New York, NY, USA: ACM, 2012. (UIST Adjunct Proceedings '12), p. 65–66. ISBN 978-1-4503-1582-1.

HARDING, P. R. G.; ELLIS, T. Recognizing hand gesture using fourier descriptors. In: **Proceedings of the 17th International Conference on Pattern Recognition (ICCPR)**. Cambridge: IEEE, 2004. v. 3, p. 286–289. ISSN 1051-4651.

HUANG, P.-K. et al. Real-time stereo matching for 3d hand gesture recognition. In: **SoC Design Conference (ISOCC), 2012 International**. Jesu Island: IEEE, 2012. p. 29–32.

HUSSAIN, S.; RASHID, A. User independent hand gesture recognition by accelerated dtw. In: **International Conference on Informatics, Electronics Vision (ICIEV)**. Dhaka: IEEE, 2012. p. 1033–1037.

ILMONEN, T.; TAKALA, T. Conductor following with artificial neural networks. In: **Proceedings of the International Computer Music Conference. (ICMC'99)**. Beijing, China: International Computer Music Association, 1999. p. 367–370.

JE, H.; KIM, J.; KIM, D. Hand gesture recognition to understand musical conducting action. In: **The 16th IEEE International Symposium on Robot and Human interactive Communication (RO-MAN).** Jeju: IEEE, 2007. p. 163–168.

JOHNSON, N.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions**: Continuous univariate distributions / johnson, norman lloyd. - new york : Wiley, 1970. 2. ed. New York: Wiley, 1995. (Distributions in statistics, v. 1). ISBN 0471584940.

JUNEJO, I. N. et al. View-independent action recognition from temporal self-similarities. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Computer Society, Los Alamitos, CA, USA, v. 33, n. 1, p. 172–185, 2011. ISSN 0162-8828.

KAISER, F.; PEETERS, G. Multiple hypotheses at multiple scale for audio novelty computation within music. In: **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Vancouver: IEEE, 2013. p. 231–235.

KAKUMANU, P.; MAKROGIANNIS, S.; BOURBAKIS, N. A survey of skin-color modeling and detection methods. **Pattern Recognition**, Elsevier Science Inc., New York, NY, USA, v. 40, n. 3, p. 1106–1122, mar. 2007. ISSN 0031-3203.

KEOGH, E. J.; PAZZANI, M. J. Derivative dynamic time warping. In: **In First SIAM International Conference on Data Mining (SDM'2001)**. Chicago: Society for Industrial and Applied Mathematics, 2001.

KIM, S.-B. et al. Some effective techniques for naive bayes text classification. **Knowledge and Data Engineering, IEEE Transactions on**, v. 18, n. 11, p. 1457–1466, Nov 2006. ISSN 1041-4347.

KIM, S.-J. The metrically trimmed mean as a robust estimator of location. **The Annals of Statistics**, v. 20, n. 3, p. 1534–1547, September 1992.

KLAPURI, A.; DAVY, M. **Signal Processing Methods for Music Transcription**. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387306676.

KOLESNIK, P. **Conducting gesture recognition, analysis and performance system**. Dissertation (Master) — McGill University, Montreal, Canada, 2004.

LAPTEV, I. On space-time interest points. **International Journal of Computer Vision**, v. 64, n. 2-3, p. 107–123, 2005.

LARROUY-MAESTRI, P. et al. The evaluation of singing voice accuracy: A comparison between subjective and objective methods. **Journal of Voice**, v. 27, n. 2, p. 259.e1 – 259.e5, 2013. ISSN 0892-1997.

LEE, E. et al. conga: A framework for adaptive conducting gesture analysis. In: **International Conference on New Interfaces for Musical Expression (NIME)**. Paris: IRCAM, 2006. p. 260–265.

LEONIDO, L. Da expressão e expressividade geral à música em particular. **Sinfonía Virtual: Revista de Música Clásica y Reflexión Musical**, Sinfonía Virtual, n. 3, April 2007. ISSN 1886-9505.

LI, X. et al. Hand gesture recognition by stereo camera using the thinning method. In: **International Conference on Multimedia Technology (ICMT)**. Zhejiang, China: IEEE, 2011. p. 3077–3080.

LICHTENAUER, J. F.; HENDRIKS, E. A.; REINDERS, M. J. Sign language recognition by combining statistical dtw and independent classification. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Computer Society, Los Alamitos, CA, USA, v. 30, n. 11, p. 2040–2046, 2008. ISSN 0162-8828.

LIN, C.-H. et al. Automatic singing evaluating system based on acoustic features and rhythm. In: **IEEE International Conference on Orange Technologies (ICOT)**. Xian: IEEE, 2014. p. 165–168.

LUCK, G.; TOIVIAINEN, P. Ensemble musicians' synchronization with conductors' gestures: an automated feature-extraction analysis. **Music Perception: An Interdisciplinary Journal**, v. 24, n. 2, p. 189–199, December 2006.

MACHADO, R. B. **Narrativas de Professores de Teoria e Percepção Musical: Caminhos de Formação Profissional**. Dissertation (Master) — Universidade Federal de Santa Maria, Centro de Educação, Programa de Pós-Graduação em Educação, Santa Maria, Rio Grande do Sul, Brasil, 2012.

MAES, P.-J. et al. The "conducting master": An interactive, real-time gesture monitoring system based on spatiotemporal motion templates. **International Journal Human Computer Interaction**, v. 29, n. 7, p. 471–487, 2013.

MAKA, T. Attributes of audio feature contours for automatic singing evaluation. In: **36th International Conference on Telecommunications and Signal Processing (TSP)**. Rome: IEEE, 2013. p. 517–520.

MANDANICI, M.; SAPIR, S. Disembodied voices: A kinect virtual choir conductor. In: SOUND; COMPUTING, M. (Ed.). **Proceedings of the 9th Sound and Music Computing Conference**. Copenhagen: Aalborg University Copenhangen, 2012. p. 271–276.

MATHEWS, M. **Three dimensional baton and gesture sensor**. Google Patents, 1990. US Patent 4,980,519. Available from Internet: <http://www.google.com.br/patents/US4980519>.

MAUCH, M. et al. Computer-aided melody note transcription using the tony software: Accuracy and efficiency. In: **Proceedings of the First International Conference on Technologies for Music Notation and Representation**. Paris: Institut de Recherche en Musicologie, 2015. Accepted.

MAUCH, M.; DIXON, S. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In: **Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)**. Florence: IEEE, 2014. p. 659–663.

MAUCH, M.; FRIELER, K.; DIXON, S. Intonation in unaccompanied singing: Accuracy, drift and a model of reference pitch memory. **Journal of the Acoustical Society of America**, v. 136, n. 1, p. 401–411, May 2014.

MIN, B.-W. et al. Hand gesture recognition using hidden Markov models. In: **IEEE International Conference on Systems, Man, and Cybernetics, 1997. "Computational Cybernetics and Simulation".** Orlando: IEEE, 1997. v. 5, p. 4232–4235.

MOLINA, E. et al. Evaluation framework for automatic singing transcription. In: **Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR, 2014**. Taipei, Taiwan: ISMIR, 2014. p. 567–572.

MOLINA, E. et al. Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In: **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Vancouver, Canada: IEEE, 2013. p. 744–748. ISSN 1520-6149.

MOLINA, E. et al. SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 23, n. 2, p. 252–263, Feb 2015. ISSN 2329-9290.

MÜLLER, M. **Information Retrieval for Music and Motion**. Berlin Heidelberg: Springer Verlag, 2007. ISBN 3540740473.

MÜLLER, M.; EWERT, S.; KREUZER, S. Making chroma features more robust to timbre changes. In: **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Taipei, Taiwan: IEEE, 2009. p. 1877 – 1880.

NAKRA, T. M. et al. The UBS Virtual Maestro : an Interactive Conducting System. In: **Proceedings of the International Conference on New Interfaces for Musical Expression**. Pittsburgh: NIME, 2009. p. 250–255.

OTTMAN, R.; ROGERS, N. **Music for Sight Singing**. 8. ed. New Jersey: Prentice Hall, 2011.

PUCKETTE, M. Pure data: another integrated computer music environment. In: **Proceedings of the Second Intercollege Computer Music Concerts**. Tachikawa, Japan: [s.n.], 1996. p. 37–41.

RAILEANU, L.; STOFFEL, K. Theoretical Comparison between the Gini Index and Information Gain Criteria. **Annals of Mathematics and Artificial Intelligence**, v. 41, n. 1, p. 77–93, may 2004. ISSN 1012-2443.

RAJKO, S. et al. Real-time gesture recognition with minimal training requirements and on-line learning. In: **IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. Minneapolis, Minnesota, USA: IEEE Computer Society, 2007.

RAMACHANDRAN, K. M.; TSOKOS, C. P. **Mathematical Statistics with Applications in R**. 2. ed. Boston: Academic Press, 2015. ISBN 978-0-12-417113-8.

RYYNÄNEN, M.; KLAPURI, A. Modelling of note events for singing transcription. In: **Proceedings of ISCA - Tutorial and Research Workshop on Statistical and Perceptual Audio**. Jeju, Korea: MIT Press, 2004.

SARASÚA, A.; GUAUS, E. Dynamics in music conducting: A computational comparative study among subjects. In: CARAMIAUX, B. et al. (Ed.). **Proceedings of the International Conference on New Interfaces for Musical Expression**. London, United Kingdom: Goldsmiths, University of London, 2014. p. 195–200.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, v. 36, p. 1627–1639, 1964.

SCHARSTEIN, D.; SZELISKI, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. **International Journal of Computer Vision**, v. 47, n. 1-3, p. 7–42, 2002.

SCHRAMM, R.; JUNG, C. Temporally coherent stereo matching using kinematic constraints. In: **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. Florence, Italy: IEEE, 2014. p. 554–558.

SCHRAMM, R.; JUNG, C. R. A tool for teaching musical metrics based on computer vision. In: **Computer Graphics International**. Petrópolis: CGI, 2007. p. 71–78.

SEGEN, J.; KUMAR, S.; GLUCKMAN, J. Visual interface for conducting virtual orchestra. **International Conference on Pattern Recognition (ICPR)**, IEEE Computer Society, Los Alamitos, CA, USA, v. 1, p. 1276, 2000.

SHOTTON, J. et al. Real-time human pose recognition in parts from single depth images. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Providence, RI: IEEE, 2011. p. 1297 – 1304.

SIGAL, L. et al. Human attributes from 3d pose tracking. In: DANIILIDIS, K.; MARAGOS, P.; PARAGIOS, N. (Ed.). **Computer Vision – ECCV 2010**. [S.l.]: Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6313). p. 243–257.

SWANWICK, K. **Musical Knowledge, Intuition, Analysis and Music Education.** Londres: Routledge, 1994.

TAN, H. L. et al. Audio onset detection using energy-based and pitch-based processing. In: **Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)**. Paris: IEEE, 2010. p. 3689–3692.

TILLMANN, B. Music cognition: Learning, perception, expectations. In: KRONLAND-MARTINET, R.; YSTAD, S.; JENSEN, K. (Ed.). **Computer Music Modeling and Retrieval. Sense of Sounds**. [S.l.]: Springer Berlin Heidelberg, 2008, (Lecture Notes in Computer Science, v. 4969). p. 11–33. ISBN 978-3-540-85034-2.

TOH, L.-W.; CHAO, W.; CHEN, Y.-S. An interactive conducting system using kinect. In: **IEEE International Conference on Multimedia and Expo (ICME)**. San Jose, CA: IEEE, 2013. p. 1–6. ISSN 1945-7871.

TRAN, C.; TRIVEDI, M. Human body modelling and tracking using volumetric representation: Selected recent studies and possibilities for extensions. In: **Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)**. Stanford: IEEE, 2008. p. 1–9.

UCHIDA, S. et al. Non-markovian dynamic time warping. In: **21st International Conference on Pattern Recognition (ICPR)**. Tsukuba: IEEE, 2012. p. 2294–2297. ISSN 1051-4651.

VIITANIEMI, T.; KLAPURI, A.; ERONEN, A. A probabilistic model for the transcription of single-voice melodies. In: **Tampere University of Technology**. [S.l.: s.n.], 2003. p. 59–63.

W., S. J.; MOON, K. S. Procrustes analysis and its application to sensor integration. **Transactions of NAMRI/SME**, v. 20, p. 347–354, May 1992.

WANG, J. et al. Mining actionlet ensemble for action recognition with depth cameras. In: **IEEE Conference on Computer Vision and Pattern Recognition, 2012**. Providence, RI: IEEE, 2012. p. 1290–1297. ISSN 1063-6919.

WANG, J.-S.; CHUANG, F.-C. An accelerometer-based digital pen with a trajectory recognition algorithm for handwritten digit and gesture recognition. **IEEE Transactions on Industrial Electronics**, IEEE, v. 59, n. 7, p. 2998–3007, 2012. ISSN 0278-0046.

WANG, L.; CHENG, L.; WANG, L. Elastic sequence correlation for human action analysis. **IEEE Transactions on Image Processing (TIP)**, v. 20, n. 6, p. 1725–1738, 2011.

WANG, L.; SUTER, D. Informative shape representations for human action recognition. In: **18th International Conference on Pattern Recognition (ICPR)**. Hong Kong: IEEE, 2006. v. 2, p. 1266–1269. ISSN 1051-4651.

WEBB, A. R. **Statistical Pattern Recognition**. 3. ed. Chichester,UK: Wiley, 2011.

WÖLLNER, C. et al. The perception of prototypical motion: Synchronization is enhanced with quantitatively morphed gestures of musical conductors. **Journal of Experimental Psychology: Human Perception and Performance**, v. 38, n. 6, p. 1390–1403, December 2012.

YU, X. et al. An audio retrieval method based on chromagram and distance metrics. In: **International Conference on Audio Language and Image Processing (ICALIP)**. Shanghai: IEEE, 2010. p. 425–428.

ZHANG, Y.; EDGAR, T. A robust dynamic time warping algorithm for batch trajectory synchronization. In: **American Control Conference**. Seattle, WA: IEEE, 2008. p. 2864–2869. ISSN 0743-1619.

ZHAO, Y.; TAUBIN, G. Real-time stereo on gpgpu using progressive multi-resolution adaptive windows. **Image and Vision Computing**, v. 29, n. 6, p. 420 – 432, 2011. ISSN 0262-8856.

ZHOU, F.; FRADE, F. D. la T. Canonical time warping for alignment of human behavior. In: **Advances in Neural Information Processing Systems Conference (NIPS)**. Vancouver, Canada: Curran Associates, Inc., 2009. p. 2286–2294.

ZHOU, F.; FRADE, F. D. la T. Generalized time warping for multi-modal alignment of human motion. In: **IEEE Conference on Computer Vision and Pattern Recognition**. Providence, RI: IEEE, 2012. p. 1282 – 1289.

ZITNICK, L. C. et al. High-quality video view interpolation using a layered representation. **ACM Transactions on Graphics**, ACM, New York, NY, USA, v. 23, n. 3, p. 600–608, August 2004.

ZUFFI, S. et al. Estimating human pose with flowing puppets. In: **IEEE International Conference on Computer Vision (ICCV)**. Sydney, NSW: IEEE, 2013. p. 3312–3319. ISSN 1550-5499.

APÊNDICE A — ARTIGOS PRODUZIDOS

Especificamente relacionados ao tema aqui abordado, foram produzidos três artigos nos últimos dois anos, sendo que dois já foram publicados (ambos Qualis A1), e outro já foi aceito para publicação em conferência internacional (Qualis A1).

1. Artigo publicado, relacionado ao Capítulo 3, apresenta um método desenvolvido para gerar mapas de profundidade com coerência temporal. Essa técnica foi substituída posteriormente pelo uso do sensor Kinect (Qualis A1):

   - Schramm, R.; Jung, C.R., **Temporally coherent stereo matching using kinematic constraints**, Acoustics, Speech and Signal Processing (ICASSP), Italy, 2014 IEEE International Conference on , vol., no., pp.554,558, 4-9 May 2014.

2. Artigo publicado, descrevendo o método proposto para classificação e avaliação dos gestos de marcação de compassos. O conteúdo do artigo é relacionado ao Capítulo 3 (Qualis A1):

   - Schramm, R.; Rosito Jung, C.; Reck Miranda, E., **Dynamic Time Warping for Music Conducting Gestures Evaluation**, Multimedia, IEEE Transactions on , vol.17, no.2, pp.243,255, Feb. 2015.

3. Artigo aceito para publicação, submetido para a conferência internacional ISMIR 2015, com conteúdo relativo ao Capítulo 4 (Qualis A1):

   - Schramm, R; Nunes, H; Rosito Jung, C; **Automatic Solfège Assessment**. International Society for Music Information Retrieval (ISMIR), Malaga, 2015, 16th International Conference on, a ocorrer em 26-30 de Outubro de 2015.