DETECÇÃO DE RÉPLICAS EM EVIDÊNCIAS DE ÁUDIO USANDO UM ESQUEMA ADAPTATIVO DE *AUDIO FINGERPRINTING*

RODRIGO GURGEL FERNANDES TÁVORA

TESE DE DOUTORADO EM ENGENHARIA DE SISTEMAS ELETRÔNICOS E DE AUTOMAÇÃO
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

DETECÇÃO DE RÉPLICAS EM EVIDÊNCIAS DE ÁUDIO USANDO UM ESQUEMA ADAPTATIVO DE *AUDIO FINGERPRINTING*

RODRIGO GURGEL FERNANDES TÁVORA

ORIENTADOR: FRANCISCO ASSIS DE OLIVEIRA NASCIMENTO

TESE DE DOUTORADO EM ENGENHARIA DE SISTEMAS ELETRÔNICOS E DE AUTOMAÇÃO

PUBLICAÇÃO: PGEA.TD-123/17
BRASÍLIA/DF: OUTUBRO - 2017

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

DETECÇÃO DE RÉPLICAS EM EVIDÊNCIAS DE ÁUDIO USANDO UM ESQUEMA ADAPTATIVO DE ÁUDIO FINGERPRINTING

RODRIGO GURGEL FERNANDES TÁVORA

TESE DE DOUTORADO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR.

APROVADA POR:

FRANCISCO ASSIS DE OLIVEIRA NASCIMENTO, Dr., ENE/UNB
(ORIENTADOR)

ALEXANDRE RICARDO SOARES ROMARIZ, Dr., ENE/UNB
(EXAMINADOR INTERNO)

JOSÉ ANTÔNIO APOLINÁRIO JUNIOR, Dr., IME
(EXAMINADOR EXTERNO)

JORGE CARLOS LUCERO, Dr., CIC/UNB
(EXAMINADOR EXTERNO)

Brasília, 10 de outubro de 2017.

FICHA CATALOGRÁFICA

TÁVORA, RODRIGO GURGEL FERNANDES
Detecção de Réplicas em Evidências de Áudio Usando um Esquema Adaptativo de Audio Fingerprinting [Distrito Federal] 2017.
xix, 143, 210 x 297 mm (ENE/FT/UnB, Doutor, Engenharia Elétrica, 2017).
Tese de doutorado - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Audio Forensics	2. Passive Authentication
3. Replica detection	4. Audio Fingerprinting
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

TÁVORA, RGF (2017). *Detecção de Réplicas em Evidências de Áudio Usando um Esquema Adaptativo de Audio Fingerprinting.* Tese de doutorado, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 143 p.

CESSÃO DE DIREITOS

AUTOR: RODRIGO GURGEL FERNANDES TÁVORA
TÍTULO: Detecção de Réplicas em Evidências de Áudio Usando um Esquema Adaptativo de Audio Fingerprinting.
GRAU: Doutor ANO: 2017

Rodrigo Gurgel Fernandes Távora
SMAS Trecho 1 Lote C
Cond. Living Park Sul Bloco E Apto 602
71.218-010 - Brasília - DF - Brasil

Como pai, compreendo o sentimento de realização com o sucesso dos filhos. Por isso e por justiça, ofereço este trabalho aos meus pais, como reconhecimento por toda dedicação a nossa educação, minha e de meus irmãos. Em especial, enalteço o exemplo, o apoio e a confiança constantes do meu pai, meu maior motivador.

AGRADECIMENTOS

Agradeço ao colega André Luiz Morisson da Costa, por ter me incentivado, há alguns anos, quando chefiava o Setor de Perícias Audiovisuais no Instituto Nacional de Criminalística, a desenvolver um método para detecção de réplicas curtas de áudio. Agradeço também ao Professor Francisco Assis de Oliveira Nascimento, pela valiosa orientação neste trabalho. Sobretudo, agradeço-lhe pela humanidade e pela confiança em mim depositada, quando precisei vir a Brasília em 2013 por questões familiares, tendo naquela circunstância aceitado me orientar em um Projeto de Pesquisa, que acabou se tornando a base para o presente trabalho. Por fim, agradeço ao meu irmão Bruno, por dividir comigo seu tempo e gosto pela matemática em frutuosas discussões.

RESUMO

Este trabalho aborda o problema de autenticação passiva de áudio e objetiva propor um método automático para detecção de edições fraudulentas produzidas através da replicação de trechos curtos de sinal dentro de uma mesma evidência de áudio. O método proposto é baseado em um esquema adaptativo de *Audio Fingerprinting*. Diversos sistemas de *Audio Fingerprinting* existentes são analisados, e, conforme os requisitos estipulados para a aplicação forense, de elevada robustez e usabilidade, uma abordagem de *Audio Fingerprinting* binária baseada na distribuição do espectro de Fourier é escolhida. Um sistema adaptativo é proposto, o qual é ajustado teoricamente e empiricamente para cada evidência de áudio. As simulações mostram uma robustez do método contra distorções no domínio do tempo e da frequência. A capacidade de discriminar áudios correspondentes a um mesmo texto e diferenciá-los de réplicas também é analisada. Novas modificações são propostas, como o emprego de um critério de dupla detecção, e o sistema final obtido demonstrou ser aplicável a áudios de longa duração e robusto contra mascaramentos por inserção de ruído.

Palavras-chave: Autenticação passiva de áudio, *Audio Fingerprinting*, detecção de réplicas curtas, análise forense de áudio.

ABSTRACT

This work addresses the problem of passive audio authentication and aims to propose an automatic method to detect forgeries produced by the replication of an audio signal within the same audio evidence. The proposed method uses an adaptive Audio Fingerprinting system. Several existing systems are analyzed, and, according to the defined requirements of usability and robustness against masking distortions, an adaptive binary Audio Fingerprinting scheme based on the Fourier spectrum distribution is chosen. An adaptive system is proposed, which is theoretically and empirically adjusted for each audio evidence. Simulations show that the designed system is robust against time and frequency-domain distortions. The power to discriminate repeated text speech and distinguish it from audio replicas is also analyzed. Further adjustments are suggested, such as the use of a double detection criteria, and the final scheme was able to detect short replicas, distorted by noise insertion, even within long audio evidences.

Key words: passive authentication, Audio Fingerprinting, short replica detection, audio forensic analysis.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AAC	Advanced Audio Coding
AF	Audio Fingerprint
AMR	Adaptive Multi-Rate
BER	Bit Error Rate
CQT	Constant Q Transform
CMN	Cepstral Mean Normalization
DCT	Discrete Cosine Transform
DFA	Detrented Fluctuation Analysis
DFT	Discrete Fourier Transform
DTW	Dinamic Time Warping
DWT	Discrete Wavelet Transform
DML	Distance Metric Learning
EMS	Exact Match Search
ENF	Electrical Network Frequency
FBE	Filter Bank Energies
FIR	Finite Impulse Response
GCC-PHAT	Generalized Cross Correlation Phase Transform
GMM	Gaussian Mixture Models
GSM	Global System for Mobile Communications
HMM	Hidden Markov Models
IIR	Infinite Impulse Response
K-NN	K- Nearest Neighbors
LPC	Linear Predictive Coding
MCLT	Modulated Complex Lapped Transform
MDCT	Modified Discrete Cosine Transform
MFCC	Mel Frequency Cepstral Coefficients
MP3	MPEG-1/MPEG-2 Audio Layer III

MPEG	Motion Picture Experts Group
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PEAQ	Perceptive Evaluation of Audio Quality
PLP	Perceptual Linear Prediction
PSOLA	Pulse Syncronization Overlap and Add
PS-ZCPA	Pitch Syncronous Zero Crossing Peak Amplitudes
RPDE	Recurrence Probability Density Entropy
ROC	Receiver Operating Characteristic
SNR	Signal to Noise Ratio
SCM	Spectral Crest Measure
SFM	Spectral Flatness Measure
SVD	Singular Value Decomposition
STFT	Short Time Fourier Transform
SURF	Speeded Up Robust Features
WMA	Windows Media Audio
WPD	Wavelet Packet Decomposition
ZCR	Zero Crossing Rate

LISTA DE SÍMBOLOS

Símbolos Latinos

$s[i]$	Sinal discreto no tempo e na amplitude
R	Taxa de amostragem
$S[n,k]$	Transformada de Curto Termo de $s[i]$
D	Duração
N	Número Inteiro
L_i	Limite da i-ésima sub-banda
Fl, Fh	Limites inferior e superior da banda de frequência
W, V, T	Variáveis intermediárias no cálculo de $F[n,m]$
F	Vetor de *Audio Fingerprinting*
FN	Falso Negativo
FP	Falso Positivo
FNQ	Falso Negativo de Quadro
FPQ	Falso Positivo de Quadro
FNR	Falso Negativo de Réplica
FPR	Falso Positivo de Réplica
$\mathbf{M}$	Matriz de autossimilaridade
$Pr\{\}$	Probabilidade
$\mathcal{N}$	Ruído

Símbolos Gregos

Δ	Espaçamento
Ω	Fator de sobreposição/interseção
α	Expoente usado na soma de coeficientes de sub-bandas
δ	Distância/Métrica

Subscritos

A	Áudio
R	Réplica
F	Quadros-*Frames*
FP	Falso Positivo
bit	Bits

1- INTRODUÇÃO

> Absence of evidence is not evidence of absence.
>
> Carl Sagan
>
> Information is the resolution of uncertainty.
>
> Claude Shannon

O presente trabalho apresenta uma ferramenta automática de análise de autenticidade de áudio forense que permite a detecção de réplicas curtas em evidências de áudio. Apesar de serem facilmente realizadas através de software, as edições em áudio dificilmente são identificadas pelo examinador com base em um único método de análise. Dessa forma, este trabalho apresenta uma nova técnica automática para análise de autenticidade de áudio, que pode subsidiar o perito forense com mais informação no seu exame, aumentando sua eficácia e reduzindo sua duração. Mesmo considerando o crescente interesse da comunidade acadêmica de processamento de sinais em aplicações forenses como análise de imagem e áudio, as análises e os desafios enfrentados pelos peritos em áudio forense são ainda pouco conhecidos do público externo a este meio. Dessa forma, fazemos no Apêndice A uma breve introdução das áreas de fonética e acústica forense e apresentamos um resumo dos trabalhos referentes à autenticação de áudio forense. Apesar dos avanços, a maioria dos métodos de autenticação de áudio não é robusta a todos os tipos de mascaramentos de edições, como compressão de áudio com perda, inserção de ruído ou filtragem em frequência. A descrição de alguns métodos pode ser vista no Apêndice A. As evidências de áudio podem ser manipuladas de várias formas, como pela supressão, inserção ou emenda por inversão de ordem de trechos, mas um tipo específico de edição fraudulenta através da replicação de locuções curtas, como o advérbio de negação "não", pode inverter completamente o sentido original da sentença. A motivação e a proposta deste trabalho são apresentadas na Seção 1.1. A organização da tese é descrita na Seção 1.2.

1.1 MOTIVAÇÃO

A **verificação perceptual** da presença de réplicas de trechos em uma evidência de áudio é feita pela oitiva atenta do áudio, em conjunto com a análise visual de gráficos da forma de onda, da envoltória do sinal ou do espectrograma, pesquisando trechos de

áudio semelhantes ou descontinuidades decorrentes das emendas. Em geral, para facilitar a análise audiovisual, são empregados gráficos deslizantes sincronizados com a reprodução do áudio. Emprega-se uma janela de visualização com alguns segundos de duração, tipicamente de 5s, longa o suficiente para possibilitar o reconhecimento perceptual do conteúdo, e curta o suficiente para que os detalhes da variação temporal da envoltória de amplitude e das características do espectrograma sejam observados.

Para exemplificar a dificuldade da análise visual, a Figura 1.1 ilustra a forma de onda e o espectrograma de um trecho de 5s de sinal de voz com alta relação sinal/ruído, onde o trecho inicial de 1s, destacado com fundo escuro, foi replicado para um trecho na parte final, também destacado. Pela análise visual, um examinador atento poderia identificar tanto uma descontinuidade da distribuição espectral nos pontos de emenda, com descasamento dos harmônicos da frequência fundamental, quanto a semelhança da imagem do espectrograma e da envoltória de sinal entre o trecho original e o trecho replicado. Cabe destacar que em um caso real de edição fraudulenta por replicação de trechos de áudio, dificilmente a defasagem entre um trecho original e o replicado seria menor que os 5s da janela de visualização comumente empregada na análise perceptual. Portanto, para uma janela de análise curta, de 5s, em geral, apenas as descontinuidades nas emendas do trecho replicado podem ser identificadas pelo examinador. Por outro lado, o emprego de janelas de visualização longas inviabilizaria a comparação visual de detalhes do espectrograma de trechos tão curtos quanto 1s. Ademais, o áudio editado pode ser mascarado, o que dificulta ainda mais a identificação visual dessas emendas.

Para réplicas com duração tão curtas quanto 100ms, que foi estimada como a duração mínima de uma locução do advérbio de negação "não" para uma taxa rápida de elocução, mesmo considerando que os trechos original e replicado estejam contidas dentro da janela de visualização de 5s, a identificação visual da réplica é bem mais difícil. Essa dificuldade é exemplificada na Figura 1.2, que ilustra a forma de onda e o espectrograma de um sinal de voz de 5s, com um trecho inicial de 100ms replicado para um trecho em uma posição aleatória, ambos em destaque com fundo escuro.

Portanto, a pesquisa através da análise visual de trechos repetidos de 100ms em evidências com alguns minutos de duração é muito difícil mesmo para um examinador atento. Em verdade, o sucesso deste tipo de análise depende muito mais da percepção de sons similares através da memória auditiva. Pela experiência do autor na realização de exames dessa natureza, observa-se que a análise de oitiva é bem mais eficiente em identificar trechos de áudio repetidos, mesmo que defasados de alguns segundos. Curiosamente, esta maior facilidade em reter a informação de padrões de áudio, comparada à persistência da informação de padrões de imagens, é compartilhada por alguns cientistas renomados, como Jacques Hadamard: "Eu tenho uma pior memória de fisionomias e sou mais exposto a esquecimentos ou falsos reconhecimentos, ao contrário, sou muito sensível ao som das palavras [...], eu sou menos sensível à semelhança de faces e mais sensível à semelhança

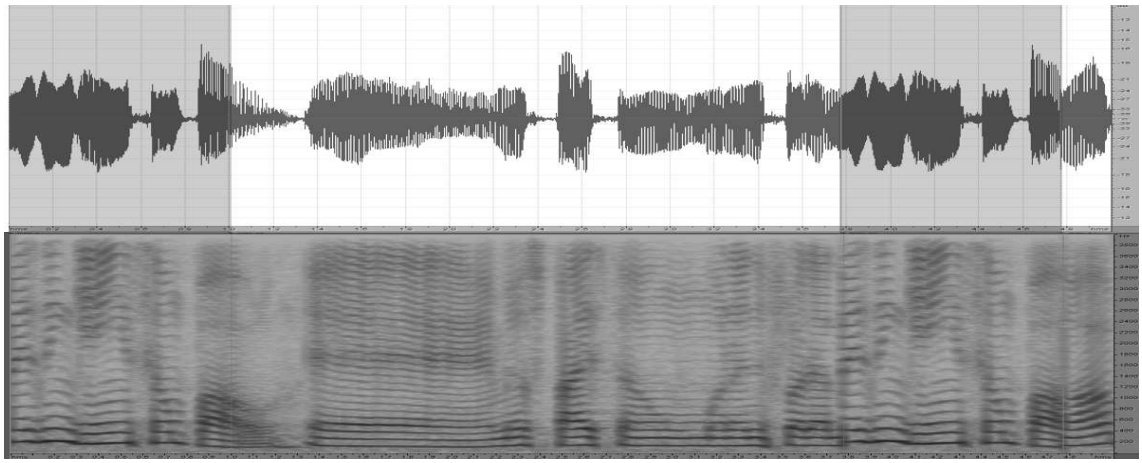

Figura 1.1: Forma de onda e espectrograma de trecho de voz de 5s, com o trecho inicial de 1s replicado para outro trecho em posição aleatória, ambos em destaque.

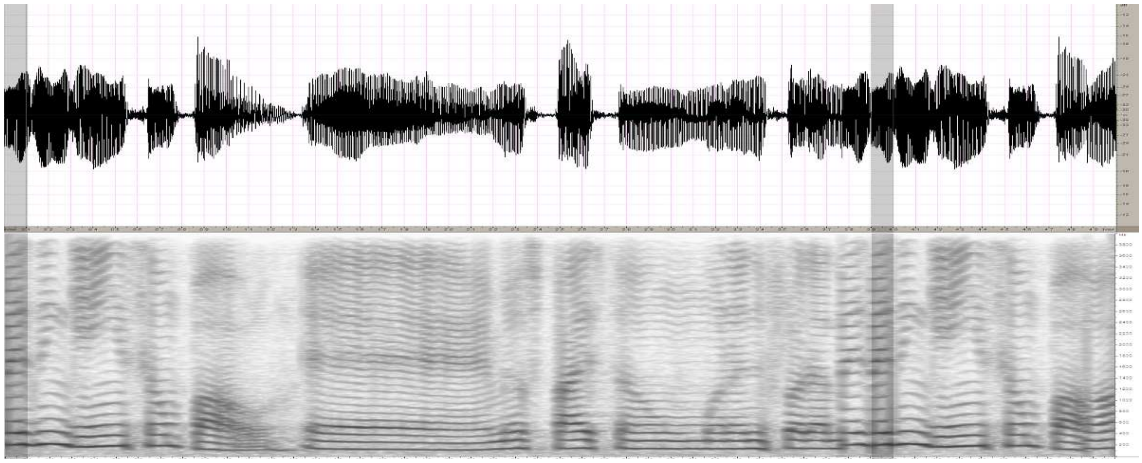

Figura 1.2: Forma de onda e espectrograma de trecho de voz de 5s, com o trecho inicial de 100ms replicado para outro trecho em posição aleatória, ambos em destaque.

de vozes."[3]. Em [4] a persistência da memória de curto prazo de padrões de áudio é analisada.

Para estimar a dificuldade de detecção de réplicas sem o emprego de um método automático, um experimento de análise perceptual foi realizado com um grupo de vinte e seis Peritos Criminais Federais participantes de um treinamento de autenticação de áudio no Instituto Nacional de Criminalística em 2016. Nos testes foi usado um áudio com alta qualidade, com SNR (*Signal-to-noise Ratio*) estimada de 70dB, oriundo de coleta de padrão de voz para exame de comparação de locutor, contendo a voz referente a texto não-controlado de um único falante, e cujos trechos de silêncio com duração superior a 100ms foram suprimidos. Foram preparados áudios com durações de um minuto e de três minutos. Cada áudio continha três réplicas de trecho de áudio referente a uma ou mais palavras, com durações de 100ms, 300ms ou 1s, feitas manualmente para minimizar inconsistências semânticas, mas sem nenhum mascaramento posterior. Aos participantes foi

concedido um tempo para análise perceptual dez vezes maior que a duração do áudio, ou seja, dez minutos para o áudio de um minuto, e trinta minutos para o áudio de três minutos. Sejam o Verdadeiro Positivo (VP) a detecção de pares de intervalos correspondentes às réplicas, o Falso Negativo (FN) a não detecção de pares de intervalos correspondentes às réplicas, Verdadeiro Negativo (VN) a não detecção de intervalos distintos das posições das réplicas, e Falso Positivo (FP) a detecção de pares de intervalos de áudios, notadamente intrasentença, não correspondentes às réplicas. A Taxa de Verdadeiro Positivo (*Recall*=VP/(VP+FN)) e a Taxa de Falso Positivo(TFP=FP/(FP+VN)) são ilustrados na Tabela 1.1.

Tabela 1.1: Detecção perceptual de réplicas em áudios de um ou três minutos de duração.

Duração do Áudio	*Recall* Réplica 1s	*Recall* Réplica 300ms	*Recall* Réplica 100ms	TFP
1 minuto	38,4%	7,7%	0%	19,2%
3 minutos	19,2%	15,4%	0%	23,1%

Os resultados ilustram a dificuldade da análise perceptual, notadamente para detecção de réplicas muito curtas, como as réplicas de 100ms que não foram detectadas por nenhum participante. Mostram ainda que a dificuldade da análise aumenta com a duração da evidência de áudio. Observou-se ao longo do experimento que a maioria dos participantes usava apenas a oitiva atenta, que fornece em média uma melhor persistência que a memória visual, para tentar identificar trechos semelhantes de áudio. As falsas detecções de réplica revelam ainda que, mesmo para evidências de áudios não muito longas, pode ser difícil para o examinador identificar perceptualmente se trechos de voz semelhantes consistem de réplicas de áudio ou de locuções distintas, referentes a uma mesma sentença e produzidas pelo mesmo locutor (voz intrasentença e intralocutor). Esta análise confirma na prática a relevância do emprego de métodos automáticos para detecção de réplicas curtas dentro de evidências longas de áudio.

Para a **verificação automática** da existência de réplicas tão curtas quanto 100ms em uma evidência de áudio, uma comparação sequencial de trechos do sinal usando a diferença das amostras no domínio do tempo como medida de similaridade não conseguiria detectar réplicas mascaradas por distorções. O emprego de técnicas de reconhecimento de voz, seguidas de busca por texto repetido, também não teria um bom desempenho para áudios com baixa relação sinal/ruído (SNR) e detectaria erroneamente áudios referentes a uma mesma sentença (intrasentença) como sendo réplicas.

A detecção de máximos locais da autocorrelação do sinal de áudio para a detecção de réplicas de áudio não seria aplicável a réplicas com duração muito mais curta que a duração da evidência de áudio. Trechos curtos de réplicas de 100ms em áudios de 60s não geram máximos locais detectáveis na autocorrelação. Na prática, evidências de áudio

costumam ter duração bem superior a 60s. Em [5] é usada a segmentação dos quadros antes da análise de autocorrelação. Portanto, uma abordagem intuitiva é representar o áudio através da segmentação em quadros, seguida da extração de um conjunto reduzido de atributos perceptuais robustos para representar cada intervalo de áudio, reduzindo assim a dimensão dos dados usados para representar o áudio em relação ao número de amostras do sinal de áudio. Na identificação de trechos semelhantes é feita uma busca, com base numa métrica, por atributos próximos. Esta abordagem, denominada de *Audio Fingerprinting*, descritor de áudio, detecção de áudio por conteúdo ou *hashing* robusto de áudio, tem sido aplicada principalmente à identificação e à verificação de integridade de música, onde trechos de áudio de alguns segundos são usados para representar uma música [1, 2].

Alguns trabalhos abordam a identificação de cópias de áudio para aplicações comerciais referentes a controle de direitos autorais. Em [6], cujo título sugere a detecção de cópias de áudio, somente réplicas com duração superior a 10s são detectadas. Em [7] é proposto um esquema para detecção de réplicas de eventos em áudios de longa duração, para a aplicação de análise de áudio de longa duração, que registram o dia-a-dia de um indivíduo. Uma dupla detecção é usada com uma janela de 2s, logo o método não seria capaz de detectar trechos de 100ms. Ademais, o desempenho relatado é bom apenas para eventos de áudio estruturado, com componentes espectrais bem definidos, como música ou tons de discagem de telefonia. Para trechos de áudio com voz, o desempenho reportado para o método foi muito baixo. Em [8], para identificação de trechos ou estruturas repetidas dentro de uma música, é usada uma dupla detecção de quadros com mesma defasagem na matriz de autossimilaridade. Em [9] são analisados diversos atributos para detecção de estruturas repetidas em músicas, e filtros morfológicos de erosão, dilação, abertura e fechamento são aplicados à matriz de autossimilaridade. A detecção de trechos iguais entre duas músicas também foi proposta, como em [10] onde o uso de AF é proposto para alinhamento de versões de áudios com distorção em escala de tempo, ou em [11, 12, 13, 14, 15] onde esquemas de AF são usados para sincronização de duas mídias. Alguns desses métodos geram matrizes de autossimilaridade com elevado número de falsos positivos de quadros. Dessa forma, alguns critérios de dupla detecção são aplicados sobre a matriz de autossimilaridade, empregando estatísticas de coeficientes de linha, transformada de Hough ou filtros morfológicos de imagem. Entretanto, todos os métodos estudados empregam granularidades superiores a 1s, e, portanto, não são capazes de detectar réplicas curtas.

Surpreendentemente, apesar dos métodos para detecção de réplicas em imagem terem sido propostos há algum tempo [16, 17], nenhum trabalho acerca da identificação de réplicas de áudio forense fora encontrado na literatura científica até a primeira publicação do trabalho desta tese em (TAVORA, 2014) [18]. Posteriormente, alguns trabalhos abordaram o assunto. Em [19] um método proposto para identificação de réplicas cur-

tas segmenta o áudio, calcula a similaridade entre os segmentos e aplica uma limiar de detecção. Em [20], a detecção de réplicas em evidências utiliza a curva de pitch como único atributo. O áudio é dividido em segmentos vozeados e as sequências são comparadas usando o coeficiente de correlação de Pearson. Os testes de robustez usam uma base de áudios curtos de 30s e réplicas com duração a partir de 600ms. Curiosamente, o desempenho é comparado com o uso do esquema de *Audio Fingerprinting* [2], sem citar [18], que propusera essa abordagem para a aplicação forense. É reportado um melhor desempenho para uso de pitch, mas não é informado sequer o nível de ruído inserido. A metodologia usada na comparação não nivela o número de falsos positivos e não verifica se o número de falsos positivos para áudios mais longos é muito elevado, o que inviabilizaria a análise perceptual dos resultados. Ademais, o pitch é escolhido pela robustez conta inserção de ruído de mascaramento, muito embora o próprio áudio original seja comumente severamente comprimido ou possua relação sinal ruído abaixo de 10dB, o que dificulta a detecção correta do pitch e de trechos vozeados. Em [21], para a detecção de duplicação de trechos de áudio ou de partes de imagens, é proposto o uso Momentos de Chebichef após a segmentação. Apesar da aplicação forense não possuir um requisito de uso em tempo real, o trabalho apresenta um algoritmo rápido de cálculo dos atributos. O desempenho é medido, com 76% de detecção para uma taxa de falso positivo alta de 29% para áudios mascarados com SNR acima que 15dB, e com taxas de compressão MP3 não informadas. Entretanto, nem a duração dos áudios nem a duração das réplicas são informadas.

Para a especificação de um esquema de *Audio Fingerprinting* (AF) adequado à aplicação de detecção de réplicas curtas dentro de uma evidência de áudio, alguns requisitos principais são definidos:

- Boa localização temporal dos atributos de AF. O esquema de AF deve ser capaz de detectar réplicas tão curtas quanto 100ms, que corresponde à duração mínima estimada de uma locução do advérbio de negação "não", para uma taxa de elocução rápida.

- Boa precisão. Alguns esquemas de AF utilizam trechos longos, com vários quadros de AF, para identificar um áudio. Neste casos, como vários quadros podem identificar o áudio, mesmo com uma baixa taxa de detecção de quadros é possível se obter uma boa taxa de detecção de áudio. Na aplicação forense em tela, a taxa de detecção de quadros de AF deve ser alta, uma vez que as réplicas podem ser tão curtas quanto um único quadro segmentado do áudio. O método automático de detecção de réplicas viabiliza a análise de áudios mais longos, mas não substitui totalmente a análise perceptual do examinador, mesmo porque outros métodos automáticos podem ser empregados em paralelo, e a decisão final acerca da hipótese de adulteração deve se basear na composição dos resultados. Dessa forma, como

a decisão depende de intervenção do perito, a usabilidade do método deve ser um requisito importante. Neste sentido, o método deve apresentar uma baixa taxa de falsa detecção de réplicas, já que cada um dos pares de trechos de áudio detectados como possíveis réplicas deve ser analisado perceptualmente. Neste trabalho procuramos ajustar os parâmetros do sistema proposto com base neste critério, limitando em 10 o número esperado de falsas detecções de réplica por áudio analisado.

- Boa robustez. O método deve identificar também trechos distorcidos, uma vez que as réplicas podem ser muito curtas e mascaradas por compressão de áudio, filtragem em frequência, conversão D/A, ceifamento de sinal, escala no domínio do tempo, ou inserção de ruído.

Como as réplicas podem ser muito curtas e mascaradas, a aplicação de detecção de réplicas curtas impõe requisitos de localização temporal, robustez e precisão mais fortes que a aplicação de identificação de música por conteúdo. Com exceção de [22, 23], onde o áudio é segmentado em quadros de 64ms, todos os esquemas estudados empregam quadros mais longos que 1s e não detectam réplicas tão curtas quanto $100ms$. Com relação à precisão, observou-se que os métodos analisados em [1, 24, 25] utilizam esquemas de AF com dimensão constante, uma vez que são projetados para analisar de maneira uniforme um número indefinido de músicas. Para a detecção de réplicas em um áudio, o número total de quadros comparados é definido, logo a dimensão da AF, ao invés de constante, pode ser ajustada para limitar o número de falsos positivos de réplicas e viabilizar a análise perceptual dos resultados. Portanto, esta revisão evidenciou que os esquemas de AF existentes não são ajustados adequadamente à aplicação de detecção de réplicas curtas.

Dessa forma, o presente trabalho propõe um novo sistema adaptativo de *Audio Fingerprinting* com atributos binários, baseado na análise de sub-bandas do espectro de Fourier, robusto contra mascaramentos e projetado para a detecção de réplicas curtas de áudio em evidências longas de áudio. Para o ajuste dos parâmetros, através da otimização do desempenho de detecção, também são propostos novos métodos de cálculo de probabilidade de detecção baseada na integração binária com janela móvel.

Apesar de não abordar outras aplicações de esquemas de *Audio Fingerprinting*, uma ampla revisão dos modelos, etapas e aplicações desses esquemas é feita no Capítulo 2. Algumas dessas aplicações, como o uso de AF para melhoria de áudio pela sincronização de ruído de fundo; a detecção de trechos repetidos em músicas; ou o uso de AF em marca d'água digital de áudio, nas quais um único arquivo de áudio é analisado, possuem potencial de emprego do método adaptativo proposto, pela possibilidade de ajuste dos parâmetros de AF para cada áudio.

1.2 ORGANIZAÇÃO DESTA TESE

Para subsidiar a escolha de um modelo de *Audio Fingerprinting* para a aplicação forense, no Capítulo 2 é feita uma ampla revisão da literatura científica, onde são descritas as etapas dos sistemas de extração e busca de *Audio Fingerprinting*. O modelo geral é descrito, e diversas abordagens são citadas. Algumas aplicações de *Audio Fingerprinting* são descritas. A leitura deste capítulo é recomendada, pois apresenta a terminologia empregada no restante da tese. Como não é considerada uma leitura essencial à compreensão deste trabalho, a descrição mais detalhada de alguns esquemas de *Audio Fingerprinting* é feita no Apêndice B. De acordo com a análise comparativa de propriedades, como robustez e desempenho de detecção, a abordagem proposta pela PHILIPS é escolhida para emprego na aplicação forense. Portanto, o esquema binário baseado na análise de sub-bandas do espectro de Fourier proposto por Hatisma em [2] é descrito, e um modelo teórico proposto para este esquema é revisado.

Nos capítulos subsequentes, é apresentado passo a passo o desenvolvimento do sistema de detecção de réplicas curtas para a aplicação forense, para manter a consistência com os resultados apresentados para o método inicialmente proposto por (TAVORA,2015) [26].

No Capítulo 3, que detalha o trabalho apresentado em (TAVORA, 2015) [26], é proposta a primeira versão do sistema, baseado na abordagem da PHILIPS, onde parâmetros como a duração dos quadros e a dimensão são adaptados conforme os requisitos da aplicação de detecção de réplicas curtas de áudio. A dimensão é ajustada com base numa análise teórica e empírica, onde o sistema é ajustado para limitar o número de falsas detecções. Com o objetivo de melhorar a unicidade da representação, é proposta uma divisão de sub-bandas adaptada para cada evidência de áudio. Simulações são realizadas, usando inicialmente o *corpus* do Instituto Nacional de Criminalística, que contém amostras de voz com alta relação sinal/ruído estimada em 65dB e com repetições de locuções intrasentença. São feitos testes para verificar a capacidade de discriminação de locuções intralocutor e intrasentença. A robustez do sistema para a detecção de réplicas mascaradas com diversos tipos de distorções também é avaliada. O sistema apresentado é robusto contra diversas distorções, entretanto, apresenta uma robustez regular contra inserção de ruído.

No Capítulo 4, é proposta uma nova metodologia de testes e análise de unicidade, precisão e robustez. Com o objetivo de melhorar a robustez do sistema inicialmente proposto contra inserção de ruído, é proposta a adaptação para cada áudio do expoente dos componentes espectrais nos somatórios das sub-bandas, além do ajuste dos limites da banda de frequência utilizada. Em seguida, parâmetros, como duração de quadro, fator de sobreposição e a distância dos deltas entre quadros ou sub-bandas, são otimizados para maximizar a taxa de detecção de réplica. O sistema obtido é robusto contra inserção de ruído, mas os testes mostram que o ajuste da dimensão não viabiliza a redução dos falsos

positivos de réplica para áudios longos. As simulações foram feitas usando um *corpus* de acesso livre, CHAINS (*CHaracterizing INdividual Speakers*) [27], contendo vozes de 36 falantes com alta qualidade, SNR estimada em 70dB, e contendo repetições de locuções intrasentença. Para se construir um conjunto de teste com SNR baixa, foi inserido artificialmente ruído obtido do *corpus* também de acesso livre DEMAND (*Diverse Environments Multichannel Acoustic Noise Database*) [28].

No Capítulo 5, é feita a última modificação do método. Para reduzir o número de falsas detecções de réplicas e permitir a análise de áudios longos, é proposto um novo critério de decisão com base no integrador binário de janleas móveis para a detecção de réplica. A probabilidade de detecção de réplicas é obtida indiretamente pela taxa de detecção de quadros, através de um novo método aproximado de cálculo de probabilidade, proposto no Apêndice C. Todos os parâmetros do sistema são ajustados para otimizar a probabilidade de detecção de réplicas, para áudios curtos, de 60s de duração, e para áudios mais longos, de 240s de duração.

No Capítulo 6, os resultados são resumidos e analisados, e algumas linhas de pesquisa, além da aplicação forense, são sugeridas.

No Apêndice A, é feita uma breve introdução das áreas de fonética e acústica forense, e é apresentado um resumo dos trabalhos referentes à autenticação de áudio forense.

No Apêndice B é feita uma revisão e análise de desempenho de alguns sistemas de *Audio Fingerprinting*, para subsidiar a análise de adequabilidade à aplicação de detecção de réplicas.

No Apêndice C são propostos métodos, um exato e outro aproximado, para o cálculo da probabilidade de detecção pelo critério do integrador binário, que são empregados na análise do Capítulo 4.

O Apêndice D apresenta uma lista das publicações do autor referentes a este trabalho.

2- UMA REVISÃO DOS SISTEMAS DE *AUDIO FINGERPRINTING*

> May not music be described as the mathematics of the sense, mathematics as music of the reason? The musician feels mathematics, the mathematician thinks music: music the dream, mathematics the working life.
>
> James Joseph Sylvester

Além da música poder ser descrita *como* a matemática do sentido, tanto música quanto outros tipos de áudio podem ser descritos *pela* matemática, e de uma forma mais compacta, para fins de reconhecimento de padrões em diversas aplicações. Os sistemas de *Audio Fingerprinting*(AF) extraem uma assinatura compacta a partir da informação acústica perceptual mais relevante de um áudio não rotulado. O objetivo final é obter uma representação compacta por meio de um ou mais vetores obtidos a partir de atributos acústicos. A representação com uma dimensão baixa deve caracterizar univocamente o áudio, e ser robusta ao ponto de identificar até mesmo versões distorcidas do mesmo áudio que possuam informação perceptual auditiva semelhante.

Uma revisão não muito recente dos sistemas de *Audio Fingerprinting* é fornecida em [1, 29], onde muitas aplicações comerciais são descritas, como identificação e verificação de integridade de música. Desde então, diversos novos sistemas e novas aplicações foram propostos, conforme discutido na Seção 2.3.

Em [30] é feita uma revisão mais recente dos esquemas de *Audio Fingerprinting*, com detalhamento sobre as abordagens de extração e de modelamento de atributos. Apesar de ser recente, esta revisão não é tão abrangente.

Em [25] é feita uma revisão mais abrangente dos atributos acústicos empregados nas aplicações de identificação de música, transcrição perceptual de música, classificação de áudio ambiental, segmentação e reconhecimento de voz ou locutor. Apesar de mais amplo, devido ao maior número de aplicações pesquisadas, o levantamento não cobre todos os atributos propostos na literatura para as aplicações de *Audio Fingerprinting*. Uma interessante taxonomia, por níveis, dos atributos de áudio foi proposta nesse trabalho.

As propriedades básicas de um esquema de *Audio Fingerprint* são:

1. Dimensão compacta: O tamanho do vetor de AF, comumente denominado de *dimensionalidade*, deve ser compacto, o que reduz os requisitos de memória do sistema

e facilita o processo de busca, reduzindo o problema da *Maldição da Dimensionalidade*, descrito na Seção 3.1, que afeta o desempenho de métodos de busca em intervalo.

2. Precisão de detecção: Um bom esquema deve fornecer uma alta taxa de detecção verdadeira (TDV-*recall*) de réplicas não distorcidas, e deve garantir uma baixa taxa de falsas detecções (TFD) de trechos de áudio com conteúdo perceptual distinto. Uma baixa taxa de falsa detecção significa que a representação permite discriminar satisfatoriamente trechos de áudio distintos, uma propriedade comumente denominada de *unicidade*. A representação vetorial obtida deve apresentar uma baixa correlação entre seus componentes de forma que se obtenha uma representação compacta com elevado poder de discriminação. Em resumo, as distribuições devem possuir uma baixa variação intraclasse, para sons perceptualmente equivalentes, e uma alta variância interclasse, para sons perceptualmente distintos.

3. Robustez: Sinais de áudio com conteúdo perceptual acústico similar devem ser mapeados em uma representação equivalente. Logo, a representação da AF deve garantir uma certa invariância a manipulações comuns, como equalização ou conversão D/A-A/D, ou compressão de áudio, que não modifiquem o seu conteúdo perceptual.

4. Baixa complexidade: O método deve evitar processos de extração de atributos com um custo computacional muito elevado, sobretudo em aplicações que demandem processamento em tempo real.

5. Granularidade: Equivale à duração mínima de sinal de áudio necessária para que o esquema de *Audio Fingerprint* possa identificar o arquivo de áudio com uma alta taxa de detecção correta. Este parâmetro depende tanto da robustez da representação, quanto do método de busca utilizado.

Existe, obviamente, uma solução de compromisso entre as propriedades acima listadas. Como exemplo, a redução da dimensionalidade reduz a complexidade computacional, pode melhorar a robustez, mas uma representação muito compacta pode reduzir a precisão da detecção, com a elevação da taxa de detecções falsas.

Dessa forma, a escolha do modelo ou o ajuste destas propriedades deve ser feito de acordo com a aplicação. Em [31, 32] os cenários de aplicação de reconhecimento de áudio por conteúdo são classificados de acordo com a especificidade do áudio ou com a granularidade. Por exemplo, a identificação de música por meio de assobios, como feito pelo aplicativo SoundHound (TM) [33], ou performances distintas ou de diferentes intérpretes, requer o emprego de esquemas de AF mais robustos. Para a identificação de uma música inteira, uma alta granularidade é satisfatória, mas para a identificação de trechos de áudio uma pequena granularidade é necessária.

Como dito anteriormente, os requisitos de um esquema de AF para a detecção de réplicas curtas em evidências de áudio são: uma baixa granularidade com uma boa localização temporal dos atributos de áudio, além de uma precisão e uma robustez de detecção elevadas. Com o intuito de analisar os esquemas existentes para a escolha de uma abordagem que atenda a estes requisitos, foi realizada uma revisão abrangente e atual dos esquemas de AF.

O emprego de um esquema de AF divide-se, em geral, em dois processos: 1) Extração da AF, para um áudio ou trecho de áudio; 2) Busca por AF, para verificar a existência ou identificar um áudio ou trecho de áudio. Uma visão geral das etapas dos métodos de extração de AF é apresentada a seguir.

2.1 EXTRAÇÃO DE *AUDIO FINGERPRINT*

Um esquema genérico para extração de AF pode ser dividido nas etapas de pré-processamento, segmentação de quadros, extração de atributos, pós-processamento e modelamento de atributos. Estas etapas são descritas em detalhe a seguir, onde diversas abordagens são citadas.

2.1.1 Pré-processamento

Para a identificação de música através do conteúdo de áudio, a AF deve ser invariante a diversos tipos de distorções que preservem a informação perceptual auditiva, tais como distorções de canal e mudanças de formato que podem reduzir a largura efetiva de banda de frequência do sinal. Portanto, a filtragem passa-banda é comumente aplicada, como por exemplo de 50Hz a 4kHz, para uniformizar os sinais analisados. Em esquemas que empregam segmentação em quadros periódicos, a filtragem passa-banda por filtros LPC (*Linear Predictive Coding*) pode ser usada para remover as dependências de curto termo [5].

Quadros de áudio ao longo de trechos de silêncio podem ser mapeados para valores de AF aleatórios, devido à presença de ruído. Logo, para reduzir a taxa de falsa detecção de AF, um limiar de potência é em geral aplicado para identificar e remover os trechos de silêncio.

A maior parte dos esquemas de AF é aplicada à música, que é comumente codificada com modelos psicoacústicos de compressão que buscam remover a informação imperceptível ao ouvido humano, em virtude de limiares de percepção ou de mascaramentos temporais ou frequenciais. Portanto, o pré-processamento com modelos psicoacústicos de representação do sinal é uma alternativa tanto para redução da dimensionalidade quanto para o aumento da robustez contra transcodificação para codificadores perceptuais. A

extração dos atributos a partir do sinal codificado pode ainda reduzir a complexidade computacional do processo de extração de atributos. Um atributo denominado F1CC, baseado no modelo psicoacústico do CODEC Vorbis, é proposto em [34], para modelar melhor os efeitos de mascaramento do ouvido humano, aumentando, assim, a robustez contra distorções decorrentes de codificações perceptuais. Em [35, 36] é aplicado um limiar de potência, que simula a percepção em dB do ouvido humano. Em [37], é feita uma revisão de padrões de compressão usados na indexação de áudio.

2.1.2 Segmentação de quadros

Para representar a voz com uma boa resolução temporal capturando o seu comportamento dinâmico, vários métodos de extração segmentam o áudio em quadros, e para cada quadro é extraído um vetor compacto de atributos. Entretanto, para gerar uma representação mais compacta do áudio, os esquemas de AF empregam em geral quadros mais longos que quadros usados em codificadores de áudio, já que as aplicações dos esquemas de AF não incluem a identificação de fala.

Os esquemas de AF usam segmentação do áudio em quadros de tamanho variável ou de tamanho constante, o que é mais comum. Em [38] o áudio é segmentado em intervalos disjuntos de tamanho variável, delimitados pela posição temporal de picos de energia em sub-bandas espectrais. A posição de quadros de tamanho constante pode ser periódica ou aperiódica. Para evitar o desalinhamento entre a AF questionada e a AF previamente obtida do áudio rotulado, quadros aperiódicos podem ser sincronizados com a posição de picos locais de potência ou de envoltória do sinal, definidos em descritores de baixo nível descritos no padrão MPEG-7 (*Moving Picture Expert Group*-7), Potência de Áudio (*Audio Power* - AP) e Forma de Áudio (*Audio Waveform* - AW). Os quadros também podem ser sincronizados pelos pontos de ativação (*onset*), que correspondem aos inícios dos picos de energia espectral [39, 10, 40]. Atributos robustos contra ruído, como o PS-ZCPA (*Pitch Syncronous Zero Crossing Peak Amplitudes*) empregado em reconhecimento de voz [41], também podem ser aplicados para a sincronização de quadros em esquemas de AF.

Quadros periódicos de tamanho constante são sobrepostos para reduzir o desalinhamento entre a AF questionada e a AF do áudio rotulado, durante o processo de busca por AF, como proposto em [2]. Para se obter uma boa localização temporal do atributo de AF, o tamanho do quadro deve ser menor que a menor duração do áudio que se queira detectar. Entretanto, o uso de fatores de sobreposição elevados e quadros mais curtos aumenta o número total de quadros e, consequentemente, o custo computacional do processo de busca.

Em [25] os sistemas de identificação de áudio por conteúdo são classificados como intraquadros, quando um vetor de atributos é extraído a partir da informação do sinal em um único quadro como os coeficientes MFCC(*Mel Frequency Cepstral Coefficients*),

ou interquadros, quando são extraídos de um conjunto maior de quadros, capturando a variação dinâmica do sinal de longo termo, como em atributos que representam ritmo e informação de modulação de frequência. Na prática, a representação interquadro é gerada a partir de uma representação intraquadro. Ademais, também existem atributos globais, extraídos pela análise de todo o sinal de áudio, como os descritores MPEG-7 de centróide temporal e tempo de ataque do som.

Esquemas de AF que empregam uma transformação para domínio espectral aplicam funções de janelas, tais como Hamming, Blackman ou Hann, para limitar o vazamento espectral. Uma análise teórica é feita em [42] para otimizar o uso de janelas em sistemas de AF.

2.1.3 Extração de atributos

Esquemas de identificação de áudio devem empregar atributos adequados à aplicação e com alta variância para o tipo de áudio da aplicação. Atributos aplicáveis na identificação de trechos de voz, como pitch, podem ser invariantes e com baixo poder de discriminação quando aplicados à música instrumental. O intervalo dinâmico dos atributos empregados deve cobrir toda a faixa de variação possível da propriedade acústica usada, para evitar truncamentos de valores, o que elevaria a taxa de falsas detecções de AF.

Em [30] é feita uma revisão dos atributos de baixo nível empregados em esquemas de AF. Em [25] é feito um amplo estudo dos atributos acústicos empregados em aplicações de identificação, segmentação e classificação de áudio, com uma revisão da literatura científica com mais de setenta atributos acústicos heterogêneos, onde é proposta uma taxonomia de atributos. É interessante citar que em [43] diversos atributos usados em sistemas de AF e em recuperação de informação de música (MIR) foram implementados em uma *toolbox* do Matlab (TM) disponibilizada gratuitamente para fins de pesquisa científica. Em [44] outra *toolbox* do Matlab (TM), que implementa atributos perceptuais de áudio, é apresentada.

Nesta revisão de esquemas de AF, empregamos a mesma taxonomia por níveis de classificação de atributos de áudio usada em [25], onde os atributos são agrupados, conforme ilustrado na Figura 2.1, de acordo com o domínio, as características perceptuais e computacionais. Ressaltamos, entretanto, que não foram encontrados na literatura esquemas de AF que empreguem alguns dos atributos citados em [25], usados em outras aplicações. Dessa forma, algumas das categorias de atributos citadas na taxonomia proposta em [25] foram suprimidas. A Tabela 2.1 fornece uma lista de referências de esquemas de *Audio Fingerprint*, categorizados conforme a taxonomia dos atributos proposta na Figura 2.1.

Tabela 2.1: Atributos empregados em esquemas de *Audio Fingerprinting*.

Classificação	Atributo	Referências
Temporal-Físico	ZCR (*Zero Crossing Rate*)	[45, 46]
Temporal-Físico	Batida(*Beat*)	[47, 48, 49]
Frequência-Físico	DFT (*Discrete Fourier Transform*)	[2, 50, 51, 52, 53, 40, 39, 23, 54, 55, 22, 38]
Frequência-Físico	DFT-Timbre	[56, 57, 58]
Temporal-Físico	DFT-Loudness	[59]
Temporal-Físico	Autocorrelação/LSPE(*Least-Square Periodicity Estimator*)	[5]
Frequência-Físico	DFT MPEG-7	[60, 61, 55, 62, 63, 64, 65]
Frequência-Físico	MCLT(*Modulated Complex Lapped Transform*)	[66, 36, 35, 67]
Frequência-Físico	WPT(*Wavelet Packet Transform*) DWT(*Discrete Wavelet Transform*)	[68, 69, 70, 71, 72, 73, 74, 75, 76, 77]
Frequência-Físico	CQT(*Constant Q Transform*)	[78, 79]
Frequência-Físico	Fourier-Mellin	[80]
Frequência-Físico	Haar	[53]
Frequência-Físico	DCT(*Discrete Cosine Transform*)	[81, 53]
Frequência-Físico	Hadamard	[53]
Frequencial-Perceptual	Croma	[82, 83, 6, 84, 85, 86, 32]
Frequencial-Perceptual	Cocleograma	[87]
Frequencial-Perceptual	Harmonicidade	[63, 64]
Cepstrum-Perceptual	MFCC(*Mel Frequency Cepstral Coefficients*)	[88, 55, 89, 5, 90, 91, 92]
Modulação em frequência-Perceptual	Ritmo	[93, 94]

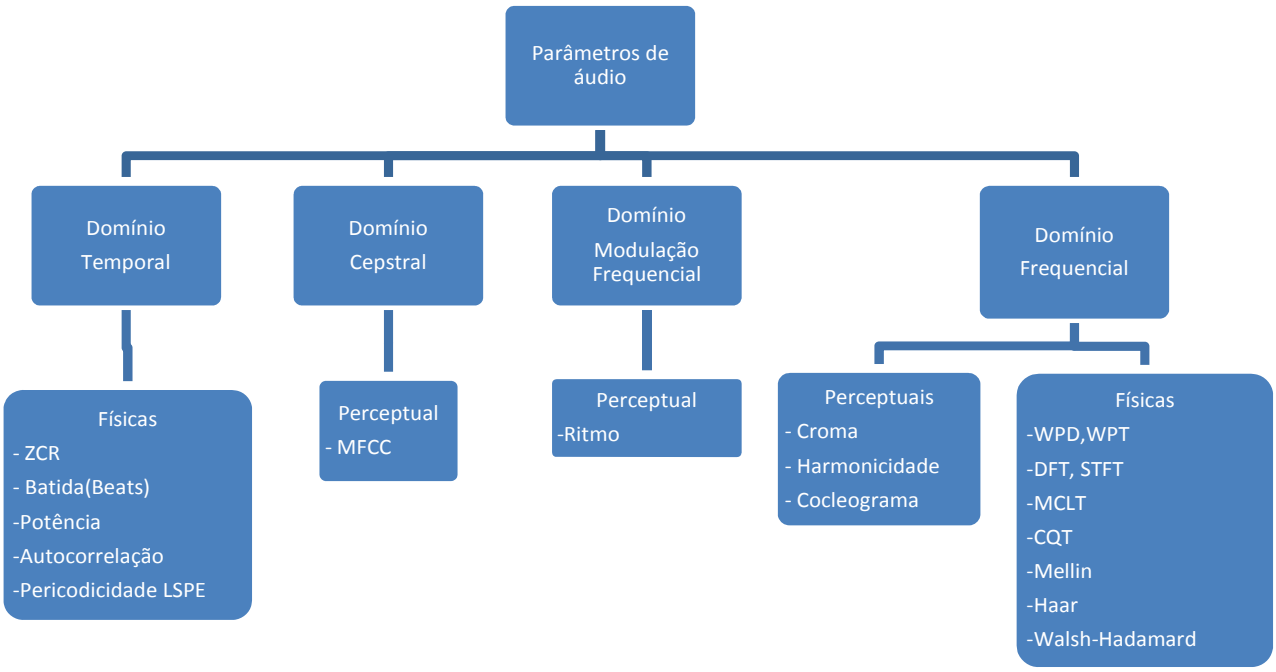

Figura 2.1: Taxonomia de atributos de áudio empregados em esquemas de *Audio Fingerprinting*.

2.1.3.1 Primeiro nível: Domínio de transformação

A classificação em [25] considera apenas o domínio final, quando mais de uma transformação é aplicada ao sinal de áudio. Dessa forma, transformações aplicadas na decomposição em autovetores e valores singulares, tais como PCA (*Principal Component Analysis*) e SVD (*Singular Value Decomposition*), são definidas como autodomínio. Ademais, operações como derivadas, normalização e quantização são consideradas complementares às transformadas e definidas como *filtros* em [25].

Nesta revisão, entretanto, optamos por considerar apenas a primeira transformação usada para capturar uma propriedade acústica. Transformadas subsequentes, empregadas em geral para descorrelacionar os dados, além de derivadas, normalizações ou quantizações, são consideradas operações de pós-processamento e discutidas da Subseção 2.1.4. Os domínios considerados são, portanto:

1. Domínio temporal: A amplitude do sinal é representada em função do tempo. São usados atributos com referência temporal, como máximos ou mínimos de energia, ou de taxa de cruzamento de zeros (*Zero Crossing Rate* - ZCR) do sinal. Atributos no domínio do tempo, tais como máximos locais significantes, podem satisfatoriamente caracterizar o áudio. Descritores de baixo nível previstos no padrão MPEG-7, referentes à potência ou forma de onda, como MPEG-7-AP (*Audio Power*) e MPEG-7-AW (*Audio Waveform*), respectivamente, podem ser empregados.

   Atributos de longo termo no domínio do tempo, referentes ao ritmo ou à potência

de batidas (*beats*), que consistem de atributos de alto nível para caracterização de música, são bastante robustos contra variações de escala temporal [49].

O Ritmo é um importante descritor de música ou de voz e caracteriza a mudança de padrões de energia ou do timbre ao longo do tempo. O descritor de Ritmo é associado a quatro componentes: *timing* (quando o evento ocorreu), *tempo* (qual a frequência do evento), *meter* (qual a estrutura temporal do evento), e *grouping* (como os eventos são agrupados).

2. Domínio frequencial: Usa a informação da amplitude e/ou da fase de componentes espectrais. Algumas transformações em frequência são a Transformada Discreta de Fourier (DFT-*Discrete Fourier Transform*) e a DCT (*Discrete Cosine Transform*). Atributos baseados na transformada de Fourier são bastante aplicados. A DFT é aplicada em [64], onde descritores MPEG-7 SFM (*Spectral Flatness Measure*) e SCM (*Spectral Crest Measure*), referentes à relação entre componentes harmônicos e o ruído do sinal, de 16 sub-bandas são usados como atributos. Os esquemas propostos em [22, 23] usam picos dominantes do espectro de Fourier em cada quadro para representar o áudio, obtendo uma representação robusta contra ruído aditivo e contra mudanças de escala temporal. Em [38] picos dominantes do espectro de Fourier são codificados de forma a permitir a detecção de sinais com variações na escala temporal. A Transformada de Fourier de Tempo Curto (STFT-*Short Time Fourier Transform*) é usada em [2, 95]. Em [52] a DFT é obtida, dividida em sub-bandas, mas apenas as sub-bandas que contenham picos de energia espectral são empregadas para representar o sinal. O algoritmo usado na aplicação Shazam [50] processa a STFT como uma imagem 2D, detectando e representando os picos de energia aos pares, para caracterizar o áudio. Os sistemas propostos em [40, 39] aplicam duas DFT's em cascata para extrair os atributos de áudio. Cabe ressaltar que em [25] as transformadas Wavelet, Transformada Q Constante (CQT- *Constant Q Transform*) e a Transformada Modulada Complexa Sobreposta (MCLT-*Modulated Complex Lapped Transform*) são também classificadas como domínio da frequência. Em [53] são comparados os desempenhos da DFT, DCT, Transformada de Haar e da Transformada Walsh/Hadamard na indexação de áudio. A Transformada Wavelet Discreta (DWT-*Discrete Wavelet Transform*) é aplicada na extração de AF's nos esquemas propostos em [72, 68, 74].

   Decomposições adaptativas também são aplicadas para a classificação de áudio, como em [71]. São empregados critérios de base de discriminação local (*Local Discriminant Bases*- LDB) avaliando o maior poder de discriminação dos coeficientes para definir a melhor decomposição, usando a variância e a energia normalizada.

   Os esquemas propostos em [35, 36] e [67] usam a MCLT. O sistema em [96] aplica a MDCT(*Modified Discrete Cosine Transform*) para extrair os atributos de áudio. A transformada Fourier-Mellin pode ser empregada [80] para obter atributos robustos

contra distorções de escala no tempo. Em [78], a CQT é usada com uma divisão logarítmica do espectro, então os atributos são obtidos a partir de diferenças de tempo e de amplitude entre picos locais sucessivos para obter uma robustez contra distorções de escala temporal.

3. Domínio Cepstral: Os Coeficientes Mel-Ceptrais de Frequência (MFCC) são obtidos pela aplicação da DFT sobre o logaritmo da magnitude do espectro do sinal, ponderados por filtros triangulares espaçados de acordo com a escala Mel, seguido da aplicação da DCT para descorrelacionar os dados. Os coeficientes MFCC são comumente usados para estimar a envoltória do espectro do sinal e podem capturar a informação sobre o timbre do sinal. Os coeficientes MFCC são usados no sistema AudioDNA [88], e nos sistemas propostos em [57, 97, 98, 13]. Implementações do MATLAB dos coeficientes MFCC também são disponíveis em [99, 100].

4. Domínio da Modulação em Frequência: Representa a estrutura temporal do sinal em termos de modulações de baixa frequência, e pode ser empregado para representar o áudio, como em [93]. A informação das modulações temporais contidas no sinal pode ser aplicada na análise de ritmo e na análise de voz robusta contra ruído, como no esquema proposto em [101]. Esta representação permite uma grande redução de dimensionalidade do sinal, sendo mais usada para representar música ou capturar propriedades suprasegmentais da voz.

2.1.3.2 Segundo nível: Interpretação semântica referente à percepção humana.

Atributos perceptuais, como pitch, croma (*chroma*), sonoridade (*loudness*), timbre, tonalidade e harmonicidade, são relacionados à percepção humana do som, e são mais empregados na identificação de músicas.

A sonoridade é uma sensação subjetiva relacionada a variações da pressão acústica, e emprega uma unidade denominada *sone*, mas também é influenciada pelo conteúdo frequencial e pela duração do sinal. Sons mais curtos geram uma menor percepção de sonoridade, e a percepção de sonoridade varia com a frequência do sinal, de acordo com a curva de percepção do ouvinte.

O pitch espectral é definido como um atributo perceptual referente aos componentes espectrais do sinal, que pode ser ordenado em uma escala *mel*. Nesta definição psicoacústica, o pitch depende também da duração e da intensidade do sinal. Entretanto, o conceito de pitch também é associado na literatura científica à frequência fundamental de vibração das cordas vocais. Cabe destacar que mesmo sem a presença da componente da frequência fundamental, pode haver a percepção do pitch, com base nos harmônicos dessa componente, o que é denominado de pitch virtual.

O croma é uma atributo associado ao pitch. O espectro é dividido em oitavas, e dentro

de cada oitava 12 sub-bandas definem classes de pitch. As classes de pitch equivalentes, mesmo de oitavas distintas, produzem uma sensação auditiva semelhante. Dessa forma, o croma é útil na descrição perceptual de músicas. O cromagrama, representação a partir das dimensões tempo-croma, representa variação temporal da energia espectral de cada uma das classes de pitch, mapeando todo o espectro em uma única oitava. Em [6] o cromagrama é usado na identificação de música e uma representação robusta contra distorções escala de tempo é obtida pela detecção dos pontos de máximos locais. Em [84] o croma é aplicado, obtendo uma robustez suficiente para identificação de músicas com diferentes intérpretes. Em [14] o croma é usado na aplicação de sincronização de diferentes performances de música clássica. Em [87] o cocleograma é usado para extrair os atributos de áudio.

O timbre é um atributo perceptual complexo, que permite a identificação de instrumentos musicais, detectando semelhanças entre áudios com sonoridade, pitch e duração distintas, bem como diferenças entre sons com sonoridade, pitch e duração equivalentes. O timbre é um atributo multidimensional influenciado tanto por padrões estacionários quanto não-estacionários, e depende da distribuição espectral do sinal nas bandas críticas. O timbre pode possuir aspectos, como tonalidade, modulação frequencial ou largura. O padrão MPEG-7 possui descritores de áudio referentes ao timbre: Centróide Espectral Harmônico (HSC), Desvio Espectral Harmônico (HSD), Espalhamento Espectral Harmônico (HSS) e Variação Espectral Harmônica (HSV).

A tonalidade está associada ao percentual de componentes senoidais no sinal. Sinais ruidosos possuem baixa tonalidade, enquanto o áudio vozeado possui uma alta tonalidade. Essa característica pode ser estimada através dos atributos SFM e SCM definidos no padrão MPEG-7.

A harmonicidade está associada à presença ou não no sinal de componentes harmônicos situados em múltiplos de uma frequência fundamental. O padrão MPEG-7 possui descritores de áudio referentes à harmonicidade: Razão de Harmonicidade e Coeficiente de Harmonicidade. Da mesma forma, sinais ruidosos possuem baixa harmonicidade, enquanto o áudio vozeado possui uma alta harmonicidade.

Uma descrição detalhada dos modelos matemáticos para a estimação destes atributos perceptuais foge ao escopo deste trabalho. Modelos matemáticos de alguns desses atributos são descritos em [44], onde também é analisada a dependência entre eles.

Atributos físicos, como a distribuição do espectro do sinal obtido pela DFT, DWT, MCLT, CQT, entre outras, descrevem características matemáticas e acústicas do som.

Outra classificação pode ser feita, entre esquemas psicoacústicos, que empregam bancos de filtros que simulam a resolução em frequência da audição humana, considerando efeitos de mascaramento, contornos de percepção de volume e de pitch [102, 93], ou modelos não-psicoacústicos. Em [103] uma pós-seleção dos picos espectrais locais é feita

aplicando-se uma curva com limiar de detecção, baseado no mascaramento temporal, semelhante àquela aplicada no sistema SHAZAM.

Os subgrupos de atributos podem ser divididos de diversas formas, como pela aplicação de bancos de filtros psicoacústicos, como escalas Mel, ERB e Bark, descritas em [104], que buscam simular a resolução em frequência da membrana basilar no ouvido interno. Cabe destacar que na escala ERB as sub-bandas mais baixas têm uma largura menor, quando comparadas às escalas Mel e Bark.

2.1.3.3 Terceiro nível: Características computacionais

Neste nível, os atributos são agrupados conforme a semelhança das operações empregadas no cálculo.

Os atributos obtidos pela representação escolhida podem ser agregados em subgrupos, para redução da dimensionalidade, ou mapeados para valores escalares, pela aplicação de operadores como soma linear, soma quadrática, média, variância, entropia, mínimo ou máximo. Estas operações são denominadas *agregadores* em [25]. Em [55, 98], por exemplo, além do emprego de coeficientes MFCC, são calculadas a entropia de Shannon e de Renyi sobre os coeficientes DFT. No esquema de AF proposto em [95] são usados momentos espectrais normalizados para cada sub-banda, que conforme se argumenta são robustos contra equalização do áudio.

Atributos que fornecem uma referência temporal, mas não são aplicáveis apenas no domínio do tempo, como máximos ou mínimos locais, picos de harmônicos e pontos de cruzamento de zero, são denominados *detectores* em [25], e podem ser usados na definição dos pontos de segmentação em quadros.

2.1.4 Pós-processamento

O pós-processamento é empregado para capturar o comportamento dinâmico do sinal de áudio, descorrelacionar os coeficientes reduzindo a dimensionalidade, ou para aumentar a robustez contra distorções.

O método em [105] aplica a DCT para descorrelacionar os atributos obtidos a partir do esquema proposto por Haitsma [2]. No esquema baseado no domínio Cesptral, em [88, 106], a DCT é aplicada para descorrelacionar os dados.

Outras abordagens usam técnicas de decomposição multidimensional, como Fatoração em Matriz Não-nula (*Non-negative MAtrix Factorization* - NMF), pra reduzir a dimensão dos dados (dimensionalidade) e a redundância dos vetores de atributos, como em [107], onde a NMF é aplicada à STFT para reduzir a dimensionalidade. Em [87], a NMF é aplicada aos atributos obtidos pela aplicação do método SURF ao cocleograma do sinal.

Em [35, 36], a filtragem passa-baixa é aplicada aos coeficientes do logaritmo espectral da DCT. Uma rede neural multicamadas, baseada em PCA, é aplicada com janelas mais longas nas camadas superiores, para reduzir a dimensionalidade. Em [5] a decomposição SVD é empregada para descorrelacionar os coeficientes MFCC. Em [73], a DWT é aplicada a uma imagem 2D obtida do espectrograma gerado pelo método proposto em [2].

Apesar de ampliarem o percentual de ruído no sinal [108], operações delta (derivada) e delta-delta (aceleração) de coeficientes espectrais podem ser usadas para filtrar distorções produzidas por canais com variação lenta, como empregado nos coeficientes MFCC [109]. Ademais, os deltas temporais (entre quadros) e espectrais (entre sub-bandas) podem representar melhor o comportamento dinâmico do sinal e descorrelacionar os atributos de AF, conforme proposto por Haitsma [2].

Em [91] a Normalização em Média Cepstral (CMN-*Cepstral mean normalization*) é usada para reduzir a distorção de canais com variação lenta. A normalização dos intervalos de variações dos atributos também permite equalizar o peso relativo de cada componente no cálculo da similaridade.

A quantização, mapeando os atributos para valores discretos, permite uma representação mais compacta dos dados, aumenta a robustez contra distorções e viabiliza o uso de códigos corretores de erro em métodos eficientes de busca [110, 111, 112, 38, 67].

2.1.5 Modelamento dos atributos

Os atributos acústicos podem ser modelados de diversas formas. Em [31] os sistemas são classificados conforme três modelos:

1. Modelos de estados: Modelos Ocultos de Markov (*Hidden Markov Models*-HMM) são obtidos através de um conjunto de atributos de treinamento. Os atributos acústicos são mapeados para estados discretos [106, 88].

2. Modelos de conjunto de quadros (*Bag-of-Frames*): Uma única estatística ou representação, como Modelo de Misturas de Gaussianas (*Gaussian Mixture Models* - GMM), é usado para representar os atributos de diversos quadros de áudio [55, 98]. Métodos de agrupamento por quantização vetorial usando regra de média-K também são aplicados [63, 84].

3. Modelo de sequências (*Audio Shingle*): O áudio é representado pela concatenação de atributos acústicos de um ou vários quadros de áudio adjacentes, como em [2].

2.2 BUSCA POR AF'S SEMELHANTES

Após a extração da AF, uma métrica de similaridade entre AF's é empregada e um critério de detecção é aplicado em uma busca de AF em uma estrutura de dados. Para facilitar a leitura, a terminologia utilizada para métodos de busca é descrita:

1. *Busca Sequencial ou Busca Exaustiva*: Processo de busca onde todos os registros armazenados são comparados com o registro questionado, o que garante um resultado inequívoco. Entretanto, para estruturas de dados muito grandes e aplicações em tempo real este método de busca pode ser proibitivo. O tempo de processamento é linearmente proporcional ao tamanho do conjunto de dados e ao tempo de cálculo de distância entre dois registros.

2. *Busca Exata* (*Exact Match Search*) [113]: Processo de busca onde apenas os registros idênticos ao registro pesquisado, sem tolerância a erro, são retornados. O problema é descrito como: Dado um conjunto de pontos $P = \{p_1, p_2, ..., p_n\}$, de algum espaço métrico, para uma busca de um ponto q, preprocesse P de forma a verificar a existência e encontrar de forma eficiente pontos em P tal que $\delta(q, p_i) = 0$ para uma dada métrica de similaridade $\delta(.,.)$. Logicamente, para se encontrar pontos idênticos, esse processo se aplica apenas a dados definidos sobre um espaço discreto, como vetores binários. Métodos eficientes, como árvores de busca, otimizam o processo de busca, reduzindo o percentual dos registros comparados a uma pequena parte da estrutura de dados.

3. *Busca com tolerância a erro*: Quando a aplicação utiliza dados ruidosos, como nos esquemas de AF, ou definidos em um domínio contínuo, a busca deve ser feita com tolerância a erros entre os registros, para uma dada métrica de similaridade $\delta(.,.)$. Duas abordagens podem ser usadas. A busca por Intervalo (*Range Search*) [113] encontra todos elementos $(p_i \in P | \delta(q, p) \leq \tau)$, para um limiar de detecção τ. A busca de vizinho mais próximo (*Nearest Neighbors Search*-NNS) encontra o elemento $p_i \in P | \delta(q, p_i) < \delta(q, p_j) \forall p_j \in P, p_j \neq p_i$. Os métodos eficientes preprocessam P e otimizam o processo de busca, minimizando o percentual dos pontos comparados a uma pequena parte da estrutura de dados. Se o método de busca garante uma busca perfeita [114], ou seja, com o mesmo resultado da busca sequencial, seja para a busca por intervalo ou por vizinho mais próximo, o percentual da estrutura de dados que deve ser comparada aumenta rapidamente com a dimensão de p_i, tornando o método mais demorado que a busca sequencial [113, 114, 115, 116] e inviabilizando o seu emprego para bases de dados muito grandes. Este problema é referido como *Maldição da Dimensionalidade* [117].

4. *Busca Aproximada por Intervalo/Vizinho mais próximo*: Nesta abordagem, o processo de busca é projetado para não sofrer com o aumento da complexidade, es-

pecialmente para registros com alta dimensionalidade, mas não garante o mesmo resultado da busca sequencial [114, 113]. A Busca Aproximada por vizinho mais próximo pode ser formulada como: para uma busca por q, encontre $p \in P$ tal que para qualquer $p' \in P$, $\delta(q,p) \leq (1+\epsilon)\delta(p',q)$ [117]. Os métodos de Busca Aproximada, também denominados algoritmos aproximados ou probabilísticos, se baseiam em geral na propriedade da desigualdade triangular do espaço métrico. Uma projeção satisfatória para um espaço métrico visa encontrar um mapeamento com a propriedade de preservação de proximidade, onde pares de registros com distância pequena, conforme a métrica do domínio original, são mapeados para pares de registros também próximos, conforme a métrica do novo domínio.

Para a identificação de música, um intervalo de áudio questionado A_q, com duração equivalente à granularidade do sistema de AF, é usado na identificação dentro de um conjunto C com $A_j, j = 1, 2...N$ áudios rotulados, o que pode ser formulado com base nas $N+1$ hipóteses

$$\begin{aligned} H_0 &: \nexists A_j \in C | A_j \approx A_q, \\ H_j, j = 1...N &: A_j \approx A_q, \end{aligned} \tag{2.1}$$

onde $\approx$ denota uma equivalência de conteúdo perceptual entre áudios de mesma origem. Um critério simples de identificação é associar o intervalo de áudio a um único AF, $F(A_q)$, e usar uma métrica $\delta(F(A_j), F(A_q))$ para encontrar $F(A_j)$ mais semelhante ao AF do intervalo da música questionada. Considerando que os dados são ruidosos, pois a música pode estar distorcida, trata-se de um problema de encontrar o vizinho mais próximo dentro de uma estrutura de dados de AF D que representa o conjunto de músicas C. A ocorrência de Falso Positivo(FP) e o Falso Negativo(FN) na identificação do áudio A_q é definida como

$$\begin{aligned} FP &: arg(\min_{i \in D}(\delta(F(A_i), F(A_q))) = j | H_k, k \neq j, \\ FN &: arg(\min_{i \in D}(\delta(F(A_i), F(A_q))) \neq j | H_j. \end{aligned} \tag{2.2}$$

Quanto maior é o número de AF's de músicas na estrutura de dados, maior é a probabilidade de ocorrência de Falso Positivo. Para o uso de atributos discretos com critério de similaridade de AF sem tolerância a erro, os Falsos Positivos são comumente denominados de colisões, e a probabilidade de colisão de AF cresce rapidamente com o número de áudios N. Este comportamento é conhecido como *Paradoxo do Aniversário*. A probabilidade de Falso Negativo de AF para áudios distorcidos depende da robustez da AF contra transformações que preservem o conteúdo perceptual.

Como descrito na Seção 2.2.3, critérios compostos de decisão, como a dupla detecção,

podem ser utilizados para reduzir os Falsos Positivos.

Na busca por AF ou na inserção de AF em uma estrutura de dados D, várias métricas de similaridade $\delta(F(A_j), F(A_q))$ podem ser empregadas, conforme descrito na Seção 2.2.1. Os processos de busca são descritos na Seção 2.2.2. A usabilidade de um sistema de AF depende da eficiência do processo de busca, especialmente para aplicações em tempo real. Bons métodos de busca devem:

1. Ter rapidez na busca e no cálculo de similaridade;
2. Retornar o resultado correto com baixas taxas de Falso Negativo ou Falso Positivo;
3. Fazer um uso eficiente da memória;
4. Permitir atualização rápida, por inserção ou supressão de registros na estrutura de dados.

Portanto, em um sistema de *Audio Fingerprinting*, a performance de detecção obtida pode depender tanto dos atributos usados na extração de AF para representar o áudio, quanto do desempenho do método de busca. Uma má escolha de atributos de áudio pode gerar uma distribuição ruim, com alta variação intraclasse e baixa variação interclasse, dificultando a classificação. Por outro lado, o emprego de um método inadequado de Busca por Intervalo Aproximada pode degradar ainda mais o desempenho de detecção, elevando a taxa de Falso Negativo.

Para a comparação do desempenho de diferentes sistemas, o método de avaliação deve ser padronizado. Em geral, as taxas de Falso Positivo e Falso Negativo podem ser modificadas ajustando-se parâmetros do sistema de *Audio Fingerprinting*, como limiares de detecção, o que permite construir uma Característica de Operação do Receptor (*Receiver Operating Characteristic*- ROC), que mostra a variação da Taxa de Detecção Verdadeira conforme a Taxa de Falsa Detecção (Falso Positivo). A partir da curva ROC, pode-se obter uma taxa de Erro Equivalente (*Equal Error Rate*), que corresponde ao ponto em que as taxas de Falso Positivo e Falso Negativo são iguais. Outra abordagem possível para a comparação do desempenho dos diferentes sistemas é o ajuste teórico dos parâmetros dos sistemas para limitar a taxa de Falso Positivo a um valor tão baixo quanto se queira, de acordo com a aplicação visada, seguido da análise comparativa das taxas de Falso Negativo.

2.2.1 Medidas de similaridade de *Audio Fingerprint*

Para medir a similaridade entre AF's, uma métrica de distância deve ser usada na fase de busca. A medida de similaridade depende do modelo da AF usada. Em sistemas onde há quantização binária, emprega-se uma distância de Hamming. Em [67] é proposto o

emprego de uma métrica denominada Pseudo Norma Exponencial (EPN), mais adequada para diferenciar valores de AF próximos e distantes. Em [88], o algoritmo de Viterbi é usado para calcular a passagem mais provável na base de dados, com base na representação de estados de classes de som. Em [118] a autocorrelação dos coeficientes MFCC é calculada para estimar a similaridade dos trechos de áudio. Em [84], onde a música é representada por histogramas com as frequências dos valores quantizados, pelo algoritmo Média-K, dos vetores do cromagrama, foram testadas três medidas de similaridade aplicáveis a distribuições: a similaridade por cosseno, a distância euclidiana, e a similaridade chi-quadrado que obteve melhor desempenho.

A métrica também pode ser escolhida ou adaptada conforme a distribuição dos dados. O desempenho de detecção pode ser melhorado pela escolha de uma distância de Hamming ou de Mahalanobis, conforme a estatística de AF degradada [119]. Já em [120], é mostrado que a distância de Hamming é sub-ótima e suscetível a correlações entre os bits de AF, e nestes casos um detector de logaritmo de verossimilhança fornece um desempenho de 5-20% melhor com uma precisão mais estável para diferentes correlações entre bits. Uma forma geral do algoritmo de aprendizado de métrica de distância de Mahalanobis (DML) é usada em [121] para melhorar a robustez da AF contra alguns tipos de distorções.

2.2.2 Métodos de busca de *Audio Fingerprint*

A busca por trecho de áudio com mesmo conteúdo perceptual é em geral feita através de uma Busca por Intervalo. Diversas abordagens usadas em Busca por Intervalo foram propostas para a aplicação de identificação de música por conteúdo.

1. Métodos de cálculo prévio de distâncias: Para grandes conjuntos de AF, uma estratégia é reduzir o número de cálculo de distâncias da busca sequencial. Para isso, uma estrutura de dados pode ser criada, com a divisão em classes, com base no cálculo prévio de distâncias entre as AF's. Se a medida de similaridade for uma métrica, satisfazendo as propriedades de positividade, simetria, reflexividade e desigualdade triangular [113], o cálculo da distância entre a AF questionada e entre algumas AF's de referência é suficiente para reduzir o espaço de pesquisa a uma única classe de AF's [122]. Alguns métodos, como LAESA(*Linear Approximating Eliminating Search Algorithm*) [123] usam a propriedade de desigualdade triangular para reduzir o número de comparações e evitar Falso-Negativos. N_b bases protótipos maximamente separadas são calculadas com uma complexidade linear. Um matriz de distâncias às bases é pré-calculada e distâncias mínimas são usadas com as propriedades da métrica usada, de forma que em cada busca apenas N_b distâncias são calculadas. A representação de AF em espaços vetoriais permite o emprego de alguns destes métodos de busca.

2. Uso de árvores de busca: Uma abordagem comum é a construção de uma estrutura em árvore de busca indexada, que guia a busca usando um critério de pontuação de similaridade entre ramos distintos. Na fase de busca, ramos com similaridade inferiores são descartados. Isso é repetido em vários níveis, limitando a pesquisa a uma pequena parte do espaço de AF's, que são comparados por busca sequencial. Soluções específicas, que aplicam árvores de busca, diferem na forma em que as árvores de busca são construídas e varridas na fase de busca, tais como árvores k-d, R ou B e suas variantes [115]. Em [72], na aplicação de classificação de áudio é usada uma variante da árvore B, onde a altura da árvore corresponde ao número de níveis da decomposição Wavelet, e em cada nível as estatísticas da DWT são usadas como métrica. Árvores k-d e variantes também foram propostas como métodos de busca aproximada em intervalo em esquemas de *Audio Fingerprinting* [124, 95]. Em [114], um método de busca aproximada baseado em uma árvore 256-ária de busca ajustada para o espaço binário foi aplicada para o esquema proposto em [2], entretanto o método apresentou uma taxa de detecção verdadeira baixa, de 55 %.

3. Uso de listas/arquivos invertidos: A idéia básica é a criação de listas de AF's que apontam para trechos longos de áudio que as contenham. Quanto maior o número de AF's em comum nos áudios comparados, maior a similaridade entre eles. O emprego de lista invertida para busca de áudio por conteúdo é sugerido em [84]. Em [2], para a identificação de música por conteúdo com uma granularidade de 3s, assumiu-se que pelo menos um quadro de AF dentro do trecho de música não teria bits modificados em relação a AF armazenada do áudio rotulado. Logo, a busca exata foi empregada aplicando um método de tabela *hash*, com uma detecção suave onde os valores dos bits menos confiáveis são alternados entre 0 e 1. Em [110], uma busca exata foi realizada usando uma tabela *hash* e um código *run-length* com código de Golomb para a compressão de índice. Em [125], uma busca por intervalo para AF's binárias é feita como uma combinação de buscas exatas, usando um método de tabela invertida, para todas as combinações de posições de vetores de erro com norma de Hamming inferior a um limiar.

4. Uso de códigos corretores de erro: O uso de AF's com valores discretos permite que uma busca em intervalo seja mapeada em uma busca exata, através do emprego de códigos corretores de erros, uma vez que AF's discretas próximas são mapeadas em uma mesma palavra de código. Em [112], um código binário de Golay pode ser aplicado para decodificar as AF's binárias e alcançar uma robustez contra distorções de canal. Em [67], um código Reed-Muller é usado para decodificar a AF e reduzir sua dimensionalidade. Em [111], a AF binária é convertida para uma palavra código, de forma a corrigir possíveis erros, e vários códigos lineares são testados, como (8, 4, 5), (8, 16, 4), (8, 14, 3) e (12, 45, 3). Outros trabalhos também empregam essa abordagem [110, 38]. O uso de códigos corretores de erros para identificação de

dados ruidosos é interessante, pois permite um intercâmbio de soluções aplicadas em outras áreas, como em esquemas de detecção e proteção de biometria [126].

5. Outras abordagens: Apesar de métodos de construção da estrutura de dados mais rápidos serem mais indicados, redes neurais podem também ser aplicadas para identificar até mesmo AF's de áudios distorcidos ou de quadros desalinhados [66]. O emprego do método Hashing Sensitivo a Localização (*Localy Sensitive Hashing*-LSH) para busca de áudio por conteúdo é sugerida em [84], como alternativa para evitar a Maldição da Dimensionalidade.

2.2.3 Critérios compostos de decisão para a detecção de áudio

Na aplicação de identificação de música por conteúdo, se a música for representada por um bloco de AF's, pode-se usar um método composto de detecção, onde inicialmente são detectados os quadros de AF's semelhantes nos dois intervalos, usando um limiar de detecção $\delta(F(a_j), F(a_k)) \leq \tau$. Para reduzir a taxa de Falso Positivo, outro critério de detecção pode ser aplicado.

Em [11] o desempenho de detecção do esquema de AF proposto por Haitsma [2] é melhorado com o descarte de Falsos Positivos de AF usando, como segundo atributo de áudio, a autocorrelação cruzada generalizada com transformação de fase (GCC-PHAT) dos quadros de AF detectados.

Em [2] a música é representada por 256 quadros de AF. Dessa forma, o critério de decisão pode ser a existência de um número mínimo de pares de AF's detectados com a mesma defasagem relativa. Em [2] também é proposta a decodificação suave, onde apenas os bits mais confiáveis, com base em uma estatística obtida de versões distorcidas de áudio, são empregados na comparação. Um segundo atributo de áudio pode ser usado como critério de decisão para reduzir os Falsos Positivos.

Em aplicações de detecção de trechos repetidos dentro de um áudio A, pode-se usar um limiar de detecção $\delta(F(a_j), F(a_k)) \leq \tau$ para todos os pares (j, k) de quadros em A, gerando uma matriz booleana de detecção M, comumente denominada de matriz de autossimilaridade. Em [9], para detectar estruturas repetidas em uma música, filtros morfológicos de imagem são aplicados à matriz de autossimilaridade para descartar Falsos Positivos. Em [10], para alinhamento de versões de áudios com distorção em escala de tempo, conjuntos de pontos com mesma defasagem na matriz de similaridade, com padrão diagonal, referentes a trechos repetidos são identificados pela análise do histograma dos coeficientes de reta, e um limiar de duração mínima é usado para descartar Falsos Positivos.

2.3 APLICAÇÕES DE *AUDIO FINGERPRINT*

Sistemas de *Audio Fingerprint* foram aplicados inicialmente no campo de Recuperação de Informação de Música (MIR), que cobre uma enorme gama de aplicações, como identificação de música por cópias, versões ou assobios, classificação de gênero ou de instrumento, transcrição musical, conforme descrito em [127]. Entretanto, novas aplicações foram propostas para esquemas de *Audio Fingerprint*, além das aplicações em MIR, como descrito a seguir.

2.3.1 Identificação de áudio comercial

A aplicação comercial mais popular de esquemas de *Audio Fingerprint* é a de identificação de música por conteúdo, para permitir ao ouvinte a obtenção de metadados e informações descritivas de uma música não rotulada, como título ou nome do autor. Aplicativos para smartphones, como Shazam (TM) [128] e Soundhound (TM)[33], oferecem solução de identificação de música por conteúdo através do emprego de esquemas de *Audio Fingerprinting*. O aplicativo Gracenote (TM) [129], disponível para plataformas de smartphones, também pode ser usado em conjunto com softwares de reprodução de áudio digital, como iTunes e Winamp, o que permite a identificação e categorização automática de uma lista de músicas por gênero.

O aplicativo Music Brainz Picard (TM) [130], oferece um serviço de disponibilização de metadados de uma enciclopédia aberta de músicas, alimentada pelos usuários. Diversos sistemas de *Audio Fingerprinting* de baixa complexidade e uma robustez média foram empregados pelo Music Brainz para a identificação de áudio por conteúdo. Inicialmente o sistema Relatable TRM [131] foi empregado, mas devido ao baixo desempenho com elevada taxa de falsas detecções, foi substituído pelo sistema MUSICIP's PUID/OFA (*Open Fingerprints Architecture*) e posteriormente pelo sistema AcousticID [132].

A identificação de música por conteúdo é mais comumente aplicada a cópias transcodificadas ou reproduções transmitidas por um canal da mesma música *Query by Example*. Entretanto, esquemas de AF mais robustos podem ser usados também para identificação da música pela melodia *Query by Humming*, como pelos serviços Shazam e SoundHound, ou através de diferentes versões ou performances ao vivo, com proposto em [133].

A aplicação comercial vai além do uso pessoal para identificação e gerenciamento de bibliotecas de músicas. Esquemas de *Audio Fingerprinting* podem ser usados por artistas e compositores para monitoramento automático de reprodução de músicas com reserva de direitos autorais, para fins de cobrança e para coibir a pirataria. O serviço Content ID [134], disponibilizado pelo YOUTUBE, permite que proprietários de direitos autorais monitorem o uso indevido de seu conteúdo em mídias compartilhadas no YOUTUBE, e implementa um esquema de *Audio Fingerprinting* na identificação de vídeos e áudios.

Cabe destacar que, no Brasil, esquemas de AF são empregados em sistemas com esta finalidade, como o software do ECAD (Escritório Central de Arrecadação) [135] há algum tempo, e em um sistema novo, o PLAYAX [136].

A abordagem de identificação de áudio por conteúdo usando *Audio Fingerprinting* é distinta de uma abordagem alternativa de uso de marca d'água digital em áudio (*Audio Watermarking*)[137], pela qual uma mensagem arbitrária, referida como marca d'água digital, é inserida no registro de áudio sem alterar o conteúdo perceptual do mesmo. A identificação do áudio é feita através da extração da mensagem inserida nele. Em [138] é feita uma comparação das propriedades, aplicações, vantagens e desvantagens dos sistemas de marca d'água digital e dos sistemas de *Audio Fingerprinting*. Em [29] as diferentes abordagens dos sistemas de marca d'água digital são descritas. Em resumo, comparando as duas técnicas, observa-se que a abordagem por *Audio Fingerprinting* é teoricamente menos vulnerável a ataques de distorções de áudio, uma vez que uma distorção que afete a AF deveria em tese alterar o conteúdo perceptual do áudio. Ademais, os esquemas de *Audio Fingerprinting* não requerem uma modificação do conteúdo do áudio. Os esquemas de marca d'água em áudio possuem um risco de degradação da qualidade do áudio. Uns aspectos negativos da abordagem por *Audio Fingerprinting* são a complexidade computacional em geral superior, e a necessidade de um repositório confiável de *Audio Fingerprints*. Ao contrário da marca d'água digital, os esquemas de *Audio Fingerprinting* não permitem a distinção entre cópias diferentes com mesmo conteúdo perceptual para o controle de direitos autorais, como pela identificação de uma cópia específica distribuída a um usuário.

2.3.2 Análise de integridade de áudio comercial

Outra aplicação também comercial é a verificação de integridade, para monitoramento automático da qualidade de áudios veiculados em propagandas, por exemplo. Dependendo da aplicação o esquema de AF pode indicar até mesmo o ponto onde o conteúdo perceptual do áudio foi alterado.

A Figura 2.2 ilustra os esquemas de emprego de AF nas aplicações de identificação e verificação comercial de áudio. Na identificação a AF extraída do áudio não rotulado é usada como argumento em uma busca no repositório de AF's rotuladas, com os respectivos metadados. Na verificação a AF extraída é comparada com uma outra AF de um áudio rotulado obtida de uma base de dados, para verificar se ambos possuem o mesmo conteúdo perceptual.

A aplicação em verificação de integridade de áudio é possível tanto pelo uso da abordagem por *Audio Fingerprinting* quanto pelo uso de marca d'água digital frágil. Em [139], um esquema autocontido (*Self Embedding*) é proposto para a aplicação de verificação de integridade de áudio, onde a AF extraída do áudio é inserida no mesmo áudio usando

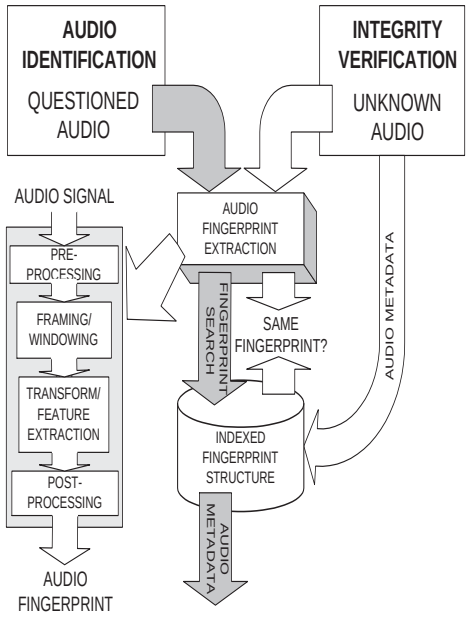

Figura 2.2: Emprego de *Audio Fingerprinting* na identificação e análise de integridade de áudio comercial.

técnicas de marca d'água de áudio. Na verificação de integridade, ambas a AF e a marca d'água de áudio são extraídas do áudio e comparadas, conforme ilustrado na Figura 2.3. Essa abordagem dispensa o uso de um repositório de AF's, já que toda informação está contida no próprio áudio.

2.3.3 Análise da qualidade de áudio comprimido

Em [140] é proposto o uso da distânica de Hamming entre as AF's do sinal e de sua versão comprimida para estimação da qualidade de áudio comprimido. Uma aplicação possível seria a seleção, sem necessidade de uso da versão original do áudio, de versões de baixa qualidade para compartilhamento de um áudio com propriedade autoral. Algoritmos baseados em modelos psico-acústicos podem ser empregados na análise da qualidade perceptual [141, 142, 143, 144], onde alguns deles empregam a medida de qualidade perceptual de áudio (*Perceptual Audio Quality*- PEAQ) [145] adotado pela ITU para avaliar codificadores de áudio. Entretanto, estes métodos requerem a referência do áudio original para a comparação. A qualidade de áudios codificados é comparada comumente usando uma escala média subjetiva de opinião (*Medium Opinium Scale*-MOS). Em [146] são listados diversos atributos no domínio da frequência, do tempo ou perceptuais, empregados na medição de qualidade de áudio. Contudo, o método proposto em [140] não visa fornecer uma estimativa precisa de qualidade subjetiva.

Em [147, 140] a taxa de erro de bits do esquema de AF proposto por Haitsma [2] é usada para estimar o efeito da compressão MP3 a taxas de 128, 80 e 32kbps. Um modelo

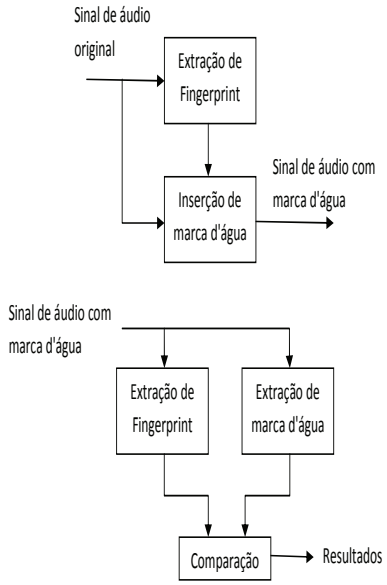

Figura 2.3: Emprego conjunto de marca d'água de áudio e *Audio Fingerprinting* na verificação da integridade de áudio comercial. Elaborada com base em [1].

com sinal descorrelacionado mostra que a taxa de erro de bits é inversamente proporcional ao quadrado da relação entre sinal e o ruído inserido. Dessa forma, quanto maior a taxa de compressão, maior é o ruído de codificação e maior é a taxa de erro de bits de AF. A análise mostra também que as regiões do espectrograma com baixa energia são mais suscetíveis a erros de bits de AF devido ao ruído de codificação, comparadas às regiões com maior energia. Isso ocorre devido à quantização com um limiar nulo, e o trabalho propõe o uso de um peso sobre os bits de AF, desconsiderando aqueles bits referentes a sub-bandas de baixa energia.

Em [24, 140], é proposto o uso de sistemas de *Audio Fingerprinting* para a extração de atributos de qualidade para áudios comprimidos. São comparados os desempenhos de três sistemas de AF, desenvolvidos pela PHILIPS por Haitsma [2], pela Microsoft [66] e pela Universidade Politécnica de Milão [62], para a avaliação de áudio comprimido com o CODEC MP3 LAME a uma taxa de 32kbps. Os esquemas foram ajustados para obterem uma mesma taxa de Falso Positivo, e então as taxas de erro de bits entre os AF's da música original e da música comprimida foram comparados. Não se observou uma diferença significativa entre as taxas de erro de bits de AF dos três sistemas.

2.3.4 Sincronização de mídias

O emprego de AF também já foi proposto para sincronização de dois áudios, ou mesmo vídeos que contenham áudio. Em [10] o uso de AF é proposto para alinhamento de versões de áudios com distorção em escala de tempo, para a anotação de áudio para inserção automática em grandes bases de áudio (*corpus*). É empregado o esquema IRCAM, com

um vetor de 38 componentes reais, com quadros de 2s espaçados de 50ms. As matrizes de autossimilaridade de música obtidas revelam um elevado número de Falsos Positivos de AF. Os elementos detectados com mesma defasagem, com padrão diagonal referente a trechos repetidos, são detectados pela análise do histograma dos coeficientes de reta, e um limiar de duração mínima é usado para descartar os Falsos Positivos.

Em [11] o áudio contido no vídeo reproduzido em uma TV, por exemplo, é usado para sincronizar em tempo real outro conteúdo relacionado ao vídeo, para permitir a visualização simultânea pelo espectador. É usado o esquema proposto por Haitsma [2], que oferece uma melhor sincronização, com resolução de 1s ou 2s, superior a esquemas que usam MFCC. O desempenho de detecção do esquema de AF é melhorado com o descarte de Falsos Positivos de AF pela análise posterior da Autocorrelação Cruzada Generalizada com Transformação de Fase (GCC-PHAT) dos quadros de AF detectados.

Em [12] o esquema de AF proposto por [50] é usado para sincronizar diferentes registros de vídeo de um mesmo evento, a partir do áudio, com uma granularidade de 1s. Em [13] é usado um esquema de AF com atributos extraídos de vetores de MFCC salientes.

Em [14] variantes do algoritmo de Ajuste Temporal Dinâmico (*Dinamic Time Warping*-DTW) são usadas para a sincronização de diferentes performances de música clássica, usando vetores de Croma de 12 bits, com granularidade de 200ms, combinados com a Estatística Normalizada da Energia de Croma (*Croma Energy Normalized Statistics*-CENS), com granularidade superior a 4s.

Outra possível aplicação é a sincronização multimodal do vídeo a partir de uma música, que também esteja contida no vídeo, mas com menor qualidade, como proposto em [15].

2.3.5 Detecção de trechos repetidos em músicas

Como alternativa aos codificadores por entropia, que exploram a frequência de ocorrência, em [148] é proposto um sistema de compressão de áudio aplicado a músicas que contenham trechos repetidos, os quais são codificados uma única vez. O método, baseado em atributos específicos para música, como batida (*beat*) ou tempo, pesquisa autossimilaridades em uma música. Um novo formato de áudio é proposto, a complexidade da busca por blocos repetidos é analisada.

A identificação de trechos repetidos em uma música também pode ser útil para a escolha de trechos característicos (*thumbnailing*). Em [8] é analisada a detecção de estruturas repetidas em uma mesma música, caracterizadas por uma linha diagonal na matriz de autossimilaridade. É proposto um modelo autorregressivo e uma distância autorregressiva de coeficientes é usada no cálculo de similaridade do áudio. O desempenho de detecção destes atributos é comparado com o uso de MFCC e timbre. Em [9] são analisados diversos atributos para detecção de estrutura de música, onde métodos de análise de imagem,

como filtros morfológicos de erosão, dilação, abertura e fechamento, são aplicados à matriz de autossimilaridade.

2.3.6 Sincronização de ruído de fundo

Em [149], é proposto o uso de *Audio Fingerprinting* na aplicação forense de melhoria de inteligibilidade de evidências de áudio contendo músicas sobrepostas à voz de interesse. Uma das dificuldades do cancelamento de ruído de fundo, mesmo quando há um sinal de referência, é a perda do alinhamento devido a variações dinâmicas de escala temporal entre o sinal de referência e o ruído de fundo. O esquema de *Audio Fingerprint* é empregado para melhorar o alinhamento temporal da música de referência com o ruído de fundo, e assim viabilizar a aplicação do algoritmo de LMS (*Least Minimum Squares*) para cancelamento do ruído de fundo.

2.3.7 Uso em esquemas de marca d'água digital

O emprego conjunto de esquemas de AF com esquemas de marca d'água de áudio já foi proposto em diversas aplicações, como na verificação de integridade no esquema autocontido (*Self Embedding*) proposto em [139].

Esquemas de marca d'água de áudio empregam chaves secretas gravadas no áudio. Um dos ataques possíveis, ataque de cópia, é feito pela extração da chave e seu reuso em outros áudios forjados. O uso de uma chave única em diversos áudios pode representar uma vulnerabilidade, uma vez que pode haver vazamento parcial da chave em cada áudio. Por outro lado, o uso de chaves individuais requer uma infraestrutura de chaves. Uma aplicação interessante, para evitar esse tipo de ataque, é o uso da AF para a geração de chaves dependentes do conteúdo do áudio. Esta chave, diferente para cada áudio, é usada na marca d'água de áudio, como proposto em [67, 5]. Ademais, o uso de AF pode prevenir ataques de dessincronização de marca d'água pela inserção/supressão de trechos, uma vez que a AF permite a localização da posição das marcas d'água no áudio.

2.4 ADEQUABILIDADE DOS ESQUEMAS EXISTENTES

De acordo com a revisão dos esquemas de AF apresentada no Apêndice B, todos os sistemas de *Audio Fingerprinting* analisados, além dos esquemas citados em [1, 24], utilizam esquemas com parâmetros fixos e com uma dimensionalidade constante, uma vez que são projetados para analisar de maneira uniforme uma grande quantidade de músicas. Para a detecção de réplicas em um único arquivo de áudio, o número total de quadros de AF a serem comparados pode ser calculado, o que não é possível na aplicação de identificação de

música por conteúdo, onde o número de músicas a serem comparadas é indefinido. Logo, para a aplicação forense de detecção de réplicas, a dimensionalidade da AF, ao invés de ser constante, pode ser adaptada para cada evidência de áudio, sendo ajustada ao menor tamanho que limite o número esperado de Falso Positivo de AF. Podemos analisar a adequabilidade de acordo com os requisitos das propriedades, estipulados anteriormente:

1. *Granularidade*: Esquemas que empregam atributos de longo termo, referentes a ritmo ou dinâmica espectral, como frequência de batimentos ou modulação em frequência, são adequados à identificação de música ou à análise de estruturas suprasegmentais da voz. Entretanto, não são aplicáveis à detecção de réplicas curtas. Atributos baseados em pares de picos de energia, como o método proposto por [50], utilizam longas granularidades de 5, 10 e 15 segundos, para que seja possível a identificação de picos nesses intervalos. O esquema proposto em [51] emprega granularidade de 10s. Em [7] este esquema é usado para detecção de trechos de áudio, onde a busca é feita com uma tabela hash, e o número de quadros de AF equivalentes em uma janela de 2s é usado como medida de similaridade. Portanto, o método não seria capaz de detectar trechos de áudio com duração inferior a 2s. Ademais, o desempenho relatado é bom apenas para eventos de áudio estruturado com componentes espectrais bem definidos, como música ou tons de discagem de telefonia. O sistema proposto em [88] emprega uma granularidade mínima de 6s, inadequada para a detecção de réplicas curtas. Ademais, a representação do sinal em classes de estados, semelhante à abordagem usada no reconhecimento de voz, pode detectar erroneamente áudio intrasentença como se fosse uma réplica. O sistema proposto em [35, 36] também emprega uma granularidade mínima de 6s, inadequada para a detecção de réplicas curtas. Em [11] o esquema proposto por Haitsma [2] é usado por possuir resolução temporal melhor que os esquemas que usam MFCC. De fato, em [98], com uso dos atributos SBE, SC, SFM e SCM, a taxa de detecção é muito baixa para uma granularidade de 2,5s. Para trechos de áudio com voz, o desempenho do método foi insuficiente. Com exceção dos esquemas propostos em [22, 23] que usam quadros de 64ms, com atributos puramente frequenciais baseados na localização temporal de picos espectrais referentes a componentes harmônicos, todos os métodos analisados empregam granularidades superiores a 1s e não são capazes de detectar réplicas tão curtas quanto $100ms$. Não há resultados que sugiram que estas representações compactas permitam a identificação de réplicas tão curtas quanto 100ms. O método proposto pela PHILIPS por Haitsma [2] usa quadros periódicos de $370ms$, superior a 100ms, mas a segmentação periódica empregada neste método permite facilmente um ajuste da duração dos quadros para a redução da granularidade.

2. *Precisão*: A precisão é afetada principalmente pelo desalinhamento entre a segmentação dos quadros de AF do áudio questionado e rotulado. O sistema proposto em

[63, 64] não emprega a sobreposição de quadros, e, portanto, não identifica 100% do áudio não distorcido, devido aos desalinhamentos entre os quadros da AF do áudio questionado e rotulado. Esquemas com elevado fator de sobreposição de quadros reduzem o erro de Falso Negativo.

3. *Robustez*: Em [147, 140] a taxa de erro de bits do esquema de AF proposto por Haitsma [2] é usada para estimar o efeito da compressão MP3 a taxas de 128, 80 e 32kbps. Quanto maior a taxa de compressão, maior é o ruído de codificação, e maior é a taxa de erro de bits de AF. A análise mostra também que as regiões do espectrograma com baixa energia são mais suscetíveis a erros de bits de AF devido ao ruído de codificação que as regiões com maior energia. Isso ocorre devido à quantização final com limiar nulo, e o trabalho propõe o uso de um peso sobre os bits de AF, desconsiderando aqueles bits referentes a regiões de baixa energia do espectrograma. Em [24, 140], é feita uma análise semelhante, comparando o desempenho de três sistemas de AF, desenvolvidos pela PHILIPS por Haitsma [2], pela Microsoft [66] e pela Universidade Politécnica de Milão [62], para a avaliação de áudio comprimido com o CODEC MP3 LAME a uma taxa de 32kbps. Os esquemas foram ajustados para uma mesma taxa de Falso Positivo, e então a taxa de erro de bits de AF das músicas original e comprimida são comparados. Não se observou uma diferença significativa entre os três sistemas. O método proposto por Haitsma [2] é bastante robusto contra diversas distorções em amplitude e em frequência, e possui uma robustez regular contra inserção de ruído ou compressão de áudio.

4. *Unicidade*: No sistema proposto em [88], a representação do sinal em classes de estados, semelhante à abordagem usada no reconhecimento de voz, pode detectar erroneamente áudio intrasentença como se fosse uma réplica. Esquemas que permitam o ajuste da dimensão do vetor de atributos são interessantes pois possibilitam a melhoria da unicidade. O método proposto por Haitsma [2] fornece uma boa unicidade, que pode ser melhorada com o aumento do número de bits.

5. *Complexidade*: Os sistemas propostos em [35, 36] e [63] empregam métodos de busca com uma fase de treinamento com alta complexidade computacional, o que é adequado a sistemas onde o número de buscas é superior ao número de inserções. Estes métodos de busca não são adequados à aplicação forense de detecções de réplicas, onde cada quadro é buscado uma única vez, para verificar a existência de réplica. Apesar da aplicação de detecção de réplica não requerer processamento em tempo real, é desejável o uso de esquemas com menor complexidade como o método proposto por Haitsma [2].

Com base nessa análise, considerando o requisito de precisão, robustez e unicidade, considerando que a segmentação periódica permite o ajuste da granularidade, e considerando ainda que o esquema proposto por Haitsma [2] possui diversas análises teóricas de

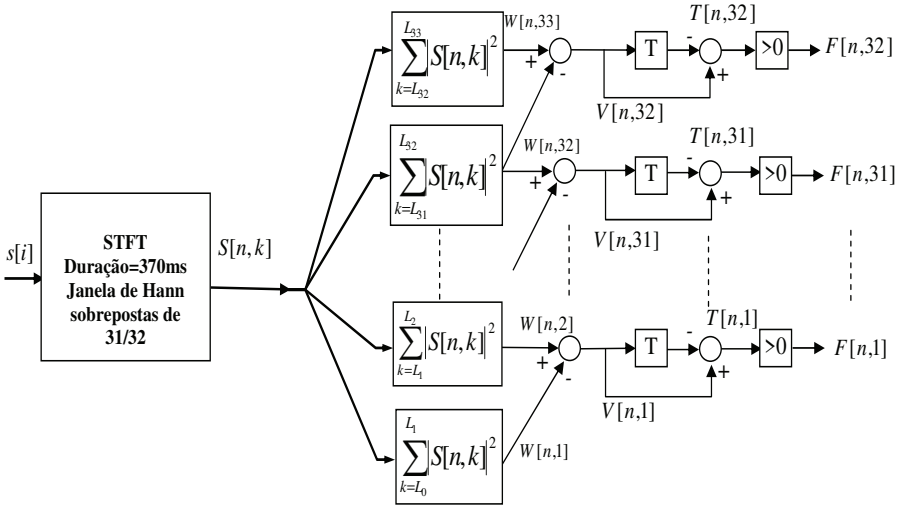

Figura 2.4: Esquema de *Audio-Fingerprinting* proposto pela PHILIPS.

desempenho, esta abordagem foi escolhida para que o esquema seja adaptado à aplicação forense de detecção de réplicas.

2.4.1 O Esquema de *Audio Fingerprinting* proposto pela PHILIPS

O esquema proposto pela PHILIPS por Haitsma[2], ilustrado na Figura 2.4, representa cada música por 256 quadros de AF, de 370ms de duração cada, o que equivale a uma granularidade total de 3s.

A transformada STFT, $S[n,k]$, obtida a partir do sinal discreto no tempo do áudio, $s[i]$, é calculada. Os quadros da STFT são ponderados com uma janela de Hann, w, com duração de $D_F = 370ms$, com um fator de sobreposição $\Omega_F = 31/32$, com quadros espaçados de 11,6ms. Portanto, o desalinhamento máximo entre o quadro de AF questionada e o quadro de AF rotulada é de 5,8ms. Seja R a taxa de amostragem e $N = RD_F$, $S[n,k]$ é dado por

$$S[n,k] = \sum_{i=-\infty}^{\infty} s[i] w \left[\frac{nD_F R}{(1-\Omega_F)} - i \right] e^{-j(\frac{2\pi}{N} ki)}, \tag{2.3}$$

onde n é o índice de quadro e k é o índice de frequência.

A banda de frequência empregada é de $F_L = 300Hz$ a $F_H = 2000Hz$. Logo os índices limites da banda são $L_0 = F_L D_F$ e $L_{33} = F_H D_F$. Os limites das 33 sub-bandas correspondentes aos índices L_1 a L_{33} são definidos por uma escala logarítmica semelhante à escala de Bark [150]. Portanto, este sistema, projetado para uma aplicação de identificação de cópias de música com mesmo conteúdo perceptual, emprega um modelo psicoacústico da resolução espectral da percepção humana, com uma divisão fixa de sub-bandas críticas. Cabe destacar que nos modelos psicoacústicos de compressão de áudio as sub-bandas críticas são usadas de outra forma, para o modelamento do efeito de mascaramento.

A informação de fase é descartada para aumentar a invariância ao deslocamento temporal, e a soma da energia da sub-banda (*Spectral Band Energy*-SBE) é usada como função-peso, ou agregador, de cada sub-banda,

$$W[n,m] = \sum_{k=L_{m-1}}^{L_m} S[n,k]^2. \tag{2.4}$$

A derivada entre sub-bandas na Eq. (2.5) é aplicada, seguida de uma derivada entre quadros na Eq. (2.6). O estágio final de quantização, definido na Eq. (2.7), mapeia $T[n,m]$ em valores binários $F[n,m]$, usando um limiar nulo, para aumentar a robustez contra distorções de áudio.

$$V[n,m] = W[n,m] - W[n,m-1] \tag{2.5}$$

$$T[n,m] = V[n,m] - V[n-1,m] \tag{2.6}$$

$$F[n,m] = \begin{cases} 1, \text{ se } T[n,m] \geq 0 \\ 0, \text{ se } T[n,m] < 0 \end{cases} \tag{2.7}$$

Ao final, cada sub-bloco de AF possui 32 bits, sendo um total de 256 sub-blocos, o que totaliza 8192 bits por bloco.

A Tabela 2.2 resume os parâmetros usados no esquema da PHILIPS, conforme descrito anteriormente.

Tabela 2.2: Parâmetros do esquema de *Audio Fingerprinting* proposto pela PHILIPS [2].

Atributos	PHILIPS system
Sobreposição de quadros	31/32
Duração de quadros	$370ms$
Número de bandas	33
Distância máxima de detecção	1
Banda de frequência	300Hz-2000Hz

No processo de busca, apenas as músicas cujos hashes contenham ao menos um sub-bloco de AF idêntico são comparados, com base em uma busca usando tabela hash. Alternativamente, todos os candidatos com sub-bloco com até k bits diferentes de um dos sub-blocos questionados são comparados. Como o número de comparações cresce exponencialmente com k, apenas os bits com maiores taxas médias de erros são alternados para 0 e 1, para viabilizar o processo. Argumenta-se que aquelas sub-bandas adjacentes

com maior diferença média de energia, e, portanto, com delta mais distante do limiar nulo, geram bits mais confiáveis. Cabe ressaltar que uma divisão de sub-bandas inadequada, com concentração de energia em poucas sub-bandas, pode gerar bits com baixa variância, e pode atribuir uma alta confiabilidade a bits com baixa variância ao longo do sinal, com baixo poder de discriminação do áudio. Uma alternativa para elevar a diferença média de energia das sub-bandas entre o áudio de referência e o áudio distorcido seria a abordagem de representação adaptativa, proposta em [52], na qual a DFT é obtida, dividida em sub-bandas, mas apenas as sub-bandas que contenham picos de energia espectral são empregadas para representar o sinal.

2.4.1.1 Análises empíricas de desempenho

Em [2], na análise de unicidade, a variância da distância de Hamming entre AF's de quadros em posições distintas, descrita como (*Bit Error Rate*-BER), é estimada em 3 vezes a variância para uma distribuição de AF i.i.d. Na análise de precisão, a taxa de Falso Negativo é estimada em 10^{-20} para uma $BER < 0,35$. Os testes de robustez são realizados para um conjunto de apenas 4 músicas.

Em [105] a robustez contra inserção de ruído branco é testada para SNR de 0dB e 5dB. A curva ROC é construída variando o limiar de erro de bits, e uma boa taxa de detecção, acima de 90% com uma taxa de falsa detecção nula, é obtida para uma SNR de 0dB.

Em [151] argumenta-se que a alta correlação da energia entre sub-bandas adjacentes gera uma distribuição de $T[n, m]$ com valores próximos de zero, e que a aplicação de delta, que equivale a um filtro passa-alta no domínio cepstrum, amplia o ruído do sinal. Outros deltas entre quadros e entre sub-bandas foram propostos em [151, 152, 153], onde os deltas entre as sub-bandas de energia (FBE- *Filter Bank Energies*) são consideradas filtros em frequência, baseados nos trabalhos de [154, 155], onde diversos filtros são aplicados às FBE's para remover efeitos de distorções lineares de canal aplicados ao reconhecimento de voz. Os diferentes filtros são descritos com base na notação da transformada Z. O limiar de erro de bits por bloco é ajustado para anular os Falsos Positivos para uma base de teste de 5000 músicas, e é feita uma análise empírica da robustez contra inserção de ruído, onde são testados diversos filtros no tempo e na frequência, com ruído inserido a uma SNR de 15dB a 0dB. O melhor desempenho, com menor taxa de Falso Negativo, é obtido para um delta entre blocos distantes duas posições, que equivale a um filtro passa-banda no domínio cepstrum.

Em [156] a robustez do método foi testada contra ataques intencionais de compressão MP3, com o objetivo de modificação da AF com a preservação do conteúdo perceptual. Argumenta-se que a robustez do método é limitada pelo fato da distribuição de $T[n, m]$ estar concentrada próximo de zero e, portanto, pequenas distorções no sinal podem alterar

o sinal de $T[n, m]$ e causar o erro do bit.

Em [103] argumenta-se que o critério de decisão e a divisão de sub-bandas rígidos do sistema proposto pela PHILIPS pode, para alguns sinais, usar sub-bandas em faixas de baixa energia ou baixa SNR gerando valores de $V[n, m]$ quase nulos, com baixa robustez contra inserção de ruído. Considerando que os picos locais de energia são mais resistentes a distorções no sinal, é proposto um esquema de AF com codificação binária e também baseado na STFT dividida em 18 sub-bandas na escala Mel, robusto e com baixa granularidade, mas que codifica a energia espectral em torno de picos espectrais, empregando também o efeito perceptual de mascaramento de frequência. Argumenta-se que codificando-se a informação da distribuição de energia em torno de cada pico local de energia, consegue-se uma representação mais localizada no tempo que o modelo usado no SHAZAM, que codifica duplas de picos locais. O esquema usa quadros de 100ms, espaçados de 10ms. Os picos são definidos como aquelas sub-bandas com energia estritamente maior que todas as sub-bandas e quadros adjacentes. Uma pós-seleção dos picos é feita aplicando-se uma curva com limiar de detecção, baseado no mascaramento temporal, semelhante àquela aplicada no sistema SHAZAM. Uma região de interesse, na adjacência de nove quadros e duas sub-bandas de cada pico detectado, é usada na extração da AF binária. A codificação dos picos emprega 4 bits para a localização, e 18 bits restantes codificam a distribuição espectral, em relação à energia do pico local, dividindo a adjacência em 3 regiões. O desempenho do método adaptativo é comparado ao do método da PHILIPS para 7 tipos de distorções obtidas das bases de dados TRECVID 2010 e 2011, sendo superior para 6 dos 7 tipos.

2.4.1.2 Análises teóricas de desempenho

Vários modelos teóricos foram desenvolvidos para a análise de desempenho desse método. Em [42], é proposto um modelo estatístico para o esquema da PHILIPS para a análise da taxa de erro dos bits de AF, decorrentes do desalinhamento de quadros ou da inserção de ruído, e é proposta a otimização da janela, em substituição da janela de Hann originalmente usada.

Em [157, 158, 140], é proposto um modelo estocástico para a AF binária. Os bits de AF mais confiáveis são escolhidos para representar cada música. Em [140] é feita uma análise teórica da distância de Hamming entre as AF's de um áudio original e sua versão distorcida, a partir de dois modelos estocásticos propostos: o primeiro considerando o áudio descorrelacionado no tempo, e o segundo considerando o áudio correlacionado. Seja o sinal de áudio $y[i]$, composto por um sinal não distorcido x e de ruído aditivo com distribuição normal, $\mathcal{N}[i]$, dado por

$$y[i] = x[i] + \mathcal{N}[i]. \tag{2.8}$$

O objetivo é caracterizar a diferença entre as AF's binárias $F_y[n,m]$, $F_x[n,m]$, de $y[i]$ e $x[i]$ respectivamente. A probabilidade de erro de bit $P_e[n,m]$ pode ser expressa como

$$\begin{aligned} P_e[n,m] &= Pr\{F_y[n,m] \neq F_x[n,m]\} = \\ &= Pr\{(T_x[n,m] \leq 0, T_y[n,m] \geq 0) \vee (T_x[n,m] \geq 0, T_y[n,m] \leq 0)\}. \end{aligned}$$

No primeiro modelo, $x[i]$ corresponde a um sinal não correlacionado no tempo. Assumindo que o sinal e o ruído são descorrelacionados e estacionários em amplo sentido, $T_x[n,m]$ e $T_y[n,m]$ possuem distribuição normal com média zero e são mutuamente descorrelacionados. Tem-se que $P_e[n,m]$ pode ser obtida em termos das variâncias de $T_x[n,m]$ e $(T_y[n,m] - T_x[n,m])$:

$$P_e[n,m] = \frac{1}{\pi} \arctan\left(\sqrt{\frac{VAR(T_y[n,m] - T_x[n,m])}{VAR(T_x[n,m])}}\right). \tag{2.9}$$

As expressões das variâncias de $T_x[n,m]$ e $T_y[n,m]$ são desenvolvidas e obtém-se:

$$P_e[n,m] = \frac{1}{\pi} \arctan\left(\sqrt{\frac{\sigma_{\mathcal{N}}^4}{\sigma_x^4} + 2\frac{\sigma_{\mathcal{N}}^2}{\sigma_x^2}}\right). \tag{2.10}$$

Para uma relação sinal-ruído relativamente alta a expressão acima pode ser aproximada para

$$P_e[n,m] \approx \sqrt{2}\frac{\sigma_{\mathcal{N}}}{\sigma_x}. \tag{2.11}$$

Nota-se que, em virtude das considerações de que $x[i]$ e $\mathcal{N}[i]$ possuem uma distribuição espectral semelhante, $P_e[n,m]$ independe da sub-banda ou do quadro (n,m). Para o segundo modelo, com $x[i]$ correlacionado no tempo, a $P_e[n,m]$ depende da sub-banda. Os resultados das simulações mostram que $P_e[n,m]$ para o primeiro modelo com sinal descorrelacionado é melhor que para o segundo modelo, o que sugere que uma equalização do erro de bit entre as sub-bandas pode melhorar o desempenho de detecção.

Para melhorar o desempenho, é proposto em [140] a escolha dos bits mais confiáveis para cada áudio. Esta mudança reduz a dimensionalidade da AF e seu poder de discriminação dos trechos de áudio. Alternativamente, para melhorar o desempenho do esquema, propomos no Capítulo 4 um esquema com equalização da energia das sub-bandas para cada áudio, que também torna mais uniforme a taxa de erro de bit entre as sub-bandas. A melhoria da unicidade da representação permite o uso de N_{bits} menores, aumentando assim a tolerância relativa de erro d_{max}/N_{bits}, e, dessa forma, também melhorando a robustez.

3- O ESQUEMA DE *AUDIO FINGERPRINTING* ADAPTATIVO PROPOSTO

> Intelligence is the ability to adapt to changes.
>
> Stephen Hawking

Como citado em [25], a escolha dos atributos de áudio de um esquema de AF é uma fase inicial e conceitual, onde os requisitos da aplicação visada são considerados. Na Seção 2.4 os sistemas propostos foram avaliados, e a análise mostrou que os esquemas existentes não atendem aos requisitos estipulados para esta aplicação de detecção de réplicas curtas: uma elevada precisão (baixas taxas de Falso Positivo e Falso Negativo), uma boa robustez para detecção de réplicas mascaradas, e uma pequena granularidade (boa localização temporal).

Por ser bastante robusta contra algumas distorções e empregar atributos com resolução temporal ajustável, uma abordagem semelhante à proposta por Haitsma [2] foi adotada. Entretanto, para atender aos requisitos da aplicação de detecção de réplicas curtas, algumas adaptações são feitas para cada evidência de áudio:

1. Para a detecção de réplicas mascaradas por edições que alterem levemente o conteúdo perceptual do sinal, o modelo psicoacústico de divisão de sub-bandas aplicado no esquema da PHILIPS não é tão relevante. Para melhorar a unicidade do sistema, é proposta uma divisão adaptativa das sub-bandas pela equalização da média temporal de $W[n, m]$.

2. Para garantir a usabilidade do método, viabilizando a análise de oitiva dos pares de quadros de AF detectados, buscou-se ajustar os parâmetros para limitar o número esperado de Falsos Positivos de Réplica em 10. Dessa forma, a dimensão do vetor de AF binário N_{bits} é ajustada em função do número total de quadros na evidência de áudio.

3. A duração dos quadros, D_F, também é ajustada para permitir a detecção de réplicas tão curtas quanto 100ms.

O esquema adaptativo proposto inicialmente em [26] é ilustrado na Figura 3.1. Ressaltamos que a adaptação dos parâmetros do esquema de AF é feita uma única vez para cada evidência de áudio. Portanto, não se trata de uma adaptação dinâmica, com ajustes de parâmetros ao longo do tempo do sinal de áudio. A transformada STFT, $S[n, k]$, obtida

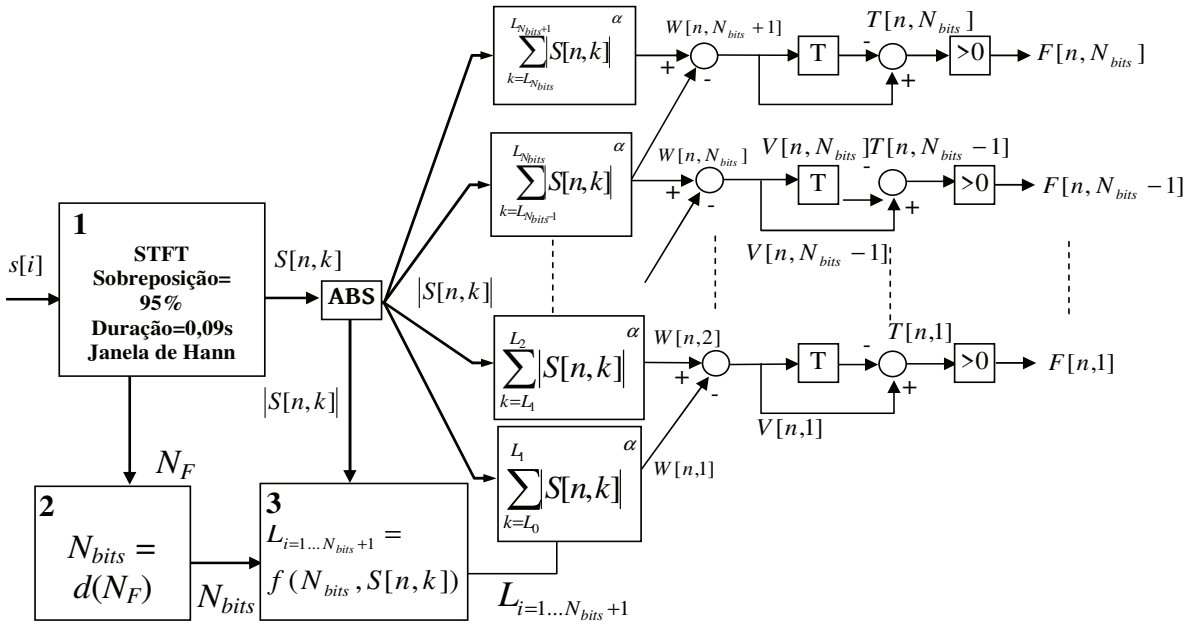

Figura 3.1: O esquema de *Audio Fingerprinting* adaptativo proposto.

a partir do sinal discreto no tempo da evidência de áudio, $s[i]$, é calculada no bloco 1. Os quadros da STFT são ponderados com uma janela de Hann, w, com um fator de sobreposição $\Omega_F = 0,95$. A duração do quadro é ajustada para $D_F = 90ms$ para tornar possível a detecção de réplicas de 100ms, logo o espaçamento entre quadros é de $\Delta_F = 4,5ms$. O ajuste destes parâmetros é detalhado na Seção 3.2. Seja R a taxa de amostragem e $N = D_F R$, $S[n,k]$ é dado por

$$S[n,k] = \sum_{i=-\infty}^{\infty} s[i] w \left[\frac{n D_F R}{(1-\Omega_F)} - i \right] e^{-j(\frac{2\pi k}{N} i)}, \tag{3.1}$$

onde n é o índice de quadro e k é o índice de frequência.

A dimensionalidade N_{bits} de uma AF binária é adaptada para cada evidência de áudio no bloco 2, ajustando para 10 a cota inferior do valor esperado para o número de Falsos Positivos de Réplica, como será detalhado na Seção 3.3.

Uma soma de componentes espectrais elevados a um expoente α, $|S[n,k]|^\alpha$ é aplicado como agregador, para todas as sub-bandas, $m = 1, 2, ...(N_{bits}+1)$. Em [2] $\alpha = 2$ é usado, mas no esquema adaptativo inicialmente proposto, como será explicado, o expoente é ajustado para $\alpha = 1$,

$$W[n,m] = \sum_{k=L_{m-1}}^{L_m} |S[n,k]|^\alpha . \tag{3.2}$$

Os índices dos coeficientes espectrais superiores de cada sub-banda $L_m, m = 1, 2, ...N_{bits}$,

como ilustrado na Figura 3.1, são definidos para cada áudio por uma equalização da média $W[n,m]$ ao longo do tempo, como será explicado na Seção 3.4. Ademais, $L_0 = F_L D_F$ e $L_{N_{bits}+1} = F_H D_F$, onde F_L e F_H são os limites inferior e superior da banda de frequência definida no esquema, respectivamente.

Uma diferença (delta) entre sub-bandas na Eq. (3.3) é aplicada, seguida de uma diferença (delta) entre quadros na Eq. (3.4).

$$V[n,m] = W[n,m] - W[n,m-1] \tag{3.3}$$

$$T[n,m] = V[n,m] - V[n-1,m] \tag{3.4}$$

O estágio final de quantização, definido na Eq. (3.5), mapeia $T[n,m]$ em valores binários $F[n,m]$, usando um limiar nulo, para aumentar a robustez contra distorções de áudio [2].

$$F[n,m] = \begin{cases} 1, \text{ se } T[n,m] \geq 0, \\ 0, \text{ se } T[n,m] < 0. \end{cases} \tag{3.5}$$

3.1 CRITÉRIO DE DETECÇÃO DE RÉPLICA

Para introduzir a simbologia e a terminologia usada, descrevemos inicialmente a última etapa do sistema que corresponde ao critério de detecção de réplica. Fazemos logo uma distinção entre o problema de identificação de música descrito na Seção 2.2.3 e o problema de detecção de réplica. Na identificação de música, o algoritmo global usa uma métrica para encontrar os registros mais semelhantes, com alguma tolerância a erros considerando que os dados são ruidosos. Portanto, trata-se de um problema de encontrar o vizinho mais próximo. Na detecção e identificação da posição de réplicas de áudio, busca-se identificar um ou mais trechos repetidos dentro de um mesmo áudio, cujas representações de AF sejam semelhantes. Dessa forma, para a verificação da existência de réplica, o áudio A é segmentado em quadros de áudio, s_j, onde $j = 1, 2, ...N_F$ é o índice de quadros, e N_F é o número de quadros contidos no áudio A. O problema pode ser formulado com base em uma combinação das hipóteses $H_{i,j}$ (e suas negações $\overline{H_{i,j}}$) não excludentes, já que pode haver vários trechos replicados, onde $\simeq$ denota uma equivalência de conteúdo perceptual de áudios com mesma origem:

$$
\begin{aligned}
H_{i,j} &: s_i \simeq s_j, i = 1, 2, ...N_F, j = i + 1...N_F, \\
\overline{H_{i,j}} &: s_i \neq s_j, i = 1, 2, ...N_F, j = i + 1...N_F.
\end{aligned}
\tag{3.6}
$$

O processo de busca por intervalo deve idealmente resultar em uma matriz $\mathbf{M}$, a qual podemos chamar de matriz de autossimilaridade, de dimensão $N_F \times N_F$, binária e simétrica, com indicação dos índices de pares de AF semelhantes, onde

$$
\mathbf{M}(i,j) = \begin{cases} 1, \text{ se } H_{i,j}, \\ 0, \text{ se } \overline{H_{i,j}}. \end{cases}
\tag{3.7}
$$

Os quadros de AF semelhantes são detectados por busca por intervalo, empregando a distância de Hamming como métrica $\delta(F[i,:], F[j,:])$ para comparar duas AF's com índices de quadro i e j:

$$
\delta(F[i,:], F[j,:]) = \sum_{m=1}^{N_{bits}} |F[i,m] - F[j,m]|,
\tag{3.8}
$$

Inicialmente, assim como em [2], usou-se um tolerância a erro de bits $d_{max} = 1$, que permite o uso de um método eficiente de detecção descrito na Seção 3.1.1. Logo, a matriz de autossimilaridade é obtida por

$$
\mathbf{M}(i,j) = \begin{cases} 1, \delta(F[i,:], F[j,:]) \leq 1, \\ 0, \delta(F[i,:], F[j,:]) > 1. \end{cases}
\tag{3.9}
$$

A ocorrência de Falso Positivo de Quadro (FPQ) e o Falso Negativo de Quadro (FNQ) na identificação de quadros de AF é definida como

$$
\begin{aligned}
FPQ &: \mathbf{M}(i,j) = 1 | \overline{H_{i,j}}, \\
FNQ &: \mathbf{M}(i,j) = 0 | H_{i,j}.
\end{aligned}
\tag{3.10}
$$

A taxa de Falso Positivo de Quadro será usada adiante para avaliar a unicidade da representação. Cabe ressaltar que as detecções de pares de quadros de áudio originários de uma mesma sentença (intrasentença), mas de elocuções distintas, são também consideradas Falso Positivo de Quadro. A taxa de Falso Negativo de Quadro será usada para avaliar a precisão e a robustez do esquema.

Inicialmente, usamos um critério simples de detecção de réplica, onde a detecção de um quadro de AF é suficiente para indicar a presença e a posição da réplica. Com base

nesse critério simples de detecção de réplica $\delta_R(\mathbf{M}(i,j)) = \mathbf{M}(i,j)$, o Falso Positivo de Réplica (FPR) tem a mesma definição do FPQ:

$$FPR : \mathbf{M}(i,j) = 1|\overline{H_{i,j}}. \tag{3.11}$$

Como para o critério de detecção simples um único par de quadros detectado corretamente é suficiente para indicar ao Perito a replicação de um intervalo de quadros $[n_1, n_1 + N]$ para outro intervalo $[n_2, n_2 + N]$, o Falso Negativo de Réplica(FNR) é definido em função de um intervalo de quadros. Temos então:

$$FNR : \mathbf{M}(i,j) = 0|H_{i,j}\forall(i,j)|(i \in [n_1, n_1 + N]) \wedge (j = i + n_2 - n_1). \tag{3.12}$$

Seja D_R a duração da réplica, D_F a duração do quadro e Δ_F o espaçamento entre quadros, o número quadros de áudio contidos na réplica é dado por $N = (D_R - D_F - \Delta_F)/\Delta_F$.

Algumas distorções, como compressão de áudio podem gerar Falsos Negativos em surto devido ao chaveamento dinâmico dos parâmetros de codificação ou a artefatos como pré-eco. Entretanto, para o mascaramento por inserção de ruído de nível constante podemos considerar que não há correlação entre as posições dos Falso Negativos de Quadros. Nesse caso, a probabilidade de detecção de réplica pode ser estimada com base na taxa média, P, de detecção de quadros medida em testes de robustez, por:

$$Pr\{\delta_R = 1|H_{i,j}\forall(i,j), (i \in [n_1, n_1 + N]) \wedge (j = i + n_2 - n_1)\} = 1 - (1 - P)^N. \tag{3.13}$$

Portanto, a probabilidade de detecção de réplica aumenta com a sua duração D_R e depende também da duração dos quadros D_F e do espaçamento entre eles Δ_F. No Capítulo 4 estes parâmetros são ajustados para otimizar o desempenho de detecção.

Para melhorar a unicidade e reduzir assim os Falsos Positivos de Quadros, o esquema aplica um delta entre sub-bandas, seguido por um delta entre quadros, definidas na Eq. (3.3) e na Eq. (3.4), respectivamente. O delta entre quadros reduz a correlação entre AF de quadros vizinhos. A Figura 3.2 ilustra, para um sinal de voz A de 60s o percentual médio da distância de Hamming entre as AF $F_A[n,m]$ e $F_A[n+k,m]$, dada por $\sum_{j=1}^{N_F-k} \delta(F_A[j,m], F_A[j+k,m])/N_F$, em função da separação temporal $k\Delta_F$. O uso do delta entre quadros permite descorrelacionar completamente as AF com separação temporal superior a $200ms$. Esta estimativa do intervalo mínimo é mais precisa que o valor obtido em [114], devido ao uso de quadros com duração mais curta.

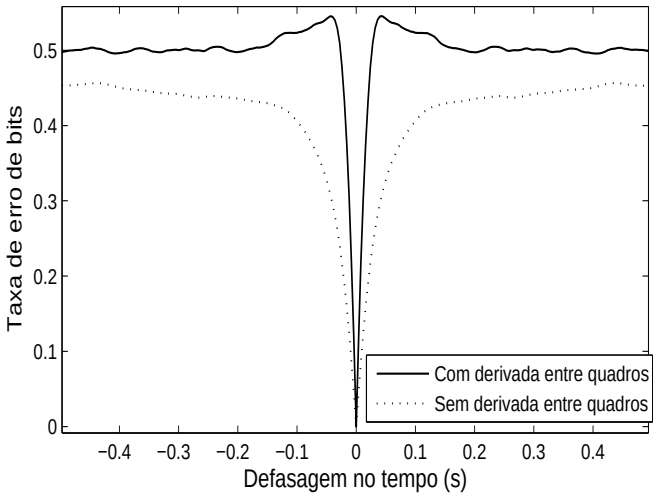

Figura 3.2: Percentual médio da distância de Hamming entre as AF's como função da separação temporal. A curva superior mostra a distância quando o delta entre quadros é aplicado.

A Figura 3.3 ilustra a posição dos elementos não nulos, $\mathbf{M}(i,j) = 1$, da matriz de autossimilaridade de AF, para um áudio com 60s contendo 2 trechos replicados. Na matriz à esquerda observa-se destacados por setas azuis os padrões lineares com defasagem constante dos pares i, j nos intervalos dos trechos originais e replicados; e destacados por setas pretas, alguns Falsos Positivos isolados. A separação mínima de $200ms$ é empregada como critério de descarte de AF's correlacionadas no tempo, detectados próximo à diagonal da matriz de autossimilaridade. Na matriz à esquerda a linha vermelha próximo à diagonal ilustra a separação mínima de 200ms. A matriz à direita mostra o resultado da aplicação deste critério de descarte.

3.1.1 O algoritmo de busca de *Audio Fingerprint* usado

Para a identificação de música por conteúdo, a granularidade longa empregada, contendo vários quadros de AF, permite assumir que pelo menos um dos quadros do intervalo de música tem AF idêntica à AF rotulada, portanto, uma busca exata pode ser empregada [2]. Entretanto, para detectar réplicas curtas que podem corresponder a apenas um quadro de AF, esta simplificação não pode ser assumida. Ademais, como as AF's podem ser modificadas por diversos tipos de distorção de áudio, uma tolerância a erro deve ser adotada, logo uma busca por intervalo deve ser usada para garantir um bom desempenho de detecção.

Para melhorar a taxa de detecção para réplicas de áudio tão curtas quanto 100ms, contendo possivelmente apenas um quadro de AF com bits modificados, métodos de busca aproximada [114] não são adequados. Portanto, a busca por intervalo deve ser realizada

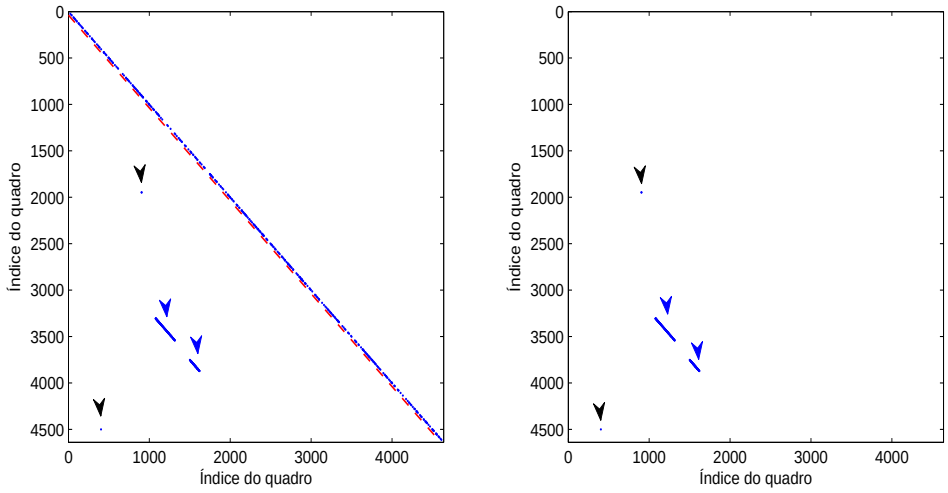

Figura 3.3: Matriz de autossimilaridade booleana. Destacadas por setas azuis, elementos detectados, $M(i,j) = 1$, em trechos replicados e com mesma defasagem. Destacados por setas pretas, alguns Falsos Positivos isolados. Na matriz à esquerda o filtro de separação mínima de 0,2s é representado pela linha vermelha próximo à diagonal. A matriz à direita mostra o resultado da aplicação deste filtro de defasagem mínima.

usando um método de busca perfeita, que garanta o mesmo resultado de uma busca sequencial.

Como a dimensionalidade da AF é alta, conforme análise da Seção 3.3, para evitar a Maldição da Dimensionalidade, o método proposto realiza uma busca por intervalo com distância de Hamming $d_{max} = 1$ por meio de uma combinação de buscas exatas.

Para isso, pré-processamos o conjunto de AF's, $F[:,m], m = 1, 2, ..., N_F$, criando uma lista ordenada lexicograficamente, com uma complexidade computacional de aproximadamente $O(N_F log(N_F))$. Cada busca exata nesta lista é feita com uma complexidade computacional $log(N_F)$.

Para todas AF's, $F[:,m], m = 1, 2, ..., N_F$, fazemos uma busca exata, e para todas as posições de bits $n = 1, 2, ..., N_{bits}$, fazemos buscas exatas para $F_n[:,m] = F[:,m] \oplus E[n]$, onde $E[n]$ é um vetor binário com zeros em todas as posições exceto em n. Portanto, $N_F(1 + N_{bits})$ buscas exatas, com complexidade $O(log(N_F))$, são necessárias.

3.2 FATOR DE SOBREPOSIÇÃO E DURAÇÃO DOS QUADROS

A taxa de detecção de uma réplica em áudio sem distorção depende apenas dos erros de bits de AF decorrentes do desalinhamento dos limites dos quadros nos trechos original e replicado, causado pela segmentação periódica do áudio. Como será mostrado em detalhe no Capítulo 4, a precisão depende apenas do fator de sobreposição Ω_F. Quanto maior

Ω_F, menor é a taxa de Falso Negativo de Quadros devido ao desalinhamento.

Para quadros periódicos sobrepostos por um fator Ω_F, com uma duração D_F e espaçados de $\Delta_F = (1 - \Omega_F)D_F$, os vetores de amostras dos quadros $\mathbf{s_k}$ são dados por:

$$\mathbf{s_k} \triangleq [s[(k-1)\Delta_F R + 1], ..., s[(k-1)(\Delta_F + D_F)R]], k = 0, 1, ..., N_F - 1. \tag{3.14}$$

Consideremos a replicação de um trecho de áudio $[s[0], s[n]]$ para $[s[j], s[j+n]]$, onde $k\Delta_F R \leq j \leq (k+1)\Delta_F R$. O desalinhamento de amostras dos quadros k e $k+1$ na réplica em relação ao quadro original é de $j - k\Delta_F R$ e $(k+1)\Delta_F R - j$. Portanto, o desalinhamento pode variar entre 0 e $\Delta_F R/2$ amostras, com uma distribuição uniforme.

Em [42] um modelo estocástico é usado para o esquema proposto por Haitsma [2], onde a taxa de erro de bits entre quadros de um áudio A e sua versão desalinhada ou distorcida por ruído aditivo A' é obtida por $\sum_{n=1}^{N_F} \delta(F_A[i,:], F_{A'}[i,:])/(N_{bits}N_F)$. A análise teórica da BER decorrente do desalinhamento é feita para um modelo de sinal i.i.d. descorrelacionado no tempo. Para o sinal i.i.d. e $\Omega_F = 31/32$, a BER pode ser de até 4% para a defasagem máxima $\Delta_F/2$. Uma janela otimizada é proposta em substituição à janela de Hann, para reduzir a BER decorrente do desalinhamento. Argumenta-se que as estimativas obtidas servem como uma cota, considerando que para sinais reais correlacionados no tempo as diferenças das distribuições espectrais dos quadros defasados são menores, e, portanto, a BER seria inferior. De fato, nas análises empíricas com músicas para $\Omega_F = 31/32$, a BER média ficou abaixo de 1%.

Considerando que, para a aplicação forense de detecção de réplicas curtas, uma alta taxa de detecção correta de AF é necessária para detectar até um trecho contendo um único quadro, foi utilizado um fator de sobreposição elevado $\Omega_F = 0,95$ no esquema proposto inicialmente. No Capítulo 4 este parâmetro é otimizado, através de simulações, para maximizar o desempenho de detecção.

Devido ao emprego do delta entre quadros subsequentes pela Eq. (3.4), e considerando ainda um desalinhamento máximo dos limites das réplicas de $\Delta_F/2$, para se detectar uma réplica com duração D_R, esta deve ter a duração maior que a duração de dois quadros sobrepostos adicionada de $\Delta_F/2$. Logo,

$$D_F + \Delta_F + \Delta_F/2 \leq D_R. \tag{3.15}$$

Para um quadro de duração $D_F = 370ms$ e $\Delta_F = 11,56ms$ como usado em [2], pela Eq. (3.15), o esquema de AF somente pode detectar réplicas com duração $D_R \geq 0,387s$, que não é curta o suficiente para a análise forense. Para detectar um quadro de AF dentro de réplicas com duração de $100ms$, uma duração de quadro de $D_F \leq 93ms$ é necessária.

Logo, a duração de quadro no esquema proposto é ajustada para $D_F = 90ms$. A redução de D_F aumenta o número de quadros de AF na réplica, o que pode aumentar as chances de detecção, conforme a Eq. (3.13). No Capítulo 4, este parâmetro também é ajustado, através de simulações, para otimizar o desempenho de detecção.

3.3 DIMENSIONALIDADE DA *AUDIO FINGERPRINT*

Para a aplicação forense, considerando que um elevado número de Falsos Positivos pode inviabilizar a inspeção dos resultados pelo examinador, priorizamos a usabilidade do método e ajustamos a dimensionalidade para limitar o número estimado de Falso Positivo de Réplica em até 10.

N_{bits} pode ser ajustado através de uma análise teórica, estimando cotas para a probabilidade de Falso Positivo de Réplica, definido pela Eq (3.11). Em [120] a análise teórica da probabilidade de Falso Positivo, para o emprego de busca perfeita por intervalo, é feita para AF binárias, tanto para modelos com bits independentes quanto para bits dependentes. Em [2], a probabilidade de Falso Positivo também é analisada teoricamente, para um modelo de AF com distribuição i.i.d.

Para uma duração do áudio D_A, o número de quadros de AF dentro da evidência de áudio, N_F, é dado por:

$$N_F = \frac{(D_A - D_F - \Delta_F)}{\Delta_F}. \tag{3.16}$$

O número de quadros N_F pode ser empregado numa análise teórica do número esperado de Falso Positivo de Quadro para ajustar N_{bits}. Para o critério simples de detecção de réplica, as taxas de Falso Positivo de Replica e Falso Positivo de Quadros são equivalentes. A análise de Falso Positivo Quadros de AF é feita assumindo-se que a AF binária possui uma distribuição i.i.d. Considerando que as distâncias de Hamming entre todos os quadros com posições distintas são independentes, o valor esperado de Falsos Positivos pode ser estimado pela soma de todas as probabilidades de detecção de AF de quadros $F[i,:]$ e $F[j,:]$ para todas as combinações de (i,j), dentro de um áudio sem quadros replicados:

$$N_{FP}(d_{max}, N_{bits}, N_F) = \begin{pmatrix} N_F \\ 2 \end{pmatrix} P_r \delta(F[i,:], F[j,:]) \leq d_{max} | \overline{H_{i,j}}. \tag{3.17}$$

Para um modelo de AF com distribuição i.i.d., a distância de $\delta(F[i,:], F[j,:])$ é binomial, de comprimento N_{bits} e probabilidade $p = 0,5$. Logo, temos

$$N_{FP}(d_{max}, N_{bits}, N_F) = \begin{pmatrix} N_F \\ 2 \end{pmatrix} \sum_{i=0}^{d_{max}} \begin{pmatrix} N_{bits} \\ i \end{pmatrix} p^i (1-p)^{N_{bits}-i} \tag{3.18}$$

$$= \begin{pmatrix} N_F \\ 2 \end{pmatrix} 0,5^{N_{bits}} \sum_{i=0}^{d_{max}} \begin{pmatrix} N_{bits} \\ i \end{pmatrix}. \tag{3.19}$$

Em [120], a (Eq. 3.18) é substituída por uma cota, com base na desigualdade de Chernoff, o que pode ser útil para a obtenção de uma fórmula fechada para a estimativa de N_{bits} em função de d_{max} e N_F.

A variância medida em [2] para uma taxa de erro de bits entre quadros foi 3 vezes maior que a esperada para um sinal i.i.d. Portanto, para o modelamento é feita uma aproximação da distribuição Binomial da Eq. 3.18 para a Normal, devido a possibilidade de ajuste da variância mantendo-se a média fixa. A função cumulativa da distribuição é aproximada para a normal, com média $\mu = np = N_{bits}/2$ e desvio padrão $\sigma = \sqrt{np(1-p)} = \sqrt{N_{bits}}/2$. Logo,

$$P_r\delta(F[i,:], F[j,:]) \leq d_{max}|\overline{H_{i,j}} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{d_{max}} e^{-(y-\mu)^2/(2\sigma^2)} dy. \tag{3.20}$$

Substituindo $x = -(y-\mu)/\sigma$, $\mu = N_{bits}/2$ e $\sigma = \sqrt{N_{bits}}/2$, temos

$$P_r\delta(F[i,:], F[j,:]) \leq d_{max}|\overline{H_{i,j}} \leq d_{max}|H1_0) = \frac{1}{\sqrt{2\pi}} \int_{(1-2d_{max}/N_{bits})\sqrt{N_{bits}}}^{\infty} e^{-x^2/2} dx. \tag{3.21}$$

Portanto, uma cota inferior para o número esperado de Falsos Positivos de Quadros, $\hat{N}_{FP}$, é dada por:

$$\hat{N}_{FP} = \begin{pmatrix} N_F \\ 2 \end{pmatrix} \frac{1}{\sqrt{2\pi}} \int_{(1-2d_{max}/N_{bits})\sqrt{N_{bits}}}^{\infty} e^{-x^2/2} dx. \tag{3.22}$$

A Figura 3.4 mostra o valor mínimo de N_{bits} para que $\hat{N}_{FP} \leq 10$, em função do número de quadros N_F para $d_{max} = 1$, calculado com base na Eq. (3.18) ou na Eq. (3.22). N_F varia de 0 a 8.10^5, que corresponde ao número de quadros para uma duração de áudio $D_A = 3600s$, para $D_F = 90ms$ e $\Omega_F = 0,95$. Observa-se que os valores de N_{bits} são baixos, comparados aos valores utilizados em [2], portanto a complexidade computacional do cálculo da Eq. (3.18) é baixa. Ademais, apesar do teste geral, $np \geq 10$ e $n(1-p) \geq 10$, apontar uma boa qualidade da aproximação da distribuição Binomial pela Normal, esta aproximação não é muito boa para probabilidades muito baixas. A Figura 3.4 mostra que

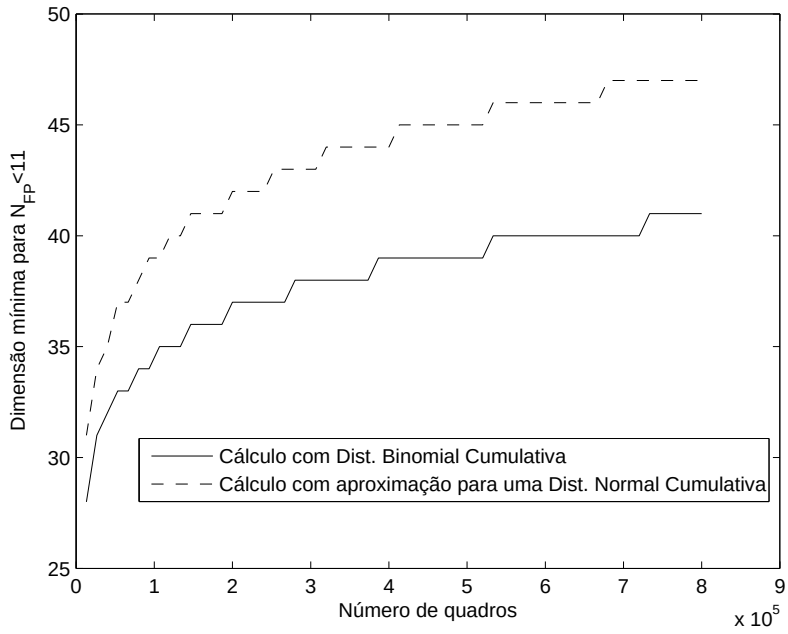

Figura 3.4: Número mínimo de bits da AF como função de número total de quadros, para um número esperado de falsas detecções $\hat{N}_{FP} < 11$, e $d_{max} = 1$, calculado com base na Eq. (3.18) e na aproximação dada pela Eq. (3.22).

a diferença relativa entre os valores de N_{bits} estimados pela Eq. (3.22) e pela Eq. (3.18) é superior a 10%. Logo optamos pelo uso da Eq. (3.18) para a análise teórica dos Falsos Positivos para o ajuste de N_{bits}.

3.4 DIVISÃO DAS SUB-BANDAS

Para trechos de áudio com envoltória estritamente crescente ou decrescente, se apenas o delta entre quadros fosse aplicado, $T[n, m]$ seria sempre positivo e quantizado para 1, ou sempre negativo e quantizado para 0, respectivamente. O emprego do delta entre sub-bandas é útil para descorrelacionar os bits de AF nestes casos. Entretanto, se o trecho de áudio possui uma distribuição espectral constante com concentração de energia em poucas sub-bandas, o uso dos dois deltas não garante uma boa unicidade. Para exemplificar esta situação, ilustramos na Figura 3.5 um sinal de teste A de 3s de duração, com envoltória estritamente crescente, contendo ruído branco e 15 harmônicos de 200Hz. A Figura 3.6 ilustra $F[n, m]$ binário correspondente ao sinal de teste, para o uso de uma divisão de sub-bandas fixa com escala logarítmica (esquerda). Pode-se observar que, para diversos trechos, alguns bits possuem um valor estático, 0 ou 1. Este exemplo ilustra como a unicidade pode ser ruim para alguns trechos de áudio.

Um análise mais precisa da unicidade é feita através do histograma da distância de

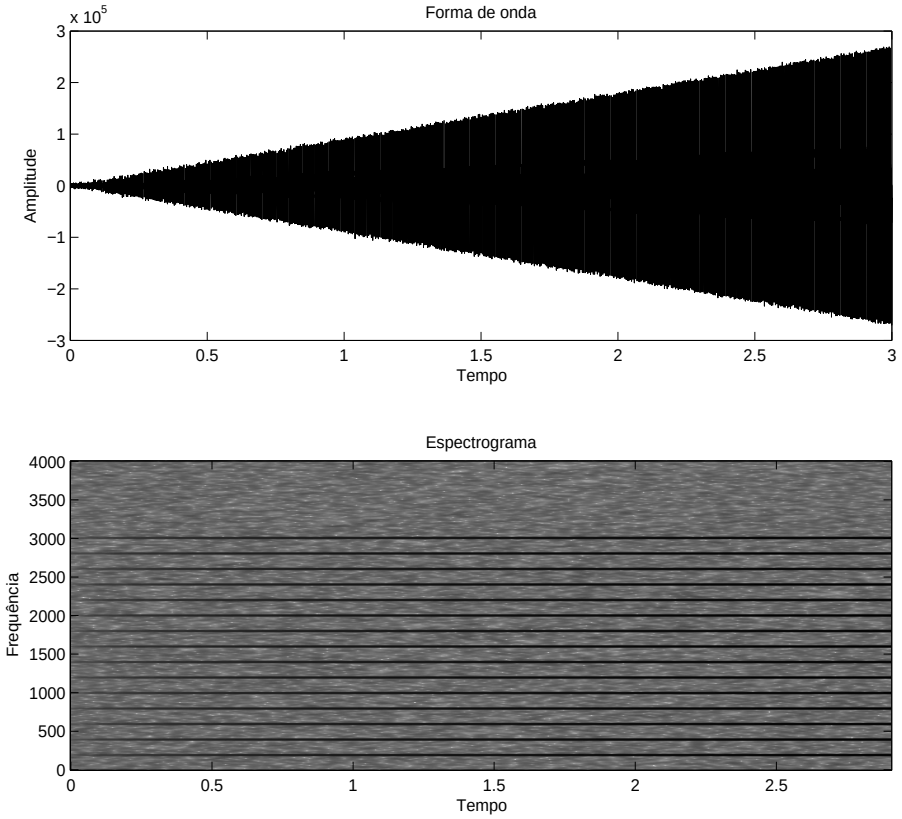

Figura 3.5: Oscilograma (acima) e Espectrograma(abaixo) de sinal de teste de 3s de duração, com envoltória de sinal estritamente crescente, contendo ruído branco e 15 harmônicos de 200Hz.

Hamming, entre AF para todas as combinações de pares de quadros de um mesmo áudio A com N_F quadros, definida por

$$Hist\left(\{\delta(F[j,:],F[k,:]), i=1,2...,N_F, j=i...,N_F\}\right). \tag{3.23}$$

A Figura 3.7 ilustra a função de densidade de probabilidade obtida a partir da distribuição da distância de Hamming entre AF's de quadros do áudio de teste, para o uso de uma divisão de sub-bandas fixa com escala logarítmica. A distribuição binomial correspondente à distância de Hamming para uma distribuição de AF i.i.d. é ilustrada em vermelho. A média da distribuição para uso de uma divisão fixa de sub-bandas é de $6,32$, bem inferior a 8, que seria esperado para distribuição de AF i.i.d.

Este exemplo sugere que a divisão fixa de sub-bandas, como proposto por Haitsma [2], pode concentrar energia em determinadas sub-bandas, o que, pela Eq. (3.4), pode gerar uma distribuição de $T[:,m]$ com média não-nula e reduzir a variância dos bits $F[:,m]$.

Portanto, para melhorar a unicidade, a divisão de sub-bandas em uma escala fixa de

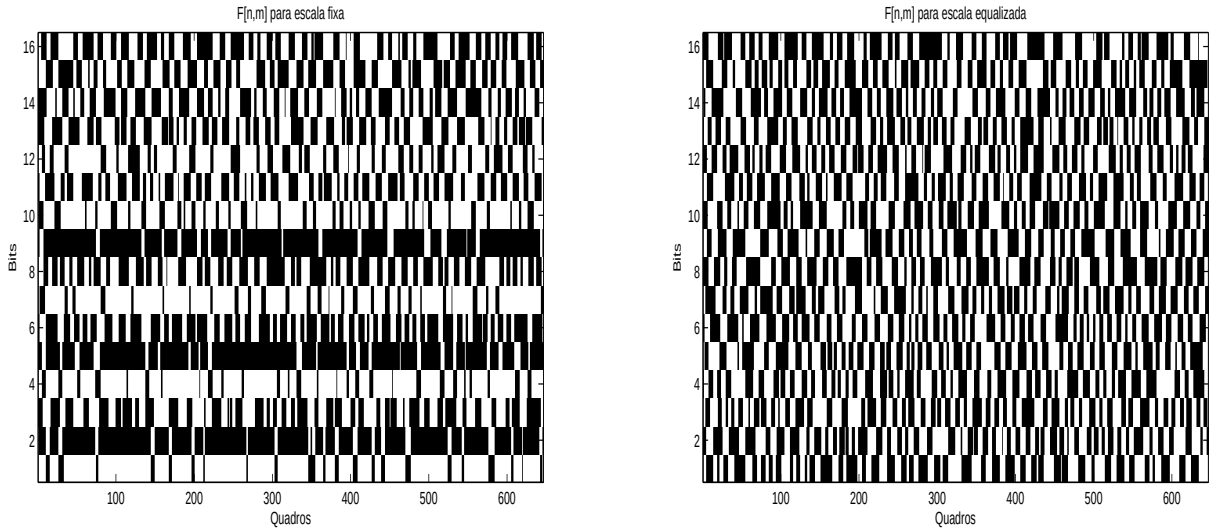

Figura 3.6: Representação de $F[n,m]$ binário, para o uso de uma divisão de sub-bandas fixa com escala logarítmica (esquerda) e para o uso de uma escala adaptativa com equalização da média temporal de $W[n,m]$ (direita).

um modelo psicoacústico é abandonada, já que a constância da informação perceptual auditiva não é uma premissa na detecção de réplicas, como o é na identificação de música. Como a quantização de $T[n,m]$ é feita em torno de zero, seria interessante obtermos uma distribuição de $V[:,m]$, $m = 1, 2, ..., N_{bits}$, com média nula para aumentar a variância de $F[n,m]$:

$$\frac{\left(\sum_{n=1}^{N_F} V[n,m]\right)}{N_F} = 0, m = 1, 2, ..., N_{bits} + 1. \tag{3.24}$$

Como $V[n,m]$ é obtido a partir de um delta sobre $W[n,m]$, temos

$$\sum_{n=1}^{N_F} W[n,m] - \sum_{n=1}^{N_F} W[n,m-1] = 0, m = 1, 2, ...N_{bits} + 1. \tag{3.25}$$

Logo, para aumentar a variância $T[n,m]$ propomos uma divisão de sub-bandas adaptada para cada evidência de áudio, por meio de uma equalização da média temporal de $W[n,m]$, para todas as sub-bandas. Para isso, definimos

$$C(i_L, i_H, S[n,k]) = \sum_{n=1}^{N_F} \sum_{k=i_L}^{i_H} |S[n,k]|^{\alpha}. \tag{3.26}$$

Sejam $L[m], m = 1...N_{bits}$ os índices dos coeficientes espectrais dos limites superiores das sub-bandas. Sejam $L_0 = F_L D_F$ e $L_{N_{bits}+1} = F_H D_F$ os índices dos coeficientes espectrais referentes aos limites da banda de frequência empregada no esquema F_L e F_H.

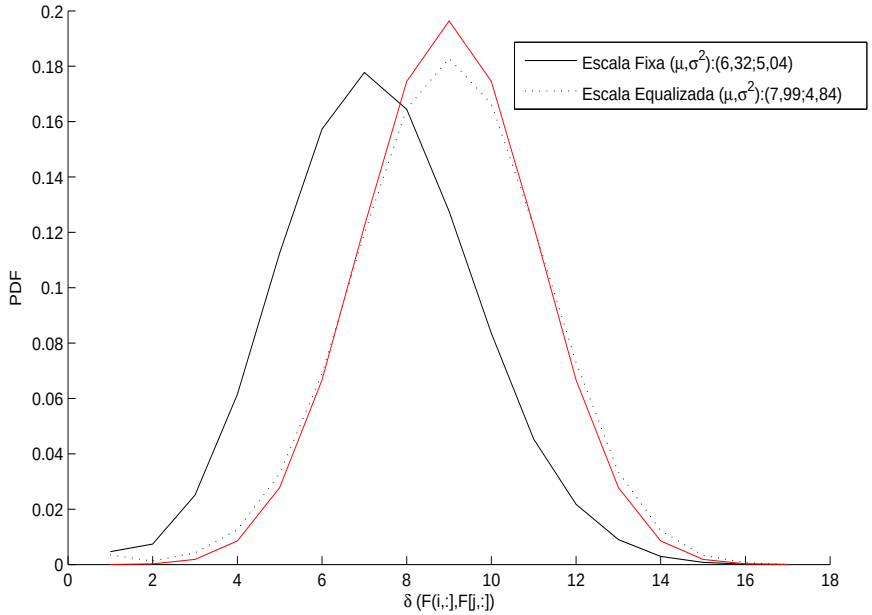

Figura 3.7: FDP's para uma divisão de sub-bandas fixa com escala logarítmica (linha sólida), e para uma escala adaptativa com equalização da média temporal de $W[n,m]$ (linha pontilhada), obtidas a partir de $Hist(\{\delta(F_{A_k}[i,:], F_{A_k}[j,:], i = 1, 2..., (N_F - 0,2/d_F), j = i + (0,2/\Delta_F)..., N_F\})$ para o áudio de teste. A distribuição binomial (N=32, P=0,5) também é ilustrada (em vermelho).

Temos

$$C(L[m-1], L[m], S[n,k]) = \sum_{n=1}^{N_F} W[n,m]. \tag{3.27}$$

Dessa forma, uma distribuição de $V[:,m]$ com média nula pode ser obtida se ajustarmos os limites das sub-bandas, tal que

$$C(L[m-1], L[m], S[n,k]) = C(L[m-2], L[m-1], S[n,k]), m = 1, 2, ..., N_{bits+1}. \tag{3.28}$$

Como existem $N_{bits} + 1$ sub-bandas, temos que

$$C(L[m-1], L[m], S[n,k]) = \frac{C(L[0], L[N_{bits}+1], S[n,k])}{N_{bits}+1}, m = 1, 2, ..., N_{bits} + 1. \tag{3.29}$$

Portanto, os limites superiores das sub-bandas $L[m], m = 1, 2, ...N_{bits}$ são então definidos implicitamente por

$$C(L[0], L[m], S[n,k]) = \frac{m}{N_{bits}+1} C(L[0], L[N_{bits}+1], S[n,k]), m = 1...N_{bits}, \quad (3.30)$$

e podem ser obtidos através do pseudocódigo listado abaixo:

for $k = F_L D_F$ to $F_H D_F$

for $n = 1$ to N_F

$C \leftarrow C + |S[n,k]|^\alpha$

end for

end for

$i \leftarrow 1$; $caux \leftarrow 0$

for $k = F_L D_F$ to $F_H D_F$

for $n = 1$ to N_F

$caux \leftarrow caux + \sum_{n=1}^{N_F} |S[n,k]|^\alpha$

end for

if $(caux > i * C/(N_{bits}+1))$

$L[i] \leftarrow k$; $i \leftarrow i + 1$

end if

end for

Com isso garantimos que $\sum_{k=1}^{N_F} W[k,m]$ é constante para toda sub-banda $m = 1, 2, ...N_{bits}$. Os limites das sub-bandas para a escala fixa logarítmica, e para a equalização da média temporal de $W[n,m]$ com $\alpha = 1$ para o áudio de teste são ilustrados na Figura 3.8. Como o sinal de teste possui harmônicos igualmente espaçados e com mesma energia, a escala de divisão de sub-bandas se aproxima de uma reta. As distribuições de $V[n,:]$, para os bits $n = 2$ e $n = 4$ são ilustradas na Figura 3.9. Para sub-bandas divididas por escala logarítmica fixa, se observam médias não-nulas. Para a divisão adaptada pela equalização da média temporal de $W[n,m]$, se observam médias nulas, conforme projetado. Há uma melhora da unicidade, como ilustrado da Figura 3.6, onde a variância dos bits de AF é mais próxima de $0,25$, que corresponde a variância para uma distribuição de AF i.i.d. A Figura 3.7 mostra ainda a função de densidade de probabilidade para a divisão adaptativa, com média e variância mais próximas da distribuição binomial para AF's i.i.d.

Para obtermos a equivalência da Eq. 3.18 através do pseudocódigo, uma condição necessária é

$$\max_k \left(\sum_{n=1}^{N_F} |S[n,k]|^\alpha \right) < \frac{C(L[0], L[N_{bit+1}], S[n,k])}{N_{bit}+1}, \quad (3.31)$$

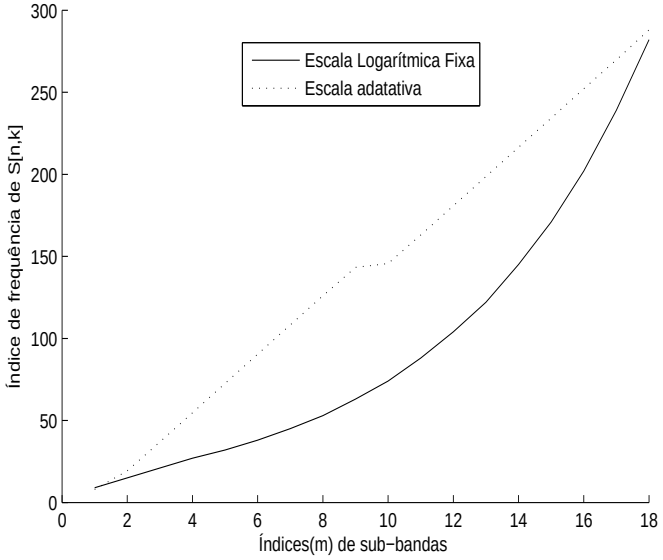

Figura 3.8: Divisão de sub-bandas para escala logarítmica fixa e para o uso de uma escala adaptativa com equalização da média temporal de $W[n, m]$ para o sinal de teste.

caso contrário, se esta condição falha para alguma sub-banda m, o respectivo bit e os bits adjacentes podem assumir valores quase constantes ao longo de todo o áudio piorando a unicidade da representação.

Para um sinal de voz de 60s, a divisão adaptativa de sub-bandas é ilustrada na Figura 3.10. Para o uso de $\alpha = 2$, como usado no esquema proposto pela PHILIPS, as sub-bandas inferiores possuem largura de apenas um bin de frequência, devido à concentração de energia nas frequências mais baixas e devido à baixa resolução em frequência para quadros de duração curta. Dessa forma, para evitar a falha da condição da Eq. 3.31, o esquema adaptativo inicialmente proposto em [26] usa um expoente mais baixo $\alpha = 1$, que expande a largura das sub-bandas inferiores, como ilustrado na Figura 3.10. No Capítulo 4, o parâmetro α é ajustado para cada áudio para melhorar a robustez.

3.5 ANÁLISE DE DESEMPENHO DE DETECÇÃO

Conforme descrito em [26], nas simulações para analisar a unicidade, a precisão e a robustez, foram gerados conjuntos de áudios de teste contendo um trecho replicado, a partir do *corpus* do Instituto Nacional de Criminalística de áudios com SNR estimada de 65dB, amostrados a 48kHz com 16 bits/amostras, com vozes de 51 locutores distintos, durações superiores a 30 minutos, com texto não-controlado e texto controlado com sentenças repetidas.

A Tabela 3.1 resume os parâmetros usados no esquema adaptativo proposto. Apesar de

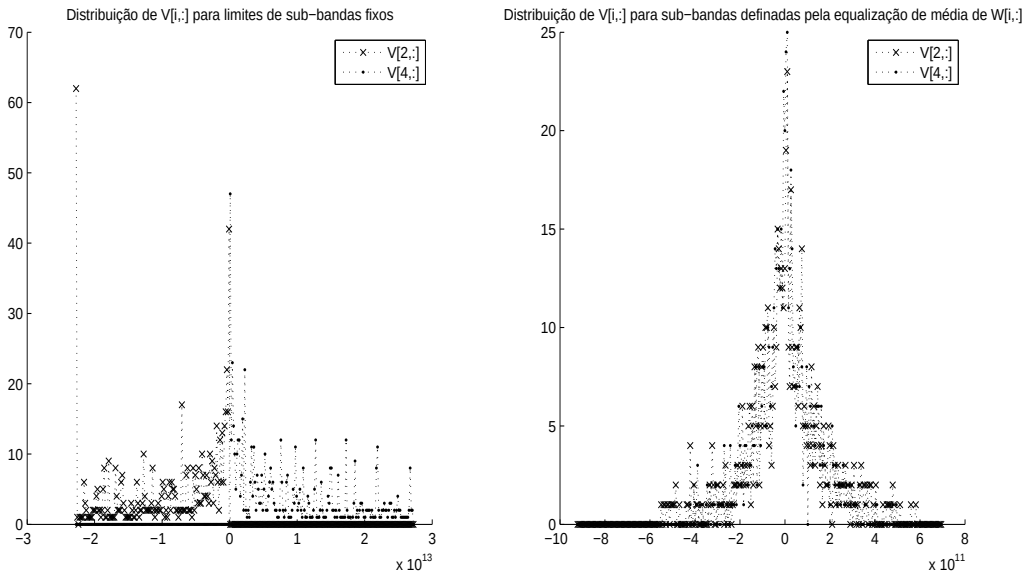

Figura 3.9: Distribuição de $V[n,:], n = 2, n = 4$ para sub-bandas divididas por escala logarítmica fixa (esquerda), e com divisão adaptada pela equalização da média temporal de $W[n,m]$ (direita).

Haitsma [2] usar $F_H = 2kHz$, o que reduz o número de coeficientes de frequência, outros métodos [63], empregam bandas mais extensas, com $F_H = 4kHz$. Dessa forma, optou-se por usar $F_H = 4kHz$ no esquema adaptativo proposto. Para comparar o desempenho do método proposto, os testes incluem outras configurações, tais como o uso de uma escala logarítmica para a divisão das sub-bandas, o uso de um limite superior de banda de frequência $F_H = 2kHz$, e de $\alpha = 2$, como proposto por Haitsma [2]. As diferenças entre os parâmetros do método adaptativo proposto e do esquema proposto pela PHILIPS podem ser observadas nas Tabelas 3.1 e 2.2.

Tabela 3.1: Parâmetros usados no método adaptativo proposto.

Parâmetros	Valores
Ω_F	$0,95$
D_F	$90ms$
N_{bits}	Ajustado para $\hat{N}_{FP} \leq 10$
d_{max}	1
Divisão de sub-bandas	Escala equalizada
Expoente	$\alpha = 1$
Limite inferior de BW	$F_L = 300Hz$
Limite superior de BW	$F_H = 4kHz$

Como o desempenho é comparado para algumas configurações, a dimensionalidade N_{bits} é ajustada previamente, assim como feito em [157], para limitar o número esperado de Falsos Positivos de Réplica, $\hat{N}_{FP} < 11$. Em seguida, nos testes de precisão e robustez,

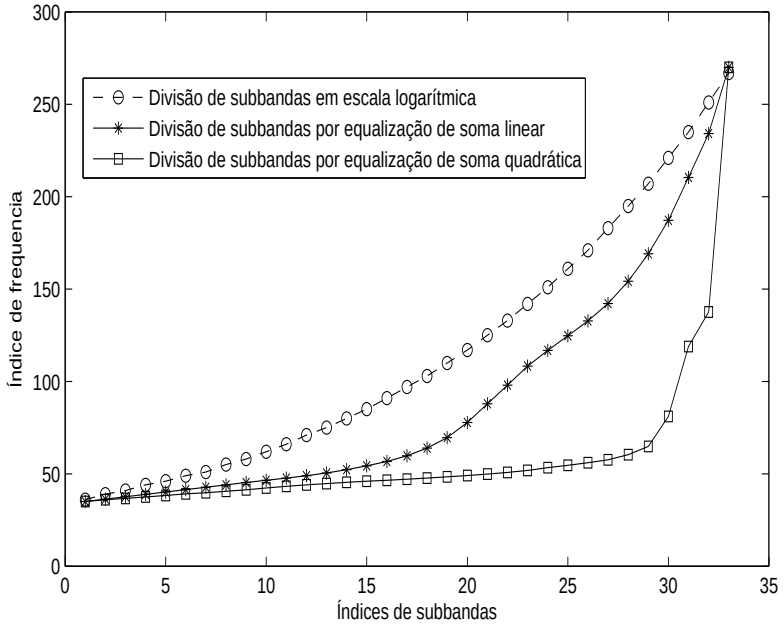

Figura 3.10: Divisões de sub-bandas para uma amostra de voz de 60s para: 1)Escala logarítmica fixa; 2) Escala adaptada para equalização da média de $W[n,m]$ com $\alpha = 1$, e 3) Escala adaptada para equalização da média de $W[n,m]$ com $\alpha = 2$.

é feita a análise de Falso Negativo de Réplica. Como descrito na Seção 3.1, aplica-se um critério simples de detecção onde um único quadro detectado $M(i,j) = 1$, com índices dentro do intervalos replicados é suficiente para indicar a existência de réplica.

3.5.1 Análise da unicidade

Na Seção 3.3 uma cota inferior do número esperado de Falso Positivo de Quadro, $\hat{N}_{FP}$, foi obtida considerando-se uma distribuição i.i.d. de AF's. Nesta seção analisamos a unicidade real do esquema adaptativo proposto, medindo a média do número de Falsos Positivos N_{FP}, para um conjunto de áudios com texto não-controlado de 60s de 20 locutores.

A dimensionalidade é ajustada inicialmente para $N_{bits} = 31$ para áudios de 60s, com base na cota inferior de $\hat{N}_{FP} \leq 10$ obtida pela Eq. (3.22), e incrementada até que $N_{FP} \leq 1$ seja obtido empiricamente. Diversas configurações de parâmetros são analisadas, combinando:

1. A divisão de sub-bandas: adaptativa pela equalização da média temporal de $W[n,m]$, ou por escala fixa logarítmica.

2. O limite superior F_H da banda de frequência: 2kHz ou 4kHz.

3. O limite inferior F_L: 0Hz ou 300Hz.

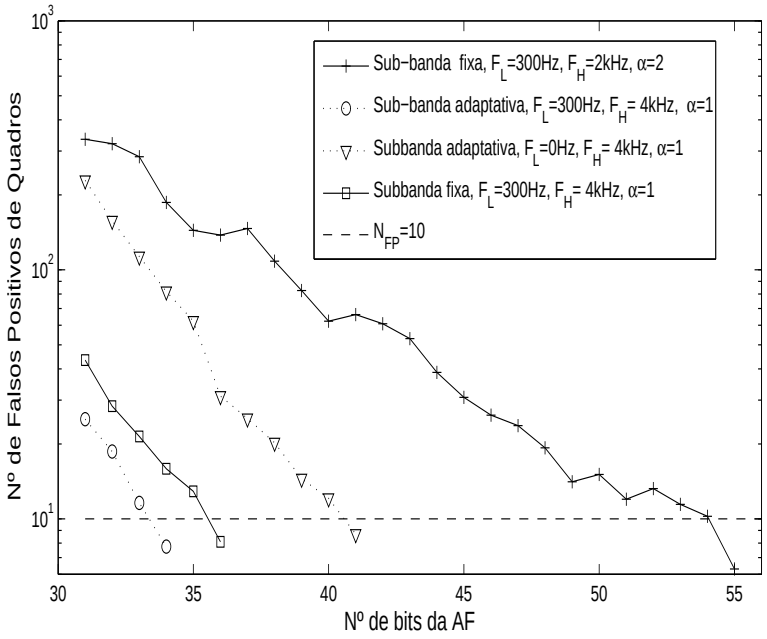

Figura 3.11: Número médio de Falso Positivos, N_{FP} para 20 áudios de 60s referentes a texto não-controlado, para $d_{max} = 1$, variando N_{bits} para várias configurações de método de divisão de sub-bandas, da banda de frequência, e de α.

4. O expoente α da função-peso de sub-bandas $W[n, m]$: $\alpha = 1$ e $\alpha = 2$.

A Figura 3.11 mostra que N_{FP}, decresce quase logaritmicamente com N_{bits} para todas as configurações. A configuração proposta por Haitsma [2], com $F_L = 300Hz$, $F_H = 2kHz$, uma escala logarítmica fixa de sub-banda e $\alpha = 2$ produz o maior número de Falso Positivos. O número de Falsos Positivos N_{FP} é reduzido significativamente para $\alpha = 1$ e a expansão da banda de frequência com $F_L = 300Hz$ e $F_H = 4kHz$.

Para o método adaptativo proposto, com uma divisão adaptativa das sub-bandas, $\alpha = 1$, e uma banda de frequência de $F_L = 300Hz$ a $F_H = 4kHz$, obtém-se o menor número de Falso Positivos N_{FP}, o que sugere que o método adaptativo fornece uma melhor unicidade, comparado às outras configurações.

Em [26], o método proposto também foi testado com uma banda de frequência de $F_L = 0Hz$ a $F_H = 4kHz$, e observou-se uma elevação significativa de N_{FP}. Este resultado inesperado, não explicado em [26], foi causado, como verificado posteriormente, pela falha da condição da Eq. (3.31) para alguns áudios, devido à concentração da energia espectral em baixas frequências. Dessa forma, alguns bits assumiram valores quase estáticos ao longo de todo o áudio, o que elevou N_{FP}. Para evitar este comportamento, é proposto no Capítulo 4 um ajuste de α para cada áudio.

3.5.1.1 Análise de Unicidade para Áudio Intrasentença

Apesar do esquema de AF não ser projetado para identificação de voz ou de locutor, as AF's obtidas para vozes de mesmo locutor (intralocutor) e de locutores distintos (inter-locutor) podem ter distribuições disjuntas, e AF's referentes ao mesmo texto (intrasentença) de um mesmo locutor (intralocutor) podem possuir uma alta correlação [90]. Portanto, a unicidade e a robustez de um esquema de AF deve ser ajustada para cada aplicação. A identificação de diferentes performances de músicas, com intérpretes distintos ou performances ao vivo, que respectivamente consistem de identificação de voz inter-locutor intrasentença e intralocutor intrasentença, requerem o uso de parâmetros mais robustos e invariantes, como descrito em [84]. Para a detecção de réplicas de voz, a AF deve ser ajustada para ser robusta contra distorções, mas também possuir uma unicidade suficiente para discriminar trechos de voz intralocutor intrasentença.

Portanto, para testar a unicidade de AF's para trechos de áudio intrasentença, dois conjuntos de teste, usando áudios de 51 falantes com duração $D_A = 60s$, foram criados, um com voz referente a texto não-controlado, e outro com voz referente a uma sentença repetida uma vez. Os parâmetros foram ajustados como ilustrado na Tabela 3.1. O número de Falsos Positivos para $d_{max} = 1$ é ilustrado na Figura 3.12. O número médio de Falsos Positivos para o conjunto contendo áudio intrasentença é notadamente maior que para o conjunto de áudios referentes a texto não-controlado, $82,4$ e $38,9$, respectivamente. Isto sugere que locuções referentes a uma mesma sentença produzem, em média, AF's semelhantes. Uma análise da matriz de autossimilaridade dos áudio com maiores números de Falso Positivos revelou um padrão linear diagonal, e a análise de oitiva dessas posições confirmou a equivalência dos alguns trechos curtos intrasentença. Observou-se também um padrão de voz com *pitch* e taxa de elocução perceptualmente estáveis para estas vozes. Ressaltamos que nos testes descritos acima não foi verificado se a condição da Eq. (3.31) foi satisfeita para cada áudio.

Esta análise demonstra que a discriminação de AF's de trechos de voz intrasentença pode ser difícil para locutores com um padrão de voz estável. Para melhorar a unicidade, no Capítulo 5, é proposto um critério de dupla detecção sobre a matriz de autossimilaridade de AF. Por fim, cabe reforçar que o método proposto deve ser aplicado em conjunto com outros métodos de autenticação passiva de áudio, seguido de uma análise perceptual dos trechos detectados, para confirmar ou descartar a hipótese de edição.

3.5.2 Análise da precisão

A precisão é definida como a capacidade de detecção de réplicas em áudios não distorcidos, o que é afetado apenas pelo desalinhamento na segmentação dos quadros do trecho original em relação ao trecho replicado. A análise de precisão em (TAVORA, 2015) [26]

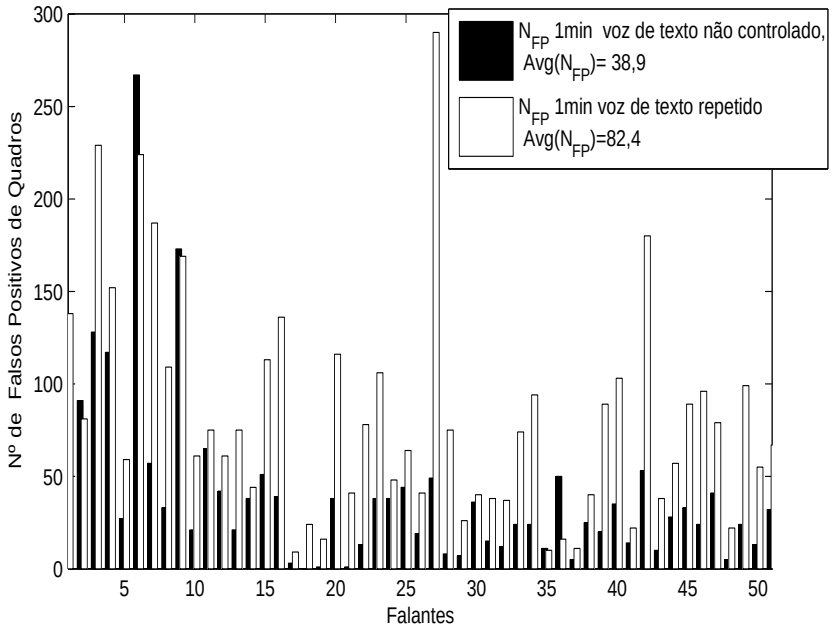

Figura 3.12: Número de Falsos Positivos de Quadros, com parâmetros conforme Tabela 3.1, para voz com texto não-controlado e para voz com texto repetido uma vez, usando um conjunto de teste com vozes de 51 locutores.

é feita a partir da criação de um conjunto áudios de teste com vozes de 20 locutores e contendo réplica sem nenhuma distorção de mascaramento posterior. O conjunto foi gerado com 2000 áudios para cada duração de réplica $D_R \in \{100ms, 200ms, ..., 1s\}$. A réplica, contendo $N = (D_R - D_F - \Delta_F)/\Delta_F$ quadros, foi gerada sem aplicação de janelas e sem mascaramento posterior, de um intervalo aleatório com índices $[i_1, i_1 + N]$ para outro intervalo também aleatório $[i_2, i_2 + N]$, com um retardo mínimo de separação temporal de $|i_1 - i_2| \geq 0,2/\Delta_F$ devido ao filtro de retardo mínimo.

Os testes incluem também algumas configurações, com variações do método de divisão de sub-bandas, dos limites da banda de frequência, e de α. Para a comparação de desempenho de cada configuração, N_{bits} é previamente ajustado para limitar em 10 o número de Falsos Positivo de Quadros.

Como ilustrado na Figura 3.13, o uso de uma escala logarítmica para divisão de sub-bandas com $\alpha = 2$ e uma banda de frequência de $F_L = 300Hz$ a $F_H = 2kHz$, como proposto por Haitsma [2], detecta 96% das réplicas com duração de $100ms$. Para uma banda de frequência de $F_L = 300Hz$ a $F_H = 4kHz$, com $\alpha = 1$, a taxa de detecção de réplica com duração $100ms$ é aumentada para 99%.

Para o método adaptativo proposto, com uma divisão adaptativa das sub-bandas, $\alpha = 1$, e uma banda de frequência de $F_L = 300Hz$ a $F_H = 4kHz$, 100% das réplicas foram detectadas para todas as durações testadas.

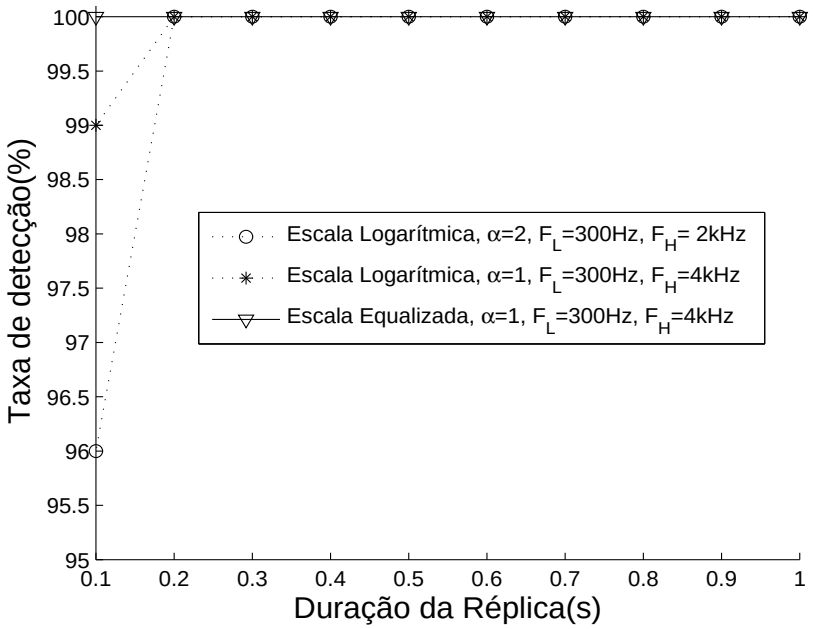

Figura 3.13: Taxa de detecção de réplicas, sem distorção de mascaramento posterior, em áudio referente a texto não-controlado, para durações de réplica $D_R \in \{100s, 200s, ..., 1s\}$, variando o método de divisão de sub-bandas, os limites da banda de frequência, e α.

3.5.3 Análise da robustez

Para testar a robustez do método proposto contra distorções, de forma a possibilitar a detecção de réplicas mascaradas, foi criado um conjunto de teste com vozes de 20 locutores contendo réplicas aleatoriamente posicionadas, com 2000 áudios para cada duração de réplica $D_R \in \{100ms, 200ms, ..., 1s\}$. Nas simulações feitas em (TAVORA, 2015) [26], as seguintes distorções de mascaramentos foram aplicadas, após a replicação:

1. Adição de ruído branco Gaussiano, com relação sinal/(ruído aditivo) SNR=30dB, 26dB, e 22dB, ao longo de todo o sinal de voz. A inserção de ruído branco apenas sobre a réplica não foi testada, pois seria um mascaramento facilmente identificável pela análise visual do espectrograma. Apesar destas SNR garantirem uma boa inteligibilidade do áudio, conforme [159], a adição de ruído branco a uma SNR de 30dB já permite, em geral, um mascaramento de uma emenda em um sinal de voz.

2. Distorções de amplitude: 1dB de ganho sobre a réplica.

3. Distorções de frequência: Atenuação de 12dB de 800Hz a 2400Hz, sobre a réplica.

4. Distorções na escala temporal: Expansão temporal de 2%, pela reamostragem da réplica a uma taxa $1,02$ da taxa de amostragem original.

5. Distorções de formato: Compressão MP3PRO CBR 16kbps e AAC CBR 16kbps, ao longo de todo o sinal.

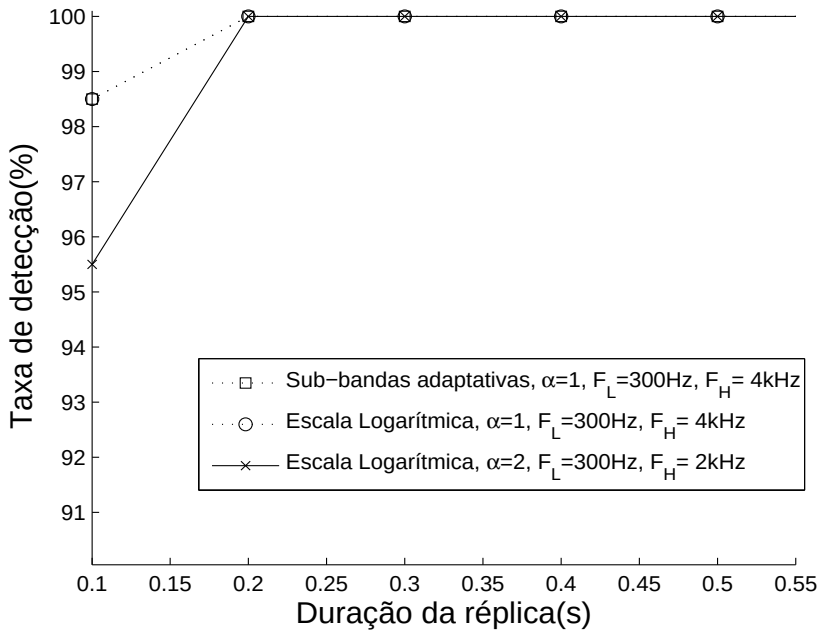

Figura 3.14: Taxa média de detecção de réplica com duração de $D_R \in \{100s, 200ms, ..., 1s\}$, com distorção de amplitude, variando o método de divisão de sub-bandas, a banda de frequência, e α.

3.5.3.1 Robustez contra distorção em amplitude

A taxa média de detecção de réplica em áudio com distorção de amplitude, descrita acima, é ilustrada na Figura 3.14. A melhor taxa de detecção de réplica, 98%, para $D_R = 100ms$ foi obtida para o método adaptativo proposto.

A robustez é alcançada através dos deltas aplicados pelas Eqs. (3.3) e (3.4), e pela quantização final aplicada pela Eq.(3.5), que captura a variação relativa de $W[n, m]$ entre sub-bandas e entre quadros. O desempenho de detecção para o uso de uma escala fixa logarítmica fixa na divisão das sub-bandas é semelhante.

3.5.3.2 Robustez contra adição de ruído

A taxa média de detecção de réplica com inserção de ruído branco Gaussiano a SNR=30dB após a replicação para o método proposto e outras duas configurações de parâmetros é ilustrada na Figura 3.15. O melhor desempenho, com uma taxa média de detecção de réplica superior a 90% para durações superiores a $500ms$, foi obtida com o método adaptativo proposto, com uma banda de frequência de $F_L = 300Hz$ a $F_H = 4kHz$. O resultado revela uma robustez média contra a inserção de ruído, o que pode ser atribuído à ampliação do ruído em relação ao sinal, pela aplicação dos deltas entre quadros e entre sub-bandas.

Para o uso de uma escala logarítmica na divisão das sub-bandas, com uma banda $F_L = 300Hz$ a $F_H = 4kHz$ e $\alpha = 1$, ou com uma banda $F_L = 300Hz$ a $F_H = 2kHz$ e $\alpha = 2$, os desempenhos foram inferiores ao desempenho do método adaptativo proposto.

O desempenho do método adaptativo proposto também foi testado com inserção de ruído branco Gaussiano a SNR=26dB e SNR=22dB. Para a adição de ruído branco Gaussiano a SNR=26dB, uma taxa média de detecção de réplicas maior que 80% foi observada para durações superiores a $600ms$. Para uma adição de ruído branco Gaussiano a SNR=22dB, uma detecção superior a 70% somente foi observada para réplicas com duração maior que $800ms$.

Devido ao uso de *corpus* com SNR=65dB, as AF's referentes a trechos de silêncio usados aleatoriamente como réplicas podem ser fortemente afetadas pela inserção do ruído branco Gaussiano, o que explica em parte a baixa robustez observada nos testes. Na prática, a quase totalidade dos áudios questionados, oriundos de gravação ambiental ou interceptação telefônica apresentam uma qualidade baixa com uma SNR máxima em torno de 30dB. No Capítulo 4, é usado um conjunto de teste com SNR=25dB, mais coerente com os áudios questionados.

Por fim, é importante discernir os testes de robustez contra adição de ruído após a replicação, dos testes de detecção de réplicas para sinais de baixa SNR, típicos de evidências de áudio, sem posterior inserção de ruído. Para exemplificar isto, o teste de detecção de réplicas em áudios obtidos pela inserção de ruído branco Gaussiano a SNR=12dB antes da replicação, sem nenhuma inserção subsequente de ruído. O desempenho do método adaptativo proposto, que não é ilustrado na Figura 3.15, foi de 100% de detecção para todas as durações de réplicas, o que é um resultado similar ao obtido para o teste com áudio de alta SNR sem distorção.

Esta robustez regular contra inserção de ruído do esquema da PHILIPS já foi descrita na literatura científica. Em [42] a análise da BER entre AF's de um áudio original e de um áudio distorcido pela inserção de ruído Gaussiano é feita. Na análise teórica é usado um sinal i.i.d. descorrelacionado no tempo. A curva teórica de $BER \times SNR$ foi comparada com as curvas obtidas para trechos de música de 5s, onde não se observou uma boa aproximação. A melhor BER obtida, dentre 3 músicas, é de aproximadamente 11%, 12,6% e 14% para SNR de 30dB, 26dB e 22dB, respectivamente, o que equivale a uma média de mais de 3 bits modificados por sub-bloco de AF com 32 bits.

Com o objetivo de melhorar a robustez, no Capítulo 4 são propostas adaptações de α e dos limites da banda de frequência.

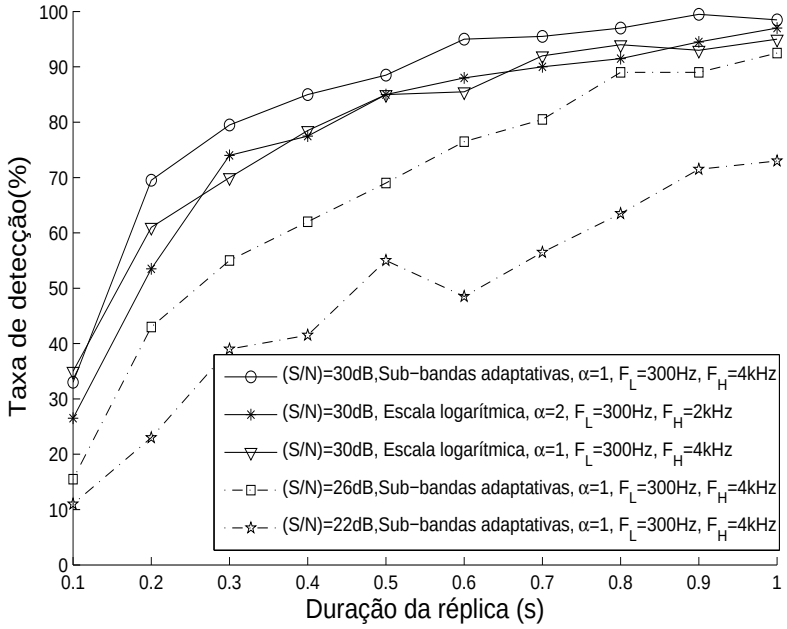

Figura 3.15: Taxa média de detecção de réplica com duração de $D_R \in \{100ms, 200ms, ..., 1s\}$, com subsequente adição de ruído branco Gaussiano, variando o método de divisão de sub-bandas, a banda de frequência, e α.

3.5.3.3 Robustez contra distorção no domínio da frequência

A taxa média de detecção para distorção no domínio da frequência após a replicação é ilustrada na Figura 3.16, que mostra o melhor desempenho para o método adaptativo proposto, 88% de detecção das réplicas com duração de 100ms e 100% para durações maiores. Este bom resultado se deve aos deltas aplicados pelas Eqs. (3.3) e (3.4), e à quantização final aplicada pela Eq. (3.5), que capturam a variação relativa da $W[n, m]$ entre quadros e entre sub-bandas.

A configuração proposta em [2] com uma divisão fixa das sub-bandas detecta 86% das réplicas com duração $100ms$, para uma banda de $F_L = 300Hz$ a $F_H = 2kHz$ com $\alpha = 2$, e 83% das réplicas para uma banda de $F_L = 300Hz$ a $F_H = 4kHz$ com $\alpha = 1$.

3.5.3.4 Robustez contra distorções na escala temporal

A taxa média de detecção para distorção com expansão na escala de tempo de 2% da réplica é ilustrada na Figura 3.17. O método adaptativo proposto fornece o melhor desempenho, detectando acima de 80% das réplicas para duração maior que $400ms$. O desempenho é inferior para o uso divisão fixa das sub-bandas para uma banda de $F_L = 300Hz$ a $F_H = 2kHz$ com $\alpha = 2$. Para um expansão temporal de 4%, a taxa de detecção do método proposto, não ilustrada na Figura 3.17, é aproximadamente nula para todas as configurações de parâmetros testadas.

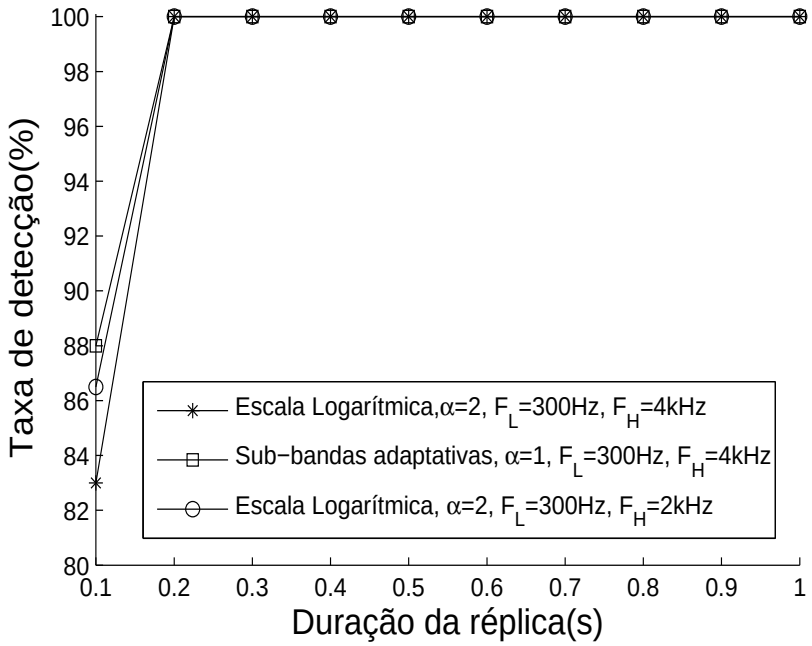

Figura 3.16: Taxa média de detecção de réplica com duração de $D_R \in \{100ms, 200ms, ..., 1s\}$, com distorção no domínio da frequência como descrito anteriormente, variando o método de divisão de sub-bandas, a banda de frequência e α.

A baixa robustez do esquema proposto por Haitsma [2] contra a escala temporal já havia sido descrita na literatura científica. Uma abordagem alternativa é proposta em [160] para aumentar a robustez contra variações de escala temporal. No Capítulo 4 é proposta a adaptação automática dos limites da banda de frequência, que pode tornar a representação mais invariante a deslocamentos da distribuição de energia espectral decorrentes da escala no tempo.

3.5.3.5 Robustez contra distorções por compressão de áudio

As taxas médias de detecção de réplica para áudios comprimidos com os CODECS MP3PRO CBR 16kbps e AAC CBR 16kbps são ilustrada nas Figuras 3.18 e 3.19, respectivamente. Algumas configurações de parâmetros são testadas, e a melhor taxa de detecção em áudio comprimido com MP3 é obtida com o uso do método adaptativo proposto. A detecção de réplica em áudio comprimido com AAC é ligeiramente melhor para o método adaptativo proposto, com uma taxa média de detecção superior a 90% para durações $D_R > 500ms$. A configuração de parâmetros usados por Haitsma [2] fornece o pior desempenho de detecção de réplicas em áudios comprimidos tanto por MP3 quanto por AAC.

A robustez baixa do método proposto pode ser atribuída a artefatos como pré-eco, ou perdas locais como lacunas espectrais gerados pelo descarte de componentes mascaradas,

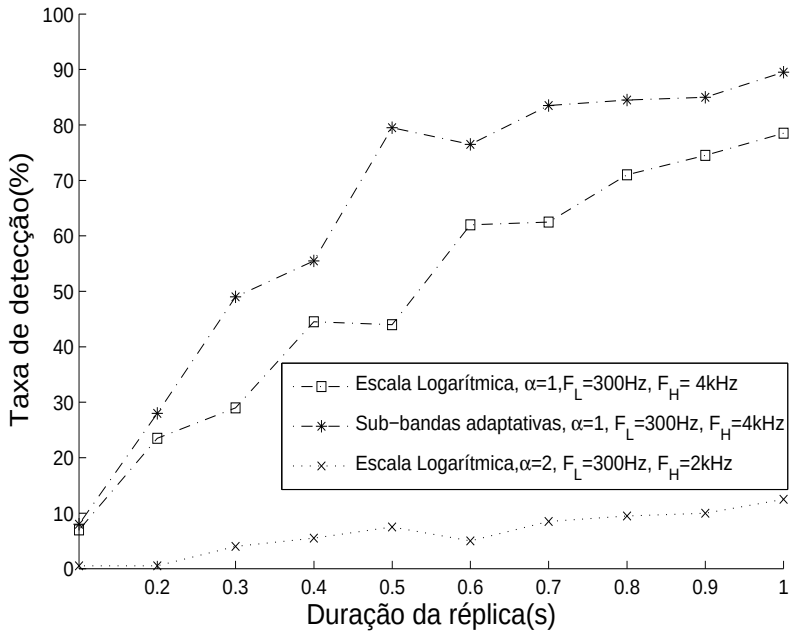

Figura 3.17: Taxa média de detecção de réplica com duração de $D_R \in \{100ms, 200ms, ..., 1s\}$, com distorção de expansão temporal como descrito anteriormente, variando o método de divisão de sub-bandas, a banda de frequência e α.

comuns em codificações perceptuais com alta taxa de compressão. A comparação destes resultados, que aplicam compressão com taxa de 16kbps, com os resultados obtidos em [2], que aplicam compressão com taxa de 32kbps, não é possível. Mas é interessante notar que no exemplo de bloco de AF de uma música comprimida a 32kbps, nenhum dos 216 sub-blocos são idênticos e apenas 5 sub-blocos tem apenas 1 bit diferente. Em [156], a robustez do método foi testada contra ataques intencionais de compressão MP3 com o objetivo de modificação da AF com a preservação do conteúdo perceptual. Argumenta-se que a robustez do método é limitada pelo fato da distribuição de $T[:, m]$, após deltas entre quadros e sub-bandas, estar concentrada próximo de zero e, portanto, qualquer modificação no sinal pode causar mudança do bit de AF pela quantização com limiar nulo. Em [147] a taxa de erro de bits do esquema de AF proposto por Haitsma [2] é observada para compressão MP3 a taxas de 128, 80 e 32kbps. Observa-se que a taxa de erro de bits é inversamente proporcional ao quadrado da relação entre sinal e o ruído de codificação. Dessa forma, quanto maior a taxa de compressão e menor a taxa de bits de codificação, maior é a taxa de erro de bits de AF. A análise mostra também que as regiões do espectrograma com baixa energia são mais suscetíveis a erros de bits de AF devido ao ruído de codificação. Isso ocorre devido à quantização final com limiar nulo. É proposto o uso de um peso sobre os bits de AF, desconsiderando aqueles bits referentes a regiões de baixa energia do espectrograma. Em [140] a análise teórica, com um sinal não correlacionado no tempo, a BER é de 1% e 3% para uma relação sinal/ruído de 30dB e 20dB, respectivamente. A BER média para as AF de 11 músicas e suas versões

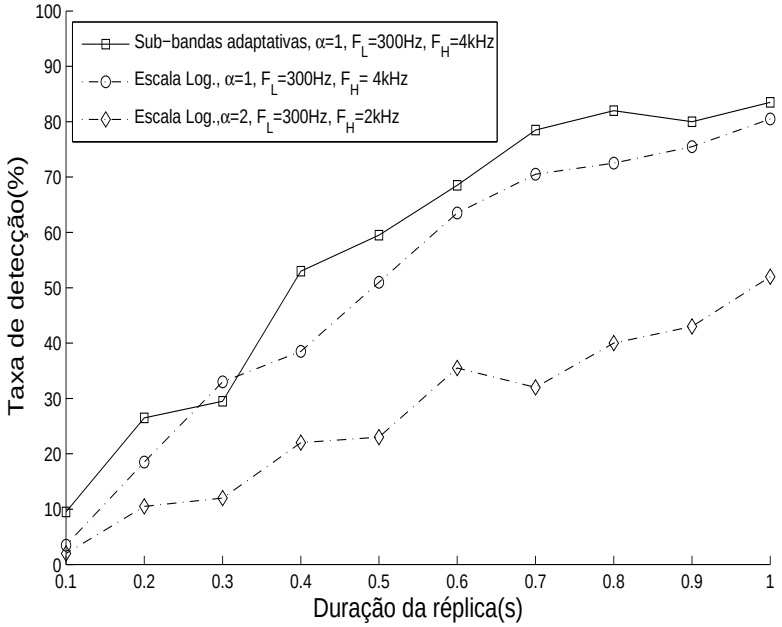

Figura 3.18: Taxa média e detecção de réplicas com duração $D_R \in \{100ms, 200ms, ..., 1s\}$, com subsequente compressão MP3PRO CBR 16kbps, para algumas configurações, variando o método de divisão de sub-bandas, a banda de frequência e α.

comprimidas com WMA (*Windows Media Audio*), por exemplo, é 1,3% e 5% para uma relação sinal/ruído de compressão de 30dB e 20dB, respectivamente. No Capítulo 4, adaptações e ajustes de parâmetros são propostas para melhorar a robustez contra o ruído aditivo, o que também melhora o desempenho de detecção de áudios comprimidos.

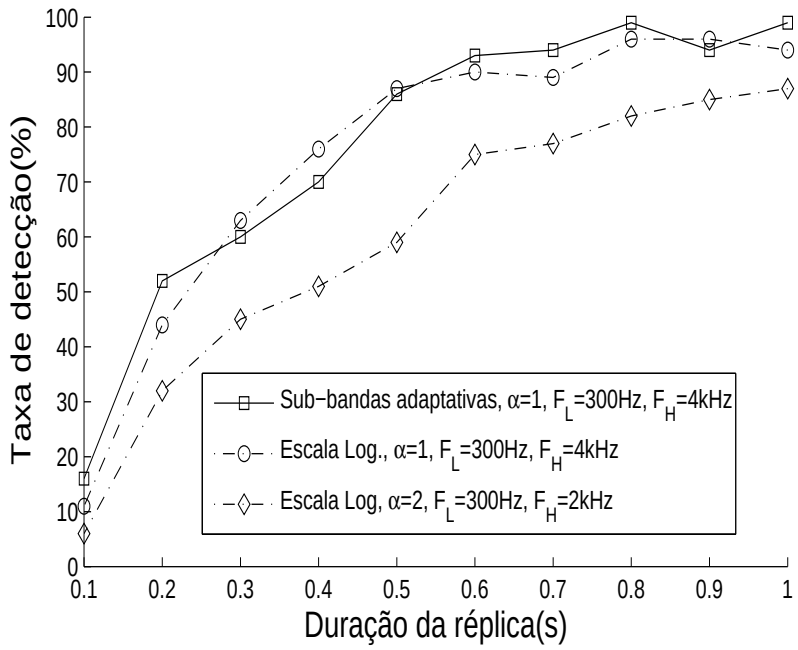

Figura 3.19: Taxa média e detecção de réplicas com duração $D_R \in \{100ms, 200ms, ..., 1s\}$, com subsequente compressão AAC 16kbps, para algumas configurações, variando o método de divisão de sub-bandas, a banda de frequência e α.

4- MELHORIA DA ROBUSTEZ CONTRA INSERÇÃO DE RUÍDO

> To improve is to change; to be perfect is to change often.
>
> Winston Churchill

Apesar de apresentar uma boa robustez contra distorções em frequência ou em amplitude, o esquema inicialmente proposto no capítulo 3 apresentou um desempenho regular para a detecção de réplicas mascaradas por inserção de ruído branco ou compressão de áudio.

Como observado, o esquema proposto por Haitsma [2] apresenta uma robustez regular contra inserção de ruído gaussiano. Uma abordagem interessante para melhorar a robustez seria reforçar o peso das componentes de maior energia do sinal, através uma representação adaptativa, como proposto em [52], na qual a DFT é obtida, dividida em sub-bandas conforme uma escala fixa, mas apenas as sub-bandas que contenham picos de energia espectral são empregadas para representar o sinal. Nesse caso há uma redução da dimensionalidade.

De outra forma, optamos por fazer uma análise da probabilidade de erro de bit de AF, com base na análise da distribuição de $T[n,m]$. Como o mascaramento por inserção de ruído distorce tanto o trecho de origem como o trecho de destino, a distorção é estimada pela diferença entre $F_{A+\mathcal{N}_1}[n,m]$ e $F_{A+\mathcal{N}_2}[n,m]$. Na análise teórica feita em [140], descrita sucintamente na Seção 2.4.1.2, a probabilidade de erro de bit de AF, pela inversão de sinal de $T[n,m]$, decorrente da inserção de ruído no sinal de voz, é dada por

$$P_e[n,m] = Pr\{F_{A+\mathcal{N}_1}[n,m] \neq F_{A+\mathcal{N}_2}[n,m]\} \tag{4.1}$$

$$= Pr[(T_{A+\mathcal{N}_1}[n,m] \leq 0, T_{A+\mathcal{N}_2}[n,m] \geq 0) \vee (T_{A+\mathcal{N}_1}[n,m] \geq 0, T_{A+\mathcal{N}_2}[n,m] \leq 0)]. \tag{4.2}$$

Assumindo que o sinal de áudio é não correlacionado no tempo, $P_e[n,m]$ pode ser obtida em termos das variâncias de $T_{A+\mathcal{N}_1}[n,m]$ e $T_{A+\mathcal{N}_1}[n,m] - T_{A+\mathcal{N}_2}[n,m]$. Para uma SNR alta, aproximamos a variância $VAR(T_{A+\mathcal{N}_1}[n,m])$ para $VAR(T_A[n,m])$, então

$$P_e[n,m] = \frac{1}{\pi} \arctan\left(\sqrt{\frac{VAR(T_{A+\mathcal{N}_1}[n,m] - T_{A+\mathcal{N}_2}[n,m])}{VAR(T_A[n,m])}}\right). \tag{4.3}$$

Como a função *arctan* é estritamente crescente, a probabilidade de erro pode ser reduzida pela redução da razão entre a variância de $T[n,,m]$ para a distorção e para o sinal não distorcido. Devido à dificuldade de modelamento do sinal de áudio, optamos por fazer uma análise empírica das distribuições de $T[n,m]$ para algumas modificações no sistema.

Uma alternativa interessante para melhoria da razão de variâncias é o emprego de expoentes α maiores. O uso de expoentes maiores de componentes espectrais para aumentar a robustez contra ruído já foi aplicado em outros esquemas, como em [161]. Na Seção 4.2 é proposta a adaptação do expoente α e dos limites da banda de frequência para cada áudio.

Nas Seções 4.3 e 4.4, também é avaliado o efeito na distribuição de $T[n,,m]$ pelo ajuste de parâmetros como o fator de sobreposição de quadros Ω_F, a duração do quadro D_F, e as distâncias entre os deltas entre quadros e entre sub-bandas.

O sistema adaptativo é análisado usando um novo *corpus*. Ademais, são usados novos métodos de estimação de Falsos Positivos, e de análise de precisão, unicidade e robustez, conforme descrito na Seção 4.1.

4.1 NOVA METODOLOGIA DE ANÁLISE DO SISTEMA

Para permitir a repetibilidade dos testes, nas novas análises de desempenho é usado um *corpus* de acesso livre, CHAINS (*CHaracterizing INdividual Speakers*) [27], contendo vozes de 36 falantes, amostradas a 44,1KHz com 16bits/amostra, com alta qualidade, SNR estimada em 70dB, e contendo repetições de locuções intrasentença. Para se construir um conjunto de teste com SNR baixa, mais coerente com os áudios questionados comumente examinados, foi artificialmente inserido ruído obtido do *corpus* de ruído também de acesso livre DEMAND (*Diverse Environments Multichannel Acoustic Noise Database*) [28]. Dessa forma, para os testes foi gerado um conjunto de 25 áudios $A_k, k = 1, 2, ...25$, para SNR=20dB.

4.1.1 Estimação do número de Falsos Positivos de Réplica

A análise teórica do número de Falsos Positivos de Quadros do Capítulo 3 considerou uma distribuição i.i.d. dos bits da AF. Sob esta hipótese as posições dos Falsos Positivos de Quadros na matriz de autossimilaridade possuem uma distribuição dispersa em $\mathbf{M}$, logo todos quadros detectados, como possível réplica, devem ser checados perceptualmente. Entretanto, as simulações mostraram que o número de Falsos Positivos de Quadros real é maior que o valor estimado pela análise teórica, o que mostra que a distribuição real dos bits de AF não se aproxima de uma distribuição i.i.d.

Uma análise visual das matrizes reais de autossimilaridade revelou que os Falsos Positivos de Quadros podem possuir uma correlação temporal e podem ser agrupados, como ilustra a Figura 4.1, para um áudio de 60s sem réplicas referente a texto não-controlado. A matriz possui 31 Falsos Positivos de Quadros distribuídos em poucos agrupamentos. A análise de oitiva revela que os agrupamentos de quadros detectados em trechos curtos de áudio podem corresponder a um mesmo fone. Como a verificação de cada agrupamento de elementos detectados, pela oitiva do trecho e análise de forma de onda, pode ser feita de uma única vez, podemos ajustar os parâmetros do sistema para limitar o número mínimo de agrupamentos observados em simulações com um conjunto de áudios do *corpus*.

Para contabilizar o número de agrupamentos, tratamos a matriz de autossimilaridade como uma imagem binária. Aplicamos uma fechamento da imagem binária com um elemento estruturante de raio 3, em seguida calculamos o número de elementos 8-conectados. A Figura 4.1 mostra o número menor de elementos 8-conectados identificados, 14.

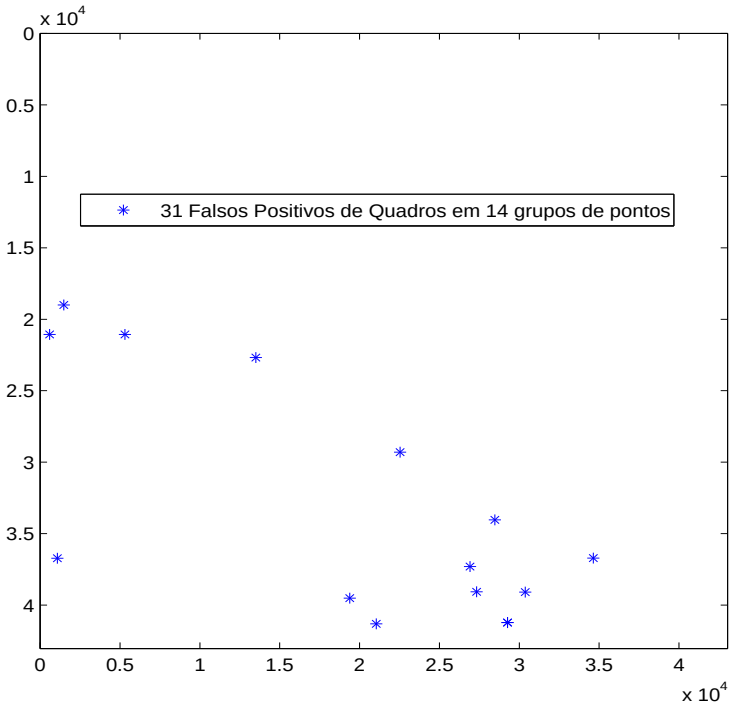

Figura 4.1: Matriz de autossimilaridade de um áudio de 60s sem réplicas referente a texto não-controlado, com 31 Falsos Positivos de Quadros, agrupados em 14 elementos 8-conectados.

4.1.2 Análise de unicidade, precisão e robustez

Nas análises subsequentes, a unicidade, a precisão e a robustez são analisadas com base nas funções de densidade de probabilidade (FDP), obtidas a partir dos histogramas da distância de Hamming entre AF's:

1. Unicidade: É analisada através do histograma da distância de Hamming entre AF para todos pares de posições de quadros em um mesmo áudio A_k de N_F quadros,

$$Hist(\{\delta(F_{A_k}[i,:], F_{A_k}[j,:]), i = 1, 2..., (N_F - 0,2/D_F), j = i + 0,2/D_F..., N_F\}). \quad (4.4)$$

   Quadros detectados separados por menos de $200ms$, com correlação temporal, são descartados pelo filtro de correlação temporal. A unicidade pode ser avaliada comparando a semelhança da distribuição binomial ($N = N_{bits}, p = 0,5$) com o histograma obtido.

2. Precisão: Visa medir a taxa de detecção de réplicas em áudios não distorcidos. Neste caso, a taxa de detecção é afetada pelo desalinhamento entre quadros. Considera-se que o desalinhamento possui uma distribuição uniforme entre $[0, \Delta_F]$. Dessa forma, para medir a taxa de detecção geramos 6 versões deslocadas de $A_k^j[i] = A_k[i + j.\Delta_F/6], j = 1, 2, ..., 6$, e calculamos o histograma da distância de Hamming entre as AF's de mesmo índice de quadro,

$$Hist(\{\delta(F_{A_k}[i,:], F_{A_k^j}[i,:]), i = 1, 2, ...N_F, j = 1, 2, ...6\}). \quad (4.5)$$

3. Robustez: Considerando que em um mascaramento mais provável o ruído seria adicionado ao longo de todo o sinal de áudio distorcendo tanto o trecho original quanto o trecho replicado, medimos a distância de Hamming entre as AF's de duas versões de um áudio A_k, distorcidas pela adição de ruído gaussiano $\mathcal{N}$, $A_k + \mathcal{N}$, para SNR=15dB, 20dB e 25dB. Obtemos o histograma da distância de Hamming entre AF's de mesmo índice de quadro,

$$Hist(\{\delta(F_{(A_k+\mathcal{N}_1)}[i,:], F_{(A_k+\mathcal{N}_2)}[i,:]), i = 1, 2...N_F\}). \quad (4.6)$$

Para que se obtenha uma estimativa real dessas distribuições, os histogramas são obtidos para a concatenação de vários áudios A_k, de diferentes locutores e de mesma duração.

4.1.3 Estimação da probabilidade de detecção de réplica a partir da taxa de detecção de quadros

A estimação da probabilidade de detecção de réplica a partir da taxa de detecção de quadros simplifica os testes de robustez, pois dispensa o criação de grandes conjuntos de áudios de testes contendo réplica, como feito no Capítulo 3. Dessa forma, o esforço computacional da otimização do sistema pelo ajuste de parâmetros é reduzido significativamente.

Seja $FDP(k), k = 0, 1, ...N_{bits}$ a função de densidade de probabilidade de erro de bits, medida nos testes de robustez, a taxa de detecção de quadros de AF é dada por $P_Q = \sum_{k=0}^{d_{max}} FDP(k)$.

A probabilidade de detecção de réplicas depende do número de quadros contidos nas réplicas, $N = (D_R - D_F - \Delta_F)/\Delta_F$, e, portanto, aumenta com a duração da réplica D_R. Se considerarmos que os erros de Falso Negativo de Quadros de AF são independentes entre si, a probabilidade de detecção de réplica para o critério simples de detecção é dada por

$$Pr\{\delta_R = 1 | H_{i,j} \forall (i,j), (i \in [n_1, n_1 + N]) \wedge (j = i + n_2 - n_1)\} = 1 - (1 - P_Q)^N. \quad (4.7)$$

Ressaltamos que a hipótese de correlação temporal nula entre os Falsos Positivos de quadros não é válida para todos os tipos de distorção. Distorções como compressão de áudio podem gerar Falsos Negativos de Quadros em surto devido a perdas ou artefatos de compressão localizados.

4.2 ADAPTAÇÃO DE α E DA BANDA DE FREQUÊNCIA PARA CADA ÁUDIO

Para melhorar a razão de variâncias da Eq. (4.3) e reduzir a probabilidade de erro de bit, sugerimos o emprego de expoentes α maiores. O uso de expoentes maiores de coeficientes espectrais para aumentar a robustez contra ruído já foi adotado em outros sistemas, como em [161].

O efeito do aumento de α, de $\alpha = 1$ para $\alpha = 1,5$, é ilustrado na Figura 4.2, com desvios padrões das distribuições de $T_A[n, m], m = 1, 2, ...N_{bit}$, para um áudio A (a esquerda), e os desvios padrões de $T_{A+\mathcal{N}_1}[n, m] - T_{A_1+\mathcal{N}_2}[n, m]$, entre trechos mascarados pela adição de ruído branco gaussiano a SNR=20dB (direita). Observa-se que com o aumento de α, os desvios padrões de $T_A[n, m], m = 1, 2, ...N_{bit}$ aumentam entre 3 a 8 vezes, enquanto o desvio padrão dos erros aumentam entre $1,5$ e 2. Dessa forma, o aumento de α tende a reduzir a razão das variâncias para a inserção de ruído branco gaussiano em patamar abaixo do nível do áudio.

Entretanto, o uso de expoentes muito elevados pode mascarar a informação de componentes com amplitude intermediária. Ademais, se uma escala fixa for empregada, subbandas com maior conteúdo harmônico podem produzir bits com baixa variância como visto anteriormente.

Com base nessas observações, propomos um esquema adaptativo onde α é ajustado

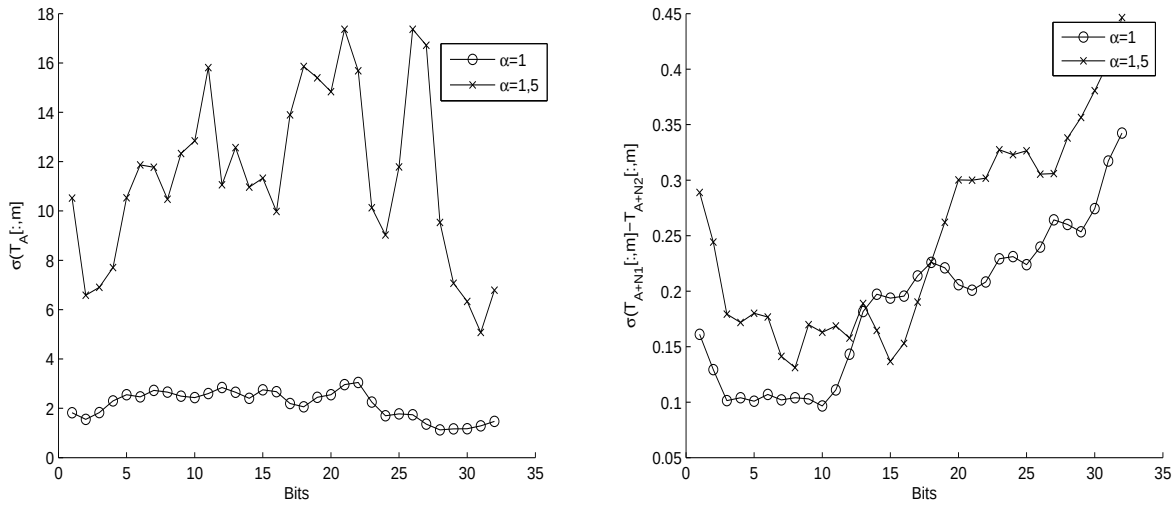

Figura 4.2: Desvios padrões das distribuições de $T_A[n,k]$ para as sub-bandas $k = 1, 2, ...N_{bit}$ para o áudio A (a esquerda), e desvios padrões de $T_{A+\mathcal{N}_1}[n,k] - T_{A+\mathcal{N}_2}[n,k]$ (direita), devido à adição de ruído branco gaussiano a SNR=20dB, para $\alpha = 1$ e $\alpha = 1,5$.

automaticamente para cada áudio. Para aumentar a robustez, com base na condição imposta pelo método de divisão de sub-bandas, Eq. (3.31), o valor máximo de α é definido implicitamente por

$$\max_k \left(\sum_{n=1}^{N_F} |S[n,k]|^\alpha \right) = \frac{\left(\sum_{n=1}^{N_F} \sum_{k=L_{bit+1}}^{L[0]} |S[n,k]|^\alpha \right)}{N_{bit}+1}. \tag{4.8}$$

Dessa forma, a componente espectral mais forte ao longo do sinal é empregada na definição da sub-banda mais estreita com apenas um bin de frequência.

O critério de maximização de α pode ser útil também para a definição dos índices dos coeficientes espectrais limites da banda $L[0]$ e $L[N_{bits}]$. Definimos uma banda mínima de $[300Hz, 3800Hz]$, e o expoente como uma função $\alpha(L[0], L[N_{bits}])$ pode ser obtido pela Eq. (4.8), para todas as combinações de $L[0] \in [0, 300D_F]$, e $L[N_{bit}] \in [3800D_F, (R/2)D_F]$. Logo, obtemos o valor máximo de $\alpha(L[0], L[N_{bits}])$ ao tempo em que definimos os limites $L[0]$ e $L[N_{bits}]$:

$$\begin{aligned} &(L[0], L[N_{bit}]) = \arg\max_{L[0], L[N_{bit}]} \left(\alpha(L[0], L[N_{bit}]) \right), \\ &L[0] \in [0, 300D_F], L[N_{bit}] \in [3800D_F, (R/2)D_F]. \end{aligned} \tag{4.9}$$

Dessa forma, expandindo a banda de frequência podemos, para alguns áudios, empregar valores ainda maiores de α, melhorando a robustez.

A Figura 4.3 ilustra, para um conjunto 20 áudios, os resultados dos testes de unicidade, precisão e robustez contra ruído gaussiano aditivo branco a uma SNR= 25dB, para três

configurações: 1)o uso de escala fixa de sub-bandas, $\alpha = 2$, $F_L = 300$ e $F_H = 3800$ (acima); 2) a adaptação de α e $L[i], i = 1, 2, ...N_{bit}$, com $F_L = 300Hz$ e $F_H = 3800Hz$(meio); 3) para a adaptação de α, dos limites da banda ($L[0], L[N_{bit+1}]$) e sub-bandas $L[i], i = 1, 2, ...N_{bit}$ (abaixo).

Observa-se uma redução da média de erro de bits no gráfico de robustez para a adaptação de α e $L[i], i = 1, ...N_{bit}$ (direita-meio), comparada à média de erro para o esquema proposto por Haitsma [2] (direita-acima). A robustez é ainda melhor para o esquema com ajuste da banda que permite um aumento de α (abaixo-direita). Os graficos à direita mostram ainda uma curva pontilhada da taxa média de erro de bit por sub-banda, onde se observa para o esquema adaptativo proposto com ajuste de banda (direita-abaixo) uma distribuição mais uniforme, comparada à taxa de erros de bits do esquema proposto por Haitsma [2] (direita-acima). Não se observaram variações significativas das distribuições de erro nas análises de unicidade (esquerda) e precisão (centro) para as três configurações testadas.

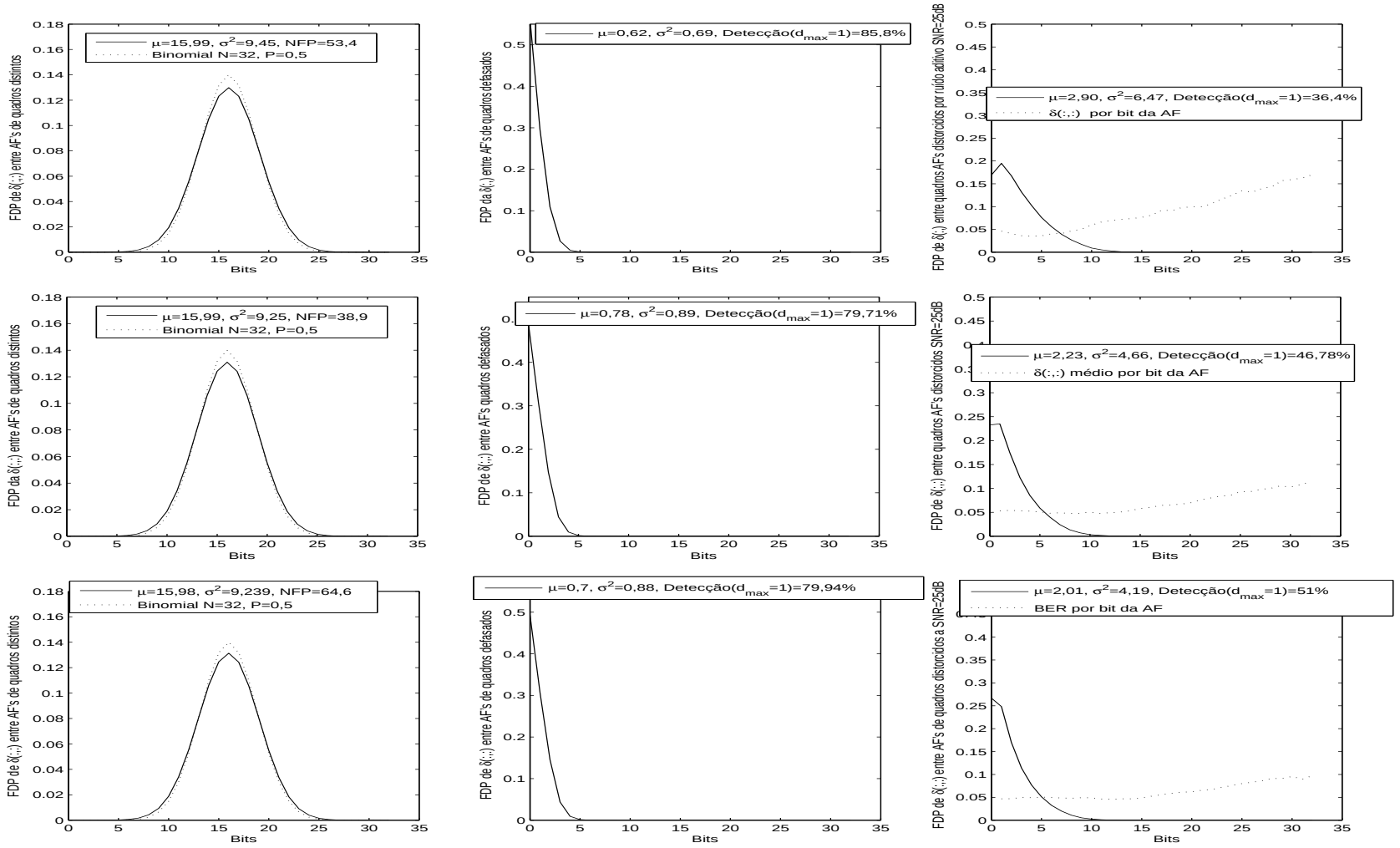

Figura 4.3: FDP de $\delta(:,:)$ entre AF's para testes de unicidade (esquerda), precisão (centro) e robustez contra ruído branco aditivo a SNR= 25dB (direita), para o uso de escala fixa de sub-bandas, $\alpha = 2$, $F_L = 300$ e $F_H = 3800$ (acima); para a adaptação de α e $L[i], i = 1, ...N_{bit}$, com $F_L = 300Hz$ e $F_H = 3800Hz$ (meio); e para a adaptação de α, dos limites $(L[0], ..., L[N_{bit+1}])$ (abaixo).

A taxa de detecção de quadros é medida para diversos níveis de ruído aditivo, usando o esquema com ajuste dos índices limites $L[i], i = 0, 1, ...N_{bit} + 1$ e de α, conforme ilustra a Figura 4.4. A taxa de detecção de quadros aumenta quase linearmente com a SNR no intervalo entre 15dB a 35dB. Para $\Omega_F = 95\%$ e $D_F = 90ms$, uma réplica de $100ms$ contém apenas um quadro de AF, logo a taxa de detecção de réplica de $100ms$ equivale à taxa de detecção de quadros ilustrada na Figura 4.4.

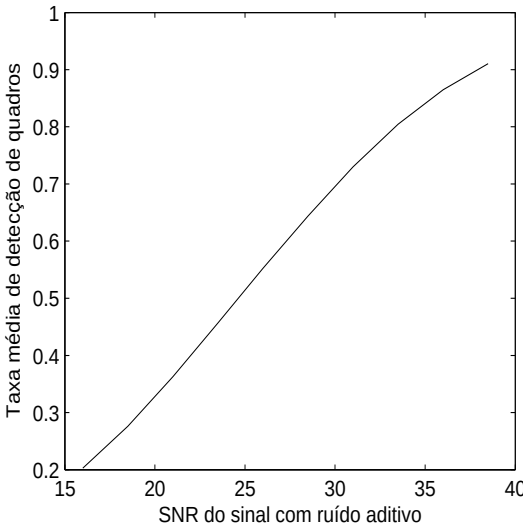

Figura 4.4: Taxa de detecção de quadros (e réplica de 100ms) para diversos níveis de ruído gaussiano branco aditivo, usando $D_F = 90ms$ e $\Omega_F = 95\%$ no esquema com ajuste dos limites $(L[0], ..., L[N_{bit+1}])$ e de α.

4.3 AJUSTE DE Ω_F E D_F

Para analisarmos o efeito do ajustes de D_F, e considerando que o esquema aplica um delta entre quadros, a Figura 4.5 ilustra as posições dos quadros subsequentes n e $n+1$ para o aumento de D_F, mantendo-se Ω_F fixo. Os trechos hachurados em cinza escuro no eixo das abscissas indicam a diferença entre os intervalos dos quadros subsequentes. O aumento de D_F, mantendo-se Ω_F fixo, aumenta a separação entre os trechos diferentes, tendendo a reduzir a correlação das distribuições espectrais nesses intervalos, e, pela aplicação do delta entre quadros, tende a aumentar o desvio padrão de $|T[n, k], k = 1, ...N_{bit}|$, melhorando a robustez. Para Ω_F constante, a razão entre a média das defasagens de quadros $[0, \Delta_F/2]$, que afetam o erro em $V[n, k]$, e D_F, que afeta o valor de $V[n, k]$, se mantém fixa. Logo, é razoável supor que a precisão não seja afetada pela variação de Δ_F, para Ω_F constante.

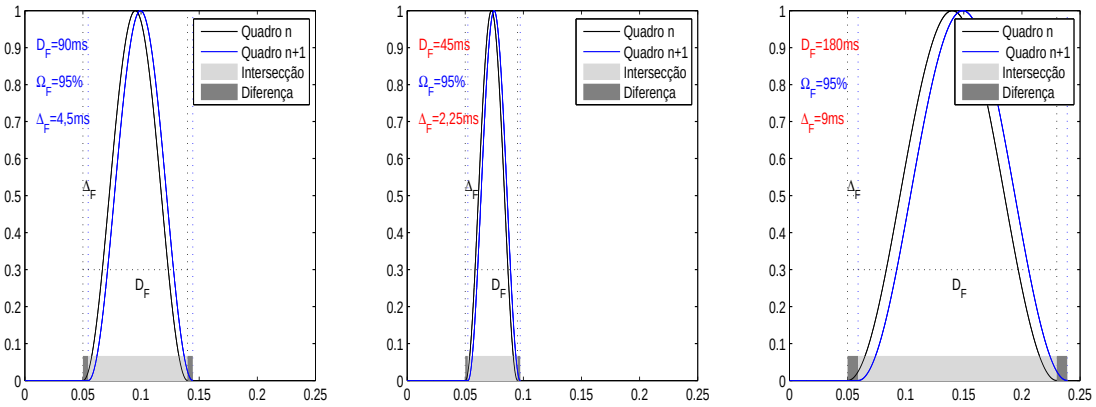

Figura 4.5: Variação de D_F, mantendo Ω_F fixo. Os trechos hachurados em cinza escuro no eixo das abscissas indicam a diferença nos intervalos entre os quadros n e $n+1$, e o trechos hachurados em claro indicam a intersecção. O aumento de D_F aumenta a separação entre as posições das diferenças entre os intervalos.

Na comparação de desempenho de detecção, N_{bits} é previamente ajustado com base em testes com o *corpus* para cada configuração de parâmetros, de forma a limitar o número de agrupamentos da matriz de autossimilaridade em no máximo 10. Os testes de precisão, robustez, e o teste conjunto de precisão e robustez são realizados para diversos valores de $D_F(ms) \in \{40, 50, 60, 70, 80, 90, 100\}$, mantendo-se $\Omega_F = 0,95$ fixo, para o método proposto na Seção 4.2, com adaptação de α e dos limites $L[i], i = 0, 1, ...N_{bit} + 1$. As curvas de densidade de probabilidade são ilustradas na Figura 4.6. Observa-se que a precisão é pouco afetada se Ω_F é mantido fixo. Também observa-se uma melhoria da taxa de detecção de quadros com o aumento de D_F.

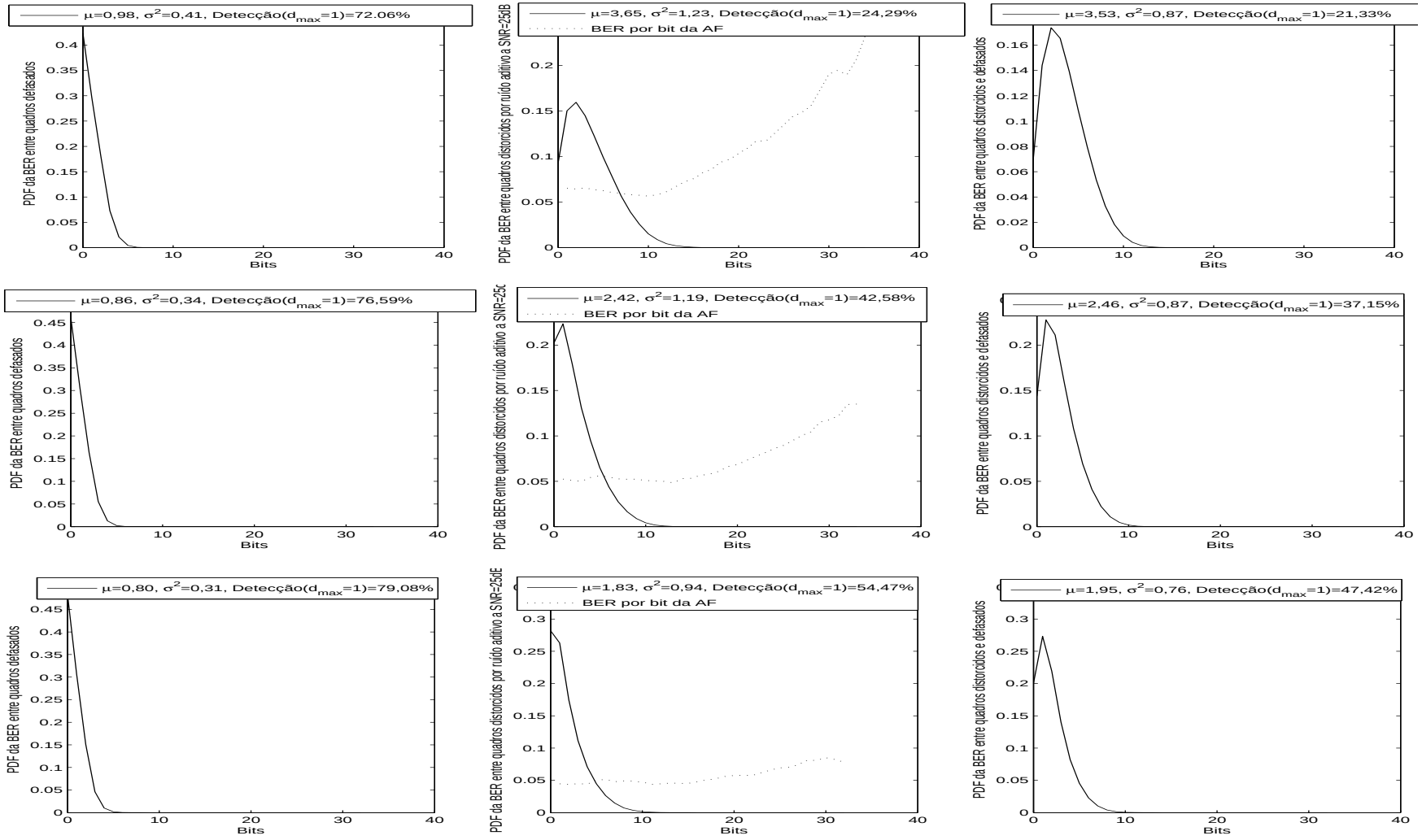

Figura 4.6: FDP para testes de precisão (esquerda), robustez contra ruído branco aditivo a SNR= 25dB (centro) e conjunto de robustez e precisão (direita), $\Omega_F = 95\%$, para com $D_F = 50ms$ (acima), $D_F = 70ms$(meio), e $D_F = 90ms$ (abaixo).

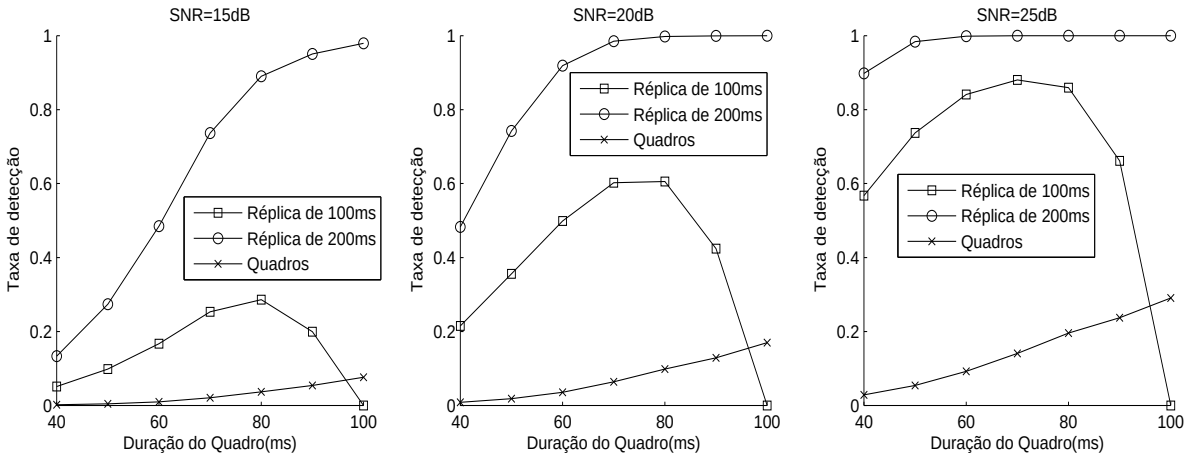

Figura 4.7: Taxa de detecção de réplica para $D_F \in [40ms, 100ms]$, para áudio mascarado com ruído branco a SNR= 25dB (esquerda), 20dB (centro) e 15dB (direita), para o esquema adaptativo, para $\Omega_F = 0,95$.

A Figura 4.7 mostra as probabilidades de detecção de réplica do esquema adaptativo para áudio mascarado com ruído branco com SNR= 25dB (esquerda), 20dB (centro) e 15dB (direita), calculadas a partir da taxa de detecção de quadros pela Eq. (4.7), para $D_F(ms) \in [40, 100]$ e $\Omega_F = 95\%$. Observa-se que para réplicas de 200ms a melhor probabilidade de detecção é obtida para quadros mais longos, para todas as SNR analisadas. Para o mascaramento mais severo com SNR=15dB, a soma das probabilidades de detecção de réplica de $100ms$ e $200ms$ é maximizada para $D_F = 80ms$. Dessa forma, ajustamos $D_F = 80ms$ para otimizar o desempenho de detecção.

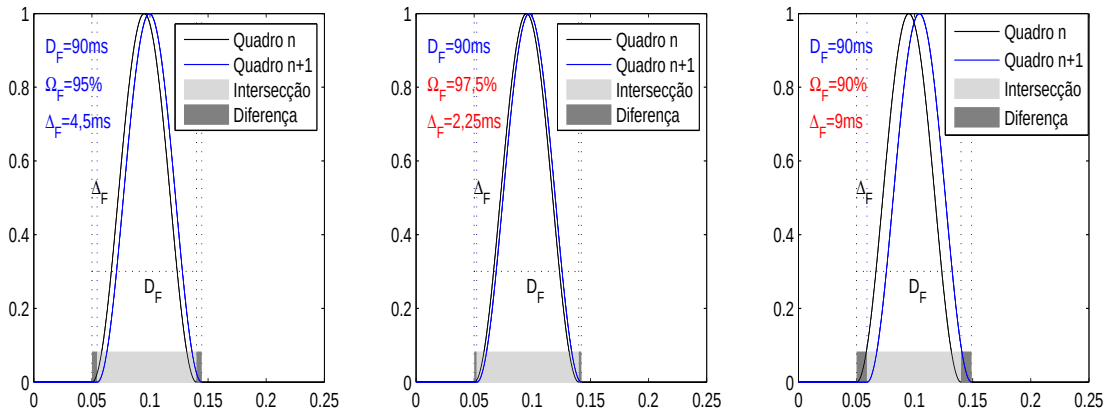

Figura 4.8: Variação de Δ_F, mantendo D_F fixo. Ao trechos hachurados em cinza escuro no eixo das abscissas indicam a diferença nos intervalos entre os quadros n e $n+1$, e o trechos hachurados em claro indicam a intersecção.

Para analisarmos o efeito do ajustes de Δ_F, a Figura 4.8 ilustra as posições dos quadros subsequentes n e $n+1$, para D_F fixo, variando Ω_F. O aumento de Δ_F, mantendo-se D_F

fixo, reduz o espaçamento Ω_F, portanto aumenta o erro de bits por desalinhamento de quadros na segmentação entre os limites de trechos replicados, piorando a precisão. A robustez deve melhorar com o aumento de Δ_F, devido ao efeito do delta entre quadros em $|T[:,m], m = 1, 2, ...N_{bit}|$, com a redução do intervalo de intersecção entre os quadros n e $n+1$ e com o aumento da separação temporal dos intervalos disjuntos.

Na comparação de desempenho de detecção, N_{bits} é previamente ajustado para limitar o valor esperado do número de agrupamentos da matriz de autossimilaridade em no máximo 10. Os testes de precisão, robustez, e o teste conjunto de precisão e robustez são realizados para diversos valores de $\Omega_F(\%) \in \{94, 95, 96, 97, 98\}$, mantendo-se $D_F = 80ms$ fixo, para o método proposto na Seção 4.2, com adaptação de α e dos limites $L[i], i = 0, 1, ...N_{bit}+1$. A função de densidade de probabilidade na Figura 4.9 mostra que a precisão piora com a redução de Ω_F. Entretanto a redução de Ω_F leva a uma melhoria da robustez, com aumento da taxa de detecção de quadros, pois há um aumento da relação entre as durações dos intervalos disjuntos e do intervalo comum das janelas vizinhas, com um aumento da variância de $T[n, m]$. A FDP obtida com análise conjunta de precisão e robustez mostra que, para esse nível de ruído (SNR=25dB), a robustez aumenta com o aumento de Ω_F.

A Figura 4.10 mostra as taxas de detecção de réplicas do esquema adaptativo para áudio mascarado com ruído branco aditivo de SNR=25dB (esquerda), 20dB (centro) e 15dB (direita), calculadas a partir da taxa de detecção de quadros pela Eq. (4.7), para diversos valores de $\Omega_F(\%) \in \{94, 95, 96, 97, 98\}$, com $D_F = 80ms$. Observa-se que a taxa de detecção de réplicas aumenta com Ω_F. Dessa forma, ajustamos $\Omega_F = 98\%$, que será usado nos testes subsequentes.

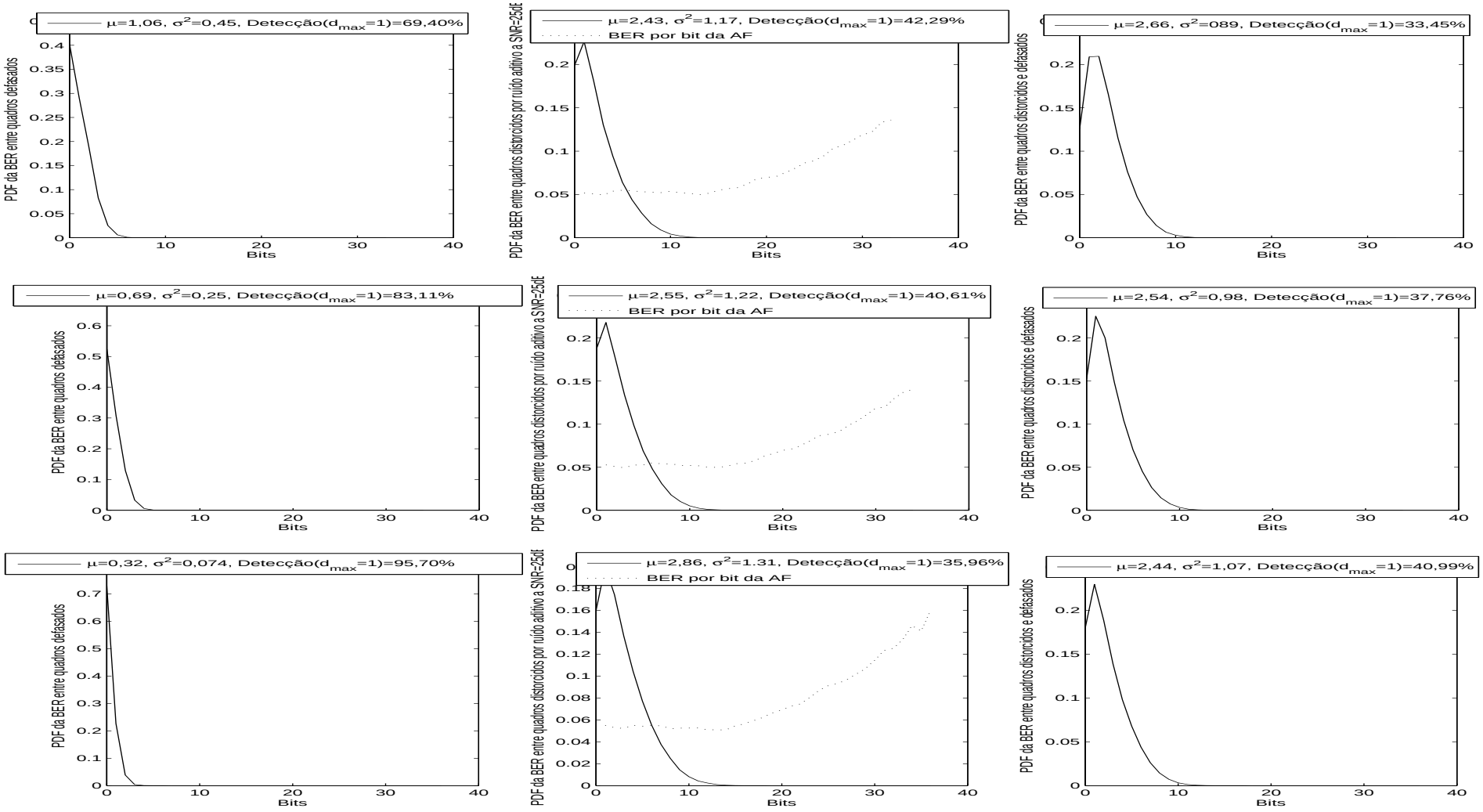

Figura 4.9: FDP para testes de precisão (esquerda), robustez contra ruído branco aditivo a SNR= 25dB (centro) e conjunto de robustez e precisão (direita), com $D_F = 80ms$, para $\Omega_F = 94\%$ (acima), $\Omega_F = 96\%$ (meio), e $\Omega_F = 98\%$ (abaixo).

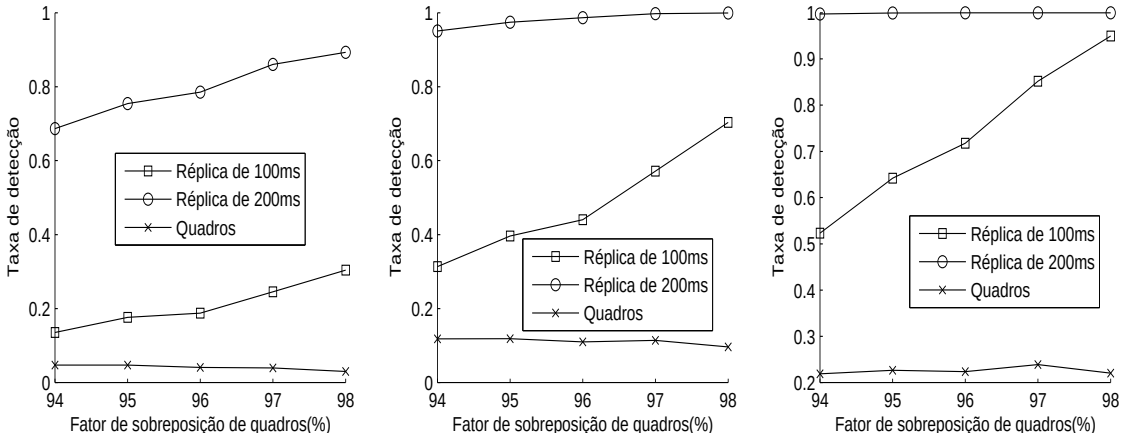

Figura 4.10: Taxa de detecção de para quadros e réplicas de 100ms e 200ms, para $\Omega_F(\%) \in [94, 99]$, para áudio mascarado com ruído branco a SNR=25dB (esquerda), 20dB (centro) e 15dB (direita), para o esquema adaptativo, para $D_F = 80ms$.

4.4 AJUSTE DAS DISTÂNCIAS DOS DELTAS ENTRE SUB-BANDAS E ENTRE QUADROS

Analisamos nesta seção se o emprego de deltas distantes entre sub-bandas ou entre quadros pode aumentar a variância das distribuições de $T[n, m]$, e, consequentemente, elevar a robustez contra inserção de ruído.

Em [151, 152, 153] argumenta-se que, devido à correlação das energias de bancos de filtro (FBE- *Filter Bank Energies*) de sub-bandas vizinhas ou de quadros vizinhos, a análise de sinal pela aplicação de deltas é pouco robusta contra a inserção de ruído. Entretanto, nenhuma análise de correlação de $W[n, m]$ é feita. Com base nos trabalhos de uso de filtros em FBE's para remover efeitos de distorções lineares de canal na aplicação de reconhecimento de voz [154, 155], além dos filtros na frequência e no tempo usados em [2], descritos com base na transformada Z, são testados dois outros filtros alternativos no tempo, vide Eqs. (4.10) e (4.10). Os filtros T2 e T3 correspondem a uma fórmula de regressão típica e ao filtro RASTA [153], respectivamente. Nos testes empíricos em [152] para o conjunto de teste mais ruidoso, com $d_{max} = 1$, o desempenho de detecção é melhorado em até 3% para o uso de T3.

$$H_{T2}(Z) = \sum_{k=1}^{K} k(Z^k - Z^{-k}) \tag{4.10}$$

$$H_{T3}(Z) = \frac{2 + Z^{-1} - Z^{-3} - 2Z^{-4}}{10Z^{-4}(1 - \alpha Z^{-1})} \tag{4.11}$$

Nos codificadores de voz, baseados no modelo fonte-filtro, a distribuição espectral do

sinal é considerada quasi-estacionária para curtos intervalos de tempo. Portanto, para quadros com duração muito curta, o delta entre quadros adjacentes $T[n,m] = V[n,m] - V[n-1,m]$ pode gerar valores de $T[n,m]$ próximos do limiar nulo de quantização e menos robustos contra distorção por inserção de ruído. Dessa forma, uma opção interessante seria o uso de delta entre quadros mais distantes, com afastamento de $N_{\Delta F}$ quadros, conforme

$$T[n,m] = V[n,m] - V[n - N_{\Delta_F}, m]. \tag{4.12}$$

O espaçamento temporal entre os quadros é dado por $\Delta_F N_{\Delta_F}$. A Figura 4.11 ilustra como a intersecção dos intervalos de quadros usados nos deltas cai com o aumento de $N_{\Delta F}$. A intersecção é nula para $N_{\Delta F} > 1/\Omega_F$. Além da redução do intervalo de intersecção entre os trechos, o aumento da distância entre os trechos não comuns dos quadros reduz a correlação da distribuição espectral e, com isso, aumenta a variância de $T[n,m]$. A Figura 4.11 ilustra janelas dispostas de forma semelhante à Figura 4.8, entretanto, ao contrário da variação de Ω_F onde Δ_F também varia, no caso da variação de N_{Δ_F}, Δ_F e D_F são constantes e, portanto, não é esperado o aumento dos erros decorrentes da defasagem de quadros.

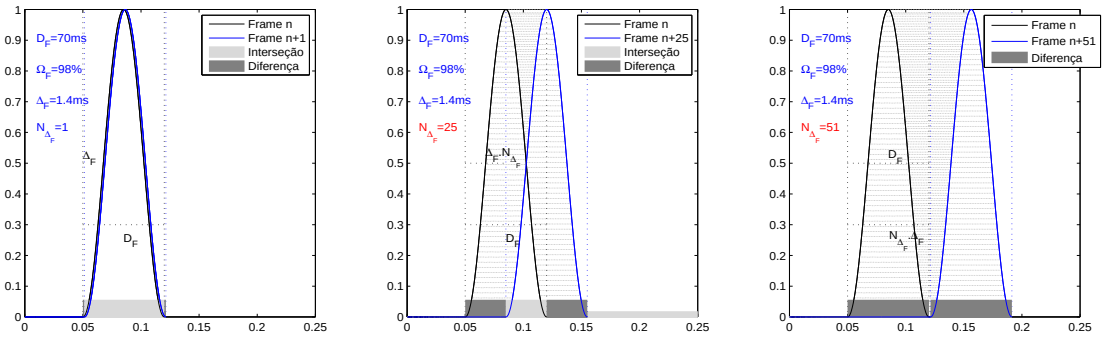

Figura 4.11: Posição dos quadros empregados no cálculo de $T[n,m]$, para $N_{\Delta F} = 1$, $N_{\Delta F} = 25$ ou $N_{\Delta F} = 51$. Os intervalos de intersecção entre os quadros são hachurados em cinza claro, e diminuem com o aumento de $N_{\Delta F}$, até se tornarem nulos para $N_{\Delta F} > 1/\Omega_F$.

A Figura 4.12 mostra para um áudio A, que o desvio padrão de $T[:,m]$ aumenta com a distância N_{Δ_F} até determinado ponto em que a correlação entre os intervalos disjuntos seja quase nula. O efeito do ruído aditivo a SNR=25dB é medido pelo desvio padrão de $(T_{A+\mathcal{N}_1}[:,m] - T_{A+\mathcal{N}_2}[:,m])$. Observa-se que as razões $\sigma(T_A[:,m])/\sigma(T_{A+\mathcal{N}_\infty}[:,m] - T_{A+\mathcal{N}}[:,m])$ aumentam com N_{Δ_F} até certo ponto, o que sugere que a robustez da detecção de quadros contra ruído branco aditivo aumenta com N_{Δ_F}.

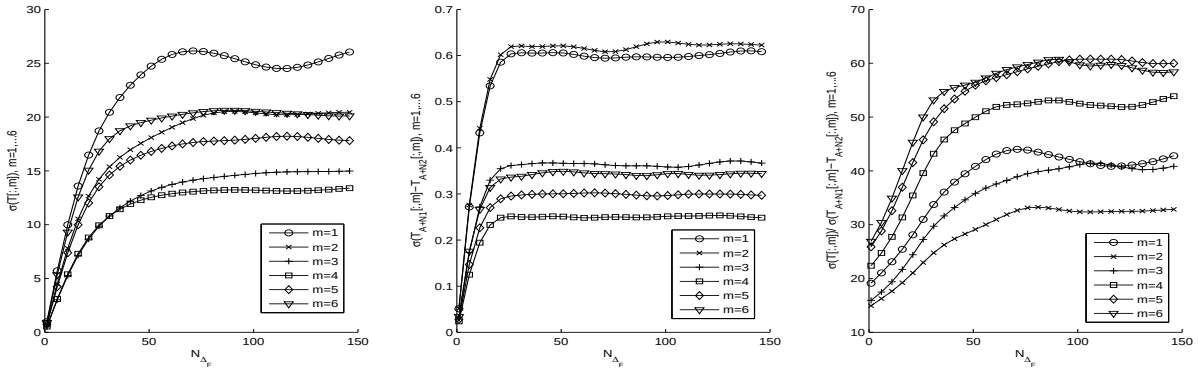

Figura 4.12: Desvio padrão de $T_A[:,m]$ (esquerda), $T_{A+\mathcal{N}_1}[:,m] - T_{A+\mathcal{N}_2}[:,m]$ (centro), e a razão entre os desvios padrões $\sigma(T_A[:,m])/\sigma(T_{A+\mathcal{N}_1}[:,m] - T_{A+\mathcal{N}_2}[:,m])$ (direita), para os bits $m = 1, 2, ...6$, para $N_{\Delta_F} \in \{1, 50\}$, $D_F = 80ms$, $\Omega_F = 98\%$.

Os testes de precisão, robustez, e o teste conjunto de precisão e robustez são realizados para diversos valores de $N_{\Delta_F} \in [1, 140]$, com $D_F = 80ms$, $\Omega = 98\%$, para o método proposto na Seção 4.2 com adaptação de α e dos limites $L[i], i = 0, 1, ...N_{bit}+1$. As curvas de densidade de probabilidade são ilustradas na Figura 4.13. Observa-se que a precisão, que esperávamos que se mantivesse constante, apresenta até uma pequena melhora com o aumento de N_{Δ_F}. A robustez contra inserção de ruído branco aditivo a SNR=25dB aumenta com N_{Δ_F}, conforme era esperado.

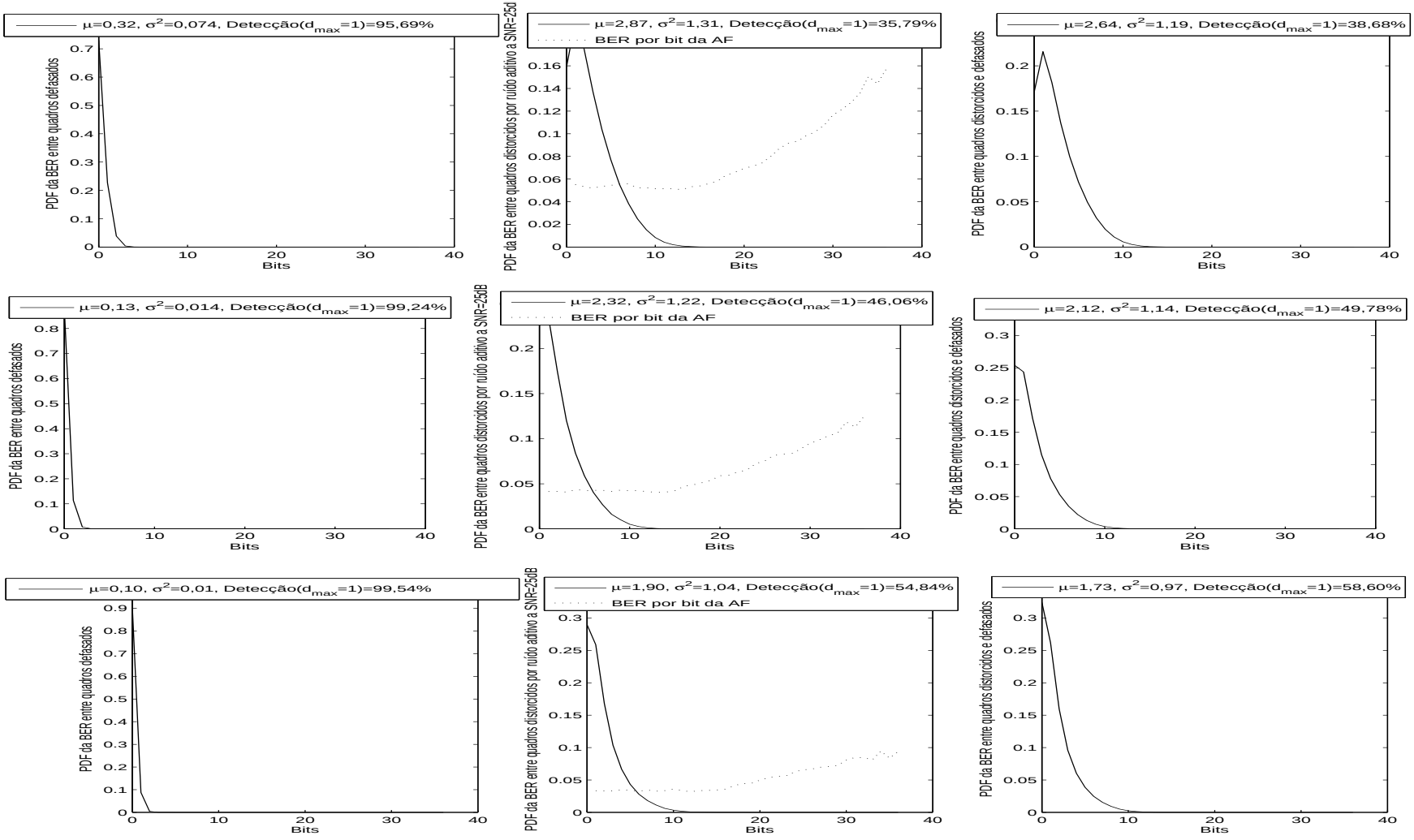

Figura 4.13: FDP para testes de precisão (esquerda), robustez contra ruído branco aditivo a SNR= 25dB (centro) e conjunto de robustez e precisão (direita), com $\Omega_F = 98\%$, $D_F = 80ms$, para $N_{\Delta_F} = 1$ (acima), $N_{\Delta_F} = 26$ (meio) e $N_{\Delta_F} = 51$ (abaixo).

Na comparação de desempenho de detecção, N_{bits} é previamente ajustado para limitar o valor esperado do número de agrupamentos da matriz de autossimilaridade em no máximo 10. A taxa de detecção de réplicas, dada pela Eq. (4.7), depende do número de quadros contidos nas réplicas, que para esse caso é dado por $N = (D_R - D_F - \Delta_F N_{\Delta_F})/\Delta_F$. A Figura 4.14 mostra as taxas de detecção de réplicas para diversos valores de $N_{\Delta_F} \in [1, 140]$, para áudio mascarado com ruído branco aditivo de SNR= 25dB, 20dB e 15dB, para o esquema adaptativo com $D_F = 80ms, \Omega_F = 98\%$. Observa-se que, como sugerido pela análise de $T[:, m]$, há uma melhora na taxa de detecção de quadros com o aumento de N_{Δ_F}. Entretanto, com a redução do número de quadros dentro das réplicas, não há uma melhora na taxa de detecção de réplicas. Dessa forma, ajustamos $N_{\Delta_F} = 1$ para otimizar o desempenho de detecção.

Cabe destacar a possibilidade de emprego de $N_{\Delta_F} > 1$ para melhorar a taxa de detecção de quadros em aplicações como identificação de música por conteúdo, já que nessa aplicação são empregadas granularidades bem maiores e a variação percentual de $N = (D_R - D_F - \Delta_F N_{\Delta_F})/\Delta_F$ decorrente do aumento de N_{Δ_F} não afeta tanto a taxa de detecção.

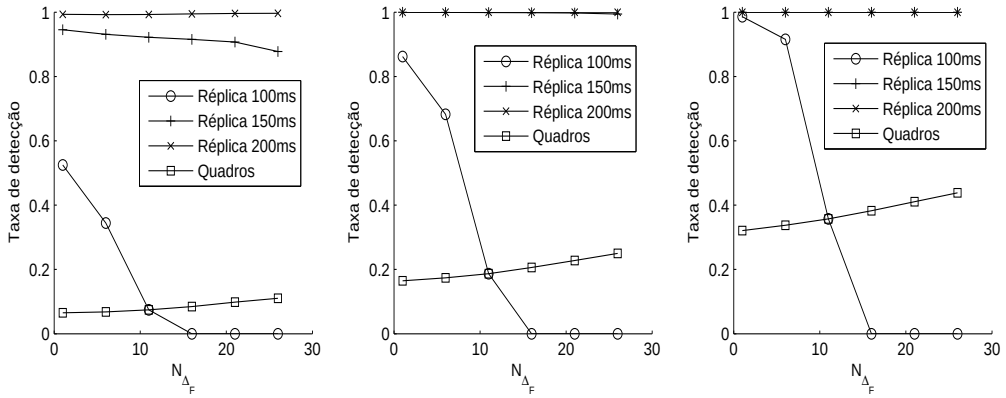

Figura 4.14: Taxa de detecção de quadros e réplicas de 100ms, 150ms e 200ms, para $N_{\Delta_F} \in [1, 50]$, para áudio mascarado com ruído branco aditivo a SNR=15dB (esquerda), 20dB (centro) e 25dB(direita), para o esquema adaptativo com $D_F = 80ms, \Omega_F = 98\%$.

Em [151, 152, 153], são testados dois outros filtros alternativos na frequência, vide Eqs. (4.13) e (4.14). Os filtros F2 e F3 correspondem a um filtro IIR passa-alta e um filtro FIR passa-banda, respectivamente. Nos testes empíricos para o conjunto de teste mais ruidoso, com $d_{max} = 1$, o desempenho de detecção é melhorado em até 5% para o uso de filtro F3.

$$H_{F2}(Z) = \frac{\eta(1 - Z^{-1})}{(\eta + 1)(1 + \frac{\eta - 1}{\eta + 1})Z^{-1})}, \eta = 0,5 \tag{4.13}$$

$$H_{F3}(Z) = Z - Z^{-1} \tag{4.14}$$

Para o esquema adaptativo, que equaliza a média de $W[n, m]$ para todas as sub-bandas,

o emprego de deltas entre sub-bandas distantes, N_{Δ_S}, descrito na Eq. (4.15), não deveria aumentar a variância de $V[n,m]$ para o sinal de áudio. Por outro lado, para um ruído aditivo com distribuição espectral aproximadamente uniforme, $W[n,m]$ é proporcional à largura de sub-banda, portanto o uso de deltas entre sub-bandas distantes N_{Δ_S}, conforme a Eq. (4.15), tende a elevar a variância final de $V[n,m]$. Por esta análise, o uso de deltas distantes não melhoraria a robustez contra ruído aditivo.

$$V[n,m] = W[n,m] - W[n,m-N_{\Delta_S}] \tag{4.15}$$

A Figura 4.15 mostra para um áudio A, que o desvio padrão de $T[:,m]$ aumenta com o incremento de $N_{\Delta_S} = 1$ para $N_{\Delta_S} = 2$, porém não aumenta significativamente para $N_{\Delta_S} > 2$. O efeito do ruído aditivo a SNR=25dB é medido pelo desvio padrão de $T_{A+\mathcal{N}_1}[:,m] - T_{A+\mathcal{N}_2}[:,m]$, que é aproximadamente constante para $N_{\Delta_S} = 2,3,4$. Dessa forma, observa-se que a razão entre os desvios padrões de $T[:,m]$ para A e dos erros $T_{A+\mathcal{N}_1}[:,m] - T_{A+\mathcal{N}_2}[:,m]$ decorrentes do ruído aditivo aumentam com o incremento de $N_{\Delta_S} = 1$ para $N_{\Delta_S} = 2$, porém não aumentam significativamente para $N_{\Delta_S} > 2$.

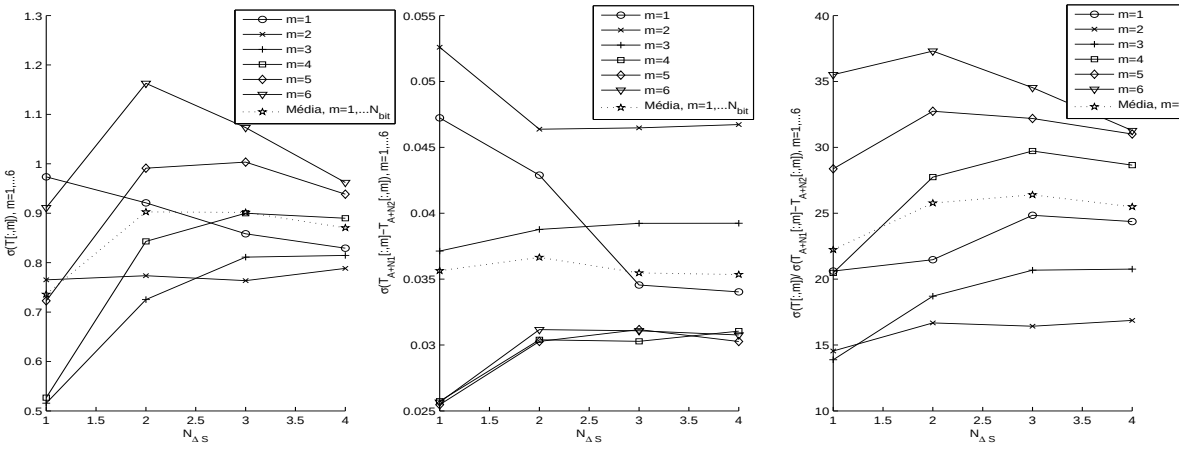

Figura 4.15: Devios padrões $\sigma(T_A[:,m])$ (esquerda), $\sigma(T_{A+\mathcal{N}_1}[:,m] - T_{A+\mathcal{N}_2}[:,m])$ (centro), e $\sigma(T_A[:,m])/\sigma(T_{A+\mathcal{N}_1}[:,m] - T_{A+\mathcal{N}_2}[:,m])$ (direita), $m = 1,2,...6$, para $N_{\Delta_S} \in \{1,2,3,4\}$, $N_{\Delta_F} = 1$, $D_F = 80ms$, $\Omega_F = 98\%$.

Na comparação de desempenho de detecção, N_{bits} é previamente ajustado para limitar o valor esperado do número de agrupamentos da matriz de autossimilaridade em no máximo 10. A média da taxa de detecção de quadros para 10 áudios é medida para $D_F = 80ms$, $\Omega_F = 98\%$, $N_{\Delta_F} = 1$ e $N_{\Delta_S} \in \{1,2,3\}$ usando o esquema com ajuste dos limites $(L[0],1,...,L[N_{bit+1}])$ e de α, para diversos níveis de ruído aditivo, SNR de 15dB a 25dB, conforme ilustra a Figura 4.16. A taxa de detecção de quadros para $N_{\Delta_S} = 2$ é superior à taxa para $N_{\Delta_S} = 1$. O incremento para $N_{\Delta_S} = 3$ não altera significativamente a taxa de detecção de quadros. Dessa forma, ajustamos $N_{\Delta_S} = 2$ para otimizar o desempenho de detecção.

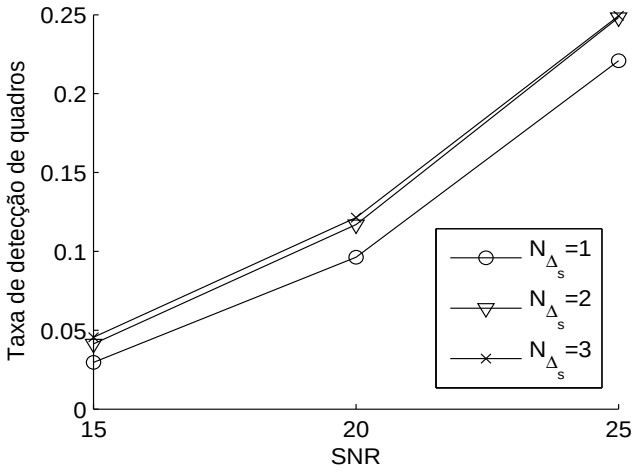

Figura 4.16: Taxa de detecção de quadros, variando a SNR, para $D_F = 80ms$, $N_{\Delta_F} = 1$, $\Omega_F = 98\%$ e $N_{\Delta_S} \in \{1, 2, 3\}$ no esquema com ajuste dos limites $(L[0], L[1], ..., L[N_{bit+1}])$ e de α.

4.5 TESTES COM ÁUDIOS LONGOS

A otimização do desempenho de detecção do esquema proposto pelo ajuste dos parâmetros foi feita apenas para áudios de 60s de duração. Entretanto, nos testes com áudios de 150s observamos que para $N_{bits} > 55$ há na verdade uma piora da unicidade com um aumento dos Falsos Positivos de Réplicas, e o ajuste de N_{bits} não é suficiente para limitar os Falsos Positivos de Réplicas abaixo de 10. A Figura 4.17 ilustra o número Falsos Positivos de Quadros e o número de agrupamentos 8-conectados de $\mathbf{M}$ (após operação de fechamento). Para $D_F = 80ms$ a resolução espectral é de aproximadamente 13Hz, logo o número de coeficientes espectrais na banda empregada é de 300. A piora da unicidade para $N_{bits} > 55$ pode ser explicada pelo fato do aumento de N_{bits} gerar sub-bandas com poucos coeficientes espectrais, o que pode aumentar a correlação dos bits de AF.

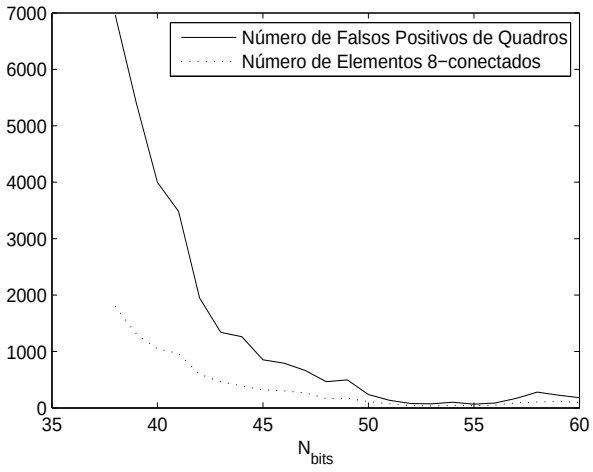

Figura 4.17: Número de Falsos Positivos de Quadros e de agrupamentos 8-conectados de $\mathbf{M}$, variando N_{bits}, para áudios de 150s.

5- USO DE DUPLA DETECÇÃO PARA APLICAÇÃO EM ÁUDIOS LONGOS

Para o esquema proposto até o momento, a unicidade pode ser ajustada pelo limiar de erro na detecção de quadros, d_{max}, ou pelo número de bits da AF, N_{bits}. Um aumento de d_{max} acima de 1 elevaria bastante a complexidade do método de busca perfeita empregado. Por outro lado, observamos que para áudios mais longos o aumento de N_{bits} pode não ser eficaz para a limitação dos Falsos Positivos de Réplica. Propomos então o emprego de um novo critério de detecção de réplicas para limitar os Falsos Positivos.

Em [9], para detecção de estruturas repetidas em uma música A, para descartar pares de quadros Falsos Positivos, filtros morfológicos de imagem são aplicados à matriz de autossimilaridade booleana $\mathbf{M}$ obtida após o uso de um limiar de detecção de quadros $\delta(F(a_j), F(a_k)) \leq \tau$ para todos os pares (j, k) de quadros em A. Em [10], para o alinhamento de versões de áudios com distorção em escala de tempo, os segmentos de AF na matriz de autossimilaridade com padrão diagonal referentes a trechos repetidos são identificados pela análise do histograma dos coeficientes de reta, e um limiar de duração mínima é usado para descartar os AF's Falsos Positivos.

Para a aplicação de réplicas curtas de áudio, a Figura 3.3 mostra que Falsos Positivos de Quadros ocorrem em geral isolados ou em pequenos agrupamentos na matriz de autossimilaridade. Por outro lado, os elementos detectados dentro de réplica possuem um padrão diagonal, com mesma defasagem temporal. Dessa forma, podemos definir um critério de detecção de réplica mais rígido que o critério $\delta_R(\mathbf{M}(i,j)) = \mathbf{M}(i,j)$ empregado no Capítulo 3, para limitar o número de Falsos Positivos de Réplica, sem necessidade de aumento de N_{bits}.

Propomos um critério que condiciona a detecção de uma réplica na posição (i, j) de um par de quadros detectados, $\mathbf{M}(i,j) = 1$ à existência de N_Q elementos não-nulos de $\mathbf{M}$ numa vizinhança de comprimento N_J na direção diagonal. O critério, portanto, dizima os elementos isolados de $\mathbf{M}$.

Para isso, usando a terminologia de operadores morfológicos, definimos um elemento estruturante $\mathbf{J}$, ilustrado na Figura 5.1, com N_J elementos não-nulos na diagonal principal:

$$\mathbf{J}(i,i) = 1, i = 1, 2, ...N_J. \tag{5.1}$$

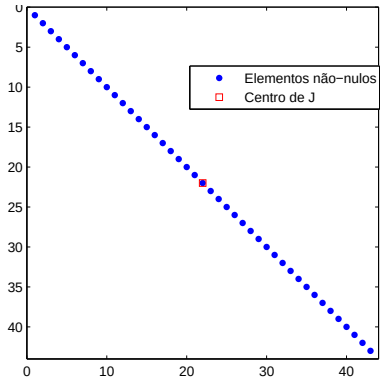

Figura 5.1: Elemento estruturante $\mathbf{J}$ com $N_J = 43$ elementos não-nulos.

Usamos a notação $(\mathbf{J})_{i,j}$ para definir a translação com deslocamento (i, j) de $\mathbf{J}$ a partir do seu centro. O critério de detecção é definido por $\delta'_R(\mathbf{M}(i,j) = \mathbf{M}(i,j).\mathbf{M_2}(i,j)$, onde $\mathbf{M_2}(i,j)$ é não-nulo para (i,j) com ao menos N_Q elementos não-nulos na vizinhança definida por $(\mathbf{J})_{i,j}$. $\mathbf{M_2}$ pode ser obtida com uso de uma matriz auxiliar $\mathbf{M_1}$ que representa para cada ponto (i,j) a soma dos bits em sua vizinhança $(\mathbf{J})_{i,j}$:

$$\mathbf{M_1} = \sum(\mathbf{J})_{i,j}, \forall(i,j)|(\mathbf{M}(i,j) = 1); \tag{5.2}$$

$$\mathbf{M_2}(i,j) = \begin{cases} 1, \text{se } \mathbf{M_1}(i,j) \geq N_Q, \\ 0, \text{se } \mathbf{M_1}(i,j) < N_Q. \end{cases} \tag{5.3}$$

Este método de dupla detecção, com pós-processamento da matriz booleana, é descrito na literatura científica como integrador binário de janelas móveis e também tem sido aplicado na detecção de RADAR [162].

Na Figura 5.2, uma matriz $\mathbf{M}$ artificialmente gerada ilustra a aplicação de $\delta'_R(\mathbf{M}(i,j))$ em uma matriz $\mathbf{M}$ para $N_Q = 2$, onde as setas escuras indicam Falsos Positivos de Quadros e as setas claras indicam quadros detectados dentro de um intervalo de réplica (esquerda superior). A matriz $\mathbf{M_1}$ (direita superior) e $\mathbf{M_2}$ (esquerda inferior), e a matriz de detecção de réplicas (direita inferior) ilustram como os Falsos Positivos de Réplica são descartados. Quanto menor o valor de N_J, menor será o número de Falsos Positivos de Réplica e maior será o número de Falsos Negativos. Por outro lado, quanto maior a duração N_J da janela, maior a taxa de detecção. O emprego deste critério permite, portanto, limitaçar os Falsos Positivos de Réplica sem a necessidade de ajuste de d_{max} ou N_{bits}. A complexidade computacional da aplicação de $\delta'_R(\mathbf{M}(i,j))$ é proporcional ao número de elementos não-nulos de $\mathbf{M}$.

O uso de N_J maiores que N_R, o número de quadros dentro do intervalo da réplica existente no áudio, não melhoraria a taxa de detecção. Para detecção de réplicas curtas

de 100ms, usaríamos, portanto, uma janela de 100ms. Logo, temos que $(N_J - 1) = (0,1 - D_F)/\Delta_F$. Para $D_F = 70ms$ e $\Omega_F = 0,98$, temos $N_J = 21$. Entretanto, na análise de áudios reais, a existência e a duração da réplica são desconhecidas. Portanto, vários valores de N_J a partir de 21 podem ser testados na análise.

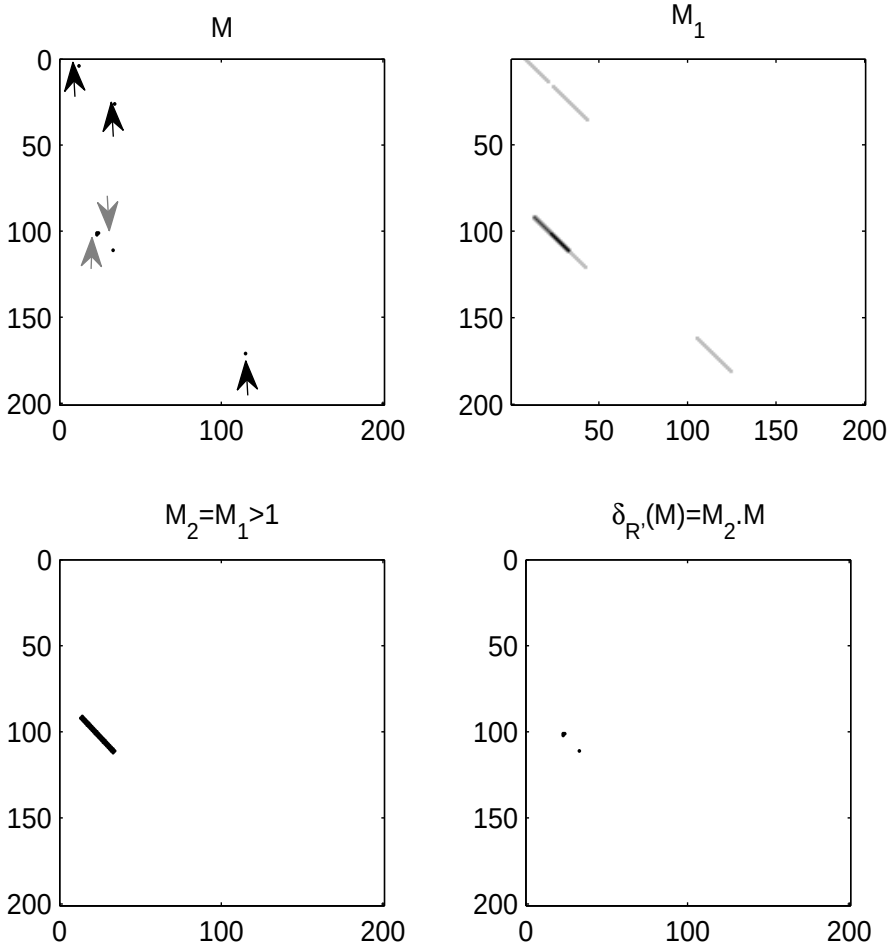

Figura 5.2: As setas escuras indicam Falsos Positivos de Quadros e as setas claras indicam quadros detectados dentro de um intervalo de réplica (esquerda superior). A matriz $\mathbf{M_1}$ (direita superior) e $\mathbf{M_2}$ (esquerda inferior), e a matriz de detecção de réplicas (direita inferior) ilustram como os Falsos Positivos de Réplica são descartados.

Para limitar o número de Falsos Positivos de Réplica, N_Q é ajustado em função de N_{bits} e N_J. Para um conjunto de 5 áudios, variamos $N_{bits} \in [25, 40]$ e $N_J \in \{13, 23, 43, 83, 163, 323\}$, e medimos o valor mínimo de N_Q para o qual o número agrupamentos 8-conectados seja inferior a 10 para todos os áudios. Os valores mínimos de N_Q obtidos para áudios de 60s (esquerda) e 240s (direita) são ilustrados na Figura 5.3. Observa-se, como esperado, que N_Q mínimo cai com um aumento de N_{bits}. O valor de N_Q mínimo aumenta com o aumento de N_J para $N_{bits} \leq 33$. Para áudios com duração de 240s, mais próxima da duração média

de áudios questionados reais, observa-se que o ajuste de N_Q é suficiente para a limitar os Falsos Positivos de Réplica. O valores de N_Q mínimo caem com o aumento de N_{bits}, e aumentam com o aumento de N_J para $N_{bits} \leq 42$. Comparando com o resultado para áudios de 60s, observa-se um aumento geral dos valores mínimos de N_Q, o que tende a reduzir a taxa média de detecção de réplica.

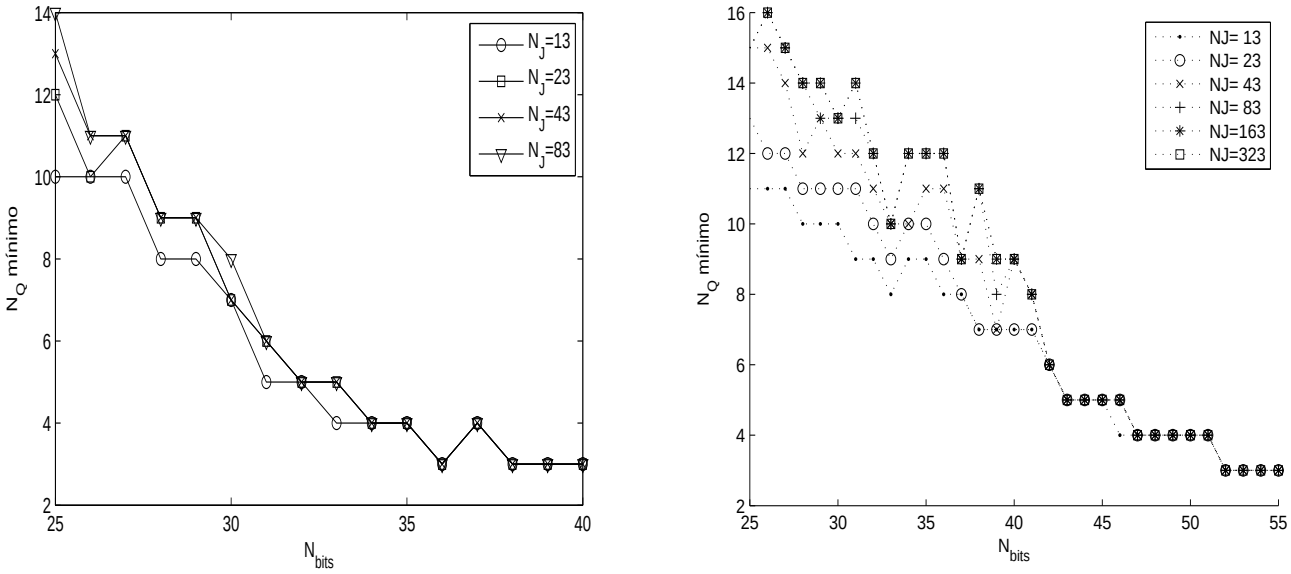

Figura 5.3: Valor mínimo de N_Q para conjunto de 5 áudios de 60s (esquerda) e 240s (direita), para limitar o número de agrupamentos de elementos 8-conectados em $\mathbf{M}$ até 10, variando N_{bits} e N_J.

5.1 TESTES DE ROBUSTEZ COM AJUSTE DO NÚMERO DE BITS E DE N_J

Os testes de robustez foram refeitos para o novo critério de detecção de réplica para a configuração $D_F = 70ms$, $\Omega_F = 98\%$, $N_{\Delta_F} = 1$ e $N_{\Delta_S} = 2$, usando o mesmo *corpus* usado no Capítulo 4.

Para viabilizar as simulações e o ajuste dos parâmetros, ao invés de usarmos grandes conjuntos de teste de áudios replicados, calculamos a probabilidade de detecção de réplicas indiretamente, a partir da taxa média de detecção de quadros, P_Q. Para o cálculo indireto da probabilidade de detecção de réplicas P_R a partir de P_Q, assumimos que a detecção de quadros tem uma distribuição de Bernoulli, embora alguns mascaramentos como compressão de áudio possam causar uma dependência temporal entre quadros próximos. Seja $FDP(k), k = 0, 1, ...N_{bits}$ a função de densidade de probabilidade de erro de bits para áudio mascarado com ruído, obtida do histograma definido na Eq. (4.6) para vários áudios, a taxa média de detecção de quadros de AF é dada por $P_Q = \sum_{k=0}^{d_{max}} FDP(k)$, e

a taxa média de Falso Negativo de Quadros é $\overline{P_Q} = 1 - P_Q$. A probabilidade de detecção de réplica depende ainda de N_J, N_Q e do número de quadros contidos nas réplicas, $N_R = (D_R - D_F - \Delta_F)/\Delta_F$.

Como o conjunto das sequências de N_R bits consiste de um espaço amostral discreto, as probabilidades discretas podem ser calculadas por força bruta. Entretanto, a complexidade computacional de 2^{N_R} inviabiliza esse cálculo para valores elevados de N_R. Em uma revisão da literatura científica, não foi encontrado nenhum método eficiente e exato de cálculo da probabilidade de detecção para o uso de integrador binários com janelas móveis. Em [162] é proposto um método de cálculo aproximado do desempenho do detector Integrador Binário. O método proposto aproxima para zero a interdependência de alguns termos do somatório de probabilidades discretas e desenvolve as equações apenas para o caso $N_Q = N_J - 1$. É destacado em [162] que para valores de N_Q distantes de N_J o método não fornece uma boa aproximação.

Propomos então um novo algoritmo de cálculo exato através do emprego de uma função recursiva, descrito no Apêndice C.1, onde a probabilidade de detecção de réplica em janelas de i bits, $P_R(i, N_J, N_Q, P_Q)$ é calculada variando $i = N_J$ até $i = N_R$. A complexidade do cálculo é reduzida através do pré-cálculo de probabilidades, obtendo-se uma complexidade proporcional a N_R. Este método, entretanto, requer o emprego de memória proporcional a 2^{N_J}. Para o ajuste dos parâmetros do esquema de detecção de réplicas proposto, valores elevados de $N_J = 43$ devem ser testados, logo o uso do método recursivo de cálculo exato seria inviável. Dessa forma, um novo método de cálculo aproximado de $P_R(i, N_J, N_Q, P_Q)$ foi proposto, com base no cálculo das probabilidades das somas dos bits de cada janela móvel de comprimento N_J, conforme descrito no Apêndice C.2. O método aproximado oferece uma boa estimativa de P_R, mesmo para valores de N_Q distantes de N_J, como mostra a Figura C.4.

A taxa de detecção de quadros média para 10 áudios de 60s de diferentes falantes, P_Q, é medida para diversos níveis de ruído branco aditivo, SNR= 15dB, 20dB e 25dB, conforme ilustrado na Figura 5.4. Observa-se que P_Q é reduzida com o aumento de N_{bits}, o que é esperado já que a tolerância a erro $d_{max} = 1$ é mantida fixa, logo a tolerância percentual de erro, $1/N_{bits}$, cai com N_{bits}.

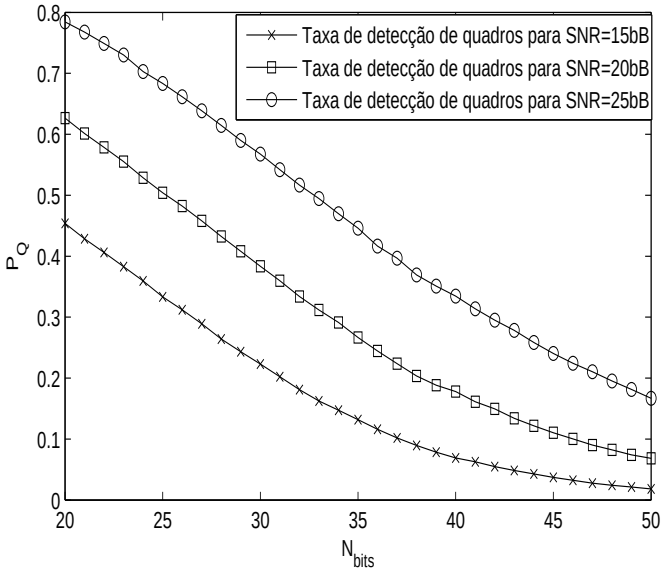

Figura 5.4: Taxa média de detecção de quadros, P_Q, para 10 áudios mascarados com ruído branco aditivo a SNR=15dB, 20dB e 25dB, medida para diversos valores de N_{bits}.

A probabilidade de detecção de réplicas é calculada indiretamente a partir de P_Q, usando o método aproximado proposto, descrito no Apêndice C.2. Os parâmetros de N_{bits} e N_J são ajustados para maximizar a probabilidade de detecção de réplicas para o critério δ'_R. O valor mínimo de N_Q é previamente obtido em função de N_{bits} e N_J, conforme ilustrado na Figura 5.3.

Nas simulações são testados valores $N_J \in \{13, 23, 43, 83, 163, 323\}$ e $N_{bits} \in [25, 40]$, e calculamos a probabilidade de detecção de réplica indiretamente a partir de P_Q, através do método de cálculo aproximado descrito no Apêndice C.2.

A Figura 5.5 mostra a probabilidade de detecção de réplicas de 100ms (esquerda) e 130ms (direita), em um áudio de 60s mascarado com ruído branco aditivo a SNR=15dB. Observa-se que para a réplica de 100ms, o melhor desempenho de detecção de réplica, acima de 40%, é obtido para $N_{bits} = 36$ e $N_J \geq 23$. Para a réplica de 130ms, observa-se também que o aumento de N_J aumenta a probabilidade de detecção da réplica até $N_J = N_R$, ou seja, $N_J = 43$.

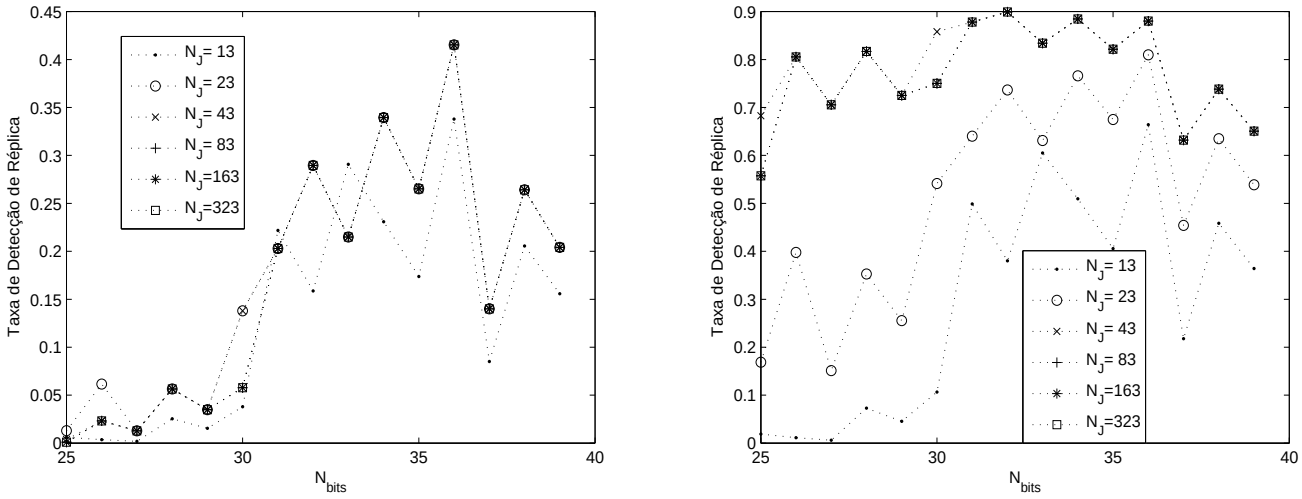

Figura 5.5: Probabilidade de detecção de réplicas de 100ms (esquerda) e 130ms (direita) em um áudio de 60s mascarado com ruído branco aditivo a SNR=15dB, calculada indiretamente para valores $N_J \in \{13, 23, 43, 83, 163, 323\}$ e $N_{bits} \in [25, 40]$.

Para um áudio de 240s mascarado com ruído branco aditivo a SNR=15dB, com réplicas entre 100ms e 190ms, o ajuste de N_J e N_{bits} para maximizar a taxa média de detecção de réplicas é ilustrado na Figura 5.6. A taxa média de detecção de réplicas de 100ms é quase nula para todas as configurações. Ajustamos N_{bits} e N_J de forma a maximizar a soma das probabilidades de detecção de réplicas de 100ms, 130ms, 160ms e 190ms. As melhores taxas de detecção são obtidas para $N_J \geq (NR = 83)$ e $N_{bits} = 25$. Ressaltamos que, apesar do aumento de N_F no áudio longo, o que na detecção simples elevaria N_{bits}, a melhor taxa de detecção é obtida para $N_{bits} = 25$, abaixo do valor obtido para áudios de 60s.

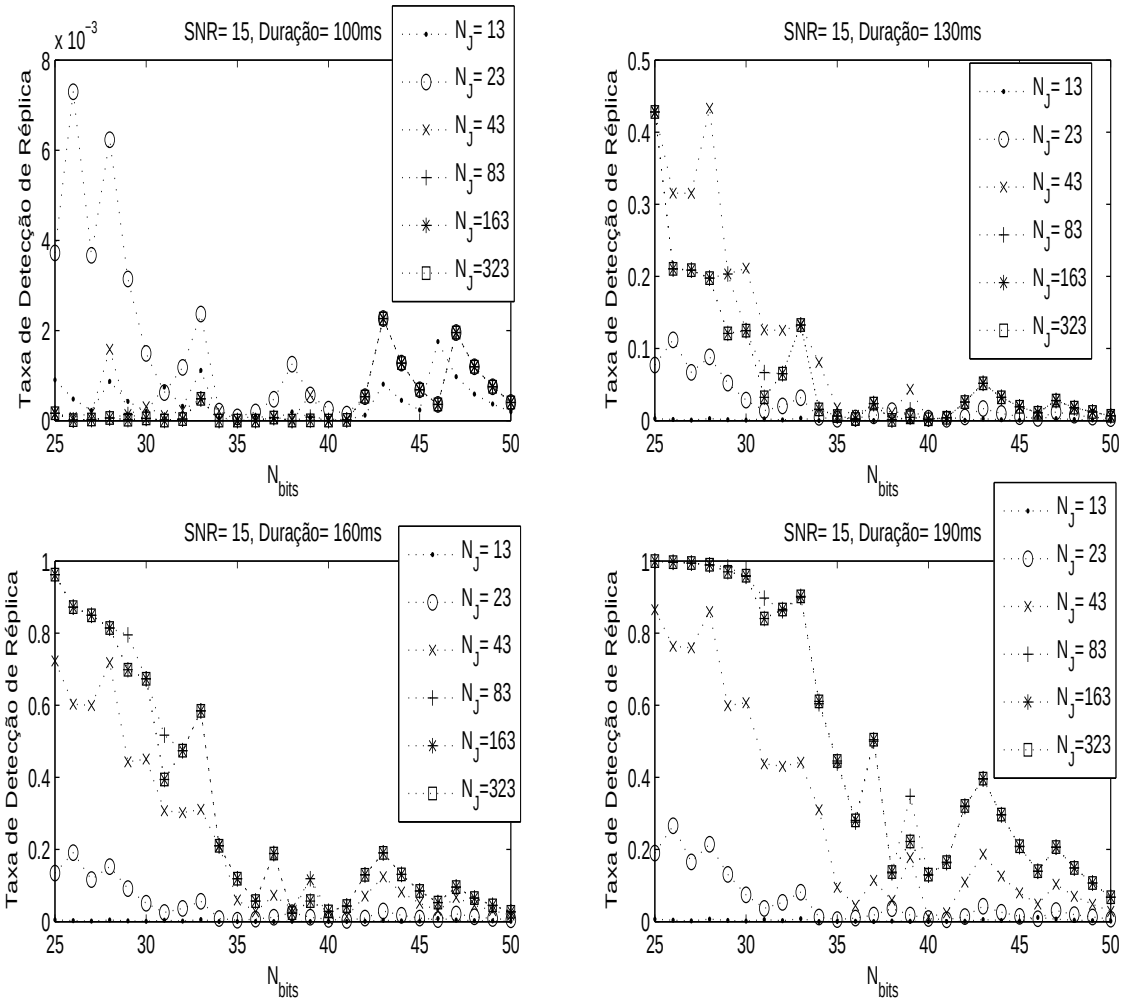

Figura 5.6: Probabilidade de detecção de réplicas de 100ms a 190ms, em um áudio de 240 s mascarado com ruído branco aditivo a SNR=15dB, calculada indiretamente a partir de P_Q, para valores $N_J \in \{13, 23, 43, 83, 163, 323\}$ e $N_{bits} \in [25, 50]$.

A taxa detecção de réplicas em um áudio de 60s mascarado com ruído branco aditivo a SNR=15dB, 20dB e 25dB é ilustrada na Figura 5.7 (esquerda), para os parâmetros $N_{bits} = 36$ e $N_J = 43$. Comparando este resultado com o desempenho do método de detecção simples, vide Figura 4.10, concluímos que a aplicação da dupla detecção melhora a robustez contra inserção de ruído aditivo. Para áudios de 240s, o desempenho de detecção de réplica para os parâmetros $N_J = 83$ e $N_{bits} = 25$ é ilustrado na Figura 5.7 (direita), com uma probabilidade de detecção próxima de 100% para réplicas de 160ms.

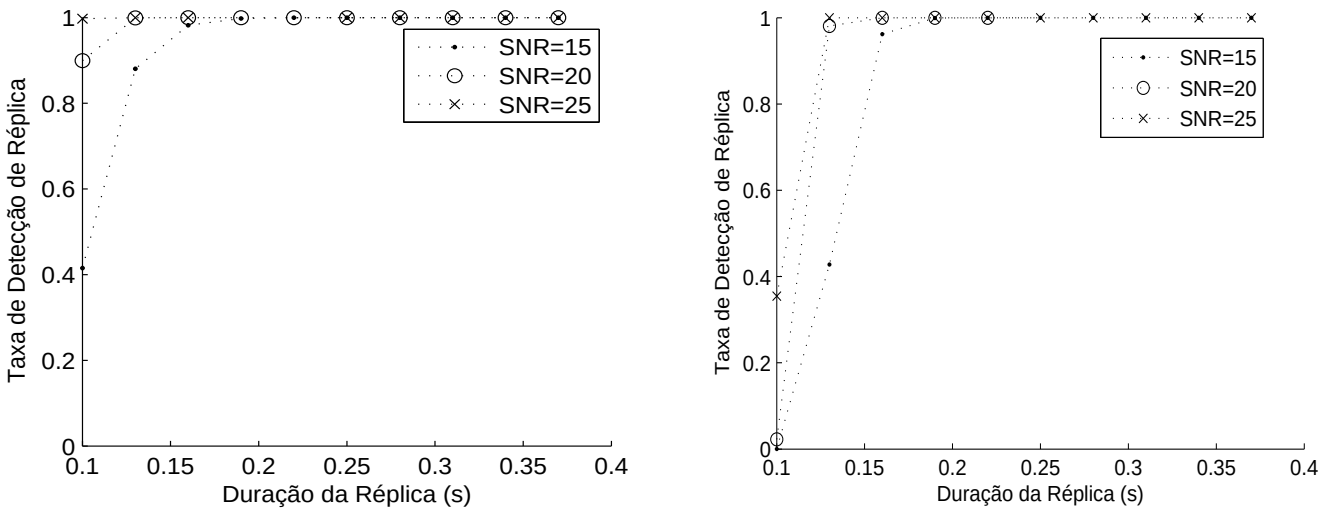

Figura 5.7: Probabilidade de detecção de réplicas de diversas durações em áudio mascarado com ruído branco aditivo a SNR=15dB, 20dB e 25dB, calculada indiretamente a partir de P_Q, com $N_{bits} = 36$ e $N_J = 43$ para áudio 60s (esquerda), e com $N_{bits} = 25$ e $N_J = 83$ para áudio de 240s (direita).

5.2 TESTES DE ROBUSTEZ COM AJUSTE DO NÚMERO DE BITS, DE Ω_F, D_F E N_J

.

Na análise anterior apenas N_{bits} e N_J foram ajustados para otimizar a probabilidade de detecção de réplica para áudios de 60s e 240s. Nesta seção, fazemos uma otimização mais ampla, com ajuste de Ω_F, D_F, N_{bits} e N_J.

Observamos que os parâmetros ótimos de detecção de réplica variam para diferentes durações de réplica. Considerando que réplicas acima de 200ms tem uma detecção próxima de 100%, definimos como uma métrica de desempenho a soma das probabilidades de detecção para réplicas de 100ms, 130ms, 160ms e 190ms. Sejam $N_R(100ms) = (0,1 - D_F)/\Delta_F$, $N_R(130ms) = (0,13 - D_F)/\Delta_F$, $N_R(160ms) = (0,16 - D_F)/\Delta_F$ e $N_R(190ms) = (0,19 - D_F)/\Delta_F$, devemos maximizar a soma de probabilidades

$$\begin{aligned} S = P_R(N_R(100ms), N_J, N_Q, P_Q) + P_R(N_R(130ms), N_J, N_Q, P_Q) + \\ + P_R(N_R(160ms), N_J, N_Q, P_Q) + P_R(N_R(190ms), N_J, N_Q, P_Q). \end{aligned} \tag{5.4}$$

Observamos que o aumento de N_J até N_R melhora a taxa de detecção de réplica, e para valores mais baixos de N_{bits} o aumento de N_J acima de N_R pode reduzir a taxa de detecção, pois causa um aumento de N_Q. Ademais, os resultados mostram uma boa taxa

de detecção para réplicas com duração maior que 200ms. Portanto, o uso de N_J maior que o número de quadros correspondentes a 200ms não é interessante. Para simplificar o processo de otimização dos demais parâmetros ajustamos $N_J = (0,2 - D_F)/\Delta_F$.

Ressaltamos que N_Q e P_Q variam com Ω_F, D_F e N_{bits}. O processo de otimização é realizado para áudios de 60s de duração, para os intervalos de parâmetros $\Omega_F(\%) \in \{94, 95, 96, 97, 98\}$, $D_F(ms) \in \{50, 60, 70, 80, 90\}$ e $N_{bits} \in \{25, 28, 31, 34, 37, 40, 43\}$. Para áudios de 240s de duração, por limitação de memória, excluímos $\Omega_F = 98\%$ dos teses, e o desempenho é maximizado para $\Omega_F(\%) \in \{94, 95, 96, 97\}$.

Tabela 5.1: Otimização do desempenho com ajuste dos parâmetros Ω_F, D_F e N_{bits}, para $N_J = N_R(200ms)$.

Duração	S	Ω_F	D_F	N_{bits}
60s	3,53	97%	50ms	25
240s	1,94	97%	70ms	25

A Tabela 5.1 mostra os resultados do ajuste dos parâmetros. Observamos que mesmo com a redução de N_{bits}, o que aumenta a probabilidade de colisão, o método de dupla detecção é eficiente no descarte dos Falsos Positivos de Réplicas. A redução de N_{bits}, com uma tolerância de erro fixa d_{max}, aumenta a probabilidade de detecção de quadros P_Q e o desempenho de detecção de réplicas. Observamos ainda ajustes distintos para o parâmetro D_F, o que confirma que os parâmetros devem ser ajustados para cada duração de áudio.

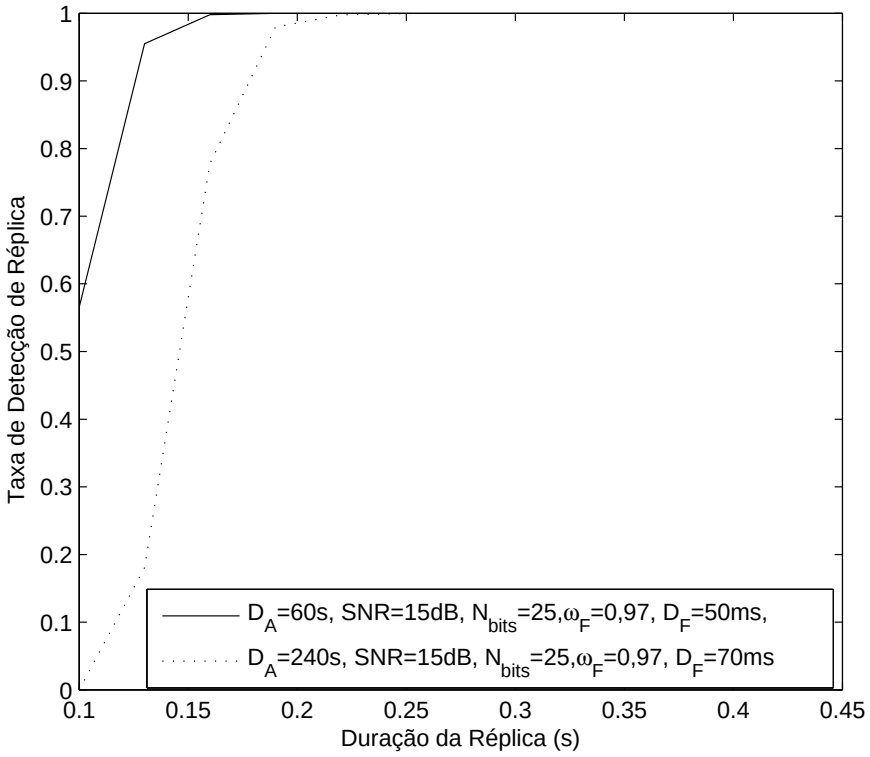

Figura 5.8: Probabilidade de detecção de réplicas de 100ms a 450ms de duração, calculada indiretamente a partir de P_Q conforme os ajustes da Tabela 5.1, para áudios com duração de 60s (linha sólida) e 240s (linha pontilhada), mascarados com ruído branco aditivo a SNR=15dB.

A Figura 5.8 ilustra a probabilidade de detecção de réplicas de 100ms a 450ms de duração, calculada indiretamente a partir de P_Q conforme os ajustes da Tabela 5.1, para áudios com duração de 60s (linha sólida) e 240s (linha pontilhada), mascarados com ruído branco aditivo a SNR=15dB. O desempenho de detecção para áudios de 60s com o ajuste dos parâmetros é melhor que o resultado anterior ilustrado na Figura 5.7. O ajuste dos parâmetros nos limites dos intervalos testados, como $N_{bits} = 25$, mostra que testes complementares com intervalos mais amplos são necessários. Os resultados comprovam a aplicabilidade deste método de detecção de réplicas, mesmo para áudios longos mascarados por inserção de ruído.

6- CONCLUSÕES

As simulações no Capítulo 3 mostraram que o esquema de AF adaptativo inicialmente proposto detecta trechos replicados de voz com duração tão curta quanto $100ms$, mesmo na presença de distorções de amplitude ou no domínio da frequência. Entretanto, para distorções como adição de ruído branco Gaussiano ou compressão de áudio, o desempenho de detecção é regular. Como descrito no Capítulo 4, a adaptação do expoente α e dos limites da banda de frequência para cada áudio, além do ajuste dos parâmetros como duração e fator de sobreposição dos quadros para otimizar a detecção de réplicas, fornecem uma boa taxa de detecção de réplicas, mesmo para réplicas de 100ms em áudios de 60s mascarados com ruído aditivo a SNR=15dB. Observou-se que o ajuste de N_{bits} não é suficiente para limitar os Falsos Positivos de Réplicas em áudios mais longos, logo, no Capítulo 5 é usado um critério de dupla detecção pela integração binária de janelas diagonais movéis, o qual permite a redução dos Falsos Positivos de Réplica sem a necessidade de ajuste de N_{bits}. A probabilidade de detecção de réplicas é calculada indiretamente a partir da taxa de detecção de quadros, para viabilizar a otimização do sistema pelo ajuste dos parâmetros. Os resultados mostram uma boa taxa de detecção de réplicas tão curtas quanto 160ms, em áudios de 4 minutos mascarados com ruído branco a SNR=15dB. Portanto, ao final, é obtido um método com boa usabilidade e robustez, para detecção de réplicas curtas em áudios longos.

Como foi observado no Capítulo 3, a taxa de erro de bits de AF para voz intralocutor intrasentença é menor para falantes com padrão de voz estável. Dessa forma, esse atributo poderia ser empregado nos sistemas de classificação automática para análise da estabilidade da voz para o diagnóstico remoto e não invasivo de patologias do aparelho fonador ou de distúrbios da fala. A análise do sinal de voz permite uma triagem remota, o que pode ser útil para um diagnóstico precoce. A estabilidade da voz pode ser usada também para estimação da idade do falante [163]. Em [164] é proposta a coleta do sinal pela rede telefônica, para a classificação automática de quatro grupos distintos de nível patológico. Diversos atributos já foram propostos para essa aplicação, como as medidas de estabilidade de *pitch* (*jitter* e *shimmer*), MFCC's, análise de harmonicidade [165, 166, 167], DFA (*Detrented Fluctuation Analysis*) e a RPDE (*Recurrence Probability Density Entropy*) [168]. Uma vantagem de uso de taxa de erro de bits de AF é a possibilidade do emprego de falas repetidas referentes a um texto controlado, sem necessidade de um treinamento prévio, coleta assistida de voz, ou mesmo de uma posterior segmentação de fonemas.

As simulações também mostraram que o esquema proposto no Capítulo 4, com adaptação de α e da banda de frequência, produz uma representação mais invariante a deslocamentos da distribuição espectral devido à mudança de escala no tempo, que consiste em

uma distorção comum em meios de difusão de música. Ademais, as simulações também sugerem que para áudios com granularidades maiores o emprego de deltas entre quadros distantes pode melhorar a robustez. Dessa forma, uma possível linha de pesquisa seria a análise de desempenho deste esquema em aplicações de MIR.

Dentre as aplicações para esquemas de AF descritas no Capítulo 2, podemos ressaltar o potencial de emprego do método adaptativo proposto para:

1. A análise da qualidade de áudio comprimido: os erros entre bits de AF de mesmo índice do áudio original e do codificado podem mostrar variações no tempo, decorrentes do chaveamento dinâmico da taxa ou do métodos de codificação para sinais vozeado/não vozeado/ ruído; ou variações nas sub-bandas, decorrentes da diferença de codificação de bandas alta e baixa. O esquema de [2], que usa uma divisão de sub-bandas fixa, pode gerar bits com baixa variância pouco úteis na análise de erro, e possui uma baixa resolução temporal comparado a algoritmos de medida de qualidade (PEAQ), como ressaltado em [145]. O método adaptativo proposto nivela a variância dos bits de todas as sub-bandas e melhora a resolução temporal com a redução da duração dos quadros.

2. Sincronização e cancelamento de ruído de fundo de referência: No esquema proposto, a robustez pode ser ajustada pelos parâmetros d_{max}, N_{bits} e N_Q para detecção do áudio de referência, mesmo que esteja misturado à voz. A boa resolução temporal permite uma melhor sincronização, mas outras formas do elemento **J** devem ser testadas para aumentar a tolerância contra variações da taxa de reprodução.

3. O uso de AF na geração de chaves dependentes de conteúdo para esquemas de marca d'água digital de áudio: Nesta aplicação a AF deve ser bastante robusta, pois qualquer erro de bit modificaria a chave. O ajuste do expoente α se mostrou interessante para melhoria da robustez. Ademais, o método adaptativo proposto permite gerar AF's com uma distribuição mais próxima de uma i.i.d., e, portanto, gerar chaves com maior entropia.

Como resultado, podemos também destacar a proposta de dois novos métodos de cálculo da probabilidade de desempenho do integrador binário com janela móvel, um exato e outro aproximado com baixo erro percentual. Este critério de detecção se mostrou muito eficiente na redução dos falsos positivos, o que permite um ajuste da tolerância relativa de erro do detector primário sem aumento da complexidade do método de busca. Tais métodos de cálculo podem ser usados também para estimativa de desempenho em outras aplicações do integrador binário, como na detecção de RADAR. Ademais, o desempenho da integração binária pode ser comparado ao de outros métodos de dupla detecção usados na identificação de trechos repetidos em música ou na sincronização de mídias, citados no Capítulo 2.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] CANO, P. et al. A review of audio fingerprinting. *J. VLSI Signal Process. Syst.*, v. 41, n. 3, p. 271–284, Nov 2005.

[2] HAITSMA, J.; KALKER, T. A highly robust audio fingerprinting system. In: *Proc. of ISMIR*. Paris, França: ISMIR, 2002. p. 107–115.

[3] HADAMARD, J. *Psicologia da invenção na matemática*. [S.l.]: Contraponto, 2009.

[4] SNYDER, J. S.; WEINTRAUB, D. M. Loss and persistence of implicit memory for sound: Evidence from auditory stream segregation context effects. *Attention, Perception and Psychophysics*, Springer, v. 75, n. 5, p. 1059–1074, 2013.

[5] OZER, H.; SANKUR, B.; MEMON, N. Robust audio hashing for audio identification. In: *Proc. of EUSIPCO*. Viena, ï£¡ustria: IEEE, 2004.

[6] MALEKESMAEILI, M.; RABAB, K. W. A local fingerprinting approach for audio copy detection. *Signal Processing*, v. 98, p. 1–10, 2013.

[7] OGLE, J. P.; ELLIS, D. P. Fingerprinting to identify repeated sound events in long-duration personal audio recordings. In: *Proc. of ICASSP*. Honolulu, EUA: IEEE, 2007. v. 1, p. I–233.

[8] FOSTER, P.; KLAPURI, A.; DIXON, S. A method for identifying repetition structure in musical audio based on time series prediction. In: *EUSIPCO*. Bucareste, Romenia: EURASIP, 2012. p. 1299–1303.

[9] ONG, B. S. *Structural analysis and segmentation of music signals*. Tese (Doutorado) — Universitat Pompeu Fabra, 2006.

[10] RAMONA, M.; PEETERS, G. Automatic alignment of audio occurrences: application to the verification and synchronization of audio fingerprinting annotation. In: *Proc. of DAFX Conference*. Paris, Franï£¡a: IRCAM, 2011. p. 429–436.

[11] DUONG, N. Q.; HOWSON, C.; LEGALLAIS, Y. Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation. In: *Consumer Electronics-Berlin (ICCE)*. Berlim, Alemanha: IEEE, 2012. p. 241–244.

[12] COTTON, C. V.; ELLIS, D. P. Audio fingerprinting to identify multiple videos of an event. In: *Proc. of ICASSP*. Dallas, EUA: IEEE, 2010. p. 2386–2389.

[13] BAGRI, A. et al. A scalable framework for joint clustering and synchronizing multi-camera videos. In: *EUSIPCO*. Marrakech, Marrocos: EURASIP, 2013. p. 1–5.

[14] MÜLLER, M.; MATTES, H.; KURTH, F. An efficient multiscale approach to audio synchronization. In: *Proc. of ISMIR Conference*. Victoria, Canadï£¡: ISMIR, 2006. p. 192–197.

[15] MACRAE, R. et al. Real-time synchronisation of multimedia streams in a mobile device. In: *Proc. of ICME*. Barcelona, Espanha: IEEE, 2011. p. 1–6.

[16] POPESCU, A. C.; FARID, H. *Exposing digital forgeries by detecting duplicated image regions*. [S.l.], 2004.

[17] FRIDRICH, J.; SOUKAL, D.; LUKAS, J. Detection of copy move forgery in digital images. In: *Digital Forensic Research Workshop*. Cleveland, EUA: AFRL, 2003.

[18] TAVORA, R.; NASCIMENTO, F. A. Detecting replicas within audio evidence using an adaptive audio fingerprinting scheme. In: . Seul, Corï£¡ia do Sul: IAFS, 2014.

[19] XIAO, J.-n. et al. Audio authenticity: Duplicated audio segment detection in waveform audio file. *Journal of Shanghai Jiaotong University (Science)*, Springer, v. 19, p. 392–397, 2014.

[20] YAN, Q.; YANG, R.; HUANG, J. Copy-move detection of audio recording with pitch similarity. In: *Proc. of ICASSP*. Brisbane, Austrï£¡lia: IEEE, 2015. p. 1782–1786.

[21] CHEN, B. et al. Fast computation of sliding discrete tchebichef moments and its application in duplicated regions detection. *IEEE Transactions on Signal Processing*, v. 63, n. 20, p. 5424–5436, 2015.

[22] DUPRAZ, E.; RICHARD, G. Robust frequency-based audio fingerprinting. In: *Proc. of ICASSP*. Dallas, EUA: IEEE, 2010. p. 281–284.

[23] BETSER, M.; COLLEN, P.; RAULT, J. B. Audio identification using sinusoidal modelling and application to jingle detection. In: *Proc. of ISMIR*. Viena, ï£¡ustria: ISMIR, 2007.

[24] DOETS, P. J. O.; GISBERT, M. M.; LAGENDIJK, R. L. On the comparison of audio fingerprints for extracting quality parameters of compressed audio. v. 6072, 2006.

[25] MITROVIC, D.; ZAPPELZAUER, M.; BREITENEDER, C. Features for contend-based audio retrieval. *Advances in Computers*, v. 78, p. 71–150, 2010.

[26] TAVORA, R.; NASCIMENTO, F. A. Detecting replicas within audio evidence using an adaptive audio fingerprinting scheme. *Journal of the Audio Engineering Society*, v. 63, p. 451–462, 2015.

[27] CUMMINS, F. et al. The chains corpus: Characterizing individual speakers. In: *Proc. of SPECOM*. Sï£¡o Petersburgo, Rï£¡ssia: META, 2006. v. 6.

[28] JOACHIM, T.; ITO, N.; VINCENT, E. Diverse environments multichannel acoustic noise database. *The Journal of Acoustical Society of America*, p. 3591–3591, 2013.

[29] CANO, P. *Content-based audio search: from fingerprinting to semantic audio retrieval*. Tese (Doutorado) — FABRA University, 2007.

[30] DUONG, N. Q.; DUONG, H.-T. A review of audio features and statistical models exploited for voice pattern design. *arXiv preprint arXiv:1502.06811*, 2015.

[31] CASEY, M. A. et al. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, v. 96, n. 4, p. 668–696, 2008.

[32] GROSCHE, P. et al. Structure-based audio fingerprinting for music retrieval. In: *Proc. of ISMIR*. Porto, Portugal: ISMIR, 2012. p. 55–60.

[33] SOUNDHOUND. Acessado em:11/12/2015. Disponível em: <http://www.soundhound.com/soundhound>.

[34] GUNDERSON, S. H. *Musical descriptors: An assessment of psychoacoustical models in the presence of lossy compression*. Dissertação (Mestrado) — Norwegian Univ. of Science and Tech., 2007.

[35] BURGES, J. C.; PLATT, J.; JANA, S. Distortion discriminant analysis for audio fingerprinting. *IEEE Transactions on Speech and Audio Processing*, v. 11, n. 3, p. 165–174, 2003.

[36] BURGES, C. J. C.; PLATT, J. C.; JANA, S. Extracting noise-robust features from audio data. In: *Proc. of ICASSP*. Orlando, EUA: IEEE, 2002. p. 1021–1024.

[37] WANG, H. et al. *Survey of Compressed-Domain Features used in Audio-Visual Indexing and Analysis*. 2003. 150-183 p.

[38] BARDELI, R. Robust identification of time-scaled audio. In: *Proc. of AES 25th International Conference*. Londres, Inglaterra: AES, 2004.

[39] RAMONA, M.; PEETERS, G. Audioprint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In: *Proc. of ICASSP*. Vancouver, Canadï£¡: IEEE, 2013. p. 818–822.

[40] RAMONA, M.; PEETERS, G. Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection. In: *Proc. of ICASSP*. Praga, Rep. Tcheca: IEEE, 2011. p. 477–480.

[41] GHULAM, M. et al. A noise-robust feature extraction method based on pitch-synchronous ZCPA for ASR. In: *Proc. of 8th International Conference on Spoken Language Processing*. Durban, ï£¡frica do Sul: WASET, 2004.

[42] BALADO, F. et al. Performance analysis of robust audio hashing. *IEEE Transaction on Information Forensics and Security*, v. 2, n. 2, p. 254–266, Jun 2007.

[43] LARTILLOT, O.; TOIVIAINEN, P.; EEROLA, T. A matlab toolbox for music information retrieval. In: *Proc. Conference on Data Analysis, Machine Learning and Applications*. Friburgo, Alemanha: Springer, 2008. p. 261–268.

[44] PEETERS, G. et al. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, v. 130, n. 5, p. 2902–2916, 2011.

[45] KIM, D.-S. et al. Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments. In: *Proc. of ICASSP*. Atlanta, EUA: IEEE, 1996. v. 1, p. 61–64.

[46] GAJIC, B.; PALIWAL, K. K. Robust speech recognition using features based on zero crossings with peak amplitudes. In: *Proc. of ICASSP*. Hong Kong: IEEE, 2003. v. 1.

[47] FOOTE, J. The beat spectrum: a new approach to rhythm analysis. In: *Proc. IEEE Int. Conf. on Multimedia and Expo*. Tï£¡quio, Japï£¡o: IEEE, 2001.

[48] KIROVSKI, D.; ATTIAS, H. Beat-id: Identifying music via beat analysis. In: *IEEE Workshop Multimedia Signal Processing*. San Thomas, Virgin Islands: IEEE, 2002. p. 190–193.

[49] KURTH, F.; GEHRMANN, T.; MULLER, M. The cyclic beat spectrum: Tempo-related audio features for time-scale invariant audio identification. In: *Proc. of ISMIR Conference*. Victoria, Canadï£¡: ISMIR, 2006.

[50] WANG, A. An industrial strength audio search algorithm. In: *Proc. of ISMIR*. Baltimore, EUA: ISMIR, 2003.

[51] ZHU, B. et al. A novel audio fingerprinting method robust to time scale modification and pitch shifting. In: *ACM Multimedia*. Florenï£¡a, Itï£¡lia: ACM, 2010. p. 987–990.

[52] PAPAODYSSEUS, C. et al. A new approach to the automatic recognition of musical recordings. *Journal of the Audio Engineering Society*, v. 49, p. 23–25, 2001.

[53] SUBRAMANYA, S. R. et al. Transform-based indexing of audio data for multimedia databases. In: *Proc. ICMCS97*. Otawa, Canadï£¡: IEEE, 1997. p. 211–218.

[54] SEO, J. S. et al. Audio fingerprinting based on normalized spectral subband moments. *IEEE Signal Processing Letters*, v. 13, p. 209–213, 2006.

[55] RAMALINGAM, A.; KRISHNAN, S. Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting. *IEEE Transactions on Information Forensics and Security*, v. 1, n. 4, p. 457–463, 2006.

[56] AUCOUTURIER, J. julien et al. The way it sounds: Timbre models for analysis and retrieval of polyphonic music signals. *IEEE Transactions on Multimedia*, v. 7, p. 1028–1035, 2005.

[57] TERASAVA, H.; SLANEY, M.; BERGER, J. Perceptual distance in timbre space. In: *Proc. of Int. Conference on ICAD*. Limerick, Irlanda: ICAD, 2005.

[58] LAGRANGE, M.; BADEAU, R.; RICHARD, G. Robust similarity metrics between audio signals based on asymmetrical spectral envelope matching. In: *Proc. of ICASSP*. Dallas, EUA: IEEE, 2010. p. 405–408.

[59] GRUHNE, M. Robust audio identification for commercial applications. *Fraunhofer IIS, AEMT*, 2003.

[60] LANCINI, R.; MAPELLI, F.; PEZZANO, R. Audio content identification using perceptual hashing. In: *Proc. of ICME*. Taipei, Taiwan: IEEE, 2004. p. 739–742.

[61] MURTHY, M. K.; SEETHA, S.; PÁDUA, F. L. Generating MPEG-7 audio descriptor for content-based retrieval. In: *Proc. of RAICS Conference*. Trivandrum, ï£¡ndia: IEEE, 2011. p. 467–470.

[62] MAPELLI, F.; PEZZANO, R.; LANCINI, R. Robust audio fingerprinting for song identification. In: *Proc. of EUSIPCO*. Viena, ï£¡ustria: EURASIP, 2004. p. 2095–2098.

[63] ALLAMANCHE, E. Content-based identification of audio material using MPEG-7 low level description. In: *Proc. of ISMIR Conference*. Indiana, EUA: ISMIR, 2001.

[64] KASTNER, T. et al. MPEG-7 scalable robust audio fingerprinting. In: *Audio Engineering Society Convention 112*. Munique, Alemanha: AES, 2002.

[65] HELLMUTH, O. et al. Advanced audio identification using MPEG-7 content description. In: *Proc. of AES Convention 111*. Nova York, EUA: AES, 2001.

[66] BURGES, C. J. C. et al. Using audio fingerprinting for duplicate detection and thumbnail generation. In: *Proc. of ICASSP*. Filadelfia, EUA: IEEE, 2005. p. 9–12.

[67] MIHÇAK, M. K.; VENKATESAN, R. A perceptual audio hashing algorithm: a tool for robust audio identification and information hiding. In: *Proc. of International Workshop on Information Hiding*. Pitsburg, EUA: Springer, 2001. p. 51–65.

[68] GHOUTI, L.; BOURIDANE, A. A robust perceptual audio hashing using balanced multiwavelets. *Proc. IEEE ICASSP*, p. 209–212, 2006.

[69] ZHANG, Q.-Y. et al. An efficient speech perceptual hashing authentication algorithm based on wavelet packet decomposition. *Journal of Information Hiding and Multimedia Signal Processing*, v. 6, n. 2, p. 311–322, 2015.

[70] REIN, S.; REISSLEIN, M.; SIKORA, T. Audio content description with wavelets and neural nets. In: *Proc. of ICASSP*. Montreal, Canadï£¡: IEEE, 2004. v. 4.

[71] UMAPATHY, K.; KRISHNAN, S.; RAO, R. K. Audio signal feature extraction and classification using local discriminant bases. *IEEE Transactions on Audio, Speech and Language Processing*, v. 15, n. 4, p. 1236–1246, 2007.

[72] LI, G.; KHOKHAR, A. A. Content-based indexing and retrieval of audio data using wavelets. In: *Proc. of IEEE International Conference on Multimedia and Expo (II)*. Nova York, EUA: IEEE, 2000. p. 885–888.

[73] BALUJA, S.; COVELL, M. Content fingerprinting using wavelets. In: *Proc. of CVMP*. Londres, Inglaterra: IEEE, 2006. p. 198–207.

[74] KAMALADAS, M. D.; DIALIN, M. M. Fingerprint extraction of audio signal using wavelet transform. In: *Proc. ICSIPR*. Coimbatore, ï£¡ndia: IEEE, 2013. p. 308–312.

[75] SUBRAMANYA, S.; YOUSSEF, A. Wavelet-based indexing of audio data in audio/multimedia databases. In: *Proc. International Workshop on MultiMedia Database Management Systems*. Londres, Inglaterra: IEEE, 1998. p. 46–53.

[76] BALUJA, S.; COVELL, M. Waveprint: Efficient wavelet-based audio fingerprinting. *Pattern recognition*, Elsevier, v. 41, n. 11, p. 3467–3480, 2008.

[77] BALUJA, S.; COVELL, M. Audio fingerprinting: Combining computer vision & data stream processing. In: *Proc. of ICASSP*. Honolunu, EUA: IEEE, 2007. v. 2.

[78] FENET, S. et al. A scalable audio fingerprint method with robustness to pitch-shifting. In: *Proc. of ISMIR*. Miami, EUA: ISMIR, 2011. p. 121–126.

[79] LU, L.; WANG, M.; ZHANG, H.-J. Repeating pattern discovery and structure analysis from acoustic music data. In: *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*. Nova York, EUA: ACM, 2004. p. 275–282.

[80] SEO, J. S.; HAITSMA, J. A.; KALKER, T. Linear speed-change resilient audio fingerprinting. In: *Proc. of IEEE Benelux Workshop on MPCA*. Leuven, Bï£¡lgica: IEEE, 2002.

[81] VENKATACHALAM, V. et al. Automatic identification of sound recordings. *IEEE Signal Processing Magazine*, p. 92–99, 2004.

[82] BARTSCH, M. A.; WAKEFIELD, G. H. To catch a chorus: using chroma-based representations for audio thumbnailing. In: *Proc. of the Workshop on the Application of Signal Processing to Audio and Acoustics*. New Platz, EUA: IEEE, 2001. p. 5–18.

[83] MULLER, M.; KURTH, F.; CLAUSEN, M. Audio matching via chroma based statistical features. In: *Proc. of ISMIR*. Londres,Inglaterra: ISMIR, 2005. p. 288–295.

[84] RILEY, M.; HEINEN, E.; GHOSH, J. A text retrieval approach to content-based audio retrieval. In: *Proc. of ISMIR*. Filadelfia, EUA: ISMIR, 2008. p. 295–298.

[85] SERRA, J. et al. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 16, n. 6, p. 1138–1151, 2008.

[86] AHONEN, T. E. Combining chroma features for cover version identification. In: *Proc. of ISMIR*. Utrecht, Holanda: ISMIR, 2010. p. 165–170.

[87] CHEN, N. et al. Robust audio hashing scheme based on cochleagram and cross recurrence analysis. *Electronic Letters*, v. 49, n. 1, p. 7–8, 2013.

[88] CANO, P. et al. Robust sound modeling for song detection in broadcast audio. In: *Proc. of Audio Engineering Society Convention 112*. Munique, Alemanha: AES, 2002. p. 1–7.

[89] BATLLE, E.; MASIP, J.; CANO, P. System analysis and performance tuning for broadcast audio fingerprinting. In: *Proc. of the DAFX*. Londres, Inglaterra: IRCAM, 2003.

[90] OZER, H. et al. Perceptual audio hashing functions. *EURASIP Journal on Advances in Signal Processing*, p. 1780–1793, 2005.

[91] BATLLE, E.; MASIP, J.; GUAUS, E. Automatic song identification in noisy broadcast audio. In: *IASTED International Conference on Signal and Image Processing*. Kauai, EUA: IASTED, 2002.

[92] LIU, Y. et al. Coherent bag-of audio words model for efficient large-scale video copy detection. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*. Xian, China: ACM, 2010. p. 89–96.

[93] SUKITTANON, S.; ATLAS, L. E.; PITTON, J. W. Modulation-scale analysis for content identification. *IEEE Transactions on Signal Processing*, v. 52, p. 3023–3035, 2004.

[94] SUKITTANON, S.; ATLAS, L. E. Modulation frequency features for audio fingerprinting. In: *Proc. of ICASSP*. Orlando, EUA: IEEE, 2002. v. 2.

[95] SEO, J. S. et al. Audio fingerprinting based on normalized spectral subband centroids. In: *Proc. of ICASSP*. Filadelfia, EUA: IEEE, 2005. v. 3.

[96] QIAN, Y. Z.; DOU, H. J.; FENG, Y. A novel algorithm for audio information retrieval based on audio fingerprint. In: *Proc. of ICINA*. Kunming, China: IEEE, 2010. v. 1, p. V1–266–V1–270.

[97] LOGAN, B. et al. Mel frequency cepstral coefficients for music modeling. In: *Proc. of ISMIR*. Massachusetts,EUA: ISMIR, 2000.

[98] RAMALINGAM, A.; KRISHNAN, S. Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting. *IEEE Transactions on Information Forensics and Security*, v. 1, n. 4, p. 457–463, 2006.

[99] VOICEBOOX. Acessado em: 11/01/2017. Disponível em: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.

[100] PURDUE. Disponível em: <https://engineering.purdue.edu/ malcolm/interval/1998-010/, note = Acessado em: 11/01/2017>.

[101] GREENBER, S.; KINGSBURRY, B. D. E. The modulation spectrogram: in pursuit of an invariant representation of speech. In: *Proc. of ICASSP*. Munique, Alemanha: IEEE, 1997. p. 1647–1650.

[102] HERMANSKY, H.; MORGAN, N. RASTA processing of speech. *IEEE Transactions on Speech and Acoustics*, v. 2, p. 587–589, 1994.

[103] ANGUERA, X.; GARZON, A.; ADAMEK, T. Mask: Robust local features for audio fingerprinting. In: *Proc. of ICME*. Melbourne, Austrï£¡lia: IEEE, 2012. p. 455–460.

[104] CHOUDULE, S. V.; CHAVAN, M. S. Comparison od frequency-warped filter banks in relation to robust features for speaker recognition. *Recent Advances in Electrical Engineering*, p. 157–162, 2014.

[105] LIU, Y.; YUN, H. S.; KIM, N. S. Audio fingerprinting based on multiple hashing in DCT domain. *IEEE Signal Processing Letters*, v. 16, n. 6, p. 525–528, 2009.

[106] BATLLE, E. et al. Scalability issues in an HMM-based audio fingerprinting. In: *Proc. of ICME*. Taipei, Taiwan: IEEE, 2004. v. 1, p. 735–738.

[107] DENG, J. et al. Audio fingerprinting based on spectral energy structure and NMF. In: *Proc. of IEEE 13th ICCT*. Jinah, China: IEEE, 2011. p. 1103–1106.

[108] PICONE, J. W. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, v. 81, n. 9, p. 1215–1247, 1993.

[109] FURUI, S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, v. 34, n. 1, p. 52–59, 1986.

[110] CLAUSEN, M.; KORNER, H.; KURTH, F. An efficient indexing and search technique for multimedia databases. In: *Proc. of SIGIR Distributed Multimedia Information Retrieval Workshop*. Toronto, Canadï£¡: Springer, 2003.

[111] KURTH, F.; RIBBROCK, A.; CLAUSEN, M. Identification of highly distorted audio material for querying large scale databases. In: *Proc. of Audio Engineering Society Convention 112*. Munique, Alemanha: AES, 2002.

[112] BARDELI, R.; SCHWENNINGER, J.; STEIN, D. Presentation on audio fingerprinting for media synchronization and duplicate detection. In: *Proc. Workshop on MEDIASYNC*. Berlim, Alemanha: UPV, 2012.

[113] CHÁVEZ, E. et al. Searching in metric spaces. *ACM Computing Surveys*, v. 33, p. 273–321, 2001.

[114] MILLER, M. L. Audio fingerprinting: Nearest neighbor search in high dimensional binary spaces. In: *IEEE Workshop on Multimedia Signal Processing*. San Thomas, Virgin Islands: IEEE, 2002. p. 182–185.

[115] FONSECA, M. J.; JORGE, J. A. Indexing high-dimensional data for content-based retrieval in large databases. In: *Proc. of 8th IEEE DASFAA*. Kyoto, Japï£¡o: IEEE Computer Society, 2003. p. 267–274.

[116] WEBER, R.; SCHEK, H.-J.; BLOTT, S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: *Proceedings of the 24th International Conference on Very Large Data Bases*. Nova York, EUA: VLDB Endowment Inc., 1998. p. 194–205.

[117] INDYK, P.; MOTWANI, R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In: *Proceedings of ACM Theory of computing*. NOva York, EUA: ACM, 1998. p. 604–613.

[118] FOOTE, J. Visualizing music and audio using self-similarity. In: *ACM International Conference on Multimedia (Part 1)*. Orlando, EUA: ACM, 1999. p. 77–80.

[119] NAINI, R.; MOULIN, P. Model-based decoding metrics for content identification. In: *Proc. of ICASSP*. Kyoto, Japï£¡o: IEEE, 2012.

[120] VARNA, A. L.; WU, M. Modeling and analysis of correlated binary fingerprints for content identification. *IEEE Transactions on Information Forensics and Security*, v. 6, n. 3, p. 1146–1159, 2011.

[121] JANG, D.; YOO, C. D.; KALKER, T. Distance metric learning for content identification. *IEEE Transactions on Information Forensics and Security*, v. 5, n. 4, p. 932–944, 2010.

[122] KIMURA, A. et al. Very quick audio searching: introducing global pruning to the time-series active search. In: *Proc. of ICASSP*. Salt Lake City, EUA: IEEE, 2001. v. 3, p. 1429–1432.

[123] MICO, M. L.; ONCINA, J.; VIDAL, E. A new version of the nearest-neighbor approximating and eliminating search algorithm(AESA) with liner preprocessing time and memory requirements. *Pattern Recognition Letters 15*, p. 9–17, 1994.

[124] JANG, D. et al. Automatic commercial monitoring for TV broadcasting using audio fingerprinting. In: *Proc. of AES 29th International Conference*. Seul, Corï£¡ia do Sul: AES, 2006.

[125] CHA, G. An effective and efficient indexing scheme for audio fingerprinting. In: *Proc. Multimidia and Ubiquitous Eng.* Seul, Corï£¡ia do Sul: IEEE, 2011. p. 48–52.

[126] TUYLS, P.; ŠKORIC, B.; KEVENAAR, T. *Security with noisy data: on private biometrics, secure key storage and anti-counterfeiting*. [S.l.]: Springer Science & Business Media, 2007.

[127] SCHEDL, M.; GÓMEZ, E.; URBANO, J. *Music information retrieval: recent developments and applications*. [S.l.]: Now Publishers, 2014.

[128] SHAZAM. Acessado em: 11/12/2015. Disponível em: <http://www.shazam.com/apps>.

[129] GRACENOTE. Acessado em: 11/12/2015. Disponível em: <http://www.gracenote.com/>.

[130] MUSICBRAINZ. Acessado em: 11/12/2015. Disponível em: <https://picard.musicbrainz.org>.

[131] RELATABLE. Acessado em: 11/12/2015. Disponível em: <http://www.relatable.com/tech/trm.html/>.

[132] ACOUSTICID. Acessado em: 11/12/2015. Disponível em: <https://acoustid.org/>.

[133] RAFII, Z.; COOVER, B.; HAN, J. An audio fingerprinting system for live version identification using image processing techniques. In: *Proc. of ICASSP*. Florenï£¡a, Itï£¡lia: IEEE, 2014. p. 644–648.

[134] CONTENTID. Acessado em: 11/12/2015. Disponível em: <https://support.google.com/youtube/answer/2797370?hl=pt-BR>.

[135] ECAD. Acessado em: 11/12/2015. Disponível em: <http://www.ecad.org.br/pt/Paginas/default.aspx>.

[136] PLAYAX. Acessado em: 11/12/2015. Disponível em: <http://www.playax.com>.

[137] BONEY, L.; TEWFIK, A.; HAMDY, K. Digital watermarks for audio signals. In: *Proc. of ICMCS*. Hiroshima, Japï£¡o: IEEE, 1996. p. 473ï£¡480.

[138] GOMES, L. d. C. et al. Audio watermarking and fingerprinting: for which applications? *Journal of New Music Research*, Taylor & Francis, v. 32, n. 1, p. 65–81, 2003.

[139] GOMEZ, E. et al. Mixed watermarking-fingerprinting approach for integrity verification of audio recordings. In: *Int. Telecom. Symp.* Natal, Brasil: SBrT, 2002.

[140] DOETS, P. J. O.; LAGENDIJK, R. L. Distortion estimation in compressed music using only audio fingerprints. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 16, n. 2, p. 302–317, 2008.

[141] BEERENDS, J. G.; STEMERDINK, J. A. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, Audio Engineering Society, v. 40, n. 12, p. 963–978, 1992.

[142] BEERENDS, J. G. Audio quality determination based on perceptual measurement techniques. In: *Applications of Digital Signal Processing to Audio and Acoustics*. [S.l.]: Springer, 2002. p. 1–38.

[143] THIEDE, T.; KABOT, E. A new perceptual quality measure for bit-rate reduced audio. In: *Proc. of AES Convention 100*. Copenhagen: AES, 1996.

[144] HERRERO, C. Subjective and objective assessment of sound quality: Solutions and applications. In: *Proc. CIARM Conference*. [S.l.: s.n.], 2005. p. 1–20.

[145] THIEDE, T. et al. PEAQ- the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, Audio Engineering Society, v. 48, n. 1/2, p. 3–29, 2000.

[146] HICSONMEZ, S.; UZUN, E.; SENCAR, H. T. Methods for identifying traces of compression in audio. In: *Proc. of ICCSPA*. Sharjah, Emirados ï£¡rabes Unidos: IEEE, 2013. p. 1–6.

[147] DOETS, P. J. O.; LAGENDIJK, R. L. Extracting MP3 quality parameters from audio fingerprints. In: *ASCI*. [S.l.: s.n.], 2005.

[148] CUNNINGHAM, S.; GROUT, V. Audio compression exploiting repetition (ACER): Challenges and solutions. In: *Proc. of Third International Conference on Internet Technologies and Applications*. Wales, Inglaterra: Glyndwr University, 2009.

[149] ALEXANDER OSCAR FORTH, D. T. A. Music and noise fingerprinting and reference cancellation applied to forensic audio enhancement. In: *Proc. of AES 46th International Conference*. Denver, EUA: AES, 2012.

[150] QUATIERI, T. F. *Discrete-time speech signal processing: principles and practice*. [S.l.]: Prentice Hall PTR, 2001. (Prentice-Hall signal processing series).

[151] PARK, M. et al. Frequency filtering for a highly robust audio fingerprinting scheme in a real-noise environment. *IEICE Trans. on Information and Systems*, E89-D, p. 2324–2327, 2006.

[152] PARK, M. et al. Audio fingerprinting scheme by temporal filtering for audio identification immune to channel-distortion. In: *Asia Information Retrieval Symposium*. Jeju Island, Corï£¡ia do Sul: [s.n.], 2005. p. 528–533.

[153] PARK, M.; KIM, H.-R.; YANG, S. H. Frequency-temporal filtering for a robust audio fingerprinting scheme in real-noise environments. *ETRI Journal*, Electronics and Telecommunications Research Institute, v. 28, n. 4, p. 509–512, 2006.

[154] JUNG, H.-Y. Filtering of filter-bank energies for robust speech recognition. *ETRI Journal*, Electronics and Telecommunications Research Institute, v. 26, n. 3, p. 273–276, 2004.

[155] NADEU, C.; MACHO, D.; HERNANDO, J. Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication*, Elsevier, v. 34, n. 1, p. 93–114, 2001.

[156] THIEMERT, S. et al. Security of robust audio hashes. In: *Proc. of International WIFS*. Seattle, USA: IEEE, 2009. p. 126–130.

[157] DOETS, P.; LAGENDIJK, R. Stochastic model of a robust audio fingerprinting system. In: *Proc. of ISMIR Conference*. Barcelona, Espanha: ISMIR, 2004. p. 349–352.

[158] DOETS, P.; LAGENDIJK, R. Extracting quality parameters for compressed audio from finngerprints. In: *Proc. of ISMIR Conference*. Londres, Inglaterra: ISMIR, 2005. p. 498–503.

[159] YANG, R. Additive noise detection and its application to audio forensics. In: *Proc. Annual Summit and Conference (APSIPA)*. [S.l.]: IEEE, 2014. p. 1–5.

[160] HAITSMA, J.; KALKER, T. Speed-change resistant audio fingerprinting using auto-correlation. In: *Proc. of ICASSP*. Hong Kong: IEEE, 2003. v. 4, p. 728–731.

[161] TYAGI, V.; WELLEKENS, C. On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition. In: *Proc. of ICASSP*. Filadelfia, EUA: IEEE, 2005. p. 529–532.

[162] NOROUZI, Y.; GRECO, M. S.; NAYEBI, M. M. Performance evaluation of k out of n detector. In: *Proc. of EUPSICO*. Florenï£¡a, Itï£¡lia: EURASIP, 2006. p. 1–5.

[163] MINEMATSU, N.; SEKIGUCHI, M.; HIROSE, K. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In: *Proc. of ICASSP*. Orlando, EUA: IEEE, 2002.

[164] MORAN, R. J. et al. Telephony-based voice pathology assessment using automated speech analysis. *IEEE Transactions on Biomedical Eng.*, v. 53, p. 468–477, 2006.

[165] BELHAJ, A.; BOUZID, A.; ELLOUZE, N. Effects of a new voicing parameter on pathological voice discrimination by SVM. *International Journal of Computational and Information Technology*, v. 3, p. 1083–1095, 2014.

[166] MAJIDNEZHAD, V. A novel hybrid of genetic algorithm and ann for developing a high efficient method for vocal fold pathology diagnosis. *EURASIP Journal on Audio, Speech, and Music Processing*, Springer, v. 2015, n. 1, p. 1–11, 2015.

[167] HARIHARAN, M. et al. A hybrid expert system approach for telemonitoring of vocal fold pathology. *Applied Soft Computing*, v. 13, p. 4148–4161, 2013.

[168] BEHROOZMAND, R.; ALMASGANJ, F. Comparison of neural networks and support vector machines applied to optimized features extracted from patients' speech signal for classification of vocal fold inflammation. In: *International Symposium on Signal Processing and Information Technology*. Atenas, Grï£¡cia: IEEE, 2005. p. 844–849.

[169] MAHER, R. C. Overview of audio forensics. In: *Intelligent Multimedia Analysis for Security Applications*. [S.l.]: Springer, 2010. p. 127–144.

[170] AES27-1996 (r2007) AES recommended practice for forensic purposes- Managing recorded audio materials intended for examination. [S.l.]: AES, 2007.

[171] IKAR Lab. Disponível em: <http://speechpro.com/product/forensic analysis/ikarlab>.

[172] AUDIO Forensics Toolbox. Acessado: 26/06/2016. Disponível em: <http://www.idmt.fraunhofer.de/en/institute/projects products/ad/audioforensics.html>.

[173] KOENIG, B. Authentication of forensic audio recordings. *Journal of the Audio Engineering Society*, v. 38, p. 3–33, 1990.

[174] AES(2000). Aes standard for forensic purposes- criteria for the authentication of analog audio tape recordings. *Journal of the Audio Engineering Society*, v. 48, p. 204–214, 2000.

[175] BOSS, D.; GFROERER, S.; NEOUSTRUEV, N. A new tool for the visualization of magnetic features on tapes. *Forensic Linguistics*, 2001.

[176] READ, M. E. et al. Magnetic scanner for forensic examination of audio tapes. *Proc. SPIE*, v. 3576, p. 144–153, 1999.

[177] PAPPAS, D. P. et al. Second-harmonic magnetoresistive imaging to authenticate and recover data from magnetic storage media. *Journal of Electronic Imaging*, v. 14, n. 1, p. 013015–013015, 2005.

[178] DEAN, D. *Publication Nr 16/1991: The relevance of replay transients in the forensic examination of analogue magnetic tape recorders*. [S.l.], 1991.

[179] MAHER, R. Audio forensic examination. *IEEE Signal Processing Magazine*, v. 26, n. 2, p. 84–94, 2009.

[180] GUPTA, S.; CHO, S.; KUO, C. C. J. Current developments and future trends in audio authentication. *IEEE Multimedia*, v. 19, n. 1, p. 50–59, 2012.

[181] YANG, R.; QU, Z.; HUANG, J. Exposing MP3 audio forgeries using frame offsets. *ACM Trans. Multimedia Comput. Commun. Appl.*, v. 8, n. 2S, p. 35:1–35:20, Set 2012.

[182] YANG, R.; QU, Z.; HUANG, J. Detecting digital audio forgeries by checking frame offsets. In: *Proceedings of the 10th ACM workshop on Multimedia and security*. Inglaterra: ACM, 2008. p. 21–26.

[183] LIU, Q.; SUNG, A. H.; QIAO, M. Detection of double MP3 compression. *Journal of Cognitive Computing*, v. 2, n. 4, p. 291–296, 2010.

[184] QIAO, M.; SUNG, A. H.; LIU, Q. Revealing real quality of double compressed MP3 audio. In: *Proceedings of the 18th ACM International Conference on Multimedia*. Florenï£¡a, Itï£¡lia: ACM, 2010. p. 1011–1014.

[185] YANG, R.; SHI, Y. Q.; HUANG, J. Detecting double compression of audio signal. In: *Proc. of IS&T/SPIE Electronic Imaging*. San Jose, EUA: SPIE, 2010.

[186] HUANG, J.; SHI, Y. Q.; YANG, R. Defeating fake-quality MP3. In: *Proceedings of the 11th ACM Workshop on Multimedia and Security*. Princeton, EUA: ACM, 2009.

[187] SHEN, Y.; JIA, J.; CAI, L. *Detecting Double Compressed AMR-format Audio Recordings*.

[188] KORYCKI, R. Detection of montage in lossy compressed digital audio recordings. *Archives of Acoustics*, v. 39, n. 1, p. 65–72, 2014.

[189] KORYCKI, R. Authenticity examination of lossy compressed digital audio recordings. In: *Proc. of Forum Acousticum*. Cracï£¡via, Polï£¡nia: EAA, 2014.

[190] SHI, Q.; MA, X. Detection of audio interpolation based on singular value decomposition. In: *Proc. iCAST Conference*. Dalian, China: IEEE, 2013. p. 287–290.

[191] IKRAM, S.; MALIK, H. Digital audio forensics using background noise. In: *Proc. IEEE ICME*. Singapura: IEEE, 2010. p. 106–110.

[192] PAN, X.; ZHANG, X.; LYU, S. Detecting splicing in digital audios using local noise level estimation. In: *Proc. of ICASSP*. Kyoto, Japï£¡o: IEEE, 2012. p. 1841–1844.

[193] RODRIGUEZ, D.; APOLINï£¡RIO, J. A.; BISCAINHO, L. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Transactions on Information Forensics and Security*, v. 5, n. 3, p. 534–543, 2010.

[194] NICOLALDE, D. P.; APOLINï£¡RIO, J. A. Evaluating digital audio authenticity with spectral distances and ENF phase change. In: *Proc. of ICASSP*. Taipei, Taiwan: IEEE, 2009. p. 1417–1420.

[195] ARCHER, H. Quantifying effects of lossy compression on electric network frequency signals. In: *Proc. of AES 46th International Conference*. Denver, EUA: AES, 2012.

[196] GARG, R.; VARNA, A. L.; WU, M. Modeling and analysis of electric network frequency signal for timestamp verification. In: *Proc. of IEEE WIFS*. Tenerife, Espanha: IEEE, 2012. p. 67–72.

[197] MALIK, H. Acoustic environment identification and its applications to audio forensics. *Transactions on Information Forensics and Security*, v. 8, n. 11, p. 1827–1837, 2013.

[198] MALIK, H.; FARID, H. Audio forensics from acoustic reverberation. In: *Proc. of I CASSP*. Dallas, EUA: IEEE, 2010. p. 1710–1713.

[199] ZHAO, H.; MALIK, H. Audio forensics using acoustic environment traces. In: *Proc. of SSP Workshop*. Miami, EUA: IEEE, 2012. p. 373–376.

[200] MALIK, H.; ZHAO, H. Recording environment identification using acoustic reverberation. In: *Proc. of ICASSP*. Kyoto, Japï£¡o: IEEE, 2012. p. 1833–1836.

[201] ZHAO, H.; MALIK, H. Audio recording location identification using acoustic environment signature. *IEEE Transactions on Information Forensics and Security*, v. 8, n. 11, p. 1746–1759, 2013.

[202] ZHAO, H. et al. Audio splicing detection and localization using environmental signature. *arXiv preprint arXiv:1411.7084*, 2014.

[203] FARID, H. *Detecting Digital Forgeries Using Bispectral Analysis*. 1999.

[204] CHEN, J. et al. Exposing digital audio forgeries in time domain by using singularity analysis with wavelets. In: *Proc. IH&MMSec '13*. Monpellier, Franï£¡a: ACM, 2013. p. 149–158.

[205] KORYCKI, R. Time and spectral analysis methods with machine learning for the authentication of digital audio recordings. *Forensic Science International*, v. 230, p. 117–126, 2013.

[206] GARCIA-ROMERO, D.; ESPY-WILSON, C. Speech forensics: Automatic acquisition device identification. *The Journal of the Acoustical Society of America*, v. 127, n. 3, 2010.

[207] GARCIA-ROMERO, D.; ESPY-WILSON, C. Y. Automatic acquisition device identification from speech recordings. In: *Proc. of ICASSP*. Dallas, EUA: IEEE, 2010. p. 1806–1809.

[208] IKRAM, H. M. S. Microphone identification using higher-order statistics. In: *Proc. of AES 46th International Conference*. Denver, EUA: AES, 2012.

[209] KURNIAWAN, F.; KHALIL, M. S.; MALIK, H. Robust tampered detection method for digital audio using gabor filterbank. In: *ICIPCS*. Istambul, Turquia: [s.n.], 2015. p. 75–82.

[210] KRAETZER, C. et al. Digital audio forensics: a first practical evaluation on microphone and environment classification. In: *Proc. Workshop on Multimedia & Security*. Dallas, EUA: ACM, 2007. p. 63–74.

[211] KRAETZER, C.; SCHOTT, M.; DITTMANN, J. Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models. In: *Proceedings of the 11th ACM workshop on Multimedia and security*. Princeton, EUA: ACM, 2009. p. 49–56.

[212] BUCHHOLZ, R.; KRAETZER, C.; DITTMANN, J. Microphone classification using fourier coefficients. In: *Proc. Information Hiding Int. Workshop*. Darmstadt, Alemanha: Springer, 2009. p. 235–246.

[213] KAIN, A.; MACON, M. W. Spectral voice conversion for text-to-speech synthesis. In: *Proc. of ICASSP*. Seattle, EUA: IEEE, 1998. v. 1, p. 285–288.

[214] WU, Z. et al. Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints. In: *Proc. of INTERSPEECH Conference*. Lyon, Franï£¡a: International Speech Communication Association, 2013. p. 950–954.

[215] WU, Z.; SIONG, C. E.; LI, H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: *Proc. of INTERSPEECH Conference*. Portland, EUA: International Speech Communication Association, 2012. p. 1700–1703.

[216] ALEGRE, F.; AMEHRAYE, A.; EVANS, N. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In: *Proc. of ICASSP*. Vancouver, Canadï£¡: IEEE, 2013. p. 3068–3072.

[217] ALEGRE, F. et al. A new speaker verification spoofing countermeasure based on local binary patterns. In: *Proc. of INTERSPEECH 14th Annual Conference*. Lyon, France: International Speech Communication Association, 2013. p. 5p.

[218] ALEGRE, F.; AMEHRAYE, A.; EVANS, N. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In: *Proc. of BTAS Conference*. Arlington, EUA: IEEE, 2013. p. 1–8.

[219] INTERNATIONAL Conference on Music Information Retrieval. Acessado em: 11/12/2015. Disponível em: <http://ismir2004.ismir.net>.

[220] MUSIC Information Retrieval Evaluation eXchange. Acessado em: 11/12/2015. Disponível em: <http://www.music-ir.org/mirex/wiki/MIREX HOME>.

[221] DOWNIE, J. S. et al. Ten years of MIREX (music information retrieval evaluation exchange): Reflections, challenges and opportunities. In: *Proc. of ISMIR Conference*. Taipei,Taiwan: ISMIR, 2014. p. 657–662.

[222] CHANDRASEKHAR, V.; SHARIFI, M.; ROSS, D. A. Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications. In: *Proc. of ISMIR*. Miami, EUA: ISMIR, 2011. v. 20, p. 801–806.

[223] KE, Y.; HOIEM, D.; SUKTHANKAR, R. Computer vision for music identification. In: *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*. San Diego, EUA: IEEE Computer Society, 2005. v. 1, p. 597–604.

APÊNDICES

A- FONÉTICA E ACÚSTICA FORENSE

As análises de evidências de áudio contendo voz aplicáveis em procedimentos judiciais, que abrangem conhecimentos multidisciplinares dentro de uma área comumente denominada de Fonética Forense, engloba diversos tipos de exames, como a comparação ou identificação de locutor; a análise de conteúdo de áudio com baixa inteligibilidade, por meio de técnicas de melhoria de áudio ou da análise acústico-linguística de locuções questionadas; ou a verificação de edições ou fonte de áudio, que engloba a análise da origem e da integralidade da evidência de áudio. Em um contexto mais amplo, a análise de evidências de áudio, como a identificação de disparos de arma de fogo e classificação da munição com base nos estampidos, ou a identificação do ambiente acústico onde o áudio foi capturado a partir da análise dos componentes de reverberação, é denominada de Acústica Forense. Uma visão geral destas áreas de aplicação, com uma descrição mais detalhada de exames em Acústica Forense, é feita em [169].

A relevância da Fonética Forense decorre da grande disponibilidade da evidências de áudio contendo voz, que podem tanto ser apresentadas por partes interessadas no processo, quanto terem sido capturadas por meio de interceptação legal. Diversos grupos de trabalho, no âmbito governamental, em esfera nacional ou internacional foram criados para desenvolver ou tentar padronizar a análise forense de áudio, como:

- Grupo de Trabalho em Voz e Áudio Forense (*Forensic Speech and Audio Analysis Working Group* - FSAAWG), no âmbito da Rede de Institutos Europeus de Ciência Forense (NFSI),
- Unidade de Análise Forense de Vídeo, Imagem e Áudio (*Forensic Analysis of Video, Image and Audio Unit* - FAVIAU), no âmbito do FBI-EUA.
- Grupo de Trabalho em Áudio Forense, no âmbito do Grupo Científico de Trabalho em Mídias Digitais (*Scientific Working Group on Digital Evidence* - SWGDE), formado por diversos órgãos do governo americano.

Outros grupos de trabalho com propósito semelhante também foram criados por representantes da comunidade científica ou da indústria:

- Grupo de Trabalho em Áudio Forense AES-WG-12, no âmbito do comitê de padronização da Sociedade de Engenharia de Áudio (*Audio Engineering Society* - AES).
- Grupo de pesquisa aplicada na Alemanha *Fraunhofer Institute for Digital Media Technology*, que também desenvolve pesquisa em autenticidade de áudio.

Os desafios impostos na aplicação de análise de áudio forense, com diversos problemas em aberto, também têm atraído o interesse da comunidade científica.

A.1 ANÁLISE DE AUTENTICIDADE DE ÁUDIO FORENSE

O termo *autenticação* é comumente usado no contexto forense para descrever o estabelecimento dos fundamentos legais para admissibilidade de um registro de áudio em um processo judicial. Com base na casuística dos diversos tipos de exame de acústica forense, pode-se afirmar que a análise da autenticidade de áudio é uma das principais tarefas de peritos forenses em áudio, pois em diversos países este é um requisito para a admissibilidade da evidência de áudio digital. Esta preocupação dos tribunais a respeito da aceitação da evidência de áudio digital advém da incerteza acerca de sua autenticidade, devido à facilidade de edição e à ausência de um método automático e confiável de verificação de integridade de áudio digital.

O padrão AES27-1996-(r2007) [170], publicado pelo Grupo de Trabalho WG-12 da Sociedade de Engenharia de áudio, define uma gravação autêntica como: "*Uma gravação realizada simultaneamente aos eventos acústicos, e de forma totalmente consistente com os métodos de gravação alegados pela parte que a produziu; uma gravação livre de artefatos inexplicáveis, alterações, adições, supressões ou emendas*". A análise de autenticidade é definida pela mesma norma como: "*Um exame, usualmente com propósitos forenses, que visa a determinar se uma dada gravação foi feita de eventos acústicos alegados pela parte que a produziu, e da forma alegada pela parte, e verificar se consiste de uma cópia ou se é uma gravação original*". Dessa forma, a análise de autenticidade de áudio visa atestar, na medida do possível, se a gravação de áudio corresponde a uma representação fidedigna dos eventos acústicos capturados de uma forma, em um lugar e em um tempo específicos, e pode englobar tanto a verificação da origem, quanto da integralidade do áudio apresentado.

A análise da origem da evidência está associada às condições, como o meio, o equipamento usado na captura, o tempo e o local em que o sinal de áudio primário foi capturado. Dessa forma, a análise de origem engloba, além da análise do sinal de áudio, a análise da mídia de suporte, dos metadados e do equipamento usado na captura. Apesar de clones ou cópias íntegras transcodificadas do áudio serem admissíveis em algumas cortes judiciais, a verificação da origem da mídia questionada também pode ser útil para identificar fraudes, caso inconsistências sejam observadas entre uma evidência apontada como mídia original e as características do equipamento usado na captura.

A análise de integridade (ou verificação de edições) visa identificar a presença de edições no áudio, como mascaramentos ou descontinuidades decorrentes de supressões, inserções, remanejamentos e emendas de trechos no sinal de áudio.

Podemos também classificar as análises de acordo com os recursos empregados como: perceptual, baseada na análise visual de gráficos do sinal no domínio do tempo/frequência ou na oitiva atenta; assistida, onde softwares são empregados para visualização de atributos acústicos específicos, como a variação da fase de um componente harmônico do sinal (e.g. interferência da rede elétrica); ou automática, através do emprego de algoritmos para identificação de edições ou da origem do áudio.

Alguns softwares, como o Ikar Lab [171] ou o *Audio Forensic Toolbox* [172], possuem recursos de análise automática de autenticidade de áudio, mas não fornecem uma boa documentação acerca da abordagem nem da precisão dos métodos empregados. Na prática, devido à ausência de ferramentas comerciais de análise automática, em geral o exame de autenticidade de áudio é feito de forma perceptual ou assistida. A análise perceptual é feita pela oitiva atenta de aspectos linguísticos e suprasegmentais, em conjunto com a análise acústica visual da forma de onda, de espectrogramas, espectro de longo termo, e atributos intrínsecos da voz, como pitch e formantes. Na análise assistida, uma abordagem comum é a pesquisa de descontinuidade de fase em componentes harmônicos do sinal de áudio, como a interferência da rede elétrica, como pelo uso do espectro de fase no software Adobe Audition TM.

Os métodos automáticos de verificação de autenticidade podem ser categorizados de acordo com a abordagem como: Autenticação Ativa, onde uma informação é adicionada ao áudio original no momento da captura sem alterar seu conteúdo perceptual, para ser empregada em um processo de verificação de origem e integridade posterior; ou Autenticação Passiva, onde a análise é feita exclusivamente com base na informação contida no sinal de áudio ou em metadados. A Autenticação Ativa, pelo uso de primitivas criptográficas ou pelo emprego de marca d'água digital em áudio, permitiria a verificação de integridade e origem do áudio, entretanto não parece razoável que em pouco tempo recursos de Autenticação Ativa estejam disponíveis em gravadores de áudio comerciais. Dessa forma, os métodos de Autenticação Passiva continuam sendo a única alternativa disponível para análise de evidências de áudio, e todos os métodos disponíveis devem ser empregados para elevar a confiabilidade geral do exame.

Cabe ressaltar que conceitos distintos de autenticação passiva e ativa existem em diversos campos de aplicação. Entretanto, a definição aqui adotada é compatível com aquela empregada na autenticação de áudio comercial. No capítulo 2, são descritas diferenças entre o método de Autenticação Ativa de áudio comercial, por meio da técnica de marca d'água digital em áudio, e o método de Autenticação Passiva, por meio da técnica de *Audio Fingerprinting*. Embora a tecnologia de marca d'água em registros de áudio permita tanto verificar a existência como identificar a posição de uma edição no áudio, nenhum padrão comercial amplo foi desenvolvido até o momento. Por fim, cabe destacar que na autenticação passiva em áudio comercial, músicas são preprocessadas para a extração e armazenamento de uma assinatura acústica (*Audio Fingerprint*) em uma base de dados.

Diferentemente, na aplicação de autenticação passiva de áudio forense, em geral, não há uma base de dados com informações acessórias acerca do áudio questionado. Como citado anteriormente, a análise é feita exclusivamente com base na informação acústica, o que a torna mais difícil.

A seguir é apresentado um histórico do desenvolvimento de métodos automáticos de Autenticação Passiva de áudio forense.

A.2 REVISÃO DOS MÉTODOS AUTOMÁTICOS PROPOSTOS PARA AUTENTICAÇÃO PASSIVA DE ÁUDIO FORENSE

As primeiras pesquisas sobre autenticação de áudio forense surgiram quando a principal mídia de armazenamento era a fita magnética, e o emprego de áudio digital ainda não era difundido. Os métodos focavam principalmente na análise física da mídia [173, 174, 175, 176, 177].

Até pouco tempo, os exames em áudio forense se baseavam na análise de oitiva e na busca visual por sinais característicos de acionamentos de gravadores [178] ou por inconsistências do sinal no domínio do tempo e da frequência. A ausência de métodos automáticos limitava a análise a áudios curtos [179].

Com a difusão dos padrões de áudio digital, novos métodos foram propostos para a análise de autenticidade de áudio. Uma revisão mais recente dos métodos de análise forense de áudio é fornecida em [180]. Diversas abordagens podem ser empregadas, como:

1. Análise dos efeitos da compressão de áudio: A identificação de inconsistências na característica do *offset* de quantização de coeficientes da MDCT é proposta em [181, 182] para a identificação de edições em áudios comprimidos com MP3 e AAC. A detecção da dupla compressão MP3 é proposta em [183], pela análise do número de coeficientes MDCT nulos, e em [184], pela análise da distribuição do número de coeficientes nulos. Em [185] a detecção de dupla compressão MP3 é proposta pela análise da similaridade da distribuição dos coeficientes MDCT com a distribuição teórica baseada na Lei de Benford. Em [186] é proposto um método para detectar a dupla compressão com sobre-amostragem na segunda codificação. Em [187], um conjunto de atributos extraídos do áudio, inclusive pelo emprego de análise biespectral, é usado com um classificador SVM para detecção de dupla compressão de áudio codificado em AMR (*Adaptive Multi-Rate*). Em [188] é proposto um método baseado na análise da regularidade interquadros das posições dos mínimos da função de número de coeficientes MDCT ativos (NAC) para identificar montagens em áudio. O método é aplicável a codificadores de áudio que empregam os coeficientes MDCT, como AAC e Vorbis empregados nos testes, com um desempenho de detecção cor-

reta em torno de 99% para uma taxa aproximadamente nula de falsa detecção. Em [189], além dos coeficientes MDCT, outros atributos descritos na norma ISO/IEC 13919-3, que variam de acordo com a implementação, são usados para a identificação de compressão múltipla, para a detecção de montagens ou para a identificação do codificador usado na compressão do áudio original. Portanto, como observado, este tipo de abordagem pode ser usada tanto para verificar a originalidade quanto a integridade do áudio questionado. Em [146] são propostos um método, baseado na medição de qualidade de áudio decodificado, e outro método, baseado em estatísticas de ordem superior da taxa de bit do áudio codificado, para a detecção de compressão simples ou dupla, ou para a identificação de diversos codificadores como AAC, AMR, G.709, GSM 6.10, GSM WAV, além do codificador MP3.

2. Análise de efeitos de reamostragem do sinal: Em [190], um método, baseado na SVD e na análise do número de autovalores não nulos do sinal, é proposto para verificar a ocorrência de reamostragem e interpolação do sinal. Esta abordagem é usada, notadamente, na verificação da originalidade do áudio questionado.

3. Análise do ruído de fundo: Em [191], a correlação do ruído de fundo é usada para detectar emendas em áudio digital. A análise de curtose é usada na detecção de descontinuidades no nível de ruído para detectar emendas em [192]. Como observado, esta abordagem de análise de ruído de fundo pode ser usada tanto na verificação da integridade, quanto na verificação de originalidade do áudio questionado.

4. Análise de componentes harmônicos: Em [193, 194], a análise de fase com alta precisão é usada para recuperar a informação de frequência da interferência da rede elétrica (*Electrical Network Frequency*- ENF) no áudio questionado. Variações bruscas detectadas na ENF podem indicar edições fraudulentas de inserção, supressão ou emenda de trechos de áudio. A informação da ENF é preservada mesmo após a compressão MP3 e WMA com perdas [195], mas pode ser destruída com uma filtragem passa-banda. Além da análise de descontinuidade de fase, a série temporal de frequências da ENF pode ser usada como uma assinatura temporal do áudio, a qual pode ser comparada com registros de frequência da rede de distribuição elétrica no local e horário alegado. Em [196], a análise baseada em um modelo autorregressivo da ENF fornece uma boa detecção de trechos equivalentes para áudios com 512 s de duração. Essa abordagem se assemelha à autenticação passiva de áudio comercial, onde a *Audio Fingerprint* extraída da música é comparada com a *Audio Fingerprint* armazenada em uma base de dados confiável. Na aplicação forense, a base da dados registraria a variação temporal da frequência da rede elétrica.

5. Identificação do ambiente acústico: Técnicas de reconhecimento automático de ambiente (*Automatic Environment Identification*- AEI) podem ser empregadas na aplicação forense. A integridade do áudio pode ser verificada através da análise de coe-

rência temporal das características dos componentes de reverberação estimados do áudio, como proposto em [197, 198, 199, 200]. Em [201] é proposto um método de estimação de componentes de reverberação baseado na subtração espectral seguida de uma análise estatística. A robustez deste método é testada contra o mascaramento por compressão de áudio. Em [202] é proposto um método de detecção de inserções de trechos de áudio de origem diferente, através da análise da coerência da estimação da resposta ao impulso do canal de áudio e do ruído de fundo, entre o áudio questionado e um áudio de referência capturado nas condições alegadas.

6. Análise de descontinuidades do sinal: Edições em áudio podem ainda ser reveladas através da detecção de transições abruptas do sinal de áudio. A análise biespectral é usada para detectar não-linearidades locais [203]. Em [204], a WPD é aplicada para detectar singularidades no sinal de voz. Em [205], a predição linear residual da energia de sinal é usada para detectar pontos de edição. Esta abordagem pode ser usada na verificação da integridade do áudio questionado.

7. Detecção de ruído branco aditivo: As edições em áudio podem ser facilmente mascaradas pela inserção de ruído, o que dificulta a identificação perceptual da emenda, tanto pela oitiva quanto pela análise visual de espectrogramas. Em [159], um método baseado na análise da taxa de cruzamento de zeros do sinal diferencial é proposto para a identificação de adição de ruído branco após a edição. O autor considera que um ruído aditivo de baixa intensidade, a uma SNR de 30 dB, já é suficiente para mascarar o sinal. Portanto, o método emprega uma adição ativa de ruído branco a uma SNR de 30dB. São realizados testes com amostras de voz com ruído branco aditivo a uma SNR de até 35dB, com 5s, 1s e 500ms de duração, e uma boa detecção, acima de 99%, é obtida.

8. Identificação do equipamento utilizado na gravação: Características intrínsecas do dispositivo de gravação são obtidas por meio de um Modelo de Misturas Gaussianas obtido de coeficientes MFCC extraídos das vozes gravadas pelo dispositivo, e são usadas para identificar microfones e telefones em [206, 207]. Estatísticas de ordem superior podem ser aplicadas para modelar artefatos gerados pelo dispositivo, que podem ser usados na identificação de microfones [208]. Em [209], a detecção de inserções de trechos de 15s de duração em áudios coletados com microfones distintos, mesmo que sejam do mesmo modelo, é testada usando atributos de *Audio Fingerprint* baseados em MFCC, PLP e bancos de filtros de Gabor, e um classificador K-NN é empregado na identificação. O melhor desempenho é observado para o emprego dos bancos de filtros de Gabor. Em [210] são empregadas 7 estatísticas no domínio do tempo e 56 coeficientes MFCC na representação do áudio para identificar o microfone usado na captura, usando diversos classificadores, com desempenho razoável. Em [211], este método é melhorado pelo emprego de uma técnica de fusão

de decisão ao nível dos classificadores. Em [212], é testado o emprego de coeficientes de Fourier na representação do áudio e a identificação do microfone usado na captura. Dessa forma, a identificação do equipamento usado na captura pode ser usada tanto na verificação da originalidade, quanto da integridade do áudio questionado.

9. Detecção de transformação de voz: A transformação de pitch, pelos diversos métodos propostos como PSOLA (*Pulse Syncronization Overlap and Add*) ou Phase VOCODER, pode ser usada tanto para disfarce de voz quanto para modificar o sentido de uma locução, como converter interrogações em afirmações, ou vice-versa. A transformação de voz [213] permite a transformação de uma locução mapeando atributos acústicos para um padrão de outro locutor, o que pode ser usado tanto para disfarce de voz, quanto para tentar imputar a autoria da fala a um locutor qualquer. Diversos métodos de transformação de voz foram propostos, e os sistemas de reconhecimento automático de locutor são em geral vulneráveis a este tipo de edição fraudulenta [214]. Dessa forma, diversos métodos automáticos foram propostos para detecção de transformação de voz [215, 216, 217, 218].

B- REVISÃO DE ESQUEMAS DE *AUDIO FINGERPRINTING*

Como observado, não há um padrão único de abordagem para a identificação de áudio por conteúdo para aplicações comerciais. O desempenho dos sistemas propostos depende muito das bases de áudios analisadas, e pode ter sido otimizado para um *corpus* de áudio específico. Dessa forma, a comparação dos métodos existentes necessita da definição de métricas padronizadas de avaliação e do emprego de uma única base de dados de teste. Alguns esforços neste sentido começaram a ser feitos. Em 2004, foi realizado um concurso de métodos de identificação de música (*Audio Description Contest*-ADC) na Conferência Internacional de Identificação de Informação de Músicas (ISMIR) [219]. No ano seguinte, em 2005, foi realizado o uma competição de métodos em várias aplicações de áudio no âmbito do MIREX(*Music Information Retrieval Evaluation Exchange*- Intercâmbio de Avaliação de Identificação de Informação de Música) [220], o que vem sendo repetido anualmente, conforme descrito em [221]. Um grande salto do desenvolvimento de sistemas de MIR foi a criação do conjunto de testes MSD (*Million Music Dataset*), com milhões de música com metadados, permitindo testes em uma escala realmente comercial. Entretanto, as aplicações incluídas nas avaliações do MIREX, que vão desde a identificação de música por conteúdo até segmentação automática de áudio, não incluem a detecção de réplicas curtas.

Para construir um esquema de AF adequado à aplicação forense de detecção de réplicas, inicialmente nós analisamos a adequabilidade dos esquemas de AF existentes. Esta análise deve considerar o tipo de sinal ao qual os esquemas são aplicados, a granularidade, o desempenho de detecção, a robustez contra distorções e o método de busca empregado. Em geral, a aplicação principal dos esquemas propostos fornece uma idéia acerca dos requisitos de granularidade, do desempenho e da robustez. A maior parte dos sistemas analisados foram propostos para a identificação de música por conteúdo, e, portanto, empregam granularidades longas.

Em [222] o desempenho dos sistemas propostos por [77, 223, 50] são avaliados para um conjunto de teste gerado pela captura de áudio de 39 músicas a partir do áudio recebido em um telefone celular. A curva ROC é gerada para cada modelo e o modelo identificado como mais robusto é aquele proposto por [77].

Na análise detalhada dos esquemas de *Audio Fingerprinting* a seguir, os métodos são divididos em grupos como sugerido em [24], conforme diferenças mais significativas das abordagens.

B.1 GRUPO 1: SISTEMAS QUE USAM ATRIBUTOS EXTRAÍDOS DE MÚLTIPLAS SUB-BANDAS

O mapeamento dos coeficientes de Fourier pela aplicação de bancos de filtros seguido do cálculo de alguma estatística reduz bastante a dimensionalidade, uma vez que o número de sub-bandas é bem menor que o comprimento da DFT. O sistema mais citado deste grupo é o proposto pela PHILIPS por Haitsma [2], descrito detalhadamente na Seção 2.4.1.

O sistema proposto em [63, 64] emprega descritores de áudio de baixo nível (LLD-*Low Level Descriptors*) previstos no padrão MPEG7, que fornece vários conceitos e elementos de descrição de áudio. O esquema emprega a sonoridade (*Loudness*), que estima a percepção humana da intensidade do som, além de medidas como SFM e SCM, referentes à relação entre componentes harmônicos e o ruído do áudio, também relacionados à harmonicidade do sinal. A definição matemática destes atributos é feita em [30].

De acordo com o padrão MPEG-7, o áudio é segmentado em quadros de 30ms. Não há descrição do emprego de sobreposição de quadros. Há uma flexibilidade na escolha da banda de frequência do sinal, com um intervalo típico de 250 a 4000Hz. A DFT é obtida, e a banda é dividida em 16 sub-bandas, com uma escala logarítmica. Não há informação acerca da taxa de amostragem do sinal. Se uma taxa de 8khz for empregada, cada quadro de 30ms conterá 240 amostras, o que corresponde a uma baixa resolução em frequência de 33Hz. Portanto, a divisão em 16 sub-bandas para posterior agrupamento de coeficientes e extração da medidas de SFM pode requerer uma amostragem do sinal com taxa superior a 8kHz. Os atributos são calculados para cada sub-banda, e os vetores obtidos são concatenados em grupos de 32 quadros, com uma granularidade final de 960ms. Outras granularidades menores são testadas. O esquema emprega um método de busca baseado em quantização vetorial (VQ/*Nearest Neighboor*). Os testes para áudio sem distorção fornecem uma taxa de detecção de 99,97%. Cabe ressaltar que ao contrário da maioria dos métodos estudados, que detectam 100% do áudio não distorcido, este método não consegue uma detecção plena. Isto pode ser atribuído ao desalinhamento entre o quadro da AF questionada e da AF rotulada, já que o método não utiliza sobreposição de quadros. Não são fornecidos resultados de testes de robustez contra inserção de ruído ou de compressão severa.

O Centróide Espectral (*Spectral Centroid*-SC) é um descritor de áudio, que indica o "centro de massa"do espectro de uma sub-banda, e é usado em esquemas de AF, como em [98, 95]. A definição matemática do SC é feita em [30]. Em [98] são usados atributos SBE, SC, SFM e SCM, além das entropias de Renyi e Shannon, para codificar o AF. A taxa média de detecção para diversos tipos de distorção é acima de 99% para trechos de áudio de 5s, mas para trechos de 2,5s a taxa de detecção foi praticamente nula.

B.2 GRUPO 2: SISTEMAS QUE EMPREGAM ATRIBUTOS EXTRAÍDOS DE UMA ÚNICA BANDA DE FREQUÊNCIA

Alguns esquemas usam métodos de análise de imagem para a análise da informação bidimensional de amplitude da STFT, como no esquema bastante robusto proposto em [50], usado no aplicativo Shazam, de identificação de música por conteúdo. Neste método são detectados picos de energia espectrais (*Spectral Energy Peaks*- SEP) da STFT, onde um ponto no espaço tempo-frequência é considerado um pico se tiver amplitude superior a seus vizinhos. Esta representação por picos locais é denominada de *mapa de constelação*. Não são fornecidos parâmetros de tamanho de janela ou sobreposição no cálculo da STFT. A representação obtida é reduzida a um conjunto esparso de coordenadas, e não guarda informação sobre amplitude, o que a torna robusta contra equalização. Alguns picos locais são escolhidos como âncoras e os demais picos dentro de uma zona de influência do âncora (*Target Zone*). Os picos são combinados em pares de coordenadas tempo-frequência, representados pela defasagem temporal entre os picos locais e em relação ao início do áudio. Esta representação, por pares de picos, reduz enormemente a complexidade do algoritmo de busca, permitindo a sua execução em aparelhos celulares. Cabe destacar que a codificação da AF a partir de duplas de picos locais produz uma representação menos localizada no tempo, ou seja com maior granularidade, que os sistemas que empregam apenas um máximo local na codificação da AF.

O desempenho deste método é medido para identificar músicas, com granularidades de 15s, 10s e 5s, corrompidas com ruído aditivo e compressão GSM.Para estas granularidades, conclui-se que o esquema baseado em SEP é bastante robusto contra inserção de ruído. Entretanto, este método não é capaz de identificar áudio modificados na escala temporal.

Esta abordagem é aplicada também em [12]. Em [7] é proposto um esquema para detecção de réplicas de eventos de áudio em gravações de longa duração, para aplicações como a análise de registros de áudio do dia-a-dia de um indivíduo. A busca é feita com uma tabela hash, e o número de quadros de áudio equivalentes em uma janela de 2s é usado como critério de detecção. Portanto, o método não seria capaz de detectar trechos de áudio com duração inferior a 2s. Ademais, o desempenho relatado é bom apenas para eventos de áudio estruturado com componentes espectrais bem definidos, como música ou tons de discagem de telefonia. Para trechos de áudio contendo voz, onde picos locais de energia espectral são menos proeminentes, o desempenho do método foi muito baixo.

Em [51] é aplicado o método de análise de imagem, SIFT (Transformação Invariante a Escala- *Scale Invariant Feature Transform*), sobre a imagem 2D obtida da STFT, com o objetivo de aumentar a robustez contra distorções de escala temporal. O cálculo da STFT emprega 97 bandas de frequência, com limites definidos por uma escala logarítmica. Os quadros possuem 2048 amostras, com um sobreposição de 50%. Os descritores SIFT são obtidos a partir de picos de energia com informação de gradientes locais da imagem,

gerando um vetor de dimensão 128 para cada descritor. Portanto, o atributo usado possui uma melhor localização temporal que o esquema proposto em [50], que usa pares de picos. Os testes empregam amostras de 60s de duração amostradas a 44kHz. A granularidade usada nos testes é de 10s. Nos testes de desempenho o método detecta 100% do áudio não distorcido ou escalado em até 20%, e detecta 94% do áudio com ruído adicionado a uma SNR=18dB. A taxa de Falso Positivo não é analisada.

Em [87] o método SURF (*Speed Up Robust Features*), que consiste em uma modificação do método SIFT mais rápida, é aplicado ao cocleograma. Os testes mostram uma elevada robustez contra variações de escala temporal, compressão e inserção de ruído.

Em [223] o espectrograma do áudio é tratado como uma superimposição de imagens, e a extração dos atributos de AF é baseada no algoritmo Viola Jones, comumente aplicado para a identificação de face.

Em [93] é proposto um método que usa a modulação de frequência como atributo, mas os testes de robustez não abordam a compressão de áudio. O método proposto por [62] usa a SFM e SCF como atributos para representar o áudio, e a robustez do método é testada inclusive contra a compressão de áudio.

O método proposto pela Google [77, 76] codifica o sinal usando a DWT, identificando os picos de energia espectral, através dos maiores coeficientes do domínio Wavelet, com uma abordagem robusta, mas com elevada complexidade computacional.

B.3 GRUPO 3: SISTEMAS OTIMIZADOS POR TREINAMENTO, COM DIVISÃO POR QUADROS E POR SUB-BANDAS

O sistema AudioDNA [88] usa uma abordagem semelhante à empregada em reconhecimento de voz. O áudio é representado pela concatenação de classes de eventos de som, representados por um alfabeto finito de símbolos, como empregado na representação de fonemas. O sinal é segmentado em quadros de curta duração, tipicamente de 25ms. Com um espaçamento de 10ms entre quadros, uma sobreposição de 60% é usada. O esquema emprega coeficientes Cepstrais, mas não informa o número de coeficientes usado. As classes são estimadas através de um método de agrupamento não supervisionado e modelado (HMM). A representação por sequências de estados captura a informação da evolução temporal do áudio. Na busca é empregado um método de busca aproximada de sequências (*strings*) para identificar o áudio. A similaridade é medida por um método baseado no algoritmo de Viterbi, com uma complexidade acima da média de outros sistemas. Nos testes de detecção é empregada uma granularidade mínima de 6s. O trabalho não apresenta muitas simulações sobre a taxa de detecção, e a única distorção simulada para testar robustez é a compressão de áudio. A taxa de detecção obtida para a compressão MP3 a

24kbps, a menor taxa de bits testada, foi de 84% para uma taxa de Falso Positivo abaixo de 0,5%.

O esquema do sistema RARE (*Robust Audio Recognition Engine*), proposto pela Microsoft, com base em uma Análise Discriminante de Distorção (DDA - *Distortion Discriminant Analysis*)[35, 36], usa sinais amostrados a 11.025Hz. O sinal é pré-processado. A DCT é aplicada para descorrelacionar o sinal e um limiar de potência perceptual, que simula a percepção em dB, é aplicado. O esquema calcula o logaritmo do módulo da MCLT obtida de quadros com duração de 372ms, com 2048 amostras, sobrepostos em 50%. Filtros perceptuais são aplicados para descarte de componentes mascaradas. Redes neurais convolucionais são aplicadas, onde camadas superiores cobrem janelas temporais mais longas, para reduzir o efeito do desalinhamento de quadros. Uma Análise Orientada de Componentes Principais (OPCA) é aplicada para descorrelacionar os vetores em cada camada. Um vetor de atributos de dimensão 64 é obtido para cada quadro. Na fase de treinamento, versões distorcidas ou deslocadas de um quarto da duração dos quadros são usadas para aumentar a robustez contra distorção ou contra o desalinhamento de quadros. A granularidade total empregada é de 6s. O desempenho de detecção para áudio não distorcido foi 98% de detecção correta a uma taxa de falso positivo de $1,5 \times 10^{-8}$. Este esquema é também considerado robusto contra inserção de ruído e compressão de áudio, com taxa média de EER (*Equal Error Rate*) em torno de 2%.

C- CÁLCULO DE DESEMPENHO DO INTEGRADOR BINÁRIO COM JANELA MÓVEL

Inicialmente, formulamos o critério de detecção por integração binária com janela móvel mantendo a simbologia do Capítulo 4. O critério condiciona a detecção de uma sequência de N_R bits com distribuição de Bernoulli de probabilidade de bits não-nulos, P_Q, à existência de no mínimo N_Q elementos não-nulos em quaisquer janelas móveis de comprimento N_J dentro da sequência R. Ressaltamos que na detecção de réplicas o critério é aplicado em janelas diagonais da matriz de autossimilaridade de quadros $\mathbf{M}$.

Este critério é também aplicado em detecção de RADAR [162], onde o método é referido como integrador binário com janela móvel. Em uma pesquisa da literatura científica não se identificou nenhum método exato de cálculo da probabilidade de detecção pelo critério do integrador binário. Em [162] é proposta uma solução aproximada, baseada em um cálculo diferencial, onde a interdependência de alguns termos é aproximada para zero. O método apresentado apenas desenvolve as equações para o caso $N_Q = (N_J - 1)$. Ademais, o autor salienta que para valores de N_Q distantes de N_J o método não fornece uma boa aproximação.

O cálculo da probabilidade discreta, pela verificação da condição de detecção e cálculo da probabilidade de todas as sequências binárias de N_R bits é inviável para valores altos de N_R. Propomos, então, na Seção C.1 um algoritmo de cálculo exato com uso de uma função recursiva, cuja complexidade computacional é linearmente proporcional a N_R, mas o uso de memória é proporcional a 2^{N_J}, portanto somente pode ser usado para valores baixos de N_J. Na Seção C.2 propomos um novo método de cálculo com baixa complexidade computacional que fornece uma boa aproximação, mesmo para valores de N_Q distantes de N_J.

C.1 MÉTODO DE CÁLCULO EXATO

O algoritmo proposto calcula a probabilidade discreta através da definição de uma função recursiva $Precursiva(N_R, N_J, N_Q, P_Q)$, onde a probabilidade de detecção de réplica P_R é calculada pela soma das probabilidades discretas de todas as sequências detectáveis de N_R bits. O processo de recursão é feito passo a passo, pela análise dos bits nas posições $i = 1, 2, ...N_R - N_J$, da esquerda para a direita, dividindo o conjunto de sequência em dois grupos disjuntos, como ilustrado na Figura C.1. Para permitir a verificação da condição de detecção os bits à esquerda do i-ésimo bit são armazenados em um vetor $\mathbf{RE}$, cujo índice de posições de bits é definido da direita para a esquerda. A cada deslocamento

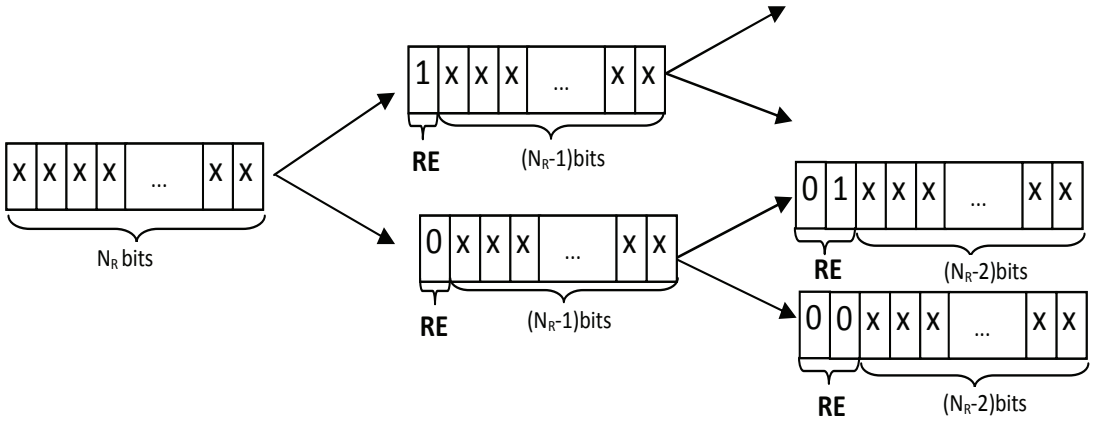

Figura C.1: Divisão do grupo de eventos discretos em dois grupos disjuntos, analisando os bits da esquerda para a direita e definindo uma variável de estado RE.

da posição analisada à direita, o vetor **RE** é também deslocado à esquerda. Seja $\mathbf{RE} = [RE(N_E), ..., RE(2), RE(1)]$, onde N_E é o comprimento de RE, definimos a operação de deslocamento à esquerda $shift(\mathbf{RE}, b) \leftarrow [RE(N_E), ..., RE(2), RE(1), b]$, para $b = \{0, 1\}$. Dessa forma, **RE** é também um argumento da função recursiva.

A probabilidade discreta de cada sequência RE é dada por:

$$P_{RE} = P_Q^{\sum_{i=1}^{N_E} RE(i)} + (1 - P_Q)^{\sum_{i=1}^{N_E}(1-RE(i))}. \tag{C.1}$$

A recursão é interrompida e a probabilidade é definida para os casos:

$$P_R(j, N_J, N_Q, P_Q, \mathbf{RE}) \leftarrow$$
$$= \begin{cases} P_{RE}\left[\sum_{i=N_Q}^{j} \binom{j}{i} P_Q^i (1-P_Q)^{j-i}\right], \text{ se } \sum_{k=1}^{N_J} RE(k) = 0, N_Q \leq j \leq N_J, \\ P_{RE}, \text{ se } \sum_{i=1}^{N_J} RE(i) = N_Q, \\ 0, \text{se } (j < N_J) \wedge (j + \sum_{i=1}^{N_J} RE(i) < N_Q). \end{cases} \tag{C.2}$$

Caso nenhuma das condições acima ocorra, o conjunto das sequências iniciadas por **RE** é sub-dividido novamente, pela chamada da função 2 vezes, conforme

$$P_R(N_R, N_J, N_Q, P_Q, \mathbf{RE}) \leftarrow$$
$$P_R(N_R - 1, N_J, N_Q, P_Q, shift(\mathbf{RE}, 0)) + P_R(N_R - 1, N_J, N_Q, P_Q, shift(\mathbf{RE}, 1)). \tag{C.3}$$

Como para a verificação da condição de detecção apenas os primeiros N_J bits de **RE** são relevantes, podemos otimizar o método evitando a repetição de cálculo de probabilida-

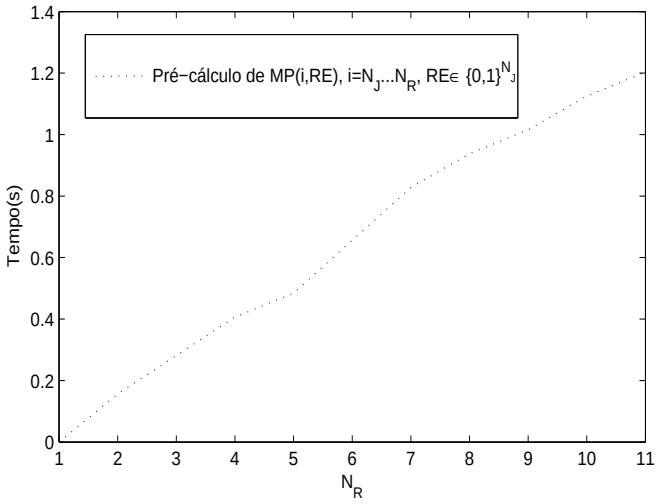

Figura C.2: Tempo de execução do método exato, com o pré-calculo de $\mathbf{MP}(j, \mathbf{RE})$, para quaisquer combinações dos primeiros N_J bits de $\mathbf{RE}$.

des intermediárias armazenando a variável de estado $\mathbf{MP}(i, \mathbf{RE}) = P_R(i, N_J, N_Q, P_Q, \mathbf{RE})$ para quaisquer combinações dos primeiros N_J bits de $\mathbf{RE}$, e para $i = 1, 2, ..., N_R$. Para evitar a repetição do cálculo, na implementação definimos uma matriz booleana $\mathbf{IMP}(i, \mathbf{RE})$ que indica se $\mathbf{MP}(i, \mathbf{RE})$ foi previamente calculado. As variáveis **IMP** e **MP** são usadas como argumentos e como resultados:

$$[P, MP, IMP] \leftarrow Precursiva(N_R, N_J, N_Q, p, RE, MP, IMP). \tag{C.4}$$

O número total de recursões é reduzido ainda mais se o pré-cálculo de $\mathbf{MP}(i, \mathbf{RE})$ for otimizado chamando-se a função $Precursiva(i, N_J, N_Q, P_Q, \mathbf{RE})$, de $i = N_J$ até $i = N_R$. Dessa forma, garantimos que, ao chamar $Precursiva(i, N_J, N_Q, P_Q, \mathbf{RE})$, $\mathbf{MP}(j, \mathbf{RE})$ é previamente calculado para todos os valores de $j < i$, evitando recursões longas que elevariam o requisito de memória.

Com estes artifícios de implementação, é possível calcular esta probabilidade com um tempo de execução aproximadamente linear com N_R, como mostra a Figura C.2. Entretanto, o uso de memória é proporcional a $N_R 2^{N_J}$. Logo, o método exato proposto resolve o problema do aumento de complexidade com N_R, mas o cálculo não é viável para valores elevados de N_J.

O pseudocódigo abaixo descreve o algoritmo de cálculo exato:

$IMP \leftarrow logical(zeros(N_R, N_J, N_J))$

$MP \leftarrow double(zeros(N_R, N_J, N_J))$

$P \leftarrow double(zeros(N_R))$

for $i = N_J$ to N_R

$[P(i), MP, IMP] \leftarrow Precursiva(i, N_J, N_Q, p, RE, MP, IMP)$

end for

Dessa forma, **MP** é pré-calculado passo a passo de forma eficiente, e a probabilidade final é dada por $P(N_R)$.

A função $Precursiva(N_R, N_J, N_Q, p, RE, MP, IMP)$ é descrita pelo pseudocódigo abaixo:

Function $[P, MP, IMP] \leftarrow Precursiva(N_R, N_J, N_Q, p, RE, MP, IMP)$

if $(IMP(N_R, RE)) = 1$

$P = MP(N_R, RE)$

else

$NE \leftarrow$ **size**(RE)

$SE \leftarrow$ **sum**(RE)

$SEJ \leftarrow$ **sum**(RE(1:J))

$PE \leftarrow p^S + (1-p)^{(NE-S)}$

if $(SEJ = 0).(R \geq Q).(R \leq J)$

$P \leftarrow 0$

for $i = Q$ to R

$$P \leftarrow P + PE\left[\begin{pmatrix} R \\ i \end{pmatrix} p^i (1-p)^{R-i}\right]$$

end for

$MP(R, RE) = P$; $IMP(R, RE) = 1$

else if $SEJ = Q$

$MP(R, RE) = PE$; $IMP(R, RE) = 1$

else if $(R + SEJ < Q).(R < J)$

$MP(R, RE) = 0$; $IMP(R, RE) = 1$

else

$[P1, MP, IMP] \leftarrow Precursiva(N_R - 1, N_J, N_Q, p, shift(RE, 1), MP, IMP)$

$[P0, MP, IMP] \leftarrow Precursiva(N_R - 1, N_J, N_Q, p, shift(RE, 0), MP, IMP)$

$P = P1 + P0$

end if

end function

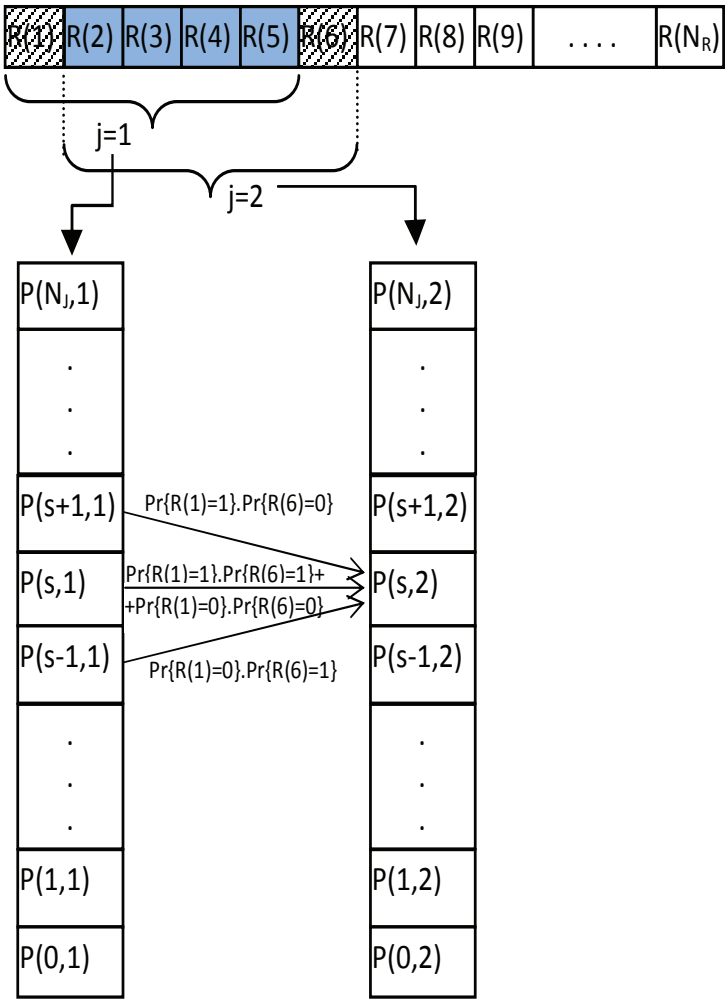

Figura C.3: Janela móvel de comprimento $N_J = 5$, deslocadas à direita sobre uma sequência R de comprimento N_R, e cálculo de $P(s, j)$ a partir de $P(s-1, j-1)$, $P(s, j-1)$ e $P(s+1, j-1)$.

C.2 MÉTODO DE CÁLCULO APROXIMADO

Propomos um cálculo aproximado da probabilidade de detecção, $P_R(i, N_J, N_Q, P_Q)$, onde as probabilidades de detecção inicial em cada posição de janela são somadas de uma forma que os conjuntos sejam disjuntos.

Usamos a análise da variação das somas dos bits de cada janela móvel de comprimento N_J ao longo da sequência R. As janelas são deslocadas à direita sobre uma sequência R de comprimento N_R, da posição $j = 1$ a $j = N_R - N_J$, como ilustra a Figura C.3 para o caso $N_J = 5$.

A probabilidade de soma dos bits de cada janela j é definida por:

$$P(s,j) = Pr\{\sum_{i=j}^{j+N_J} R(i) = s\}. \tag{C.5}$$

As somas podem assumir valores entre 0 e N_J. As probabilidades $P(s,1)$, da primeira janela, são obtidas pela probabilidade de combinação

$$P(s,1) = \begin{pmatrix} N_J \\ s \end{pmatrix} P_Q^s (1-P_Q)^{N_J - s} \tag{C.6}$$

As probabilidades $P(s,j)$ das janelas subsequentes são então calculadas, sempre a partir das probabilidades $P(s-1,j-1)$, $P(s,j-1)$ e $P(s+1,j-1)$, da janela anterior, dependendo somente das possibilidades dos bits $R(j-1)$ e $R(j+N_J)$, já que os demais bits são comuns às janelas $j-1$ e j, e, portanto, não alteram a soma de bits.

Caso $R(j-1) = R(j+N_J)$, as somas dos bits das janelas $j-1$ e j se mantêm. Caso $R(j-1) = 0$, $R(j) = 1$, a soma dos bits da janela s é incrementada, e caso $R(j-1) = 1$, $R(j) = 0$, a soma dos bits da janela s é decrementada, como ilustra a Figura C.3. Definimos $P_1(i) = Pr\{R(i) = 1\}$, e variando $s = 2$ a $s = N_R - N_J$, calculamos

$$\begin{aligned} &P(s,j) = \\ &= P(s,j-1)[P_1(j+N_J)P_1(j-1)) + (1-P_1(j+N_J))(1-P_1(j-1))] + \\ &+P(s+1,j-1)P_1(j-1)(1-P_1(j+N_J)) + \\ &+P(s-1,j-1)(1-P_1(j-1))P_1(j+N_J), s = 0,1,..,N_J. \end{aligned} \tag{C.7}$$

A probabilidade de detecção em cada janela é dada pela somas das probabilidades $P(s,j)$, para $s \geq N_Q$:

$$P_{det}(j) = \sum_{s=N_Q}^{N_J} P(s,j). \tag{C.8}$$

A probabilidade de detecção, calculada como a soma de eventos disjuntos, é obtida somando-se as probabilidades de detecção inicial (não detecção nas janelas anteriores) em cada uma das janelas de 1 a $N_R - N_J$:

$$\begin{aligned} &Pr\{\delta_R(R) = 1\} = \\ &= P_{det}(1) + \overline{P_{det}}(1)P_{det}(2) + ... + \overline{P_{det}}(1)\overline{P_{det}}(2)...\overline{P_{det}}(N_R - N_J - 1)P_{det}(N_R - N_J). \end{aligned} \tag{C.9}$$

Para viabilizar o cálculo das probabilidades pela soma de eventos disjuntos acima, a cada avanço na posição da janela, zeramos as probabilidades $P(s,j) \leftarrow 0, s = Q, Q+1, .., N_J$, o que corresponde ao descarte de todas as possíveis sequências com mais de $Q-1$ bits no intervalo $i = j, j+1, ..., j+N_J$, mantendo somente as sequências correspondentes à não detecção até a janela j. Entretanto, este descarte afeta a distribuição de $P_1(i) = Pr\{R(i) = 1\}, i = j, j+1, ..., j+N_J$.

Para atualizar as probabilidades $P_1(i) = Pr\{R(i) = 1\}, i = j, j+1, ..., j+N_J$ a cada incremento de j, conjecturamos que os bits não-nulos, 1's, são uniformemente distribuídos entre todas as sequências binárias no intervalo $i = j, j+1, ..., j+N_J$, independente da posição. Dessa forma, a redução de $P_1(i), i = j, j+1, ..., j+N_J$ com o descarte de uma sequência seria proporcional ao número de 1's contidos neste intervalo. Para atualizar $P_1(i), i = j, j+1, ..., j+N_J$, calculamos, portanto, a razão R_1 entre o número de 1's contidos em todas as sequências e o número de sequências antes do descarte, e a mesma razão após o descarte, R_2. O pseudocódigo abaixo descreve o algoritmo aproximado:

for $i = 1$ to N_R

$P_1(i) \leftarrow P_Q$

end for

for $j = 1$ to $N_R - N_J$

for $s = 1$ to N_J

$$P(s,j) \leftarrow \begin{array}{l} P(s,j-1)P_1(j+N_J)P_1(j-1)) + P_1(j+N_J)P_1(j-1) + \\ +P(s+1,j-1)P_1(j-1)P_1(j+N_J) + P(s-11,j-1)P_1(j-1)P_1(j+N_J) \end{array}$$

end for

$P_{det}(j) \leftarrow \sum_{s=Q}^{N_J} P(s,j)$;

$P_T \leftarrow P_T + P_{det}(j)$;

for $s = Q$ to N_J

$P(s,j) \leftarrow 0$;

end for

$R_1 \leftarrow \frac{\sum_{i=0}^{J} P(i,j)i}{\sum_{i=0}^{J} P(i,j)}$

$R_2 \leftarrow \frac{\sum_{i=0}^{Q-1} P(i,j)i}{\sum_{i=0}^{Q-1} P(i,j)}$

for $k = j$ to $j + N_J$

$P_1(k) \leftarrow P_1(k)R_2/R_1$

end for

end for

Ao final P_T corresponde à probabilidade total de detecção, que no caso da aplicação de detecção de réplica é definida por $Pr\{\delta_R(R) = 1\}$.

Uma análise do comportamento de $P_1(i) = Pr\{R(i) = 1\}$ pela geração de todas as sequências R, para valores baixos de N_J e N_R, mostra que em alguns casos a redução relativa de $P_1(i) = Pr\{R(i) = 1\}$ é maior para as últimas posições de bits de cada janela. Logo a conjectura empregada é falsa, e o cálculo é, portanto, uma aproximação.

Entretanto, observou-se que este método aproximado oferece uma boa estimativa, mesmo para valores de N_Q distantes de N_J. A Figura C.4 ilustra a probabilidades obtidas pelos métodos de cálculo exato e aproximado propostos, para $N_J = 16$ e $N_Q = 4, 8, 12$.

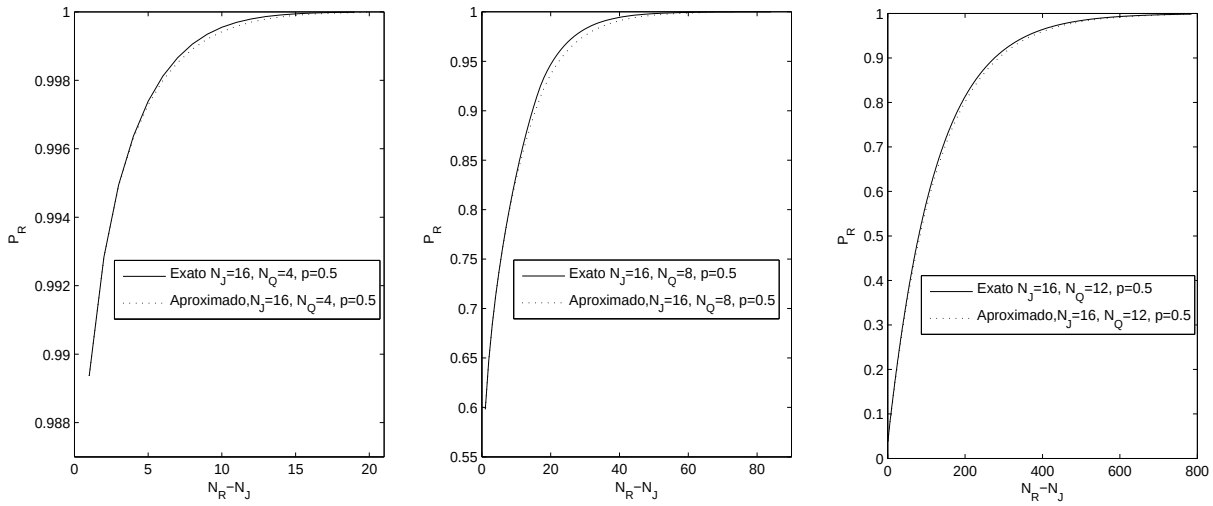

Figura C.4: Curvas de detecção de integração binária para uma distribuição Bernoulli com $p = 0,5$, $N_J = 16$, $N_Q = 4$ (esquerda), $N_Q = 8$ (centro) e $N_Q = 12$ (direita), usando os métodos exato (linha sólida) e aproximado (linha pontilhada) propostos.

D- PUBLICAÇÕES RELEVANTES PELO AUTOR

Neste apêndice, listamos as publicações feitas pelo autor, referentes ao trabalho desta tese.

D.1 ARTIGO EM PERIÓDICO

1. Távora, R.; Nascimento, F. A. *Detecting replicas within audio evidence using an adaptive audio fingerprinting scheme.* J. Audio Engineering Society. Vol. 63, 2015, pp. 451-462

D.2 RESUMO COMPLETO EM CONFERÊNCIA INTERNACIONAL

1. Távora, R.; Nascimento, F. A. *Detecting replicas within audio evidence using an adaptive audio fingerprinting scheme.* IAFS Conference, Seul, Coréia, Out. 12-18, 2014.

D.3 PREMIAÇÕES

1. Travel Award: IAFS Conference 2014, Seul, Coréia, Out. 12-18.