UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

DETECÇÃO ROBUSTA DE ADULTERAÇÃO EM ÁUDIO EXPLORANDO A FORMA ANALÍTICA E O SUBESPAÇO DE SINAIS INTERFERENTES DA REDE ELÉTRICA

PAULO MAX GIL INNOCENCIO REIS

ORIENTADOR: JOÃO PAULO CARVALHO LUSTOSA DA COSTA

DISSERTAÇÃO DE MESTRADO EM
ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPGEE.DM - 627 A/16

BRASÍLIA/DF: JULHO/2016.

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

DETECÇÃO ROBUSTA DE ADULTERAÇÃO EM ÁUDIO EXPLORANDO A FORMA ANALÍTICA E O SUBESPAÇO DE SINAIS INTERFERENTES DA REDE ELÉTRICA

PAULO MAX GIL INNOCENCIO REIS

DISSERTAÇÃO DE MESTRADO 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 MESTRE.

APROVADA POR:

JOÃO PAULO CARVALHO LUSTOSA DA COSTA, Dr., ENE/UNB
(ORIENTADOR)

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

JOSÉ ANTÔNIO APOLINÁRIO JÚNIOR, Dr., IME/RJ
(EXAMINADOR EXTERNO)

Brasília, 05 de julho de 2016.

FICHA CATALOGRÁFICA

REIS, PAULO MAX GIL INNOCENCIO

Detecção robusta de adulteração em áudio explorando a forma analítica e o subespaço do sinais interferentes da rede elétrica [Distrito Federal] 2016.

xv, 74 p., 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2016).

Dissertação de Mestrado - Universidade de Brasília.

Faculdade de Tecnologia. Departamento de Engenharia Elétrica.

1. Detecção de adulterações em áudio
2. Análise forense de áudio
3. ESPRIT
4. ENF

I. ENE/FT/UnB

II. Título (série)

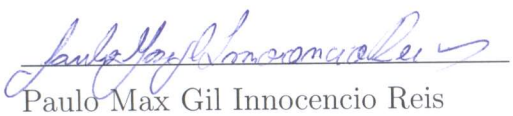
REFERÊNCIA BIBLIOGRÁFICA Reis, P. M. G. I. (2016). Detecção robusta de adulteração em áudio explorando a forma analítica e o subespaço do sinais interferentes da rede elétrica. Dissertação de Mestrado em Engenharia Elétrica, Publicação PPGEE.DM - 627 A/16, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 74p.

CESSÃO DE DIREITOS

NOME DO AUTOR: Paulo Max Gil Innocencio Reis.

TÍTULO DA DISSERTAÇÃO DE MESTRADO: Detecção robusta de adulteração em áudio explorando a forma analítica e o subespaço do sinais interferentes da rede elétrica.

GRAU / ANO: Mestre / 2016



Paulo Max Gil Innocencio Reis

SPO, Quadra 7, Lote 23, INC/SEPAEL - Sala A-209, Asa Sul

70.610-200 Brasília - DF - Brasil.

DEDICATÓRIA

Dedico este trabalho aos meus amados pais, Max (in memorian) e Zezé, que muito me apoiaram durante toda a vida para minha formação.

AGRADECIMENTOS

Agradeço primeiramente a Deus por permitir a conclusão de mais essa etapa. Agradeço à minha amada namorada, pela paciência, suporte e companhia. Agradeço à minha família pelo apoio incondicional, ainda que à distância. Agradeço ao meu orientador, Prof. João Paulo Carvalho Lustosa da Costa, pelas sugestões e correções que tanto contribuíram para esse trabalho e pela confiança depositada, e ao seu aluno Ricardo Kherle Miranda pelas sugestões e contribuições na revisão do trabalho. Agradeço ao Prof. José Antonio Apolinário Junior pela cessão da base de dados Carioca 1, sem a qual não poderia ter atingido os resultados obtidos. Por último agradeço ao Instituto Nacional de Criminalística, e em especial ao Perito Criminal Federal Getúlio Menezes Bento, cujo apoio na chefia do SEPAEL foi crucial para a conclusão deste trabalho.

A todos meus sinceros agradecimentos.

O presente trabalho foi realizado com o apoio do Departamento Polícia Federal - DPF, com recursos do Programa Nacional de Segurança Pública com Cidadania - PRONASCI, do Ministério da Justiça

RESUMO

DETECÇÃO ROBUSTA DE ADULTERAÇÃO EM ÁUDIO EXPLORANDO A FORMA ANALÍTICA E O SUBESPAÇO DE SINAIS INTERFERENTES DA REDE ELÉTRICA

Autor: Paulo Max Gil Innocencio Reis
Orientador: João Paulo Carvalho Lustosa da Costa
Programa de Pós-graduação em Engenharia Elétrica
Brasília, Julho de 2016

Arquivos de áudio digital são uma importante fonte de vestígios e evidências relacionadas aos mais diversos crimes e conflitos. Seja por meio de gravações devidamente autorizadas pela autoridade judicial ou por gravações realizadas por um dos interlocutores em um diálogo, tais arquivos têm o potencial de serem determinantes em importantes decisões, uma vez que prestam-se a, via de regra, esclarecer algum aspecto da realidade dos fatos. Dessa forma, a autenticação dessa fonte de prova é uma tarefa muitas vezes necessária e crítica, porém ainda sujeita a muitos desafios.

Com o objetivo de identificar edições em arquivos de áudio propõe-se uma técnica para detecção automática de adulterações em gravações de áudio por meio da constatação de variações anormais na frequência de oscilação de sinais interferentes da rede elétrica (ENF), eventualmente incorporados em um registro de áudio questionado. Variações anormais na ENF podem ocorrer como resultado de transições abruptas de fase decorrentes de inserções ou supressões de segmentos de áudio realizados durante o processo de edição. Dessa forma, propõe-se o estimador de ENF ESPRIT-Hilbert em conjunto com um detector de *outliers* baseado na curtose amostral da estimada ENF, do inglês *ESPRIT-Hilbert ENF estimator in conjunction with an outlier detector based on the sample kurtosis of the estimated ENF* (SPHINS). A técnica utiliza conjuntamente um estimador baseado na frequência instantânea obtida via transformada de Hilbert, e outro baseado na técnica ESPRIT. Calcula-se a curtose amostral das estimativas da ENF como medida do grau de anomalia da ENF, compondo-se um vetor de características que é aplicado a um classificador de máquinas de vetores de suporte (SVM), devidamente treinado a partir de uma base de dados conhecida para indicar a presença de edições. O método proposto tem seus resultados validados utilizando uma base de dados que contém 100 gravações telefônicas autorizadas de áudios não editados, e 100 gravações telefônicas de áudios editados. Os resultados obtidos são comparados com trabalhos correlatos anteriores.

ABSTRACT

AUDIO TAMPERING ROBUST DETECTION EXPLOITING THE ANALYTICAL FORM AND SIGNAL SUBSPACE OF ELECTRIC NETWORK INTERFERING SIGNALS

Author: Paulo Max Gil Innocencio Reis
Supervisor: João Paulo Carvalho Lustosa da Costa
Programa de Pós-graduação em Engenharia Elétrica
Brasília, Julho of 2016

Digital audio recordings are an important source of evidences related to various crimes and conflicts. Whether through recordings duly authorized by a judicial authority or made by one of the parties in a dialogue, such files have the potential to be crucial in important decisions since they contribute to clarify some aspects of reality. Thus, the authentication of this source of evidence is often a necessary and critical task, but still subject to many challenges.

In order to identify audio tampering we propose a technique to detect adulterations in audio recordings by exploiting abnormal variations in the Electric Network Frequency (ENF) signal eventually embedded in a questioned audio recording. These abnormal variations may be caused by abrupt phase discontinuities due to insertions and suppressions of audio segments during the tampering task. Thus, we propose the ESPRIT-Hilbert ENF estimator in conjunction with an outlier detection based on the sample kurtosis of the estimated ENF (SPHINS). The technique uses a joint estimate of ENF by two methods, one based in the Hilbert Transform, and the other in the ESPRIT approach. It calculates the sample kurtosis of the estimates as a measure of outlierness, computing a feature vector applied to a Support Vector Machine (SVM) classifier to indicate the presence of tampering. The proposed scheme is validated using an audio database with 100 edited and 100 unedited authorized audio recordings of phone calls. The results obtained are further compared with previous related works.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SÍMBOLOS, NOMENCLATURA E ABREVIAÇÕES

3E: Esprit-based ENF Estimator

DET: Detection Error Trade-off

DFT: Discrete Fourier Transform

DHT: Discrete Hilbert Transform

EER: Equal Error Rate

ENF: Electric Network Frequency

ESPRIT: Estimation of Signal Parameters via Rotational Invariance Techniques

FIR: Finite Impulse Response

FNR: False Negative Rate

FPR: False Positive Rate

FFT: Fast Fourier Transform

HEE: Hilbert-based ENF Estimator

MDCT: Modified Discrete Cossine Transform

MP3: MPEG-1/2 Audio Layer 3

MUSIC: Multiple Signal Classification

OER: Overall Error Rate

PLL: Phase Locked Loop

RBF: Radial Basis Function

RTPC: Rede de Telefonia Pública Comutada

SL: Saturation Level

SNR: Signal-to-noise Ratio

SVD: Singular Value Decomposition

SVM: Suport Vector Machine

VAD: Voice Active Detector

VCO: Voltage-Controlled Oscilator

1 INTRODUÇÃO

É particularmente evidente que nas últimas décadas houve um notável crescimento na utilização de dispositivos destinados a realizar a gravação e reprodução de conteúdo audiovisual. Tal fato acarreta uma série de novos hábitos e costumes cotidianos que se relacionam com praticamente todos os aspectos da vida moderna (GUPTA; CHO; KUO, 2012; BÖHME et al., 2009). Um dos aspectos relevantes do convívio social diz respeito a forma com que os costumes influenciam a relação entre as pessoas. Com isso, esse novos hábitos e costumes geram uma série de fatores ecológicos e ambientais que influenciam na determinação de fenômenos sociais, tais como o cometimento de crimes e as maneiras com que a sociedade reprime tais delitos (MCKENZIE, 1924; HAWLEY, 1944).

Dessa forma, com o aumento exponencial no acesso a tecnologias sofisticadas de comunicação, as evidências em multimídia passaram a possuir um relevante papel na persecução penal. Verifica-se uma considerável elevação na casuística relacionada à apresentação de gravações de áudio como prova do cometimento dos mais variados crimes (BÖHME et al., 2009). A despeito de controversos posicionamentos jurídicos existentes na doutrina brasileira, gravações oriundas de dispositivos de escuta telefônica e ambiental são frequentemente apontadas como a principal, senão única, prova acerca da materialidade e autoria de um determinado fato delituoso (RODRIGUES; BERNACCHI; WIESELTHALER, 2016; FANTINI, 2013).

Como consequência direta do desenvolvimento tecnológico, o declínio constante do custo de produção de dispositivos de codificação e processamento digital de sinais facilita o acesso a gravadores de áudio. Entretanto, as mesmas tecnologias digitais de processamento de áudio que viabilizam a produção de gravadores de alta qualidade por um preço acessível, também promovem um acesso amplo a sofisticadas ferramentas destinadas à edição e adulteração de áudio no formato digital.

O desenvolvimento de tecnologias voltadas à edição de áudio atingiu níveis que permitem a um usuário criar facilmente uma versão adulterada de um arquivo de áudio. Por exemplo, usuários com pouco ou nenhum treinamento conseguem editar gravações inserindo, modificando ou retirando eventos acústicos com qualidade tal que edições

passem despercebidas pela oitiva ordinária (MAHER, 2009; BÖHME et al., 2009). Como conclusão imediata emerge a necessidade de mecanismos que permitam assegurar que uma determinada gravação de áudio, cujo conteúdo seja relevante para a persecução penal, seja devidamente autenticada (BÖHME et al., 2009).

Há na literatura diversas técnicas destinadas a detecções de adulteração em arquivos de áudio, explorando diferentes aspectos relacionados a captação, ambientação, compressão e natureza estatística do sinal (TÁVORA; NASCIMENTO, 2015). Um primeiro ramo de pesquisa diz respeito a técnicas ativas de detecção de adulterações, tais como a utilização de marcas d'água com o emprego de informação conhecida e estruturada adicionada ao áudio original. No entanto, a maioria das gravações de áudio é apresentada como prova sem qualquer tipo de marcação e, na prática, é utópico esperar que todo áudio a ser apresentado possua qualquer tipo de marca d'água. Dessa forma, a pesquisa relativa a mecanismos passivos de detecção de adulterações ganha destaque, mostrando-se mais útil e alinhada a realidade (GUPTA; CHO; KUO, 2012).

Nesse sentido, Pan, Zhang & Lyu (2012) propõem um método para detecção de composições por meio da revelação de diferenças anormais nos níveis de ruídos de fundo locais, baseando-se na observação de que amostras de sinais de áudio, filtrados por um filtro passa faixa, tendem a apresentar uma curtose aproximadamente constante. Malik (2013) propõe uma abordagem estatística para modelar e estimar o grau de reverberação e o ruído de fundo de uma gravação de áudio, resultando em uma ferramenta que busca identificar adulterações por meio de uma análise de estabilidade dessas características.

Algumas técnicas analisam efeitos produzidos por esquemas de compressão de áudio com perdas, tais como o MP3. Em Yang, Qu & Huang (2012), os autores propõem uma maneira de revelar fraudes identificando a existência de inconsistências em características de *offset* de quantização no mecanismo de compressão MP3. Em Liu, Sung & Qiao (2010), os autores detectam a existência de dupla compressão MP3 extraindo características estatísticas da transformada de cosseno discreta modificada, do inglês *Modified Discrete Cossine Transform* (MDCT), e aplicando uma máquina de vetores de suporte, do inglês *Support Vector Machine* (SVM), às características extraídas.

Outras técnicas detectam não linearidades produzidas por transições abruptas no sinal de áudio por meio de análise biespectral (FARID, 1999), bem como por meio de estatísticas de ordem superior para identificar microfones e revelar possíveis adulterações

(IKRAM; MALIK, 2012). Távora & Nascimento (2015) propuseram uma técnica de autenticação destinada a identificar adulterações produzidas por meio da cópia de segmentos de áudio de intervalo pequeno replicados dentro do mesmo arquivo.

Apesar do grande número de diferentes abordagens para detecção de edições em arquivos de áudio, muitos trabalhos utilizam a informação proveniente da componente de interferência do sinal da rede elétrica, do inglês *electrical network frequency* (ENF). A ENF é amplamente utilizada e referenciada pela comunidade forense para detecção de adulterações em diversos trabalhos, conforme detalhado na Seção 2.2. De fato, o sinal ENF é frequentemente encontrado em gravações de áudio provenientes de dispositivos de escuta. A alta disponibilidade associada às características determinísticas e de estabilidade tornam a ENF uma figura de mérito do ponto de vista forense, o que justifica o seu amplo uso. A utilização da ENF incorporada em sinais de áudio como figura de mérito para detecção cega de inserções e supressões de segmentos de áudio é o objeto de estudo desta dissertação.

1.1 JUSTIFICATIVA E MOTIVAÇÃO

Áudios adulterados podem ser utilizados como provas falsificadas em ações penais e cíveis, retratando eventos e circunstâncias em diálogos forjados que representem, em algum aspecto de interesse, uma realidade diferente do real encadeamento dos fatos. Técnicas destinadas a identificação dessas adulterações são importantes para impedir que áudios forjados sejam utilizados como prova na solução de crimes e na resolução de conflitos. Dessa forma, o constante desenvolvimento de técnicas robustas para identificação de tais adulterações deve ser objeto de atenção visando a mitigar eventuais fraudes. Mecanismos de identificação de adulterações que exploram a estabilidade de sinais interferentes em banda estreita, como a ENF, mostram-se atraentes devido a alta disponibilidade e ao comportamento estável de tais sinais. No entanto, as atuais técnicas passivas que permitem identificar a presença de edições por meio de perturbações na ENF mostram-se sensíveis a cenários desafiadores em que haja baixa relação sinal ruído, do inglês *signal-to-noise ratio* (SNR) ou alto nível de saturação, do inglês *saturation level* (SL). Na prática, situações com baixa SNR e elevado SL são uma casuística frequente o que limita a efetividade das técnicas atuais, demandando o desenvolvimento de técnicas mais robustas a estes cenários. Nesse trabalho propõe-se um mecanismo de identificação de adulterações por meio do estimador de ENF ESPRIT-Hilbert em conjunto com um detector de *outliers* baseado na curtose amostral da ENF estimada, do inglês *ESPRIT-Hilbert ENF estimator in conjunction with an outlier detector based*

on the sample kurtosis of the estimated ENF (SPHINS). A partir da curtose amostral das estimativas da ENF compõe-se um vetor de características que é aplicado a um classificador SVM, devidamente treinado a partir de uma base de dados conhecida para indicar a presença de edições.

1.2 METODOLOGIA

Esta dissertação tem como objetivo estudar mecanismos de identificação de variações anormais no sinal ENF eventualmente incorporado em uma gravação de áudio e sua relação com tarefas de edição de áudio baseadas em inserção ou supressão de trechos. Para isso, trabalhos anteriores correlacionados ao objeto dessa dissertação são estudados em uma ampla revisão bibliográfica, e duas técnicas distintas destinadas a estimar a ENF são detalhadas em uma revisão teórica a partir de um modelo de dados estabelecido. Como subproduto desse estudo propõe-se um novo método para identificação de adulterações denominado SPHINS, destinado a revelar edições em áudio por meio de uma abordagem robusta para analisar perturbações na ENF. Com esse objetivo explora-se conjuntamente estimativas da ENF por meio da técnica de estimação da frequência instantânea do sinal analítico de Hilbert, que é mais sensível a transições abruptas de fase, e da técnica ESPRIT, que é mais precisa em baixas condições de SNR. O SPHINS utiliza um vetor de características que resume os distúrbios na ENF baseado na curtose amostral das estimativas, aplicado a um classificador SVM treinado para identificar a presença de edições. Como forma de validar a técnica desenvolvida, são realizados experimentos em diferentes cenários de SNR e SL. O mecanismo proposto é validado em áudios de uma base de dados que contém 100 gravações telefônicas autorizadas de áudios não editados, e 100 gravações telefônicas de áudios editados, computando-se as diferentes taxas de erro para efeito de comparação com as técnicas atuais.

1.3 RESULTADOS ESPERADOS E OBTIDOS

Como resultado esperado para este estudo tem-se a proposição de um novo mecanismo para detecção de adulterações em arquivos de áudio questionados por meio da informação constante de sinais ENF eventualmente incorporados, robusto a condições adversas de SNR e SL. Como resultado obtido tem-se a proposição do SPHINS, um mecanismo de identificação de adulterações baseado na estimação da ENF, cujo desempenho foi medido a partir de uma base de dados conhecida, obtendo uma taxa de erro de classificação inferior aos trabalho anteriores correlatos em condições mais severas de SNR e SL.

1.4 ORGANIZAÇÃO DA DISSERTAÇÃO

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta conceitualmente o exame de verificação de edição em áudio e realiza uma revisão bibliográfica do estado da arte na detecção de adulterações por meio da ENF. O Capítulo 3 apresenta a formulação teórica relativa ao modelo de dados considerado neste trabalho para análise e extração da ENF a partir de um sinal de áudio questionado, e descreve dois estimadores do sinal ENF. O Capítulo 4 descreve a proposta deste trabalho, o SPHINS, abordando sua concepção e dividindo-o em três etapas: pré-processamento, extração do vetor de características e classificação. O Capítulo 5 apresenta os resultados de experimentos onde o método SPHINS foi aplicado a uma base de dados conhecida, visando a avaliar o desempenho na classificação de áudios editados e não editados, simulando condições desfavoráveis com diferentes graus de SNR e SL. E, finalmente, no Capítulo 6 são apresentadas as conclusões do trabalho desenvolvido e recomendações para trabalhos futuros.

2 ESTADO DA ARTE NA UTILIZAÇÃO DA ENF EM EXAMES DE VERIFICAÇÃO DE EDIÇÃO EM ÁUDIO

Este capítulo visa a contextualizar o leitor acerca da natureza do exame de verificação de edição em áudio, terminologia amplamente utilizada nos institutos de criminalística brasileiros para designar os exames periciais destinados a identificar a presença de traços de edições em arquivos de áudio. Visa também a realizar uma revisão bibliográfica com apresentação do estado da arte na utilização da ENF em ferramentas de auxílio à detecção de edições em áudio.

2.1 VERIFICAÇÃO DE EDIÇÃO EM ÁUDIO

A área das ciências forenses, ou criminalística como por muitos é chamada, diz respeito a utilização da ciência na aplicação das leis pelos órgãos de Estado dentro do sistema penal (MELOAN, 2000). Trata-se de um termo geral que abarca uma série de áreas do conhecimento destinadas a extrair fatos probatórios a partir de vestígios físicos. Um dos fundamentos gerais das ciências forenses diz repeito ao princípio da transferência, que estabelece que sempre que duas entidades de uma mesma realidade física interagem, haverá transferência, em alguma escala, de vestígios de uma para a outra (INMAN; RUDIN, 2002). Assim, por exemplo, um ladrão deixa suas impressões digitais ao arrombar um cadeado, um homicida retém em suas mãos a pólvora proveniente de um disparo de arma de fogo e um projétil retém em seu corpo as marcações compatíveis com o raiamento do cano de um armamento.

O princípio da transferência estabelece relações fortes entre as evidências materiais e a realidade dos fatos. Não se trata de tarefa simples para um criminoso apagar todos os vestígios de sua atuação delituosa ou, o que seria pior, forjar vestígios com objetivo de acusar outrem. Mesmo o mais sofisticado dos criminosos não poderia ter certeza que sua tarefa em modificar a realidade é totalmente consistente com a realidade esperada caso não tivesse cometido determinado crime (BÖHME et al., 2009). Nesse sentido, o princípio da transferência se aplica a tarefa de realizar edições em arquivos de áudio. Ainda que áudios digitais sejam representações simbólicas de referentes físicos, guardam relação direta com a realidade material, e portanto estão sujeitos ao princípio da transferência. Para estabelecer as relações entre a representação digital e a realidade

física, são necessários modelos. De uma forma geral, a qualidade das técnicas forenses voltadas a identificar eventuais edições em um arquivo de áudio depende dos modelos empregados e, naturalmente, é tão boa quanto a qualidade do modelos utilizados (BÖHME et al., 2009).

Dentre os diversos exames que podem ser elencados na área de perícias em registros de áudio e imagens tem-se o exame de verificação de edição, cujo objetivo é verificar se nas gravações questionadas são encontrados elementos indicativos de alterações que possam, de algum modo, modificar o seu conteúdo original. Para tal são analisados o maior número possível de elementos, buscando-se acumular o maior número de evidências capazes de sustentar a hipótese de que o material analisado está editado. Na existência de indicativos fundamentados por evidências corretamente interpretadas, é possível apontar de forma categórica um resultado positivo para a existência de uma eventual edição.

Por outro lado, por uma condição lógica decorrente da natureza empírica da prova material (BÖHME et al., 2009), o resultado negativo acerca da existência de edição não pode ser diretamente determinado pela não observação de evidências que sustentem a hipótese de edição. Em vez de uma implicação direta há uma implicação indireta onde, a medida em que diversos elementos são analisados sem que se observem evidências de edição, obtém-se gradativamente maior plausibilidade da hipótese de que a gravação questionada não fora editada. Dentre o conjunto de técnicas distintas para se realizar a busca por evidências de edição, encontra-se a utilização da informação da ENF eventualmente incorporada no sinal de áudio como fonte de características que permitam, entre outras coisas, identificar se houve alguma adulteração no referido sinal.

2.2 ESTADO DA ARTE

O sinal ENF é frequentemente encontrado em gravações de áudio produzidas por dispositivos de interceptação telefônica e ambiental, fruto da interferência da rede elétrica. Mesmo dispositivos de interceptação alimentados por baterias podem produzir áudios com proeminente sinal de interferência da rede elétrica. Dispositivos de captação ambiental, por exemplo, quando colocados em uma sala ou gabinete, devido à necessidade de ocultação, acabam sendo dispostos dentro de paredes, forros ou assoalhos, próximos a cabos de energia, o que favorece a presença da ENF. Além disso, a necessidade de miniaturização de dispositivos de escuta muitas vezes sacrifica um correto isolamento dos circuitos. Um cenário comum ao uso de equipamentos de escuta é a utilização de

longos fios para conexão entre o microfone e o circuito de gravação com o intuito de aproximar o microfone à fonte de áudio, melhorando a captação. Porém, esses cabos longos tendem a ser mais facilmente induzidos pela tensão da rede elétrica adjacente, favorecendo o aparecimento da ENF.

A Figura 2.1 ilustra o espectrograma de um típico sinal de áudio de 2 minutos de duração, obtido por meio de escuta telefônica contendo componente frequencial correspondente a sinal interferente da rede elétrica em 60 Hz[1].

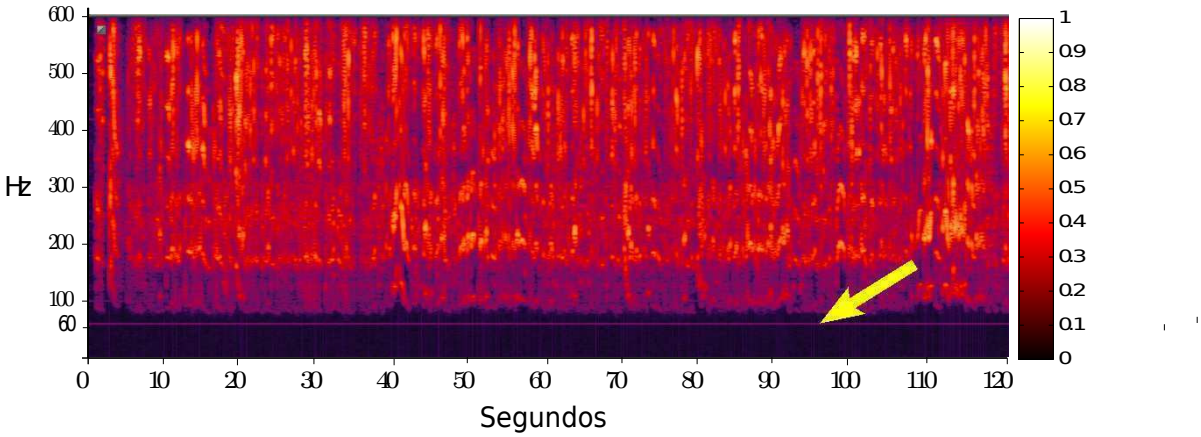


Figura 2.1: Espectrograma de sinal de áudio obtido por escuta telefônica contendo a ENF. A seta indica a componente ENF em 60 Hz.

O sinal da rede elétrica de potência é um sinal padronizado com características bem determinadas. Trata-se de um sinal *quasi*-senoidal[2], com frequência de oscilação nominal igual a 50 Hz ou 60 Hz, a depender da região do planeta. Os países europeus, Austrália e a maior parte dos países da África e Ásia utilizam 50 Hz. Os países da América do Norte e América Central utilizam 60 Hz. É importante notar que na América do Sul, alguns países utilizam 50 Hz e outros 60 Hz. O Japão, por seu lado, utiliza ambos os valores em sua malha (RODRÍGUEZ; APOLINÁRIO JR.; BISCAINHO, 2010). Trata-se de sinais muito bem comportados, que não mostram transições abruptas de frequência ou fase ao longo do tempo. Esta característica estável no comportamento dos sinais ENF é mandatória para o correto funcionamento de grande parte dos equipamentos elétricos e eletrônicos conectados à rede, fazendo com que sejam utilizados mecanismos rígidos para controle dos valores de frequência de oscilação, mantendo o mesmo dentro de limites bem definidos e estreitos (ANEEL, 2013). A Figura 2.2 ilustra em detalhe o espectrograma dos primeiros 30 segundos do mesmo sinal de áudio retratado na Figura 2.1, porém em detalhe numa escala frequencial entre 59.1 e 60.6 Hz, ilustrando uma

[1] O áudio foi subamostrado a taxa de 1200 Hz, utilizando um filtro anti-aliasing do tipo FIR.

[2] Na prática o sinal não é perfeitamente senoidal, apresentando distorção harmônica, variações na amplitude e na frequência de oscilação.

excursão confinada e bem comportada entre, aproximadamente, 59,9 e 60,0 Hz.

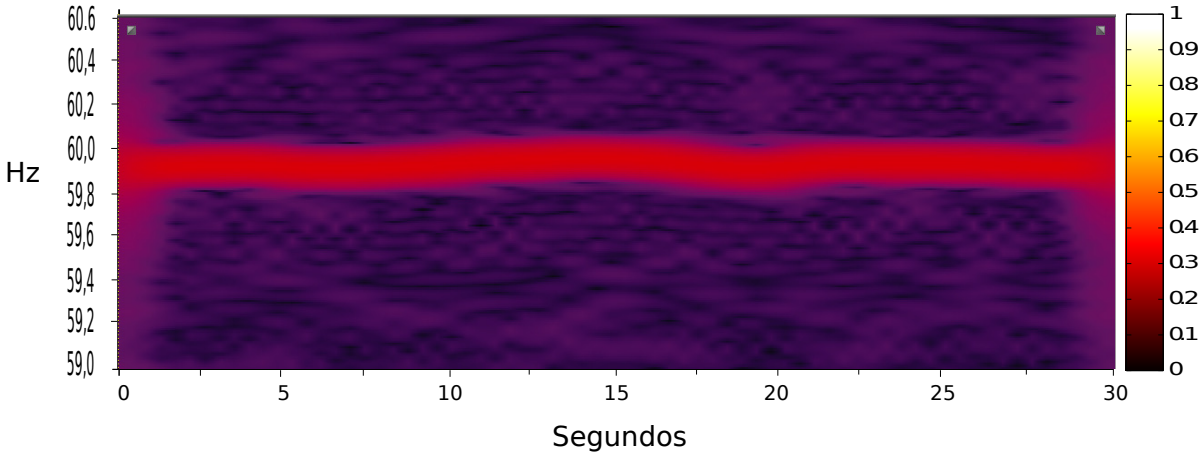


Figura 2.2: Espectrograma em detalhe de sinal de áudio obtido por escuta telefônica contendo a ENF.

O sinal ENF eventualmente encontrado e gravações é útil do ponto de vista forense uma vez que pode ser utilizado como meio de prova para autenticar áudios, identificar a localização de uma gravação e detectar edições (GUPTA; CHO; KUO, 2012). Uma vez que o sinal ENF apresenta uma regularidade natural, a existência de variações abruptas na ENF estimada a partir de uma gravação de áudio pode indicar a existência de alterações em seu conteúdo, mormente aquelas provocadas por inserções , supressões e remanejamentos. De fato, uma inserção ou supressão de um segmento de áudio em uma gravação apresenta uma grande chance de produzir descontinuidades nas informações de fase instantânea e gerar variações abruptas e anormais na ENF estimada. As Figuras 2.3 e 2.4 ilustram por meio de oscilogramas os processos de edição por inserção e supressão de segmentos de áudio, respectivamente. Remanejamentos, por sua vez, são tarefas de edição que consistem na aplicação em sequência de uma ação de supressão de um determinado trecho, com a subsequente inserção do trecho suprimido em outro ponto do mesmo arquivo de áudio.

Realizando-se supressão de amostras, por exemplo, a menos que a região suprimida tenha um comprimento de aproximadamente um número inteiro de ciclos da ENF, haverá uma quebra na evolução da fase do sinal interferente da rede elétrica, o que acarreta variações anormais no sinal ENF. A Figura 2.5 ilustra dois oscilogramas. O superior diz respeito a um sinal senoidal puro onde a área hachurada corresponde a uma região a ser suprimida, e o oscilograma inferior corresponde ao mesmo sinal, reconstruído após a supressão de amostras, onde se observa a transição abrupta de fase.

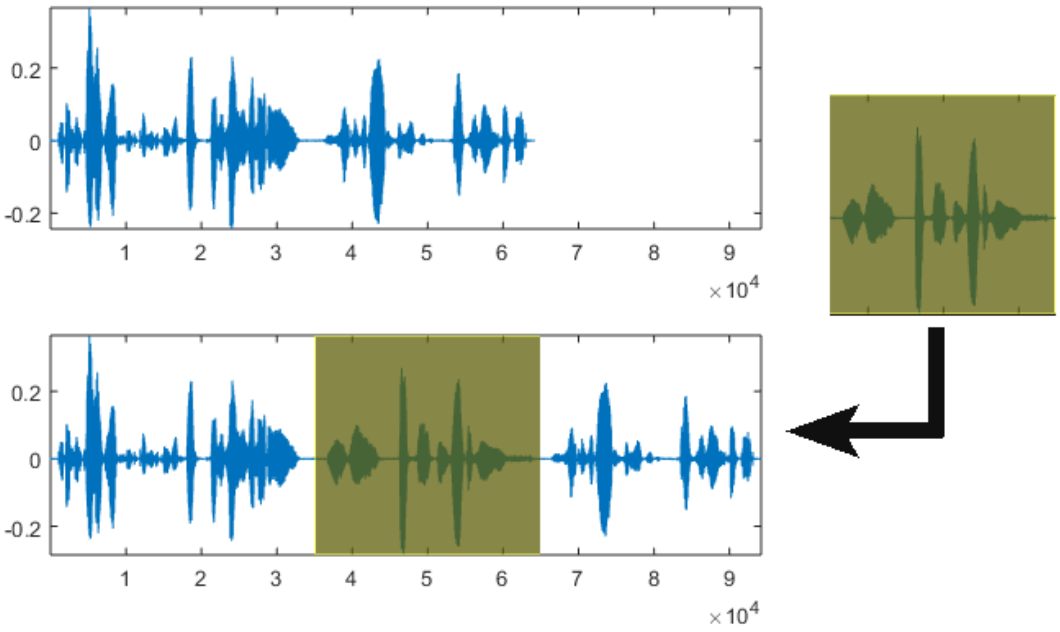


Figura 2.3: Procedimento de edição por inserção. Nos oscilogramas, a área hachurada corresponde ao segmento de áudio inserido.

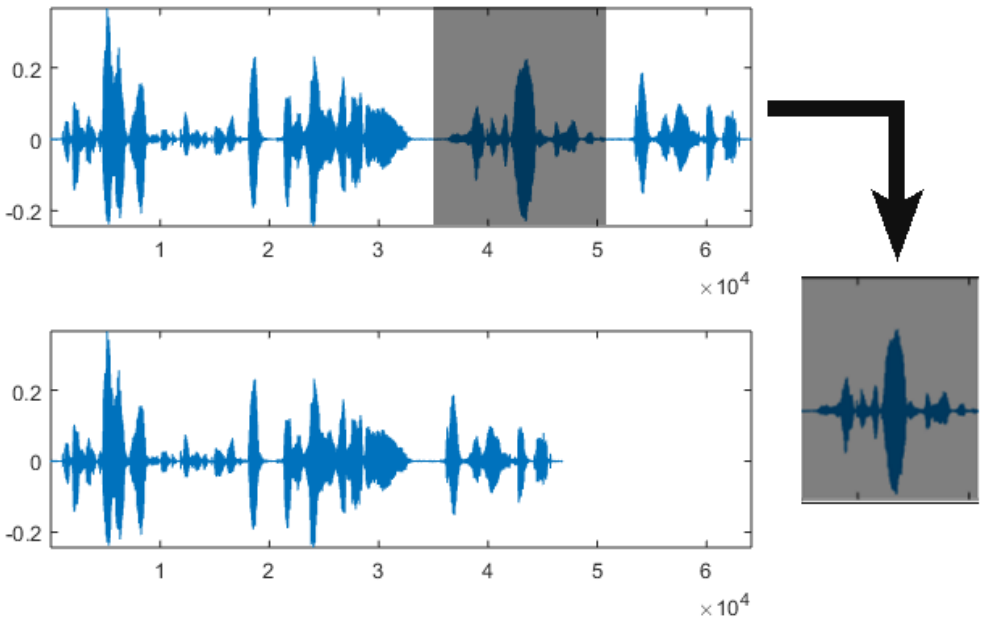


Figura 2.4: Procedimento de edição por supressão. Nos oscilogramas, a área hachurada corresponde ao segmento de áudio suprimido.

Em Cooper (2008), é proposto um mecanismo automático para extrair a informação da ENF por meio da interpolação quadrática dos valores de pico no espectro de Fourier. A informação extraída pode então ser comparada com uma base de dados armazenada da variação da ENF na localidade de interesse a fim de autenticar o arquivo de áudio. No

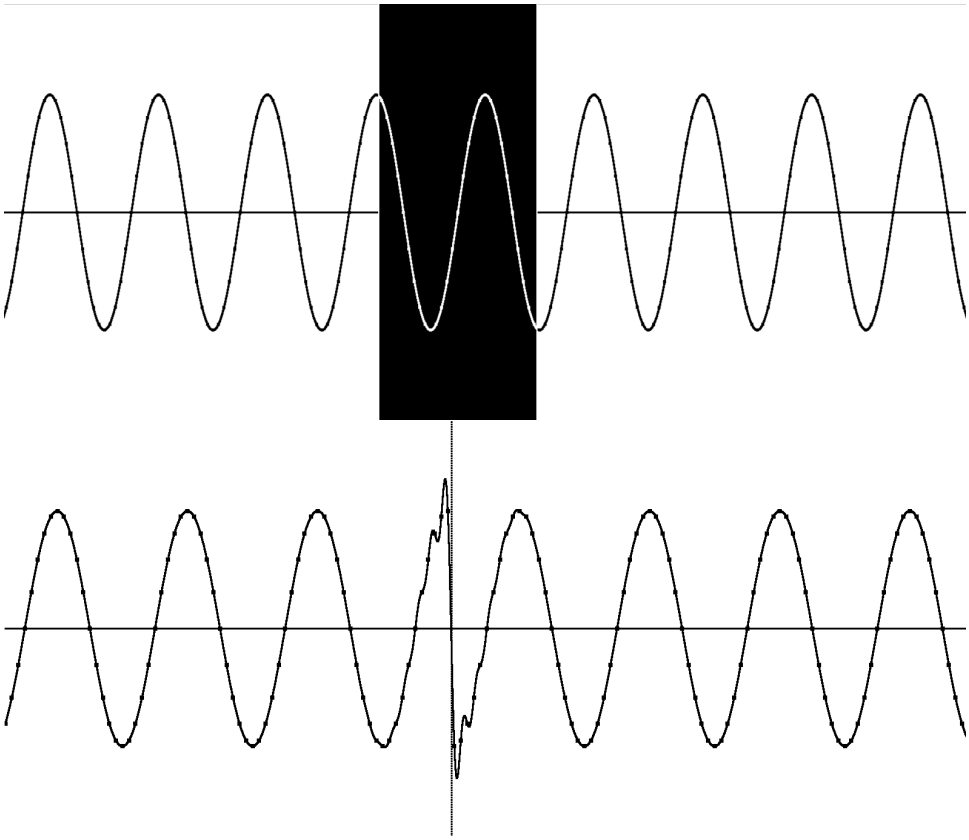


Figura 2.5: Transição abrupta de fase gerada em um sinal perfeitamente senoidal após a supressão de um segmento.

entanto, essa técnica de autenticação demanda que esteja disponível uma base de dados que armazene o sinal ENF ao longo do tempo, por longos períodos para a região de interesse com o propósito de permitir a autenticação. Rodríguez, APOLINÁRIO JR. & Biscainho (2010) propõem um método baseado na identificação de descontinuidades de fase na ENF incorporada utilizando uma análise de Fourier de alta precisão para estimar as variações de fase. Se a variância das estimativas de fase excedem um limiar definido empiricamente, caracteriza-se um comportamento anormal, tipicamente derivado de supressões e inserções de segmentos de áudio. Em Rodríguez, APOLINÁRIO JR. & Biscainho (2013) os autores estendem o método para análise de descontinuidade de fase no terceiro harmônico da ENF.

Em Hajj-Ahmad, Garg & Wu (2012), os autores avaliam o desempenho de dois diferentes métodos para estimação da ENF: *MUltiple Signal Classification* (MUSIC) (SCHMIDT, 1986) e o *Estimation of Signal Parameters via Rotational Invariance Techniques* (ESPRIT) (ROY; KAILATH, 1989). Ambos os métodos apresentam maior precisão que o utilizado em Cooper (2008), tendo o ESPRIT se mostrado o mais preciso.

Em Esquef, APOLINÁRIO JR. & Biscainho (2014), os autores utilizam o método de

estimação da frequência instantânea por meio do sinal analítico de Hilbert para obter a informação da ENF. Os autores também estabelecem um limiar adaptativo para a excursão da ENF. Para isso utilizam um limite superior para flutuações normais observadas em arquivos de áudio não editados. A partir desse limiar implementam um critério automático para classificar um arquivo de áudio como editado. Para validar seus resultados utilizam bases de dados públicas com arquivos de áudio de conversações telefônicas. Mais recentemente esse método foi modificado por melhorias no processo de detecção, realizando-se comparações por correlações cruzadas com padrões de perturbação na ENF resultantes de edições em sinais senoidais produzidos sinteticamente (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2015). Em Fuentes et al. (2016) os autores utilizam uma malha de captura de fase, do inglês *Phase Locked Loop* (PLL), para estimar a fase instantânea da ENF e com isso implementar um sistema automático de decisão a partir das perturbações existentes na frequência de oscilação do oscilador controlado por voltagem, do inglês *Voltage-Controlled- Oscilator* (VCO), empregado no PLL.

Os resultados obtidos nos trabalhos de Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) e Rodríguez, APOLINÁRIO JR. & Biscainho (2010) para detecção de alterações em arquivos de áudio por meio de perturbações de fase e frequência do sinal ENF avaliam o desempenho dos métodos com base exclusivamente na EER (*Equal Error Rate*). A EER medida corresponde a taxa de erro do classificador durante o treinamento, no ponto de operação em que a quantidade de falsos positivos é igual a quantidade de falsos negativos. Para seu cálculo, os autores aferem o número de classificações incorretas ocorridas em todos os áudios existentes na base de dados que serve como critério para definição dos limiares de decisão. Não empregam técnicas de validação cruzada com o objetivo de mensurar de forma mais efetiva a capacidade de generalização dos classificadores. Já no trabalho de Fuentes et al. (2016) foram realizados testes de validação cruzada utilizando-se duas bases distintas. Observa-se que os três métodos apresentam considerável degradação em condições de baixa SNR e elevada SL.

2.3 SUMÁRIO

Neste capítulo foram apresentados conceitos importantes sobre os exames de verificação de edição em áudio e como as técnicas que utilizam a ENF podem ser úteis na detecção e identificação de eventos de edição em áudio. Para isso foi apresentado o estado da arte na autenticação de áudio utilizando a ENF por meio de uma revisão bibliográfica dos trabalhos anteriores.

Verificou-se que o exame de verificação de edição em áudio é uma subclasse de exame dentro da criminalística, sujeita à aplicação do princípio da transferência. De forma geral, observou-se que existe uma série de abordagens que se valem do princípio da transferência para averiguar a presença de edições. Assim, detectam-se traços de manipulação correspondentes a perturbações anormais no sinal ENF eventualmente incorporado no áudio e associados a ações de inserção ou supressão de segmentos de áudio. Para tal, o estado da arte emprega basicamente estimadores da ENF e mecanismos de detecção automática de edição que verificam variações anormais na frequência do sinal interferente da rede elétrica. O próximo capítulo apresenta o modelo de dados utilizado e duas técnicas de estimação da ENF empregadas no método proposto.

3 MODELO DE DADOS E ESTADO DA ARTE DOS ESTIMADORES DA ENF

Este capítulo aborda conhecimentos teóricos relevantes para os métodos de estimação da ENF estudados nessa dissertação e sua aplicação no classificador proposto. Na Seção 3.1 é apresentado o modelo de dados utilizado para caracterizar a ENF presente em um sinal de áudio questionado. Na seção 3.2 são apresentados dois estimadores da ENF, um por meio da frequência instantânea do sinal analítico de Hilbert e outro por meio da técnica ESPRIT.

3.1 MODELO DE DADOS

Para maior compreensão dos mecanismos de obtenção da ENF é preciso um modelo matemático que permita representar o fenômeno físico a ser estudado, ou seja, um modelo matemático que permita caracterizar a ENF presente em um sinal de áudio questionado.

Considere $s(n) \in \mathbb{R}$ o sinal de áudio questionado, ou seja, aquele que se deseja saber se passou por algum processo de edição, com um total de N_s amostras. O sinal questionado pode ser representado pela seguinte superposição de sinais:

$$s(n) \;=\; v(n) + x(n) + e(n), \tag{3.1}$$

onde $v(n)$ é o sinal de áudio contendo a conversação de interesse e livre de qualquer ruído ou interferência, $x(n)$ é a componente interferente do sinal da rede elétrica e $e(n)$ é uma parcela de ruído aditivo. No modelo de dados estabelecido, considera-se que os sinais e o ruído são mutuamente descorrelacionados.

Idealmente, o sinal da rede elétrica pode ser considerado um sinal senoidal que oscila com frequência nominal f_{nom}. Entretanto, observam-se variações na frequência instantânea da ENF devido a descasamentos entre a quantidade de energia produzida pelas unidades de geração e a demanda total de energia elétrica determinada pelo consumo dos usuários da rede. Apesar dos rígidos mecanismos de controle na frequência de rotação das turbinas realizados na unidades de geração de energia elétrica, na prática

é impossível realizar um casamento perfeito entre oferta e demanda ao longo do tempo (SHORT; INFIELD; FRERIS, 2007). Como regra geral, a ENF é diretamente proporcional as variações na demanda de energia excursionando, em pequena escala, em torno da sua frequência nominal. A estabilidade da frequência de oscilação é um parâmetro de qualidade importante no funcionamento da malha de energia, uma vez que desvios acentuados para valores distantes da f_{nom} podem acarretar mal funcionamento e danos em equipamentos, e até mesmo acidentes. Os limites aceitáveis para as flutuações no valor da frequência de oscilação do sinal da rede elétrica dependem de normativos regulatórios locais. Na rede de energia elétrica brasileira, por exemplo, a ENF deve ser mantida entre 59,9 Hz e 60,1 Hz em condições normais de operação em estado estacionário (ANEEL, 2013). Por ser um parâmetro de qualidade importante cuja não adequação acarreta consequências danosas, em geral, uma transição abrupta na frequência instantânea é uma ocorrência inesperada, especialmente em redes que atendem os centros mais desenvolvidos. Essa característica torna a ENF um sinal interessante para a análise forense.

Levando-se em conta as flutuações na ENF, o sinal interferente da rede elétrica pode ser modelado por:

$$x(n) \;=\; A\cos\left(\theta_0 + \omega(n)\; n\right), \tag{3.2}$$

para $0 \leq n < N_s$, onde A é a amplitude do sinal, θ_0 é sua fase inicial, e $\omega(n)$ é a frequência angular normalizada da rede elétrica, relacionada com a ENF linear $f(n)$ por:

$$\omega(n) \;=\; \frac{2\pi f(n)}{f_{\text{s}}}, \tag{3.3}$$

onde f_{s} é a frequência de amostragem do sinal.

O objetivo deste trabalho é, dado apenas o sinal questionado $s(n)$, detectar uma possível edição explorando variações anormais em estimativas $\hat{\omega}(n)$. Para isso, é necessário obter estimativas fidedignas $\hat{\omega}(n)$, robustas a cenários desfavoráveis de SNR e SL, porém sensíveis a eventuais transições abruptas de fase. Dessa forma, são descritos dois estimadores da ENF distintos cuja atuação conjunta permite atingir o objetivo proposto, a menos de uma margem de erro de classificação.

3.2 ESTIMADORES DA ENF

Nesta seção serão apresentadas duas abordagens referentes ao estado da arte na estimação da ENF. Na Subseção 3.2.1, apresenta-se o estimador ENF baseado na frequência instantânea do sinal analítico de Hilbert (HEE), e na Subseção 3.2.2 apresenta-se o estimador ENF baseado na técnica ESPRIT (3E).

Para extrair a ENF, o sinal $s(n)$ é previamente processado por um filtro passa-banda de fase zero, com L coeficientes, seletivo, com largura de banda $\mathrm{BW_{ENF}}$ centrada na frequência nominal f_{nom}, produzindo a estimativa do sinal interferente da rede elétrica $\hat{x}(n)$:

$$\hat{x}(n) \;=\; h_{\mathrm{bp}}(n) * s(n), \tag{3.4}$$

contendo $N_{\mathrm{x}} = N_{\mathrm{s}} + L - 1$ amostras[3], onde $h_{\mathrm{bp}}(n)$ representa a resposta ao impulso do filtro passa-banda e o operador $*$ representa a operação de convolução linear.

Para que a estimativa $\hat{x}(n)$ contenha informação do tom senoidal que representa o sinal da rede elétrica, porém descartando-se degradação proveniente de outros sinais, a largura de banda $\mathrm{BW_{ENF}}$ precisa ser tão estreita quanto possível com o objetivo de rejeitar todo e qualquer sinal diferente de $x(n)$, porém suficientemente larga para cobrir toda a excursão esperada para a ENF. Considerando áudios provenientes de gravações obtidas, por exemplo, a partir de escutas realizadas na rede de telefonia pública comutada (RTPC), não há, na prática, sobreposição espectral relevante entre o sinal da rede elétrica e sinais de voz, de tal forma que pode-se escrever:

$$\begin{aligned} \hat{x}(n) = \;& h_{\mathrm{bp}}(n) * x(n) + h_{\mathrm{bp}}(n) * [v(n) + e(n)] \\ \hat{x}(n) \;\approx\; & x(n) + z(n) \end{aligned} \tag{3.5}$$

onde $z(n) = h_{\mathrm{bp}}(n) * v(n) + h_{\mathrm{bp}}(n) * e(n)$ corresponde à soma da parcela de ruído filtrada em banda passante com a parcela residual proveniente da filtragem em passa banda do sinal sob teste.

A Figura 3.1 ilustra o sinal $\hat{x}(n)$, com $0 \leq n < N_s$, para um sinal de áudio de 30 segundos proveniente de escuta telefônica, contendo componente ENF nominal de 60 Hz, com supressão de um segmento de áudio correspondente a uma descontinuidade de

[3]Como a filtragem é fase zero gera-se $(L-1)/2$ amostras situadas em $n < 0$ e $(L-1)/2$ em $n \geq N_s$.

fase de 160 graus, subamostrado a uma taxa $f_s = 1200$ Hz e filtrado por um filtro FIR, passa-banda, centrado em 60 Hz com largura de banda de 0.8 Hz[4]. Observa-se que há considerável modulação em amplitude no ponto em que há a edição com quebra de fase de 160 graus. No entanto, edições que resultem em desvios de fase menores produzem modulações mais sutis (RODRÍGUEZ; APOLINÁRIO JR., 2009) que facilmente se confundem com flutuações de amplitude no sinal ENF interferente, fazendo com que as modulações em amplitude não sejam a melhor característica para identificar uma eventual adulteração.

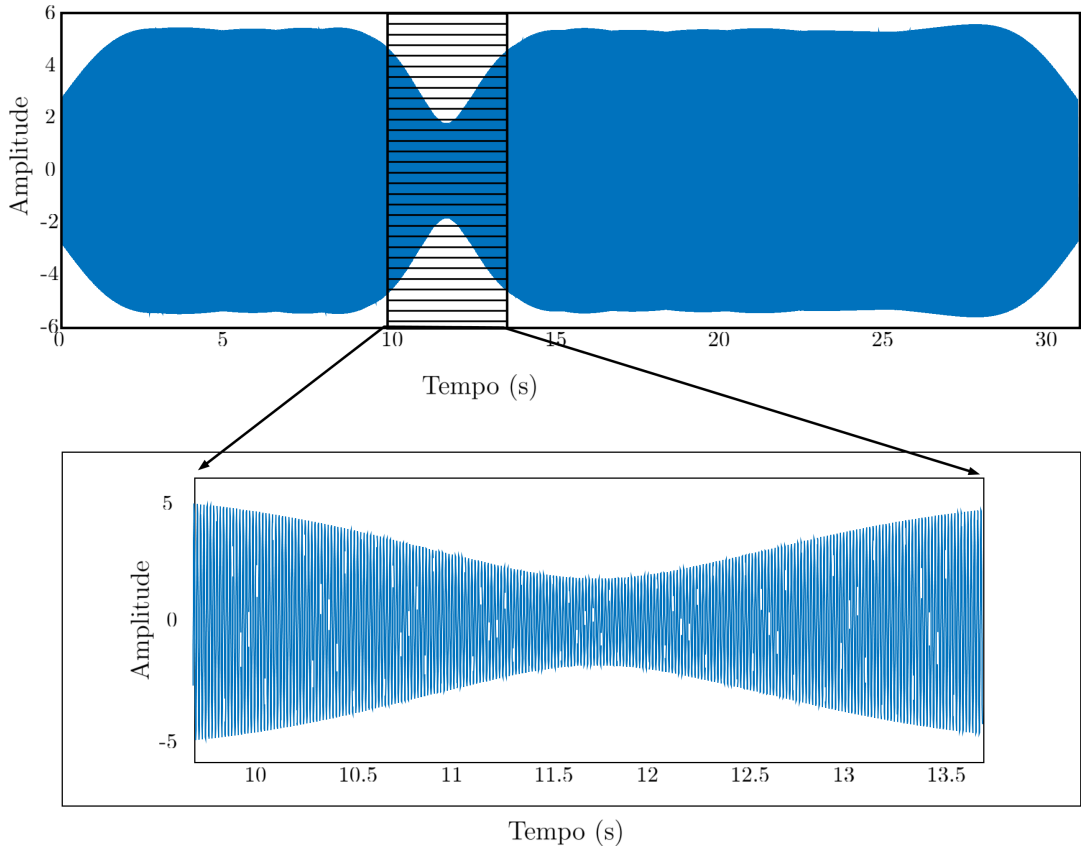


Figura 3.1: Sinal $\hat{x}(n)$, para $0 \leq n < N_s$ em um sinal de áudio de 30 segundos proveniente de escuta telefônica, contendo componente ENF nominal de 60 Hz, com supressão de um segmento de áudio correspondente a uma descontinuidade de fase de 160 graus.

3.2.1 Estimador ENF Baseado na Transformada de Hilbert (HEE)

O HEE é um estimador da ENF que utiliza a informação de fase e frequência instantânea da aproximação analítica de Hilbert $\hat{x}_{\mathrm{a}}(n) \in \mathbb{C}$ do sinal real correspondente à estimativa $\hat{x}(n)$ (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014):

[4]Foi utilizado um filtro passa banda do tipo FIR, de fase zero conforme descrito na Seção 4.1, com largura de banda de 0,8 Hz, utilizando-se as funções do Matlab ® *conv*, nos sentidos direto e reverso, e *fir1*, com 10.000 coeficientes.

$$\hat{x}_{\mathrm{a}}(n) \;=\; \hat{x}(n) + j\mathcal{H}\left\{\hat{x}(n)\right\}, \tag{3.6}$$

onde o operador $\mathcal{H}\{\}$ denota a Transformada de Hilbert Discreta (DHT)[5] e $j = \sqrt{-1}$.

A implementação discreta da DHT visando a obtenção das representações analíticas admite algumas abordagens distintas (MARPLE, 1999). Em Marple (1999) os autores propõem uma abordagem no domínio frequencial para a obtenção do sinal discreto analítico, preservando-se a mesma taxa de amostragem e a ortogonalidade entre suas componentes real e imaginária.

Admitindo-se que a quantidade N_{x} de amostras do sinal $\hat{x}(n)$ é um número par, considere que $\hat{X}(k)$ seja a sua Transformada Discreta de Fourier, do inglês *Discrete Fourier Transform* (DFT), tal que:

$$\hat{X}(k) \;=\; \sum_{n=-(L-1)/2}^{N_{\mathrm{x}}-(L+1)/2} \hat{x}(n)\exp(-j2\pi kn/N_{\mathrm{x}}) \tag{3.7}$$

O sinal analítico no domínio frequencial $\hat{X}_{\mathrm{a}}(k)$ é portanto obtido por meio da seguinte construção:

$$\hat{X}_{\mathrm{a}}(k) = \begin{cases} \hat{X}(0), & \text{para } k = 0 \\ 2\hat{X}(k), & \text{para } 1 \leq k \leq \frac{N_{\mathrm{x}}}{2} - 1 \\ \hat{X}(\frac{N_{\mathrm{x}}}{2}), & \text{para } k = \frac{N_{\mathrm{x}}}{2} \\ 0, & \text{para } \frac{N_{\mathrm{x}}}{2} + 1 \leq k \leq N_{\mathrm{x}} - 1 \end{cases} \tag{3.8}$$

Finalmente, o sinal analítico no domínio do tempo é obtido por meio da Transformada Discreta de Fourier Inversa, do inglês *Inverse Discrete Fourier Transform* (IDFT), tal que:

$$\hat{x}_{\mathrm{a}}(n) \;=\; \frac{1}{N_{\mathrm{x}}} \sum_{k=0}^{N_{\mathrm{x}}-1} \hat{X}_{\mathrm{a}}(k)\exp(j2\pi kn/N_{\mathrm{x}}) \tag{3.9}$$

[5]Neste trabalho, a representação analítica do sinal discreto foi implementada com a função *hilbert* do Matlab ®.

Substituindo a Equação (3.2) na Equação (3.5), tem-se que:

$$\hat{x}(n) \;=\; A\cos\left(\theta_0 + \omega(n)\; n\right) + z(n), \tag{3.10}$$

Aplicando-se a Equação (3.5) na Equação (3.6), tem-se:

$$\hat{x}_{\mathrm{a}}(n) \;=\; A\cos\left(\theta_0 + \omega(n)\; n\right) + z(n) + j\mathcal{H}\left\{A\cos\left(\theta_0 + \omega(n)\; n\right) + z(n)\right\}, \tag{3.11}$$

Como a transformada de Hilbert é um operador linear, pode-se reescrever a Equação (3.11) como:

$$\hat{x}_{\mathrm{a}}(n) \;=\; A\cos\left(\theta_0 + \omega(n)\; n\right) + j\mathcal{H}\left\{A\cos\left(\theta_0 + \omega(n)\; n\right)\right\} + z(n) + j\mathcal{H}\left\{z(n)\right\}. \tag{3.12}$$

Como a representação analítica de um sinal cossenoidal de fase θ é adequadamente descrita pela exponencial complexa de mesma fase, pode-se aproximar (3.12) por:

$$\hat{x}_{\mathrm{a}}(n) \;=\; A\exp\left[j\left(\theta_0 + \omega(n)\; n\right)\right] + z_{\mathrm{a}}(n), \tag{3.13}$$

onde $z_{\mathrm{a}}(n)$ é a representação analítica da parcela de ruído $z(n)$.

O método HEE obtém a estimativa $\hat{\omega}_H(n)$ da frequência angular normalizada $\omega(n)$ por meio de uma aproximação a derivada de primeira ordem da estimativa de fase instantânea:

$$\hat{\omega}_{\mathrm{H}}(n) \;=\; \dot{\hat{\theta}}(n), \tag{3.14}$$

onde $\dot{\hat{\theta}}(n)$ é a primeira derivada de:

$$\hat{\theta}(n) \;=\; \angle \hat{x}_{\mathrm{a}}, \tag{3.15}$$

onde $\angle \hat{x}_{\mathrm{a}}$ corresponde ao argumento de $\hat{x}_{\mathrm{a}}$, dado por:

$$\angle \hat{x}_{\mathrm{a}} \;=\; \arctan\left(\frac{\mathrm{Im}(\hat{x}_{\mathrm{a}})}{\mathrm{Re}(\hat{x}_{\mathrm{a}})}\right), \tag{3.16}$$

onde $\mathrm{Im}(\hat{x}_{\mathrm{a}})$ e $\mathrm{Re}(\hat{x}_{\mathrm{a}})$ são as partes imaginária e real de $\hat{x}_{\mathrm{a}}$, respectivamente.

Considerando que $\hat{\theta}(n)$ varia lentamente em comparação à taxa de amostragem do sinal $\hat{x}(n)$, uma aproximação razoável para a primeira derivada de $\hat{\theta}(n)$, e portanto para a estimativa $\hat{\omega}_{\mathrm{H}}(n)$, é dada por (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014):

$$\hat{\omega}_{\mathrm{H}}(n) = \angle\left(\hat{x}_{\mathrm{a}}(n)\hat{x}^{*}_{\mathrm{a}}(n-1)\right), \tag{3.17}$$

para $0 \leq n < N_s$[6], onde o operador * é aqui utilizado para denotar o valor complexo conjugado do operando.

A título de exemplo, a Figura 3.2 mostra um áudio de 30 segundos contendo sinal ENF com frequência nominal de 60 Hz, e a Figura 3.3 mostra o sinal ENF linear obtido a partir da estimativa HEE, conforme a Equação (3.17).[7].

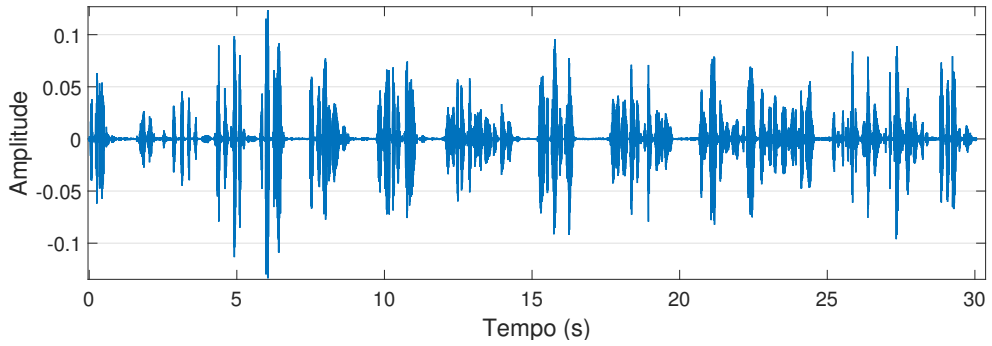


Figura 3.2: Sinal de áudio contendo componente intereferente da rede elétrica

[6]O sinal analítico é obtido para todas as $N_{\mathrm{x}} = N_{\mathrm{s}} + L - 1$ amostras do sinal filtrado em banda passante, porém, após o cálculo de $\hat{\omega}_{\mathrm{H}}(n)$ ignora-se as $L-1$ amostras correspondentes a $n < 0$ e $n \geq N_s$, obtendo-se as estimativas de frequência precisamente nos instantes em que $s(n)$ é definida.

[7]Foi utilizado um filtro passa banda do tipo FIR, de fase zero conforme descrito na Seção 4.1, com largura de banda de 0,8 Hz, utilizando-se as funções do Matlab ® *conv*, nos sentidos direto e reverso, e *fir1*, com 10.000 coeficientes.

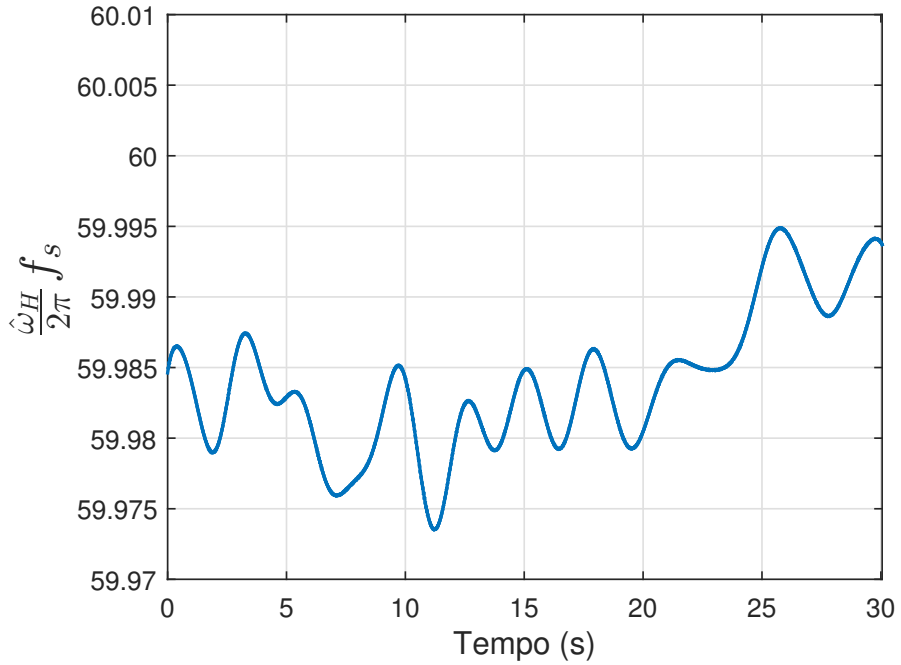


Figura 3.3: Sinal ENF linear obtido a partir da estimativa HEE da frequência angular normalizada $\hat{\omega}_{\mathrm{H}}(n)$ do áudio da Figura 3.2 (abaixo)

3.2.2 Estimador ENF Baseada em ESPRIT (3E)

O 3E é uma abordagem para estimação da ENF que se baseia na decomposição de subespaços (REIS et al., 2016a), e utiliza a propriedade de invariância a rotação para estimação da frequência explorada pela técnica ESPRIT (ROY; KAILATH, 1989; MANOLAKIS; INGLE; KOGON, 2005).

Para tal, inicialmente calcula-se a representação analítica $\hat{x}_{\mathrm{a}}(n)$ do sinal questionado $s(n)$, conforme disposto na Equação (3.6). Dessa forma, seja $\mathbf{X}$ uma matriz de dados $N \times M$ tal que:

$$\mathbf{X} = \left[\mathbf{x_a}(0) \quad \mathbf{x_a}(1) \quad \ldots \quad \mathbf{x_a}(N-2) \quad \mathbf{x_a}(N-1)\right]^{\mathrm{T}}, \tag{3.18}$$

onde $\mathbf{x_a}(n) = [\hat{x}_{\mathrm{a}}(n)\ \hat{x}_{\mathrm{a}}(n+1)\ \ldots \quad \hat{x}_{\mathrm{a}}(n+M-1)]^{\mathrm{T}}$, e $()^{\mathrm{T}}$ diz respeito à operação de transposição de matrizes.

A matriz $\mathbf{X}$ pode ser decomposta utilizando Decomposição em Valores Singulares, do inglês *Singular Value Decomposition* (SVD), resultando em:

$$\mathbf{X} \;=\; \mathbf{LSU}^{\mathrm{H}}, \tag{3.19}$$

onde $\mathbf{L}$ é uma matriz $N \times N$ contendo os valores singulares à esquerda de $\mathbf{X}$, $\mathbf{S}$ é uma matriz $N \times M$ contendo os valores singulares de $\mathbf{X}$ em sua diagonal, $\mathbf{U}$ é uma matriz $M \times M$ contendo os valores singulares à direita de $\mathbf{U}$, e $()^{\mathrm{H}}$ corresponde ao operador Hermitiano.

A matriz $\mathbf{U}$ pode ser decomposta como $\mathbf{U} = [\mathbf{u}_s|\mathbf{U}_{noise}]$, onde $\mathbf{u}_s$, primeira coluna de $\mathbf{U}$, é o vetor coluna que gera o subespaço do sinal, de dimensão $M \times 1$, formado pelo vetor singular que corresponde ao maior valor singular de $\mathbf{X}$. Os demais vetores singulares formam uma matriz cujas colunas correspondem a uma base geradora para o subespaço de ruído $\mathbf{U}_{noise}$, de dimensão $M \times (M-1)$, ortogonal ao subespaço de sinal. É importante observar que a ordem de modelo da matriz de dados $\mathbf{X}$ é igual a um, uma vez que o sinal $\hat{x}(n)$ é modelado como um tom contendo apenas uma componente frequencial de interesse e, portanto, na ausência de ruído corresponderia a uma matriz $\mathbf{X}$ de *rank* unitário.

Seja $\mathbf{u_u}$ e $\mathbf{u_d}$ vetores formados a partir dos primeiros e últimos $M-1$ elementos de $\mathbf{u_s}$, respectivamente. A propriedade de invariância a rotação, explorada pelo ESPRIT, permite escrever que:

$$\mathbf{u_u}\phi \;=\; \mathbf{u_d}, \tag{3.20}$$

onde ϕ é uma exponencial complexa cujo argumento corresponde a frequência angular desejada.

Solucionando a Equação (3.20), obtém-se a estimativa de frequência angular normalizada da rede elétrica pela técnica ESPRIT denominada por $\hat{\omega}_{\mathrm{E}}$, tal que:

$$\hat{\omega}_{\mathrm{E}} \;=\; \angle\frac{\mathbf{u_u}^{\mathrm{H}}\mathbf{u_d}}{\mathbf{u_u}^{\mathrm{H}}\mathbf{u_u}}. \tag{3.21}$$

Diferentemente da abordagem que utiliza a transformada de Hilbert, o método que utiliza a técnica ESPRIT fornece um único parâmetro fixo que corresponde à estimativa ótima sob o critério dos mínimos quadrados para a frequência de oscilação de um sinal senoidal puro que melhor se ajuste aos dados. Com isso, a partir da Equação (3.21)

não se obtém uma estimativa da frequência da rede elétrica a cada instante, mas tão somente um melhor valor que se ajusta à totalidade dos dados. Para obter-se um estimativa da ENF ao longo do tempo utilizando a técnica ESPRIT, pode-se particionar o sinal de entrada $\hat{x}_{\mathrm{a}}(n)$, para $0 \leq n < N_s$, em N_b blocos com N amostras por bloco e sobreposição de $N - M$ amostras. Dessa forma, aplica-se o método utilizando a técnica ESPRIT para cada bloco, visando estimar a frequência angular ao longo do tempo. Como resultado obtém-se uma sequência de amostras $\hat{\omega}_{\mathrm{E}}(n_b)$ que representa o sinal ENF variante no tempo para cada bloco consecutivo. A Figura 3.4 mostra o resultado da aplicação do estimador 3E para o mesmo áudio de 30 segundos correspondente a Figura 3.2. Observa-se que os resultados dos dois estimadores, HEE e 3E são muito semelhantes. No exemplo, foram utilizados os valores de $N = 200$ e $M = 20$[8].

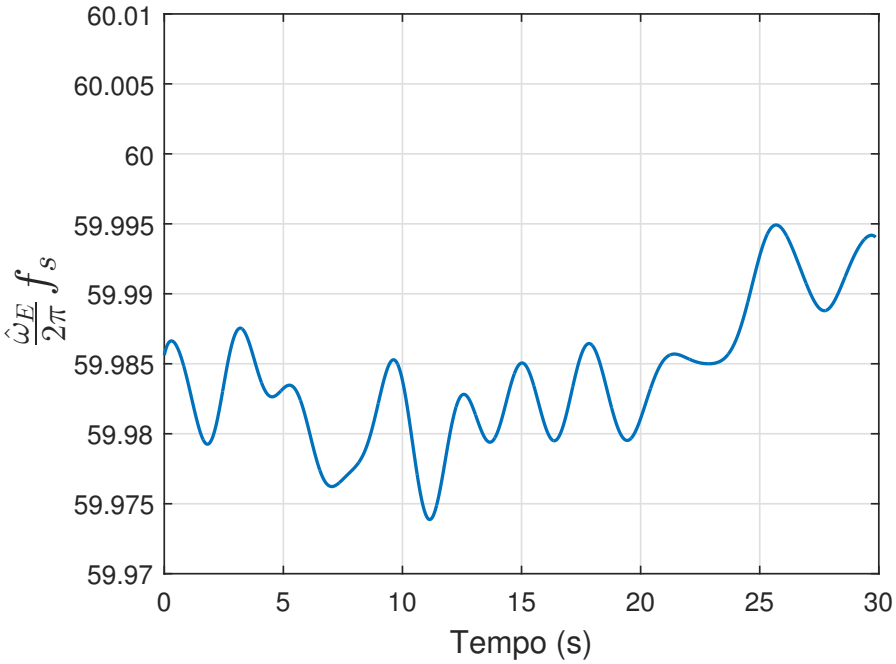


Figura 3.4: Sinal ENF obtido a partir do áudio ilustrado na Figura 3.2 por meio da estimativa $\hat{\omega}_E(n)$

3.3 SUMÁRIO

Neste capítulo foi apresentada a formulação teórica relativa ao modelo de dados utilizado para análise e extração da ENF a partir de um sinal de áudio questionado, visando a compreensão dos métodos utilizados na extração das estimativa ENF. Também foram descritos dois estimadores que correspondem ao estado da arte na extração do sinal ENF.

[8]Foi utilizado um filtro passa banda do tipo FIR, de fase zero, com largura de banda de 0,8 Hz, utilizando-se as funções do Matlab ® *conv* e *fir1*, com 10.000 coeficientes.

Foi apresentada a formulação matemática que permite obter a ENF instantânea, amostra a amostra, por meio da técnica baseada na informação de fase e frequência instantânea da aproximação analítica de Hilbert (HEE). Por último foi apresentada a formulação matemática que permite utilizar a técnica ESPRIT para obter a estimativa da ENF. Para obter uma estimativa da ENF variante ao longo do tempo, a técnica ESPRIT requer que o sinal seja adequadamente dividido em blocos superpostos de tal forma que se obtém uma sequência de amostras do sinal ENF para cada bloco consecutivo. No próximo capítulo os dois estimadores apresentados serão utilizados para propor o método (SPHINS) de detecção de edições a partir de variações anormais do sinal ENF.

4 TÉCNICA PROPOSTA PARA DETECÇÃO DE ADULTERAÇÃO DE ÁUDIO USANDO A ESTIMAÇÃO DA ENF

Neste capítulo é apresentado o SPHINS (REIS et al., 2016b), uma técnica para detecção automática de adulterações em gravações de áudio por meio da constatação de variações anormais na frequência de oscilação de sinais interferentes da rede elétrica (ENF), eventualmente incorporados em registros de áudio questionados. Para tal, o SPHINS explora conjuntamente estimativas da ENF por meio da técnica de estimação da frequência instantânea do sinal analítico de Hilbert (HEE), mais sensível a transições abruptas de fase, e da técnica ESPRIT (3E), mais precisa em baixas condições de SNR e SL.

Diferentemente de outras abordagens, o SPHINS utiliza as estimativas obtidas por ambos os estimadores conjuntamente, gerando um vetor de características que visa a resumir o grau de anomalia das variações da ENF medidas por cada estimador. Para isso, utiliza a curtose amostral das estimativas de cada estimador como elemento do vetor de características que é então aplicado a um classificador SVM devidamente treinado para identificar a presença de edições.

Similarmente aos métodos do estado da arte, o mecanismo proposto requer a validade de algumas suposições (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014):

1. O sinal interferente da rede elétrica é o sinal mais intenso presente em uma estreita região em torno do valor nominal da ENF, garantindo-se uma SNR suficiente na vizinhança espectral deste sinal;
2. Inserções e supressões de segmentos de áudio são realizadas em instantes em que não há atividade vocal;
3. Não há ruídos impulsivos ou distorções não lineares em nível significativo.

A primeira suposição guarda razoabilidade pois os valores nominais da ENF se encontram em frequências em que tal suposição é, via de regra, válida, devido a características

espectrais dos sinais e sistemas envolvidos, como por exemplo o sinal de voz e a RTPC. A segunda suposição é razoável uma vez que inserções e supressões em pontos de corte que coincidem com regiões em que há sinal de voz ativo são facilmente detectáveis, pois deixam traços observáveis por análise espectral (RODRÍGUEZ; APOLINÁRIO JR., 2009) e por alterações na coarticulação, ajuste temporal, prosódia e ritmo da fala, perceptíveis à oitiva (MORISSON; MACHADO; REIS, 2015). A Figura 4.1 mostra o espectrograma de um áudio em que uma supressão foi realizada em região de voz ativa. Pode-se observar diretamente no espectrograma o ponto de corte, onde verifica-se vazamento espectral e uma transição abrupta na evolução dos formantes. Na prática, a oitiva do audio em questão revela uma coarticulação anormal, uma quebra no ritmo e na prosódia da fala que denunciam muito facilmente a ocorrência da edição. A terceira suposição é mais crítica e advém da necessidade de evitar-se interferências temporalmente localizadas fruto de não linearidades que podem impactar em perturbações na ENF estimada.

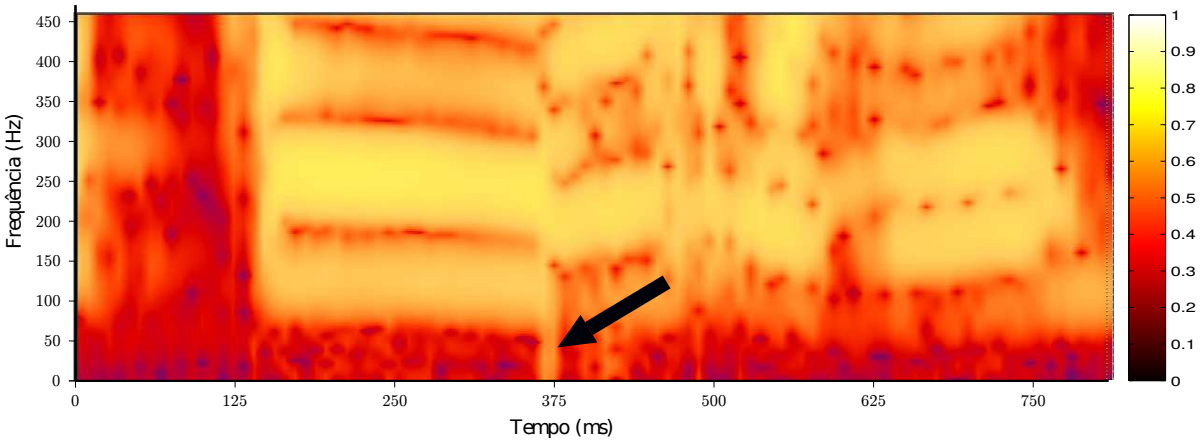


Figura 4.1: Espectrograma de sinal de áudio com ponto de supressão em região de voz ativa.

Apesar das suposições acima imporem algumas restrições, o SPHINS mostra-se eficiente para aplicação em sinais práticos fruto de gravações em que algumas destas premissas encontram-se relaxadas. Dessa forma, no Capítulo 5 o método é avaliado em uma base de dados pública e simulações são realizadas em cenários de baixa SNR e elevado SL.

No decorrer deste capítulo as diferentes etapas do método proposto SPHINS são descritas. No diagrama apresentado na Figura 4.2 ilustra-se o método dividido em três blocos, correspondentes a três etapas do mecanismo de detecção. Na Seção 4.1, descreve-se o primeiro bloco da Figura 4.2, abordando-se os procedimentos relativos ao pré-processamento do sinal questionado. Nesta etapa, o sinal questionado $s(n)$ é pré-processado, obtendo-se como saída o sinal $\hat{x}(n)$ da Equação (3.4) e o parâmetro $\hat{\text{SNR}}_{\text{ENF}}$,

ambos utilizados na como entrada para o segundo bloco. Na Seção 4.2 discute-se o segundo bloco da Figura 4.2, onde a partir do sinal $\hat{x}(n)$ e do parâmetro $\hat{\text{SNR}}_{\text{ENF}}$, extrai-se o vetor de características $\mathbf{F}$. Para isso utiliza-se a curtose amostral das estimativas ENF calculadas pelos métodos HEE e 3E do sinal $\hat{x}(n)$. Finalmente, a etapa de classificação correspondente ao último bloco da Figura 4.2 é detalhada na Seção 4.3, onde um classificador SVM é aplicado ao vetor de características $\mathbf{F}$ obtido no segundo bloco. Como resultado dessa etapa tem-se um índice de classificação c, que vale +1, caso o áudio seja classificado como tendo sido editado, e -1 caso contrário.

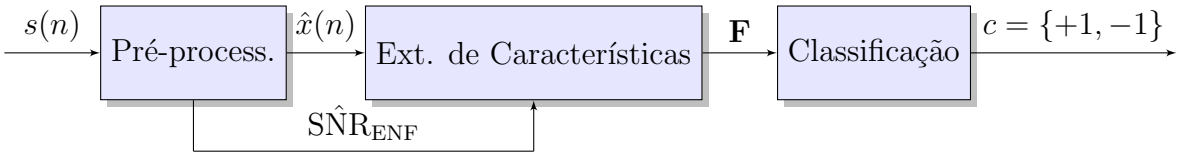


Figura 4.2: Diagrama em blocos ilustrando as três etapas do método proposto

4.1 ETAPA DE PRÉ-PROCESSAMENTO

Seja $s(n)$ na Equação (3.1) o sinal de áudio questionado em que se deseja verificar a presença de traços de edição, e f_{nom} o valor nominal, em Hz, da ENF na localidade em que o áudio foi gravado. Antes das etapas de extração do vetor de características do áudio questionado e sua posterior classificação, $s(n)$ é submetido a uma etapa de pré-processamento dividia em três estágios ilustrados no diagrama em blocos mostrado na Figura 4.3.

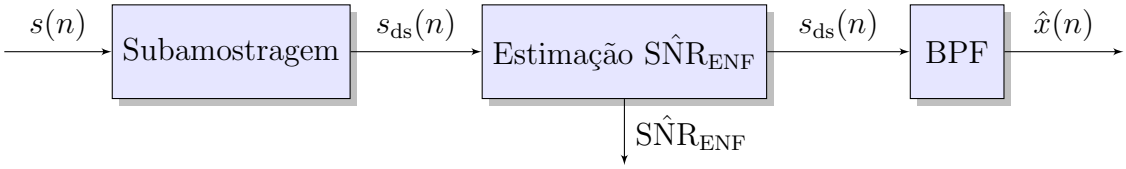


Figura 4.3: Diagrama em blocos da estapa de pré-processamento

No primeiro bloco da Figura 4.3, o sinal questionado $s(n)$ é subamostrado para uma taxa de amostragem f_{s} que seja síncrona ao valor nominal f_{nom} da ENF, tal qual em Rodríguez, APOLINÁRIO JR. & Biscainho (2010). Esta amostragem síncrona garante a existência de um coeficiente DFT exatamente na frequência correspondente ao valor nominal da ENF, o que é desejável para o processamento realizado no segundo bloco. O sinal subamostrado é denominado por $s_{\text{ds}}(n)$. Anteriormente a subamostragem, deve ser realizada a filtragem adequada, do tipo passa baixas com frequência de corte inferior a $f_{\text{s}}/2$ para prevenir a ocorrência de *aliasing*.

Em seguida, no segundo bloco da Figura 4.3, é calculado o parâmetro $\hat{\text{SNR}}_{\text{ENF}}$, que corresponde a uma estimativa simplificada da relação sinal-ruído do sinal ENF em uma vizinhança espectral estreita em torno da f_{nom}. Essa estimativa $\hat{\text{SNR}}_{\text{ENF}}$ será utilizada como indicador da presença do sinal ENF e no processo de extração do vetor de características (Seção 4.2). Caso $\hat{\text{SNR}}_{\text{ENF}}$ seja maior do que um limiar τ, considera-se que o áudio encaminhado possui componente interferente da rede elétrica. Essa abordagem é similar à apresentada em Rodríguez (2010). A determinação do limiar τ é abordada na Seção 5.3.

O cálculo da estimativa $\hat{\text{SNR}}_{\text{ENF}}$ é obtido levando-se em conta que a ENF é o sinal mais intenso confinado em uma vizinhança espectral estreita em torno da f_{nom}, e que o ruído nessa região espectral estreita pode ser bem aproximado por um ruído branco. Assim, com o objetivo de computar $\hat{\text{SNR}}_{\text{ENF}}$, primeiramente calcula-se uma estimativa da densidade espectral de potência $\hat{P}_{\text{ds}}(k)$ do sinal $s_{\text{ds}}(n)$, em dB, onde $k = 0, 1, 2..., N_{FFT}/2$, e o parâmetro N_{FFT} corresponde ao número de pontos utilizado no algoritmo da Transformada Rápida de Fourier, do inglês *Fast Fourier Transform* (FFT). A referida densidade espectral de potência é calculada pelo método baseado na média de periodogramas obtidos por meio da FFT, conforme proposto por Welch (1967). A Figura 4.4 ilustra a densidade espectral de potência $\hat{P}_{\text{ds}}(k)$ obtida em uma vizinhança limitada em torno da frequência nominal f_{nom}, para uma gravação de áudio típica proveniente de escuta telefônica.

Para computar $\hat{\text{SNR}}_{\text{ENF}}$, estima-se o patamar de ruído presente numa estreita região espectral em $\hat{P}_{\text{ds}}(k)$ que corresponde a largura de banda de 3BW_{ENF} centralizada em tono da frequência nominal f_{nom}, como por exemplo em 60 Hz. Uma vez que o sinal ENF é modelado como um sinal real senoidal com frequência instantânea que excursiona entre valores confinados em uma estreita margem em torno da f_{nom}, estima-se também o valor de pico de $\hat{P}_{\text{ds}}(k)$ numa vizinhança de largura BW_{ENF} centralizada em f_{nom}. Como foi realizada uma subamostragem síncrona, haverá coeficiente localizado exatamente na frequência correspondente ao valor nominal da ENF, o que favorece, juntamente com uma elevação do número de pontos da FFT, a uma detecção mais precisa do valor de pico da densidade espectral de potência. O valor do parâmetro $\hat{\text{SNR}}_{\text{ENF}}$ é em seguida calculado como sendo a diferença, em dB, entre o valor de pico da densidade espectral de potência e o valor correspondente ao patamar de ruído.

Formalmente, seja $\hat{P}_{\text{ENF}}$ o valor de pico da $\hat{P}_{\text{ds}}(k)$ em uma região de largura de banda estreita BW_{ENF}, e seja $\hat{P}_{\text{noise}}$ o patamar de ruído desejado. Dessa forma, pode-se

escrever que:

$$\begin{aligned} \hat{P}_{\text{ENF}} &= \max\left[\hat{P}_{\text{ds}}(k)\right],\ k \in \mathbf{\Omega_1}, \\ \hat{P}_{\text{noise}} &= \frac{1}{|\mathbf{\Omega_2} - \mathbf{\Omega_1}|} \sum_{k \in (\mathbf{\Omega_2} - \mathbf{\Omega_1})} \hat{P}_{\text{ds}}(k), \end{aligned} \tag{4.1}$$

onde max[·] corresponde a função que retorna o valor máximo em um conjunto de dados, $|\cdot|$ corresponde ao operador de cardinalidade que retorna o número de elementos de um conjunto, e $\mathbf{\Omega_1}$ e $\mathbf{\Omega_2}$ são os subconjuntos:

$$\begin{aligned} \mathbf{\Omega_1} &= \left\{ \frac{N_{\text{FFT}}}{f_{\text{s}}}(f_{\text{nom}} - \frac{\text{BW}_{\text{ENF}}}{2}) \leq k \leq \frac{N_{\text{FFT}}}{f_{\text{s}}}(f_{\text{nom}} + \frac{\text{BW}_{\text{ENF}}}{2}) \right\}, \\ \mathbf{\Omega_2} &= \left\{ \frac{N_{\text{FFT}}}{f_{\text{s}}}(f_{\text{nom}} - \frac{3\text{BW}_{\text{ENF}}}{2}) \leq k \leq \frac{N_{\text{FFT}}}{f_{\text{s}}}(f_{\text{nom}} + \frac{3\text{BW}_{\text{ENF}}}{2}) \right\}. \end{aligned} \tag{4.2}$$

Na Figura 4.4, a área hachurada corresponde ao subconjunto $\mathbf{\Omega_2} - \mathbf{\Omega_1}$ e a área não hachurada ao subconjunto $\mathbf{\Omega_1}$, para uma largura de banda $\text{BW}_{\text{ENF}} = 0,8$ Hz e uma frequência nominal $f_{\text{nom}} = 60$ Hz. Para finalmente computar o valor do parâmetro SNR_{ENF}, simplesmente calcula-se a diferença entre os dois termos da Equação (4.1):

$$\hat{\text{SNR}}_{\text{ENF}} = \hat{P}_{\text{ENF}} - \hat{P}_{\text{noise}}. \tag{4.3}$$

Para determinar o limiar τ acima do qual considera-se que há sinal interferente da rede elétrica, necessita-se de uma base de dados de áudios sem sinal ENF. A partir de tal base obtém-se os valores de $\hat{\text{SNR}}_{\text{ENF}}$ para os áudios correspondentes e calcula-se τ como sendo igual a média mais três desvios-padrão amostrais[9], ou seja:

$$\tau = \text{mean}\left(\hat{\text{SNR}}_{\text{ENF}}\right) + 3\sqrt{\text{var}\left(\hat{\text{SNR}}_{\text{ENF}}\right)}, \tag{4.4}$$

onde mean(·) e var(·) correspondem , respectivamente, à média e à variância amostrais, calculadas para o conjunto de parâmetros $\hat{\text{SNR}}_{\text{ENF}}$ da base sem sinal ENF.

Por último, no terceiro bloco da Figura 4.3, o sinal sub-amostrado $s_{\text{ds}}(n)$ é filtrado com

[9]Para sinais sem ENF, o parâmetro $\hat{\text{SNR}}_{\text{ENF}}$, em dB, tem sua distribuição modelada por uma normal, de tal forma que o valor proposto para o limiar τ corresponde a uma taxa de falso positivo de cerca de 0,3 % para a presença da ENF.

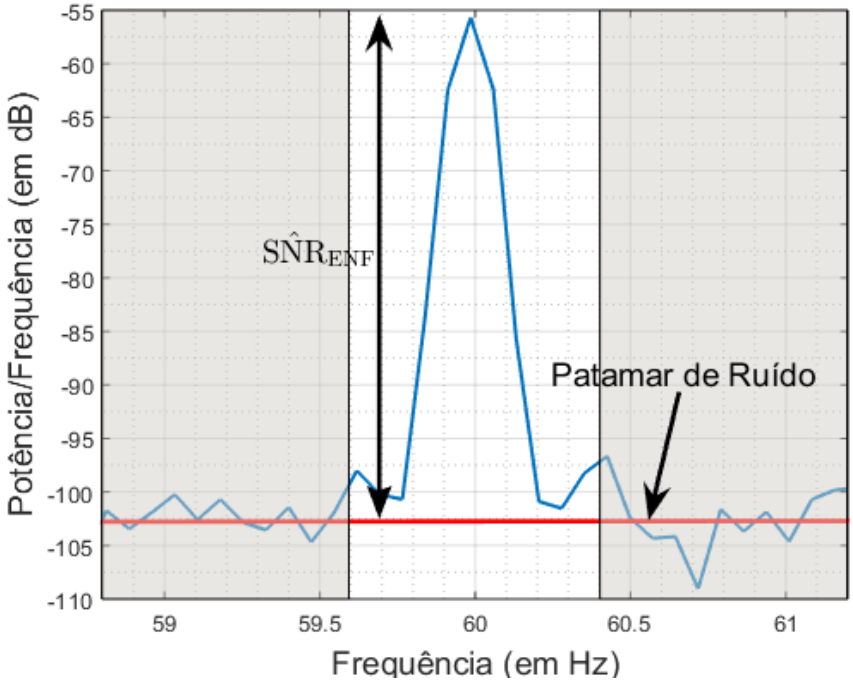


Figura 4.4: Estimativa da Densidade Espectral de Potência $\hat{P}_{\text{ds}}(k)$ em vizinhança espectral estreita em torno da frequência nominal da ENF para uma gravação de áudio tipicamente proveniente de escuta telefônica. A área hachurada corresponde ao subconjunto $\boldsymbol{\Omega_2} - \boldsymbol{\Omega_1}$ e a área não hachurada corresponde ao subconjunto $\boldsymbol{\Omega_1}$.

um filtro passa-banda de resposta ao impulso finita, do inglês Finite Impulse Response (FIR), de fase zero, com uma largura de banda estreita igual à BW_{ENF} e centrada em torno da frequência nominal f_{nom}, tal que o sinal filtrado em banda passante obtido corresponde ao sinal $\hat{x}(n)$ na Equação (3.5). O filtro em questão foi projetado utilizando o método do janelamento, valendo-se de janelas de Hamming, e utilizando a técnica da reversão temporal para encontrar uma versão de fase zero.

Inicialmente projeta-se um filtro passa banda FIR $h_{\text{FIR}}(n)$ de fase linear com $Q+1$ coeficientes, com frequências de corte normalizadas ω_{c1} e ω_{c2}, por meio do janelamento da resposta ao impulsos de um filtro ideal, tal que:

$$h_{\text{FIR}}(n) = \left(\frac{\sin[\omega_{c2}(n - Q/2)]}{\pi(n - Q/2)} - \frac{\sin[\omega_{c1}(n - Q/2)]}{\pi(n - Q/2)} \right) W(n), \tag{4.5}$$

onde $W(n)$ corresponde a janela de Hamming com $Q+1$ coeficientes , tal que:

$$W(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/Q), & 0 \leq n \leq Q, \\ 0, & \text{caso contrário.} \end{cases} \tag{4.6}$$

Os valores das frequências normalizadas ω_{c1} e ω_{c2} são calculados por:

$$\begin{aligned} \omega_{c1} &= \tfrac{2\pi}{f_\text{s}}(f_\text{nom} - \text{BW}_\text{ENF}/2) \\ \omega_{c2} &= \tfrac{2\pi}{f_\text{s}}(f_\text{nom} + \text{BW}_\text{ENF}/2) \end{aligned} \tag{4.7}$$

O correspondente filtro de fase zero é obtido por meio da técnica de reversão temporal, tal que:

$$\begin{aligned} y_1(n) &= h_\text{FIR}(n) * s_\text{ds}(n) \\ y_2(n) &= h_\text{FIR}(n) * y_1(-n) \\ \hat{x}(n) &= y_2(-n), \end{aligned} \tag{4.8}$$

de tal forma que, substituindo as equações em (4.8), tem-se:

$$\begin{aligned} \hat{x}(n) &= h_\text{FIR}(-n) * y_1(n) \\ \hat{x}(n) &= [h_\text{FIR}(-n) * h_\text{FIR}(n)] * s_\text{ds}(n) \\ \hat{x}(n) &= h_\text{bp}(n) * s_\text{ds}(n), \end{aligned} \tag{4.9}$$

onde o filtro FIR de fase zero corresponde a:

$$h_\text{bp}(n) = h_\text{FIR}(-n) * h_\text{FIR}(n). \tag{4.10}$$

Observe-se que $h_\text{bp}(n)$ é um filtro com $L = 2Q + 1$ coeficientes, não causal, contendo coeficientes diferentes de zero para para valores de $-Q \leq n \leq Q$

4.2 EXTRAÇÃO DO VETOR DE CARCATERÍSTICAS

Nessa seção discutimos a etapa correspondente ao segundo bloco da Figura 4.2. A principal contribuição desta dissertação é a proposta de um novo vetor de características para descrever as perturbações anômalas na ENF produzidas por inserções e supressões

de segmentos de áudio em gravações, bem como a incorporação de técnicas de aprendizado de máquina para resolver a detecção de distúrbios na ENF estimada que indiquem a existência de edições. Após a fase de pré-processamento, utiliza-se os dois estimadores da ENF descritos no Capítulo 3, denominados HEE e 3E, resultando em duas estimativas diferentes para a variação da ENF ao longo do tempo, $\hat{\omega}_{\mathrm{H}}(n)$ e $\hat{\omega}_{\mathrm{E}}(n_b)$, respectivamente.

A fim de promover uma comparação direta entre as duas estimativas que se permita visualizar a diferença entre a sensibilidade dos dois estimadores a eventuais perturbações na ENF, considera-se que o estimador HEE produz estimativas bloco a bloco, e não amostra a amostra, utilizando as estimativas tomadas no centro de cada bloco, como a seguir:

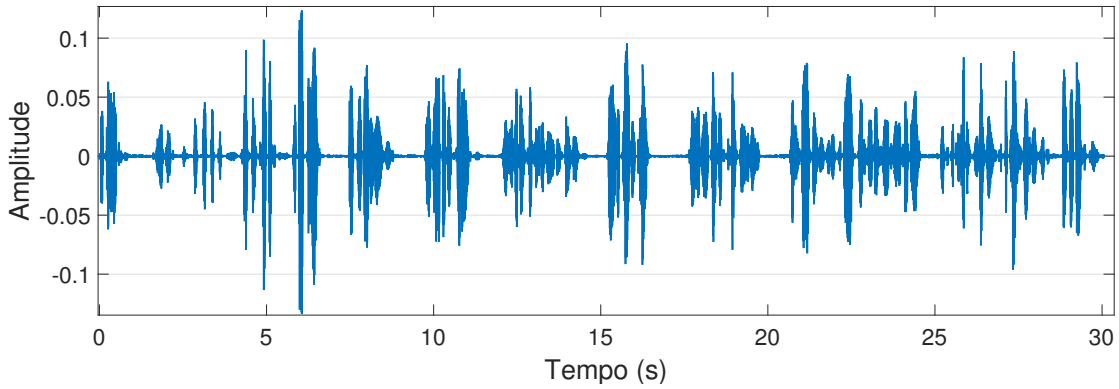
$$\hat{\omega}_{\mathrm{H_b}}(n_b) = \hat{\omega}_{\mathrm{H}}\left(\left\lceil n_b M + \frac{N}{2} - 1 \right\rceil\right), \tag{4.11}$$

onde $\lceil \cdot \rceil$ representa a função *teto* cujo valor de retorno corresponde ao menor inteiro maior ou igual ao respectivo operando.

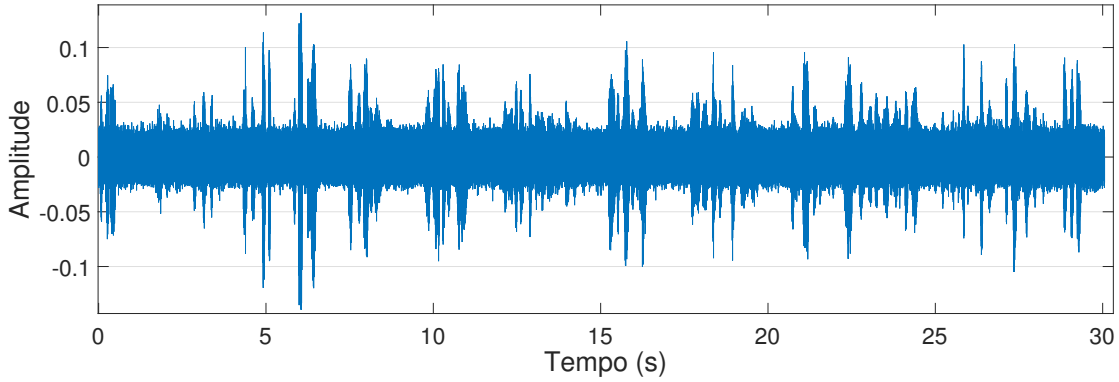
Em condições adversas de relação sinal ruído observa-se que os estimadores apresentam significativa diferença nas estimativas produzidas. Para ilustrar tal diferença o mesmo áudio retratado na Figura 3.2 foi degradado para uma SNR = 5 dB (Figura 4.5). Foi adicionado ruído branco, Gaussiano, de média nula, e para atingir a SNR de 5 dB foi utilizada a mesma abordagem e o mesmo algoritmo de detecção de atividade de voz, do inglês *Voice Activity Detection* VAD, empregado em Esquef, APOLINÁRIO JR. & Biscainho (2014) e Esquef, APOLINÁRIO JR. & Biscainho (2015).

As estimativas HEE e 3E para o áudio degradado encontram-se ilustradas na Figura 4.6 onde é possível verificar uma diferença relevante nas estimativas. Observa-se uma maior sensibilidade das estimativa HEE à presença de ruídos, e por conseguinte uma maior tendência a apresentar variações anormais e abruptas na ENF estimada para áudios ruidosos, conforme pode ser constatado na superposição ilustrada na Figura 4.7.

Em situações de inserções ou supressões de segmentos de áudio em que há descontinuidade de fase no sinal ENF, haverá uma perturbação na ENF estimada. Da mesma forma, o estimador HEE apresenta uma maior sensibilidade na detecção desses espúrios, que se traduz na produção de variações abruptas nas estimativas. Na Figura 4.8 tem-se a ilustração dos resultados dos dois estimadores superpostos (acima) e de sua diferença



(a) Áudio original



(b) Áudio degradado

Figura 4.5: Sinal de áudio da Figura 3.2 (a) degradado com ruído aditivo, branco, Gaussiano, de média nula, para uma SNR de 5 dB (b).

(abaixo) quando aplicados a um sinal de áudio de 28 segundos, com SNR estimada de 28 dB, com supressão de um segmento de áudio. Também nesse caso, os dois estimadores produzem resultados semelhantes, capturando conjuntamente as variações da ENF ao longo do tempo, exceto no ponto de supressão, onde pode ser observada uma diferença significativa.

O estimador HEE mostra-se bem mais sensível a descontinuidades bruscas de fase do que o estimador 3E. De fato, a abordagem parametrizada e por decomposição de subespaços do 3E tende a forçar a estimação em um modelo perfeitamente senoidal para cada bloco (subespaço de sinal), e descartando informações espúrias que não se encaixam nesse modelo parametrizado (subespaço de ruído). Como as estimativas baseadas na frequência instantânea por meio da transformada de Hilbert são sensíveis a ruído, tendendo a apresentar variações abruptas da ENF em cenários de baixa SNR ainda que não haja descontinuidade de fase por conta da supressão e inserção de segmentos, espera-se obter uma identificação mais robusta das perturbações da ENF por meio da aplicação conjunta de ambos os estimadores.

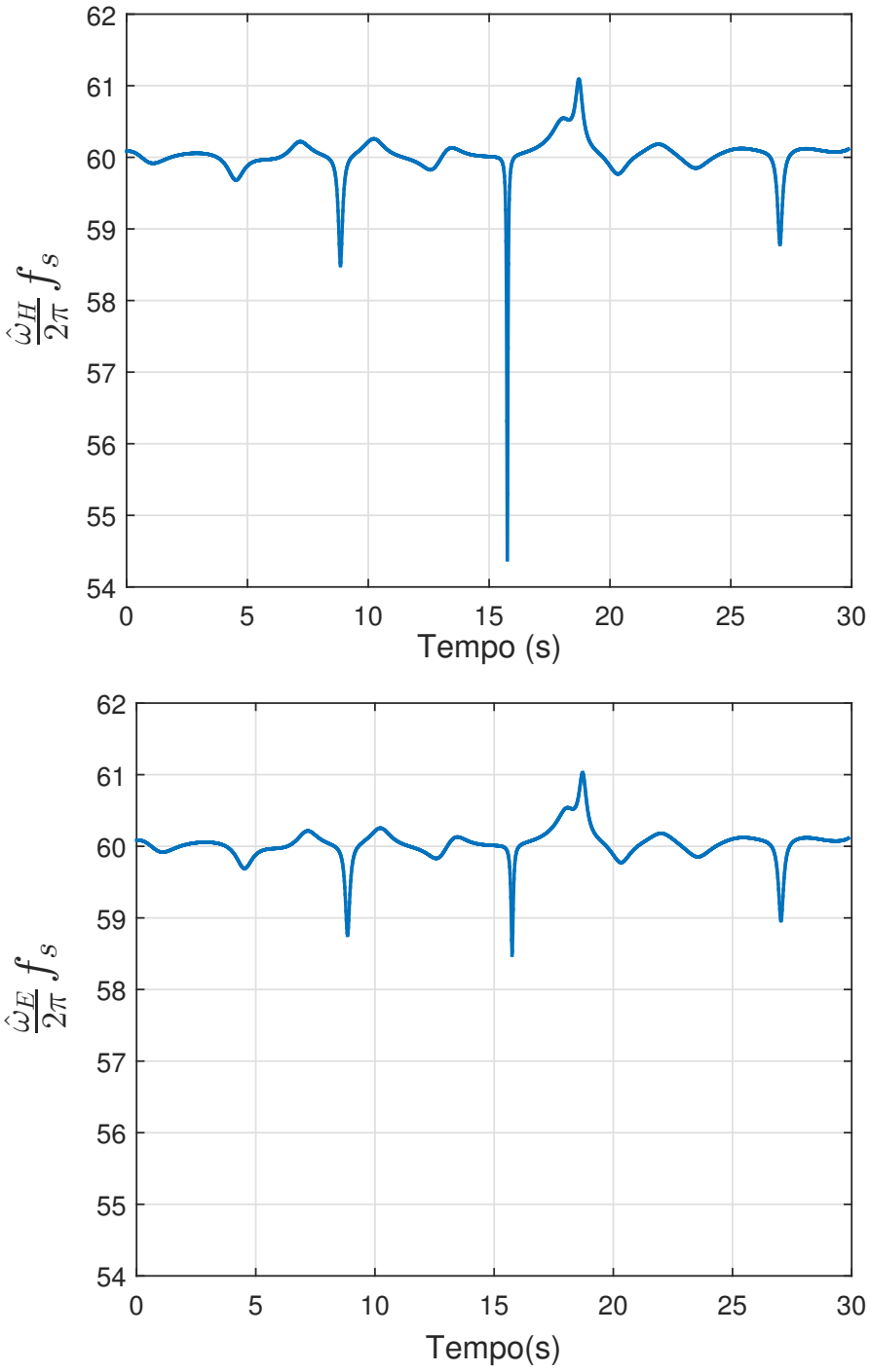


Figura 4.6: Estimador HEE (acima) vs. Estimador 3E (abaixo): comparação de estimativas para áudio degradado à SNR de 5 dB

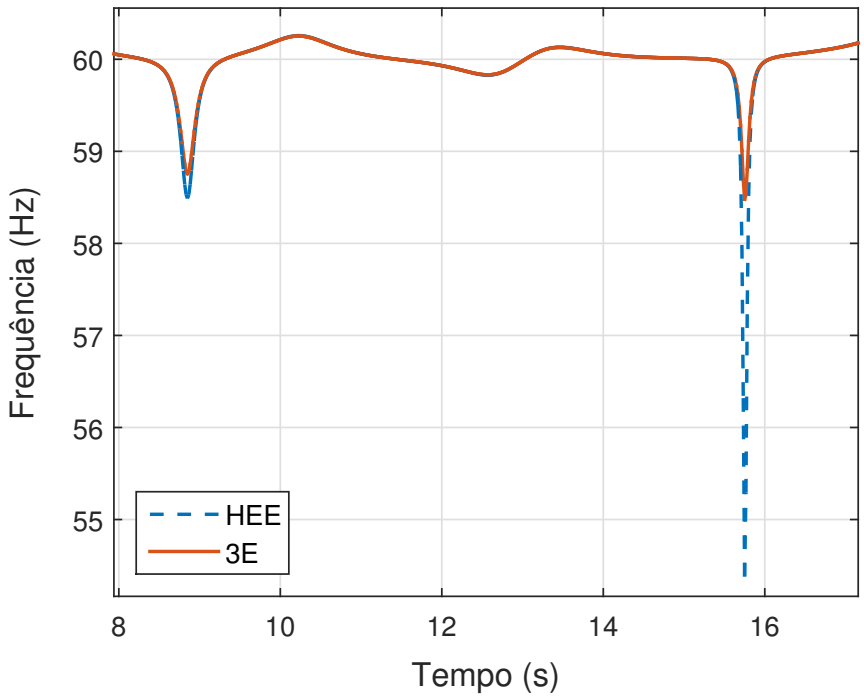


Figura 4.7: Superposição das estimativas HEE e 3E para áudio degradado à SNR de 5 dB

Para classificar automaticamente um sinal de áudio como editado ou não editado, faz necessária alguma medida característica que traga informações sobre as variações anormais na estimativa do sinal. Com este propósito, o SPHINS utiliza a curtose amostral das estimativas como uma medida do grau de anomalia produzido por valores extremos (REIS et al., 2016b, 2016a).

Diferentemente da variância, a curtose é um parâmetro independente de escala que mede o quão caudal é uma distribuição. Dessa forma, a presença de valores anômalos na distribuição de dados, simétricos ou assimétricos, tem a propriedade de aumentar a curtose original, de tal forma que seu valor fornece uma medida da influência de valores extremos na variância global dos dados (PEÑA; PRIETO, 2012).

A curtose amostral de $\hat{\omega} = \{\hat{\omega}(n_b) : n_b = 1, 2 \ldots N_b\}$ pode ser definida como a relação entre o quarto momento central amostral e variância amostral ao quadrado:

$$\kappa(\hat{\omega}) = \frac{(1/N_b)\sum_{n_b=1}^{N_b}(\hat{\omega}(n_b) - \bar{\hat{\omega}})^4}{[(1/N_b)\sum_{n_b=1}^{N_b}(\hat{\omega}(n_b) - \bar{\hat{\omega}})^2]^2}, \tag{4.12}$$

onde $\bar{\hat{\omega}} = (1/N_b)\sum_{n_b=1}^{N_b}\hat{\omega}(n_b)$.

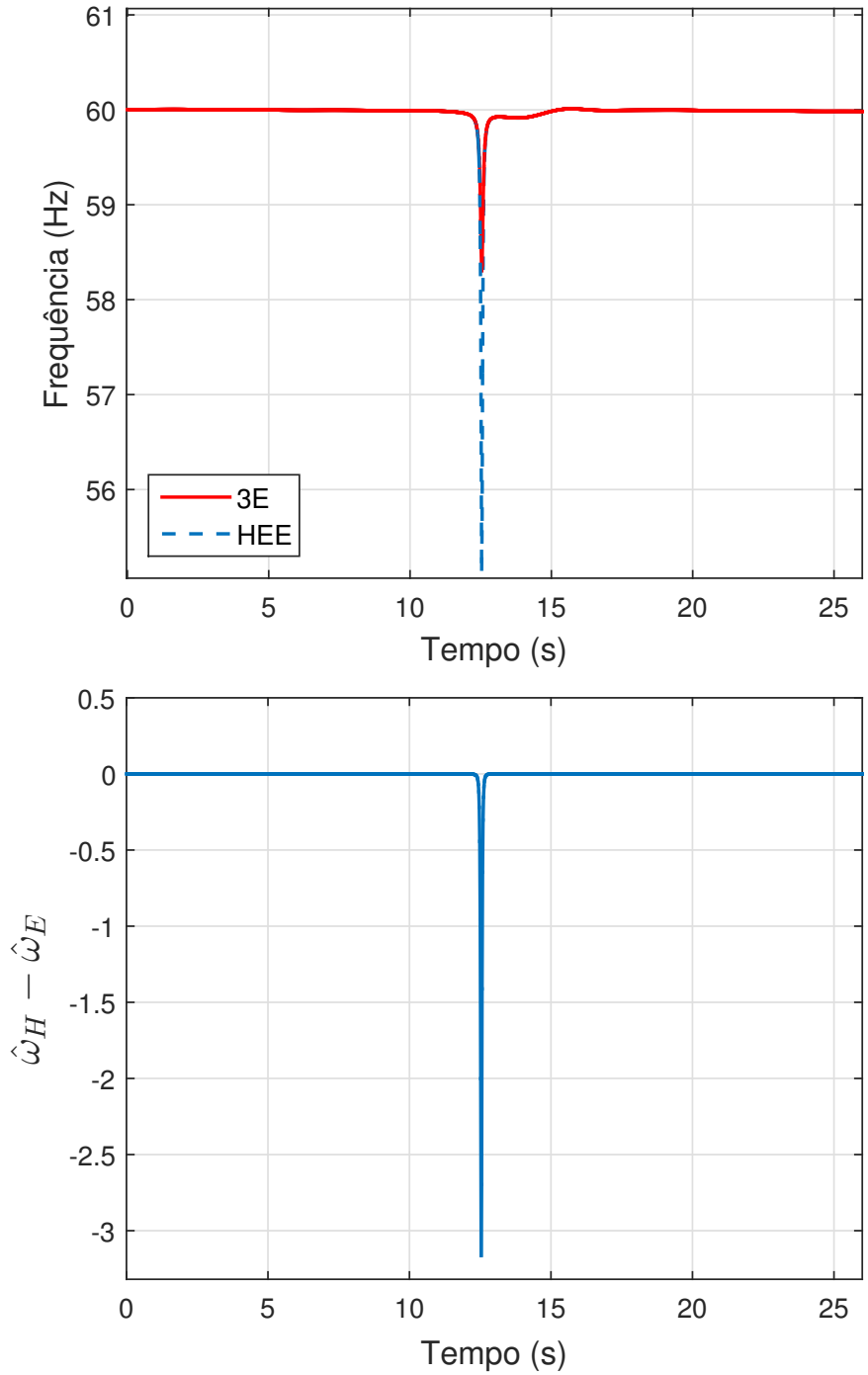


Figura 4.8: Estimador HEE vs. Estimador 3E: superposição das estimativas da ENF em áudio de 28 segundos, com 28 dB de relação sinal-ruído, com supressão de um segmento de áudio

Usando a curtose amostral como uma medida do grau de anomalia das estimativa HEE e 3E, e SNR_{ENF} como fator de ponderação, o vector de características proposto é definido por:

$$\mathbf{F} = \left[\kappa(\hat{\omega}_{\mathrm{H_b}}) \quad \kappa(\hat{\omega}_{\mathrm{E}}) \quad \log_{10}\left[\kappa(\hat{\omega}_{\mathrm{H_b}} - \hat{\omega}_{\mathrm{E}})\right] + \hat{\mathrm{SNR}}_{\mathrm{ENF}}/10\right]^{\mathrm{T}}, \tag{4.13}$$

onde $\kappa(\cdot)$ é a curtose amostral como definido em (4.12), e $\mathbf{F} \in \mathbb{R}^3$ é o vetor de características que resume as variações anômalas da ENF para uma gravação de áudio arbitrária.

4.3 CLASSIFICAÇÃO UTILIZANDO VETOR DE CARACTERÍSTICAS BASEADO NA CURTOSE

Nessa seção discutimos a etapa correspondente ao último bloco da Figura 4.2. Para classificar um áudio como adulterado, treina-se um classificador SVM (BISHOP, 2006) a partir de um conjunto de treinamento $(\mathbf{F}_i, c_i)$ com um número igual de gravações de áudio editadas e não editadas, a partir de um banco de dados conhecido, onde $\mathbf{F}_i$ é o vetor de características definido na Equação (4.13), e $c_i = \{+1, -1\}$ é o índice de classe correspondente à i-ésima gravação de áudio, para $i = 1, \ldots, N_{ar}$, com N_{ar} sendo o número total de gravações de áudio no banco de dados .

Mais especificamente, treina-se um classificador encontrando-se a função discriminante $g(\cdot) : \mathbb{R}^3 \to \mathbb{R}$ e um limiar t, tal que:

$$g(\mathbf{F}_i) = \begin{cases} \geq t, \ c_i = +1 \\ < t, \ c_i = -1. \end{cases} \tag{4.14}$$

Na fase de treinamento, o SVM calcula o hiperplano de máxima margem de separação no espaço de características, que separe os dados de duas classes com a máxima distância para o vetor mais próximo. Este hiperplano é totalmente determinado por um vetor $\mathbf{a}$, normal ao hiperplano e que regula seu posicionamento no espaço de características, e por um escalar b, que controla a distância desse hierplano em relação à origem. Na fase de treinamento, a partir de um conjunto de dados de classes distintas determina-se estes dois parâmetros (BISHOP, 2006), de tal modo que:

$$g(\mathbf{F}_i) = \mathbf{a}^{\mathrm{T}}\mathbf{F}_i + b, \tag{4.15}$$

onde a relação $\mathbf{a}^{\mathrm{T}}\mathbf{F} + b = 0$ é valida para os vetores $\mathbf{F}$ que no espaço de características eventualmente pertençam ao hiperplano de máxima margem de separação. Via de

regra, utiliza-se como limiar o valor de $t = 0$, o que corresponde a tomar o próprio hiperplano de máxima margem como superfície de decisão, uma vez que a maximização da margem de separação correlaciona-se com a minimização do erro de generalização do classificador[10]. A Figura 4.9 ilustra a região de decisão no caso de um espaço de características em duas dimensões. O hiperplano de máxima separação está representado por uma linha cheia. A linha tracejada corresponde a região de decisão para um limiar t arbitrário. Uma vez definido o hiperplano de máxima margem de separação, a região de decisão corresponderá a um hiperplano paralelo cuja distância é controlada pelo valor t.

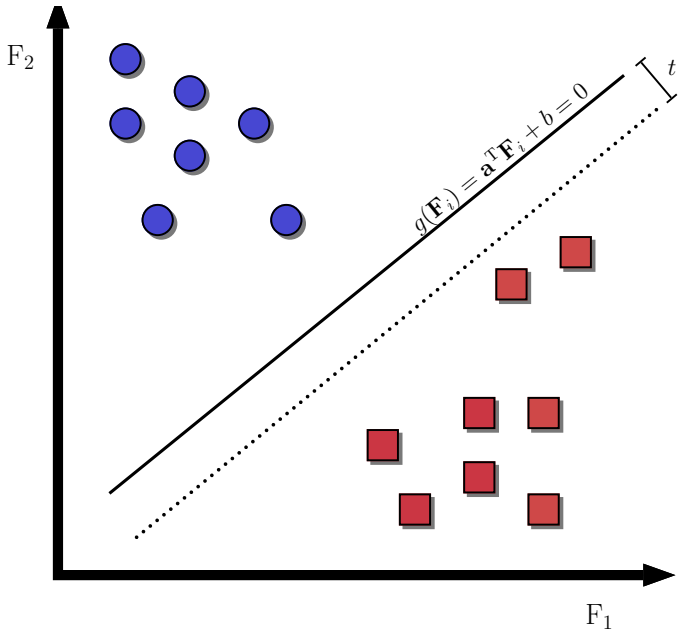


Figura 4.9: Hiperplano de máxima separação em uma classificador SVM (linha cheia) e a região de decisão correspondente para um limiar t arbitrário (linha tracejada).

A determinação do hiperplano de máxima margem de separação corresponde a um processo de minimização de $\| \mathbf{a} \|^2$, sujeito à restrição (BISHOP, 2006):

$$c_i(\mathbf{a}^{\mathrm{T}}\mathbf{F}_i + b) \geq 1, \quad i = 1, \ldots, N_{\mathrm{ar}} \tag{4.16}$$

A solução desse problema de otimização pode ser obtida pela introdução de multiplicadores de Lagrange $\lambda_i \geq 0$, minimizando-se a seguinte função de Lagrange:

[10]Generalização é a capacidade do classificador em classificar dados novos, diferentes daqueles com que foi treinado. Erro de generalização pode ser entendido como a taxa de erro do classificador ao classificar dados novos.

$$L(\mathbf{a}, b, \lambda) = \frac{1}{2} \parallel \mathbf{a} \parallel^2 - \sum_{i=1}^{N_{ar}} \lambda_i \left[c_i(\mathbf{a}^{\mathrm{T}}\mathbf{F}_i + b) - 1 \right], \tag{4.17}$$

onde $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{N_{ar}}]^{\mathrm{T}}$. Derivando-se em relação a b e $\mathbf{a}$ e igualando-se a zero, obtém-se:

$$\begin{array}{l} \mathbf{a} = \sum_{i=1}^{N_{ar}} \lambda_i c_i \mathbf{F}_i \\ \sum_{i=1}^{N_{ar}} \lambda_i c_i = 0, \end{array} \tag{4.18}$$

onde, substituindo-se em (4.17), obtém-se a seguinte função Lagrangiana:

$$\tilde{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{N_{ar}} \lambda_i - \frac{1}{2} \sum_{i=1}^{N_{\mathrm{ar}}} \sum_{j=1}^{N_{\mathrm{ar}}} \lambda_i \lambda_j c_i c_j \mathbf{F}_i^{\mathrm{T}} \mathbf{F}_j, \tag{4.19}$$

que deve ser maximizada em relação a $\boldsymbol{\lambda}$, sujeita as restrições:

$$\begin{array}{ll} \lambda_i \geq 0, & i = 1, \ldots, N_{ar} \\ \sum_{i=1}^{N_{ar}} \lambda_i c_i = 0. & \end{array} \tag{4.20}$$

Dessa forma, o processo de treinamento consiste em resolver um problema de otimização quadrática, determinando-se $\boldsymbol{\lambda}$. Para avaliar-se o valor de $g(\mathbf{F})$, aplica-se (4.18) em (4.15), onde:

$$g(\mathbf{F}) = \sum_{i=1}^{N_{ar}} \lambda_i c_i \mathbf{F}_i^{\mathrm{T}} \mathbf{F} + b. \tag{4.21}$$

É interessante notar que, na prática, tanto no treinamento na Equação (4.19), onde encontra-se os valores $\boldsymbol{\lambda}$, quanto na classificação de novos dados na Equação (4.21), são utilizados apenas escalares correspondente aos produtos internos $\langle \mathbf{F}_i, \mathbf{F}_j \rangle = \mathbf{F}_i^{\mathrm{T}} \mathbf{F}_j$ dos vetores de características.

Para o caso de dados separáveis de forma não linear, a superfície ótima de separação não corresponderá a um plano, necessitando-se de modificações no algoritmo SVM linear.

Para ajudar a entender como o SVM lida com casos não lineares, pode-se imaginar que nos casos de dados separáveis de forma não linear a superfície de máxima margem de separação pode ser encontrada a partir de novos vetores de características $\eta(\mathbf{F}_i)$ que correspondem ao mapeamento dos vetores originais para um espaço vetorial de dimensão superior, onde os dados são linearmente separáveis. A função $\eta(\cdot) : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$, com $d_1 < d_2$, é a função responsável por realizar o mapeamento dos vetores característicos originais para o espaço vetorial de dimensão superior. Nesse novo espaço vetorial, sendo os dados linearmente separáveis, pode-se proceder tal qual no caso linear e encontrar o hiperplano de máxima separação, ou seja, em última análise encontrar a função discriminante $g(\cdot)$ tal que:

$$g(\mathbf{F}_i) = \mathbf{a}^{\mathrm{T}}\eta(\mathbf{F}_i) + b, \tag{4.22}$$

onde a relação $g(\mathbf{F}) = 0$ será válida para os pontos $\eta(\mathbf{F})$ que eventualmente pertençam à superfície de máxima margem de separação. A Figura 4.10 ilustra o processo de mapeamento de um conjunto de dados separáveis de forma não linear para um espaço de dimensão maior onde os dados são linearmente separáveis.

Na prática, como o algoritmo SVM utiliza somente o produto interno entre os vetores característicos para encontrar e avaliar a função discriminante $g(\cdot)$, conforme Equações (4.19) e (4.21), não é necessário efetuar-se o mapeamento de cada vetor característico para um espaço dimensionalmente maior, o que implicaria num custo computacional elevado. Ao invés disso, utiliza-se o truque do *Kernel*, cujo objetivo é computar os escalares correspondentes ao produtos internos dos vetores característicos mapeados diretamente a partir dos vetores característicos originais, valendo-se de uma função com propriedades de métrica de similaridade, denominada função Kernel, tal que:

$$K(\mathbf{F}_i, \mathbf{F}_j) = \langle \eta(\mathbf{F}_i), \eta(\mathbf{F}_j) \rangle \tag{4.23}$$

onde $K(\cdot, \cdot)$ é a função *Kernel* (BISHOP, 2006). De fato, a função de mapeamento sequer é definida, sendo somente estabelecida qual métrica de similaridade será utilizada como função *Kernel*, o que permite reescrever as Equações (4.19) e (4.21) para o caso não linear como:

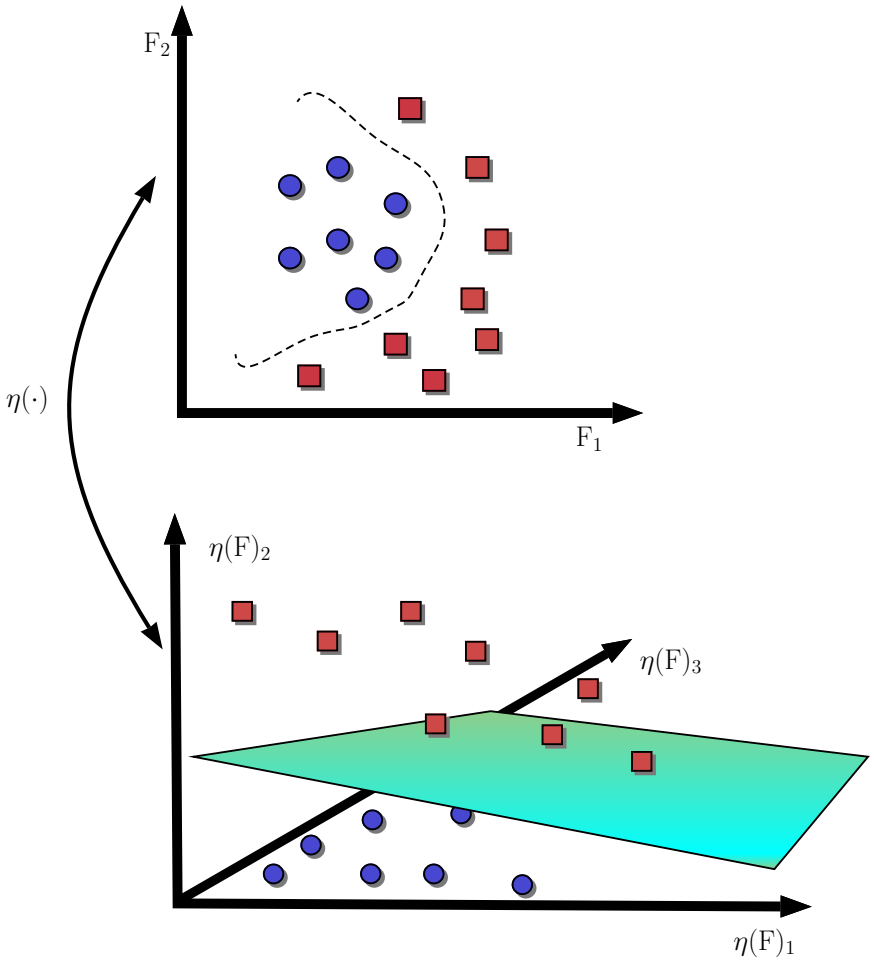


Figura 4.10: SVM para o caso não linear: mapeamento para um espaço de características com dimensão superior para aplicação do SVM linear.

$$\tilde{L}(\lambda) = \sum_{i=1}^{N_{ar}} \lambda_i - \frac{1}{2} \sum_{i=1}^{N_{\mathrm{ar}}} \sum_{j=1}^{N_{\mathrm{ar}}} \lambda_i \lambda_j c_i c_j K(\mathbf{F}_i, \mathbf{F}_j), \tag{4.24}$$

$$g(\mathbf{F}) = \sum_{i=1}^{N_{ar}} \lambda_i c_i K(\mathbf{F}_i, \mathbf{F}) + b. \tag{4.25}$$

Existem diversos tipos de função *Kernel* que podem ser utilizadas. O método proposto neste trabalho utiliza como *Kernel* uma função de base radial, do inglês *Radial Basis Function* (RBF), Gaussiana, definida como se segue:

$$K(\mathbf{F}_i, \mathbf{F}_j) = \exp\left(-\frac{||\mathbf{F}_i - \mathbf{F}_j||^2}{2\sigma^2}\right), \tag{4.26}$$

onde σ e um parâmetro a ser determinado, via de regra, empiricamente, e que dá

grande flexibilidade ao kernel empregado, ao passo em que controla o compromisso do classificador entre variância e polarização (BISHOP, 2006).

Tipicamente o valor ótimo de σ é determinado por meio de testes de validação cruzada. O compromisso entre variância e polarização é tal que valores muito baixos da variável σ tendem a fazer o classificador se ajustar demais aos dados de treinamento porém perdendo a capacidade de generalização, ocasionando *overfitting*, de tal forma que para dados diferentes dos dados de treinamento haja um mau desempenho. Por outro lado, valores muito altos de σ tendem a ajustar pouco o classificador aos dados de treinamento, ocasionando *underfitting*, o que também pode se traduzir em baixo desempenho. A Figura 4.11 ilustra a regiões de decisão para um classificador em casos de *overfitting* (acima) e *underfitting* (meio), bem como para um ajuste ótimo intermediário (abaixo).

Uma gravação de áudio que resulte em um vetor característico $\mathbf{F_i}$ é classificada como pertencente a classe $c_i = +1$ se o valor retornado pela função $g(\mathbf{F}_i) \geq t$, e como $c_i = -1$, caso contrário. O limiar t pode ser ajustado segundo algum critério de interesse. Neste trabalho, para fins de comparação com trabalhos correlatos (RODRÍGUEZ; APOLINÁRIO JR.; BISCAINHO, 2010; ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014, 2015; FUENTES et al., 2016), define-se o limiar de decisão t como sendo aquele cujo ponto de operação corresponde à taxa de erro igual, do inglês *Equal Error Rate* (EER), ou seja, $t = t_{\text{EER}}$ é o limiar que produz no conjunto de treinamento a mesma taxa de erros do tipo falso positivo e do tipo falso negativo, ou em outras palavras, produz a mesma taxa de falso positivo, do inglês *False Positive Rate* (FPR), e a mesma taxa de falso negativo, do inglês *False Negative Rate* (FNR).

Um falso positivo corresponde à classificação de um áudio como editado, quando na verdade não o foi. Um falso negativo corresponde à classificação de um áudio como não editado, quando na verdade o foi. Portanto, a FPR é a razão entre o número de falsos positivos e o total de áudios que se sabe pertencer a classe dos não editados $c_i = -1$, e a FNR é a razão entre o número de falsos negativos e o total de áudios que sabe-se pertencer a classe dos editados $c_i = +1$.

Mais especificamente, calcula-se FPR e FNR para vários valores de t e obtém-se a EER, definindo-se o limiar t_{EER} correspondente, a partir da seguinte formulação:

$$FPR(t) = \frac{1}{\sum_{i=1}^{N_{\text{ar}}} \left(\frac{1-c_i}{2}\right)} \sum_{i=1}^{N_{\text{ar}}} \frac{\lceil \text{sign}(g(\mathbf{F_i}) - t) + 0.5 \rceil}{2} \left(\frac{1-c_i}{2}\right), \tag{4.27}$$

$$FNR(t) = \frac{1}{\sum_{i=1}^{N_{\text{ar}}} \left(\frac{1+c_i}{2}\right)} \sum_{i=1}^{N_{\text{ar}}} \frac{\lceil \text{sign}(t - g(\mathbf{F_i})) + 0.5 \rceil}{2} \left(\frac{1+c_i}{2}\right), \tag{4.28}$$

$$t_{\text{EER}} = \arg\min_t ||\text{FPR(t)} - \text{FNR(t)}||, \tag{4.29}$$

$$EER = FPR(t_{\text{EER}})/2 + FNR(t_{\text{EER}})/2, \tag{4.30}$$

onde sign$(\cdot)$ é um operador que retorna o valor +1 para operandos de valor positivo, e -1 para operandos de valor negativo, $\lceil \cdot \rceil$ é função *teto*, que retorna o menor inteiro, maior ou igual ao operando, e $|| \cdot ||$ retorna o valor absoluto do operando.

Outros critérios além da EER podem ser ajustados para a obtenção do limiar t visando a garantir relações de compromissos distintas entre a FPR e FNR. Por exemplo, pode-se definir como ponto de operação t aquele em que a soma entre a FPR e a FNR é mínima, ou aquele em que o valor máximo entre a FNR e a FPR é minimizado, ou ainda fixar-se um valor desejado para a FPR (ou FNR) e minimizar-se a FNR (ou FPR).

Visando a permitir uma comparação direta com trabalhos anteriores (RODRÍGUEZ; APOLINÁRIO JR.; BISCAINHO, 2010; ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014, 2015; FUENTES et al., 2016), a mesma abordagem de utilizar a EER como critério para definição do ponto de operação do classificador é considerada, conforme descrito em testes no Capítulo 5. No entanto, propõe-se que o método seja utilizado considerando-se como ponto de operação aquele correspondente à máxima margem de separação estre as classes, que no SVM corresponde a utilizar $t = 0$, e portanto beneficiando a capacidade de generalização do método (BISHOP, 2006). Para isso foram realizadas avaliações de desempenho mais realistas que a definição da EER no conjunto de treinamento, em testes de validação cruzada, conforme descrito no Capítulo 5.

4.4 SUMÁRIO

Neste capítulo foi apresentada a proposta desta dissertação, o SPHINS, uma técnica para detecção automática de adulterações em gravações de áudio por meio da constatação de variações anormais na frequência de oscilação de sinais interferentes da rede elétrica (ENF) eventualmente incorporados em um registro de áudio questionado.

O SPHINS baseia-se na determinação de um vetor de características que descreve o grau de anomalia nas variações da ENF. Para tal vale-se de dois estimadores da ENF, o HEE e o 3E, onde o primeiro é mais sensível a transições abruptas de fase e o segundo é mais preciso em baixas condições de SNR. Extrai-se de cada áudio um vetor de características que usa a curtose amostral como medida do grau de anomalia das estimativas da ENF. Os vetores de características são aplicados a um classificador SVM para decidir se o áudio está editado.

No próximo capítulo serão apresentados os resultados de experimentos para avaliar o desempenho do SPHINS na classificação de arquivos de áudio pertencentes a um *corpus* de voz correspondente a gravações telefônicas autorizadas. O desempenho do método também foi testado em condições adversas de relação sinal-ruído e de saturação, sendo os resultados comparados com o de trabalhos correlatos.

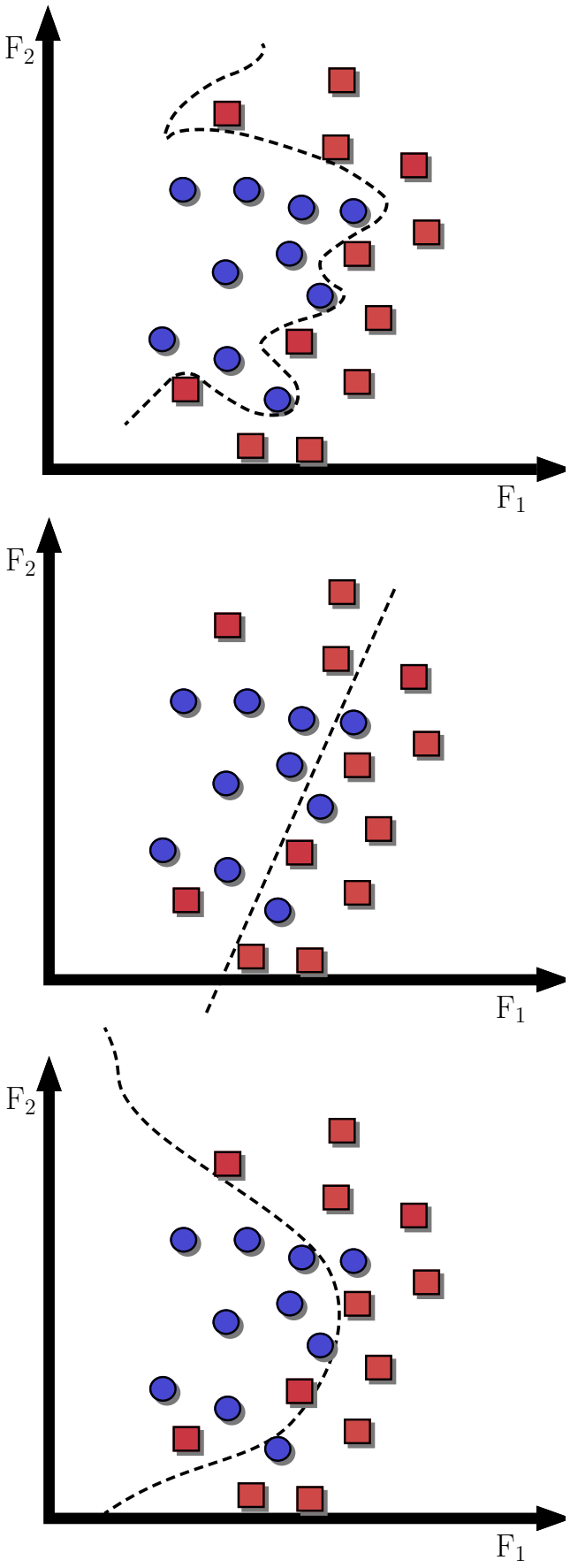


Figura 4.11: Regiões de decisão para um classificador em casos de *overfitting* (acima) e *undefitting* (meio), bem como para um ajuste intermediário (abaixo).

5 EXPERIMENTOS E RESULTADOS

Neste capítulo são apresentados os experimentos realizados bem como os resultados obtidos na avaliação de desempenho do método proposto na classificação de áudios editados e não editados. As avaliações realizadas valem-se de uma base de dados conhecida, denominada Carioca 1, introduzida nos trabalhos de Rodríguez, APOLINÁRIO JR. & Biscainho (2010) e também utilizada em Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015), Fuentes et al. (2016). Para avaliar a robustez do método proposto em condições desfavoráveis, algumas simulações foram realizadas relaxando-se algumas premissas referentes a intensidade de ruído e ao nível de distorções por saturação. Na Seção 5.1 a base de dados Carioca 1 é descrita, sendo apresentadas características acerca de sua composição. Na Seção 5.2 são descritos os parâmetros utilizados na fase de pré-processamento durante os experimentos realizados, bem como é descrito os resultados do procedimento de detecção da presença do sinal ENF na Base de dados utilizada. Na Seção 5.3 o método é avaliado na base de dados Carioca 1 em sua forma original. Na Seção 5.4 o método proposto é avaliado em diferentes condições de SNR. Na Seção 5.5, a técnica é avaliada em diferentes níveis de SL.

5.1 BASE DE DADOS CARIOCA 1

A fim de avaliar o desempenho da técnica proposta são utilizados arquivos de áudio conhecidos, não editados e controladamente editados. Para isso, empregou-se um *corpus* público de arquivos de voz denominado Carioca 1 (RODRÍGUEZ; APOLINÁRIO JR.; BISCAINHO, 2010). O corpus é o mesmo usado em Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015), Fuentes et al. (2016), permitindo que sejam realizadas algumas comparações diretas com os resultados obtidos em trabalhos correlatos anteriores.

O banco de dados Carioca 1 possui 100 gravações de áudio não editadas, autorizadas pelos respectivos interlocutores, e oriundas de conversações telefônicas, sendo 50 delas provenientes de interlocutores do sexo masculino e 50 por interlocutores do sexo feminino. Além disso, a base de dados Carioca 1 tem 100 versões editadas das mesmas conversações telefônicas, sendo 50 delas editadas por supressão de um segmento de áudio e outras 50 por inserção de um segmento de áudio, uniformemente distribuídas

entre as gravações de falantes do sexo masculino e feminino. As gravações de chamadas telefônicas são originalmente amostradas à taxa de 44,1 kHz com 16 bits de quantização e codificadas sem perdas por meio de modulação PCM linear. A duração dos sinais de áudio varia entre 19 s e 35 s, não havendo praticamente nenhuma interferência entre o sinal ENF e sinais de voz devido à largura de banda RTPC.

Os áudios contidos na base de dados Carioca 1 apresentam originalmente ruído de fundo de nível baixo a moderado, caracterizando em média uma SNR de 22,3 dB, variando entre 16 dB a 30 dB. Da mesma forma observa-se que a intensidade do sinal ENF nos sinais de áudio não é uniforme, de tal forma que os sinais com locutores femininos apresentam ENF com maior intensidade relativa (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014). Possivelmente tal fato deva-se a maior potência do sinal de voz masculina aliada a algum ajuste de ganho durante o processo de gravação.

5.2 PRÉ-PROCESSAMENTO E EXTRAÇÃO DE CARACTERÍSTICAS

Nos experimentos realizados considerou-se $f_{\text{nom}} = 60$ Hz e realizou-se a subamostragem de $s(n)$, gerando-se $s_{ds}(n)$ com uma frequência de amostragem de $f_s = 1200$ Hz[11]. Esta subamostragem síncrona garante um número exato de 20 amostras por período nominal da ENF. Foi utilizado um filtro passa banda do tipo FIR, de fase zero, com largura de banda de 0,8 Hz, utilizando-se as funções *conv* e *fir1*, com 10.000 coeficientes, do *software* Matlab ®. Escolheu-se a mesma quantidade de coeficientes no filtro FIR utilizada em Rodríguez, APOLINÁRIO JR. & Biscainho (2010), bem como a mesma relação de 20 vezes entre f_s e f_{nom} utilizada em Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015), com objetivo de processar os sinais de forma similar para melhor comparação entre os métodos. Além disso empregou-se o valor de $N_{\text{FFT}} = 2^{14}$ para o cálculo do parâmetro $\hat{\text{SNR}}_{\text{ENF}}$ visando a garantir uma mínima resolução frequencial na estimativa de densidade espectral de potência $\hat{P}_{\text{ds}}(k)$[12]. Utilizou-se $\text{BW}_{\text{ENF}} = 0,8$ Hz, uma vez que a largura de banda deve ser tão estreita quanto possível visando a rejeitar ruído e outros espúrios que não sejam o sinal ENF, porém suficientemente larga para cobrir toda a faixa de excursão do sinal ENF. Na Seção 5.4, avalia-se a robustez do SPHINS em um conjunto de diferentes valores de BW_{ENF}.

Inicialmente, antes de se proceder a classificação dos dados da base de dados Carioca

[11] Neste trabalho utilizou-se a função *resample* do software Matlab ®.

[12] Para tal foi utilizada a função *pwelch* do Matlab ®, com janelas de *Hamming* e sobreposição de 50 % das amostras entre janelas consecutivas.

1, é necessário verificar se há sinal ENF incorporado nos áudios questionados. Para isso é preciso determinar o limiar τ na Equação (4.4), sendo necessária uma base de dados sem sinal ENF. Para isso, calculou-se os valores $\hat{\text{SNR}}_{\text{ENF}}$ utilizando-se a própria base de dados Carioca 1, porém considerando-se uma frequência nominal de 50 Hz, uma vez que a base de dados possui áudios gravados em região cuja ENF nominal vale 60 Hz, obtendo-se um limiar $\tau = 5,528$ a partir do disposto na Equação (4.4).

A Figura 5.1 mostra, na parte superior, os histogramas dos valores $\hat{\text{SNR}}_{\text{ENF}}$ correspondentes a áudios sem sinal ENF e com sinal ENF obtidos para a Base de dados Carioca 1 em seu estado original, com SNR em média de 22,3 dB, variando entre 16 dB a 30 dB. Na parte inferior, a mesma comparação é feita para um versão da mesma base deteriorada por ruído aditivo, branco, Gaussiano e de média nula. Os áudios foram degradados visando a atingir uma SNR de 15 dB para todos os áudios. Para a degradação utilizou-se a mesma abordagem usada em Esquef, APOLINÁRIO JR. & Biscainho (2014) e descrita na Seção 5.4. Observa-se que o valor de τ, identificado pela linha tracejada, permite classificar todos os áudios da base Carioca 1, em sua forma original e degradada, como possuidores de sinal interferente da rede elétrica.

Após a etapa de pré-processamento, os sinais de áudio da base de dados passam à etapa de extração dos vetores de características. Dessa forma os sinais são submetidos a estimação da ENF pelos métodos HEE e 3E e têm sua curtose amostral calculada visando ao cálculo do vetor característico conforme a Equação (4.13). As Figuras 5.2 e 5.3 ilustram a distribuição das curtoses das estimativas ENF do tipo HEE para os áudios editados e não editados.

5.3 RESULTADOS NA BASE CARIOCA 1 ORIGINAL

Após a extração das características, para realizar os testes é preciso treinar o classificador SVM com a base de dados conhecida Carioca 1. No processo de treinamento, após a determinação da função discriminante não-linear, as gravações de áudio são classificadas para uma ampla gama de limiares t, conforme a Equação (4.14), e os valores correspondentes às taxas FPR e FNR são computados, tal qual às Equações (4.27) e (4.28), permitindo a construção da curva DET (*Detection Error trade-off*) (MARTIN et al., 1997). A EER corresponde ao ponto na curva de DET em que FPR = FNR, e é calculada por meio da minimização correspondente à Equação (4.30), sendo o valor de t_{EER} da Equação (4.29) estabelecido.

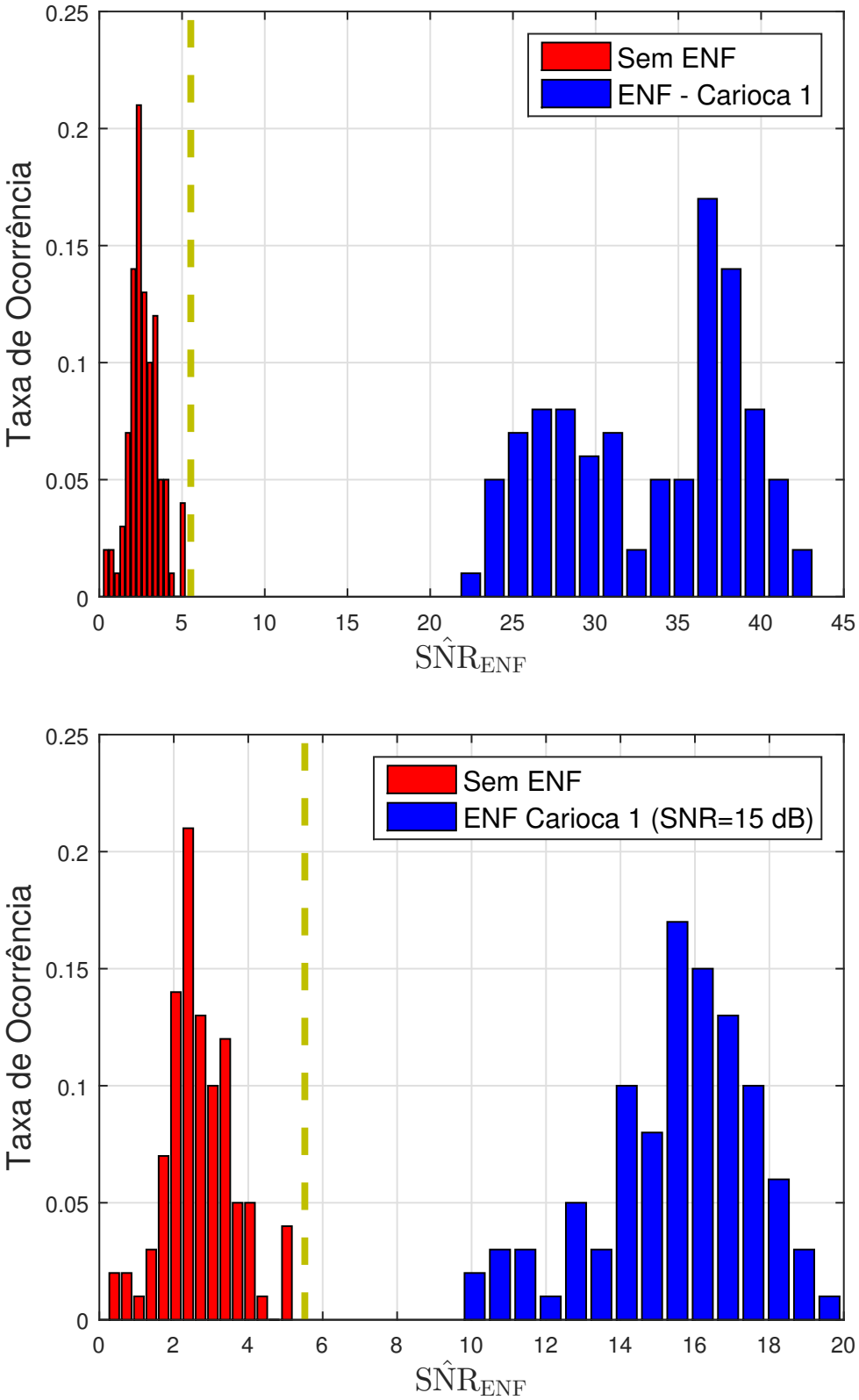


Figura 5.1: Histogramas de $\hat{\text{SNR}}_{\text{ENF}}$ correspondentes a sinais de áudio sem ENF e com sinal ENF obtidos para a base de dados Carioca 1 em seu estado original (SNR em média de 22,3 dB, variando entre 16 dB a 30 dB), e degradada por ruído aditivo, branco e Gaussiano, para uma SNR de 15 dB em todos os sinais de áudio.

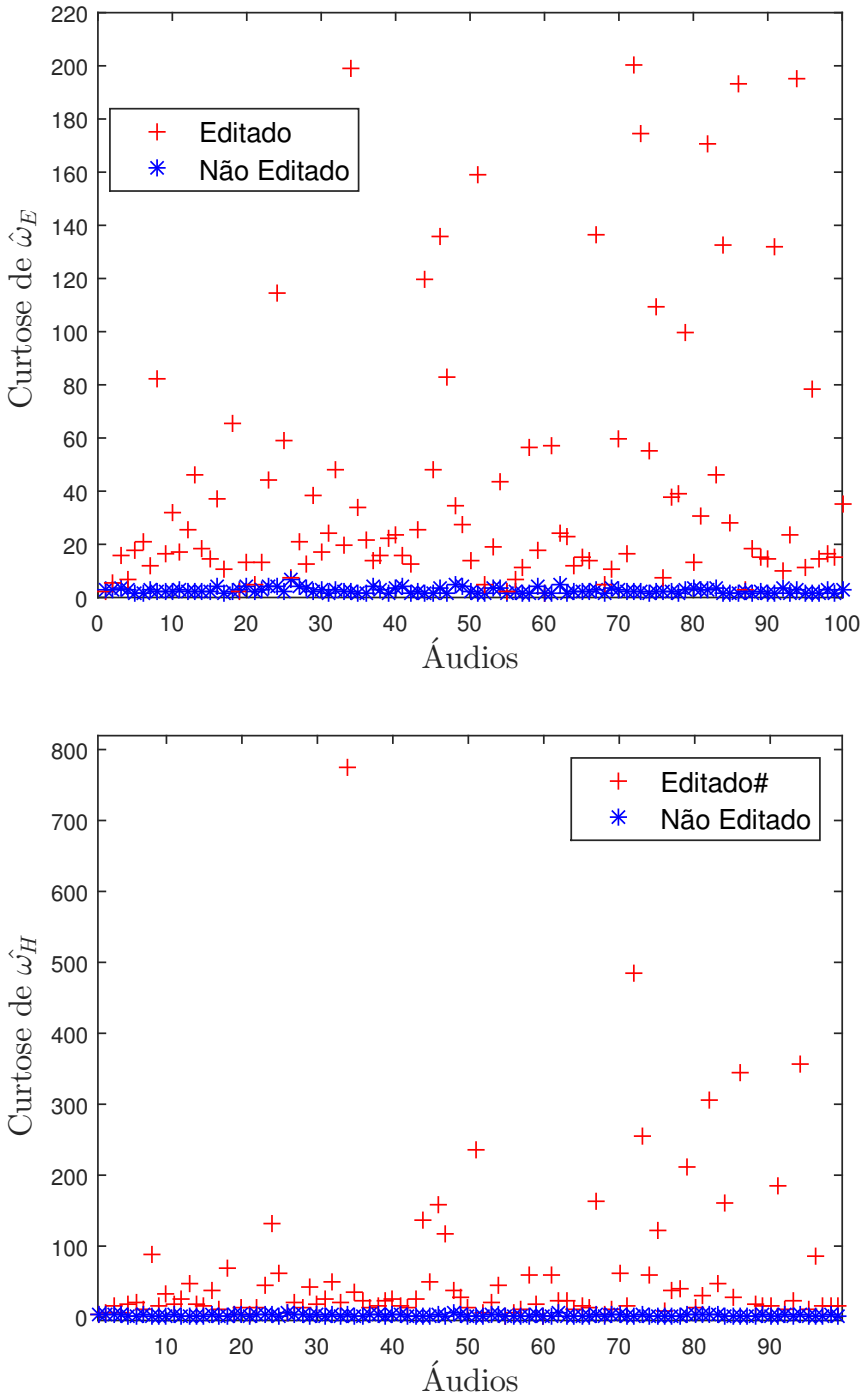


Figura 5.2: Distribuição das curtoses das estimativas ENF do tipo HEE e 3E para os sinais de áudios editados e não editados da base de dados Carioca 1 em seu estado original.

Para permitir uma comparação com trabalhos anteriores, foi calculada a curva DET do método utilizando-se a base de dados Carioca 1, como mostrado na Figura 5.4. Verifica-se que o método SPHINS apresenta uma EER de 4 %, superando a EER de 7 % do método apresentado em Rodríguez, APOLINÁRIO JR. & Biscainho (2010) e com o mesmo desempenho de Esquef, APOLINÁRIO JR. & Biscainho (2014) e Fuentes

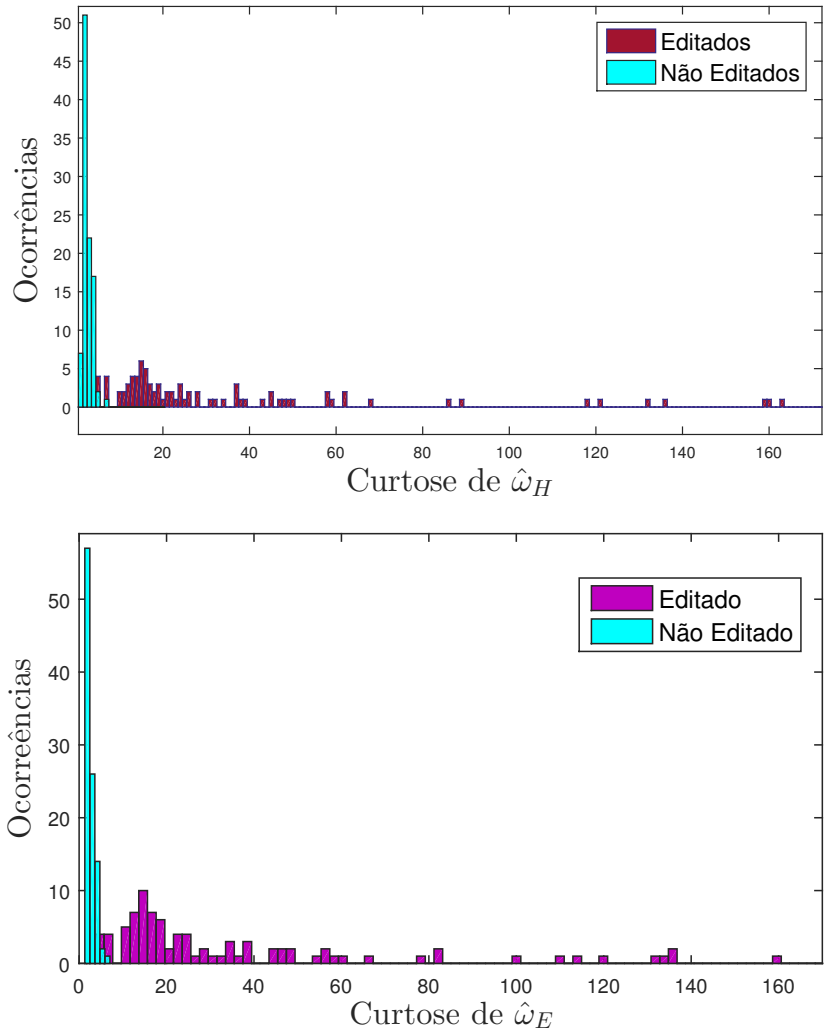


Figura 5.3: Histogramas das curtoses das estimativas ENF do tipo HEE e 3E para os áudios editados e não editados.

et al. (2016). Entretanto, em Esquef, APOLINÁRIO JR. & Biscainho (2015) os autores reportam uma EER de 2 % para a base de dados Carioca 1 original, demonstrando desempenho superior ao proposto.

A EER obtida no processo de treinamento é um parâmetro útil para caracterizar e comparar desempenhos de diferentes classificadores. No entanto, a EER é obtida após o treinamento do classificador com a determinação do limiar t_{EER}. Tal limiar é obtido a partir da base de treinamento, computando-se a quantidade de dados de treinamento pertencentes a uma classe c_i posicionados numa região do espaço de características que corresponda à classificação $-c_i$. Dessa forma, como é obtida a partir da classificação dos próprios dados utilizados para treinar e definir o ponto de operação, tende a fornecer uma estimativa otimista para a taxa de erro do sistema, sem avaliar de forma mais ampla a capacidade de generalização do método.

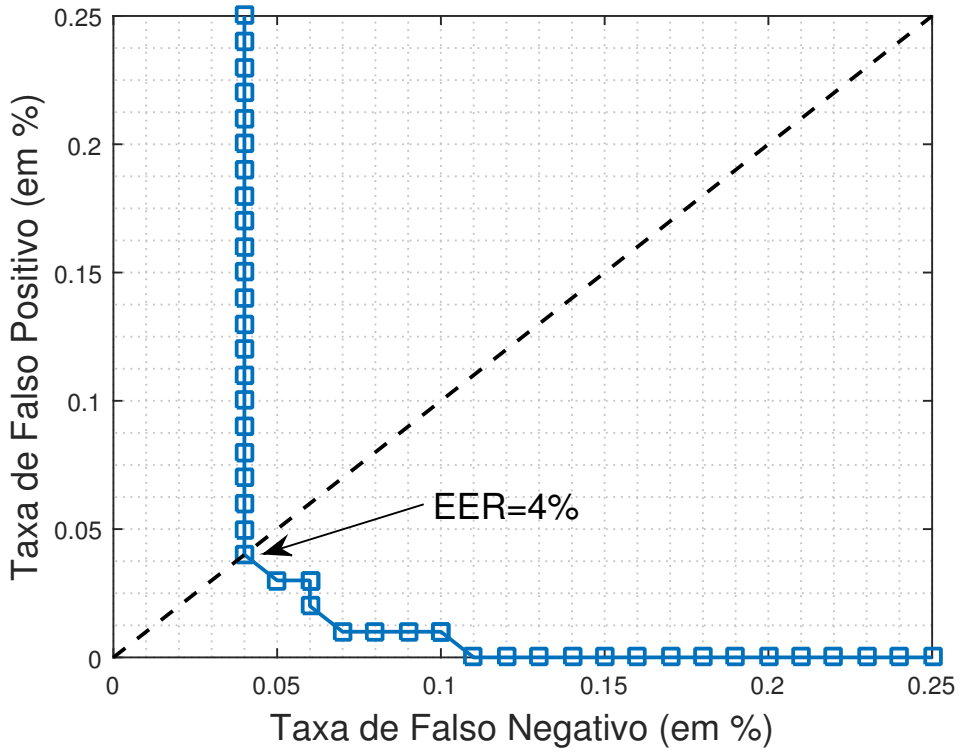


Figura 5.4: Curva DET, mostrando as taxas FPR vs. para a base de dados Carioca 1. A EER de 4 % está marcada.

Para uma avaliação mais realista é realizado um procedimento de validação cruzada utilizando estratégia do tipo *10-fold*, em que a base de dados é dividida igualmente e de forma aleatória em 10 subgrupos, em que 9 deles são utilizados para treinar o classificador e o subgrupo restante para testes. Uma vez treinado, realiza-se a etapa de testes classificando o subgrupo restante. Utiliza-se como ponto de operação do classificador o limiar $t = 0$, que corresponde a máxima margem de separação no SVM e portanto tende a diminuir o erro de generalização. Este processo é repetido dez vezes, onde a cada iteração o subgrupo de teste é trocado para um subgrupo diferente, que ainda não tenha sido testado, até que todos os áudios da base de dados tenham sido classificados. Os erros de classificação são então contados, avaliando-se a FPR, a FNR e a taxa de erro global, do inglês *Overall Error Rate* (OER).

Ressalta-se que a FPR e FNR calculadas aqui são ligeiramente diferentes das definidas nas Equações (4.27) e (4.28) para a definição do ponto de operação t_{EER}. Aqui, para cada partição de teste, são computados os erros de classificação do tipo falso positivo e falso negativo considerando-se o classificador treinando na partição de treinamento correspondente e o limiar $t = 0$. Ao final, a FPR será a razão entre o total de erros do tipo falso positivo computados em todas as partições de teste e o total das amostras

de teste pertencentes a classe $c_i = -1$, e a FNR será a razão entre o total de erros do tipo falso negativo computados em todas as partições de teste e o total das amostras de teste pertencentes a classe $c_i = +1$. Por último, a OER é calculada como sendo a razão entre o total de erros dos dois tipos computados em todas as partições de teste e o total das amostras de teste. Como no experimento em questão a quantidade de áudios editados e não editados nos conjuntos de teste e treinamento são as mesmas, a OER é a média aritmética entre a FPR e a FNR.

Para avaliar a influência do valor N, da Equação (3.18), no método proposto, foram realizados testes de validação cruzada *10-fold*. Nos teste realizados utilizou-se $N = \{100, 200, 300, 400, 500, 600, 700, 800\}$ amostras. Para cada um dos valores de N os testes de validação *10-fold* foram repetidos 10 vezes de forma a mitigar polarizações nos resultados devido a uma escolha específica de particionamento da base de dados. Foram computados os valores médios de OER, FPR e FNR considerando-se todas as 10 realizações dos testes de validação cruzada *10-fold* . Como mostrado na Figura 5.5, na base de dados considerada o valor de N tem pouco impacto no desempenho de classificação para valores de N entre 100 e 500 amostras. Observa-se um desempenho ligeiramente melhor para $N = 200$, resultando em uma taxa de erro global (OER) de 4,5 %, o que corresponde a uma acurácia de 95,5 % de precisão, com uma FPR = 1 % e uma FNR = 8 %. O valor de $N = 200$ foi escolhido para utilização nos demais testes realizados.

Os 4,5 % de OER correspondem a classificação de 9 áudios de forma incorreta no conjunto de 200 arquivos (100 editados e 100 não editados). A maior contribuição para esta taxa foi de sinais de áudio editados que não podem ser corretamente identificados, resultando na FNR de 8 %, contabilizando 8 arquivos editados não identificados. O valor da curtose amostral da estimativa HEE para cada um desses áudios é relativamente baixa, variando entre 1,73 e 5,45, caracterizando um baixa perturbação do sinal ENF. Tais valores sugerem que para esses sinais os desvios de fase provocado pelas edições é pequeno, ou a ENF apresenta flutuações mais elevadas, de tal forma que as perturbações de pico por descontinuidade de fase não caracterizam valores anômalos na distribuição da curtose amostral. Todos os falsos negativos correspondem a supressões de segmento de áudio. De fato, inserções de segmentos produzem dois pontos de descontinuidade de fase, aumentando-se a probabilidade de se inserir uma descontinuidade de valor suficientemente elevado para acusar uma detecção.

A Tabela 5.1 relaciona os arquivos de áudio editados que foram erroneamente clas-

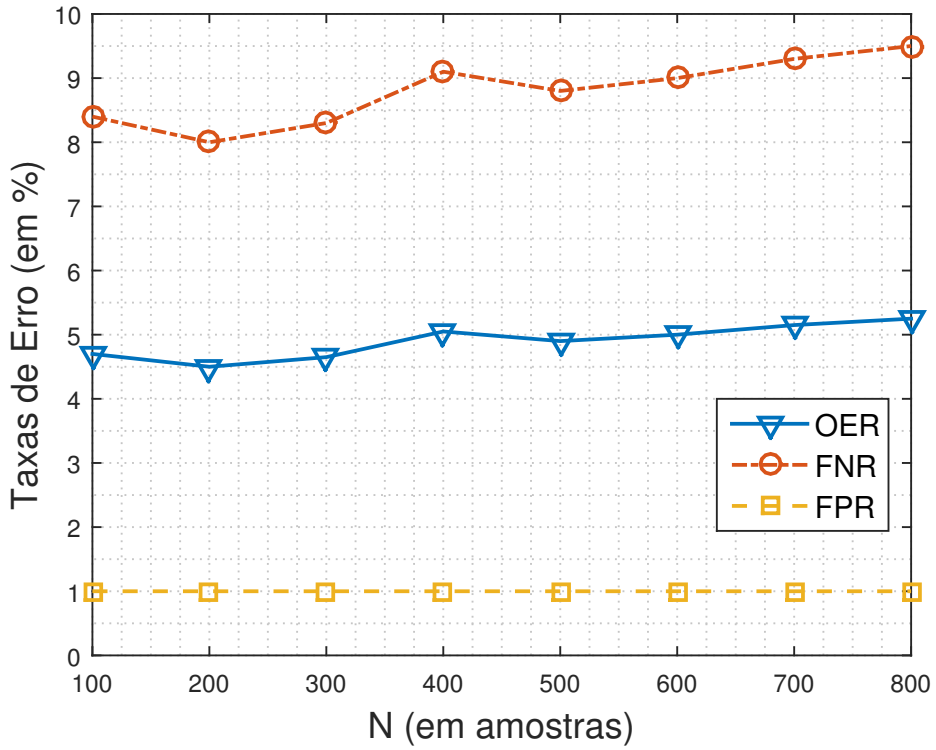


Figura 5.5: Taxa de erro global (OER) do método SPHINS aplicado a base de dados Carioca 1 para diferentes valores de N.

sificados como não editados. De fato, dos oito arquivos de áudio quatro apresentam curtose acima do valor 4. Uma inspeção visual da ENF estimada pelo método HEE para esses quatro arquivos permite esclarecer que todos os quatro apresentam ligeira perturbação na ENF estimada, caracterizada por um valor de pico que corresponde à edição realizada, com ordem de grandeza compatível com a flutuação da ENF. Na Figura 5.6, de (a) a (d), ilustra-se as estimativas ENF para cada um dos quatro áudios mencionados.

Por outro lado há apenas um falso positivo na classificação dos áudios da base Carioca 1 original, e que corresponde ao arquivo de áudio "HI01.wav". O valor da curtose amostral das estimativas HEE para esse áudio resulta no valor de 7,05. A Figura 5.6 (e) ilustra a estimativa ENF HEE do áudio, onde verifica-se que há moderada flutuação na estimativa ENF.

Os mesmos testes de validação cruzada não foram encontrados nos trabalhos de Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) que permitissem uma comparação apropriada. Em Fuentes et al. (2016) os autores testaram os áudios da base de dados Carioca 1 a partir de um esquema de detecção

Tabela 5.1: Valores de Curtose da Estimativa HEE para Áudios com Resultado do Tipo Falso Negativo

Arquivo	$\kappa(\hat{\omega}_{\mathrm{H_b}})$
HC1e.wav	2,42
HC2e.wav	5,45
HC19e.wav	2,50
HC21e.wav	4,68
MC2e.wav	4,90
MC5e.wav	2,36
MC10e.wav	1,73
MC18e.wav	4,60

treinado com uma base de áudios distinta, configurando um esquema de validação cruzada diferente, reportando uma OER de 4 %.

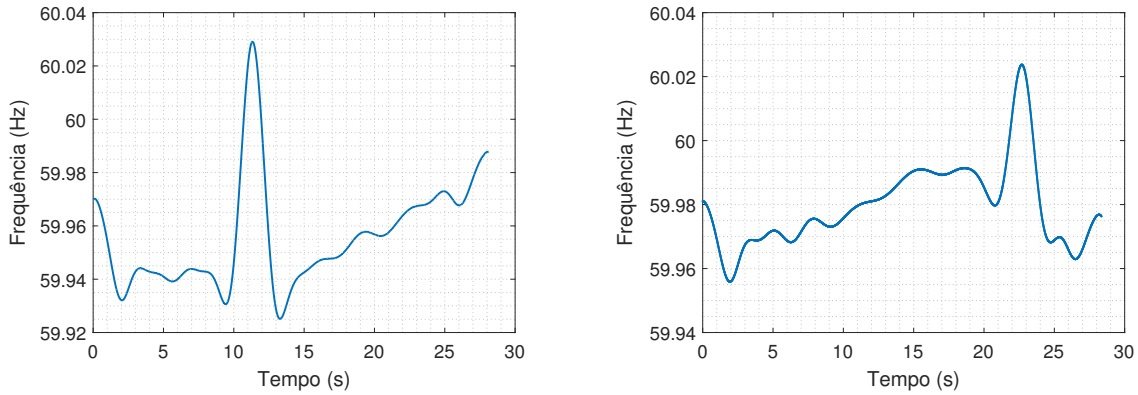
5.4 RESULTADOS EM DIFERENTES CONDIÇÕES DE RELAÇÃO SINAL RUÍDO

Com o propósito de avaliar o desempenho do método frente a condições adversas de SNR são realizados experimentos corrompendo de forma controlada o banco de dados de áudio Carioca 1 com ruído branco Gaussiano. Os experimentos foram feitos buscando-se adotar a mesma abordagem e o mesmo algoritmo VAD empregado em Esquef, APOLINÁRIO JR. & Biscainho (2014) e Esquef, APOLINÁRIO JR. & Biscainho (2015)[13], visando a permitir comparações diretas. Para tal, mede-se a SNR original de cada sinal de áudio na base de dados utilizando-se o VAD, visando separar o sinal de voz do ruído de fundo presente nas gravações. Na Figura 5.7 ilustra-se um sinal de áudio em que foi suprimido o ruído de fundo, nos instantes de voz inativa, por meio do algoritmo VAD. Dessa forma, a SNR do sinal de áudio, em dB, é computada como:

$$\mathrm{SNR_{db}} = 10\log_{10}\left\{\frac{\left[\frac{1}{|\mathbf{N}_{\mathrm{ativo}}|}\sum_{n\in\mathbf{N}_{\mathrm{ativo}}} s(n)^2\right] - \left[\frac{1}{|\mathbf{N}_{\mathrm{inativo}}|}\sum_{n\in\mathbf{N}_{\mathrm{inativo}}} s(n)^2\right]}{\left[\frac{1}{|\mathbf{N}_{\mathrm{inativo}}|}\sum_{n\in\mathbf{N}_{\mathrm{inativo}}} s(n)^2\right]}\right\}, \qquad (5.1)$$

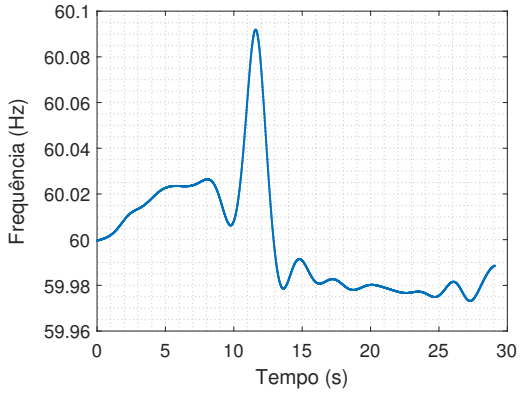
onde $\mathbf{N}_{\mathrm{ativo}}$ é o subconjunto de índices n correspondentes a amostras detectadas como provenientes de atividade de voz, $\mathbf{N}_{\mathrm{inativo}}$ é o subconjunto de índices n correspondentes

[13]São utilizados os mesmos parâmetros de mínima duração da atividade de voz de 60 ms e de mínima duração de inatividade de voz de 150 ms.

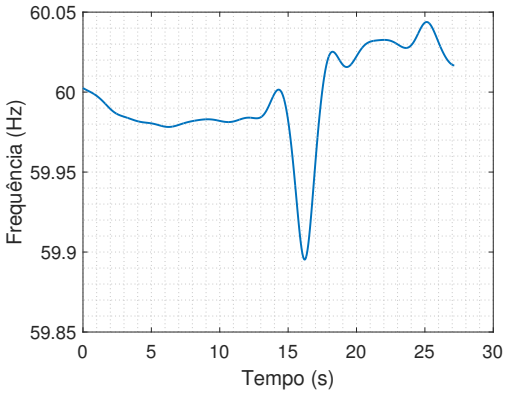


(a) HC2e.wav

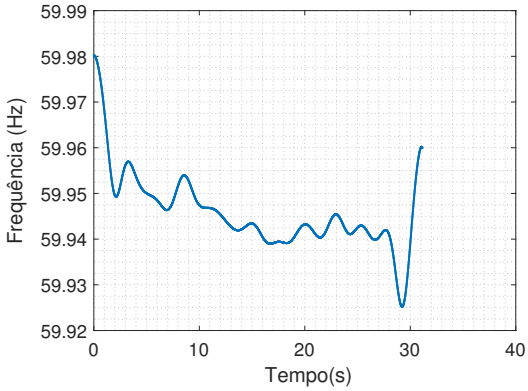
(b) HC21e.wav



(c) MC2e.wav



(d) MC18e.wav



(e) HI1.wav

Figura 5.6: Estimativas $\hat{\omega}_{\mathrm{H_b}}$ para os áudios correspondentes aos erros do tipo falso negativo (a) a (d), e falso positivo (e).

a amostras detectadas como provenientes de inatividade de voz, e $|\cdot|$ é o operador que retorna a cardinalidade do conjunto.

Em seguida, para aqueles arquivos de áudio em que a SNR original é maior do que um valor prescrito, uma quantidade extra de ruído de fundo é adicionada visando a atingir a SNR pretendida. Nos testes realizados utiliza-se um conjunto de valores alvo

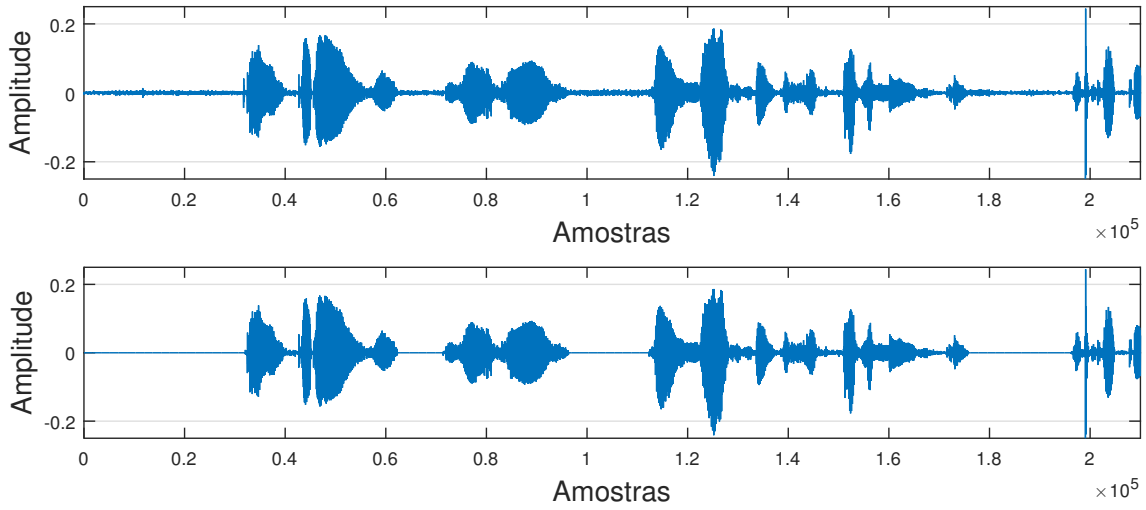


Figura 5.7: Sinal de áudio (acima) em que foi suprimido o ruído de fundo (abaixo) nos instantes de voz inativa por meio do algoritmo VAD empregado em Esquef, APOLINÁRIO JR. & Biscainho (2014).

para a SNR, em dB, variando de 5 a 30, com passos de 5 dB. Para garantir uma maior variabilidade esta tarefa é repetida 10 vezes, produzindo 10 versões ruidosas da base de dados Carioca 1 para cada uma dos 6 valores SNR prescritos, totalizando 60 versões ruidosas da base de dados original.

Após a produção das bases de dados ruidosas, o SPHINS é aplicado em cada uma das 10 versões ruidosas do banco de dados para cada SNR prescrito, extraindo-se os vetores de características baseados nas estimativa ESPRIT-Hilbert, treinando-se o classificador SVM correspondente e computando-se o valor da EER. Por fim, a EER correspondente a um determinado valor de SNR é calculada como sendo a média aritmética das EER computadas para cada uma das 10 versões ruidosas associadas a SNR desejada.

A Figura 5.8 mostra os valores de EER obtidos, comparando o desempenho do SPHINS com a abordagem empregada em Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015). De acordo com a Tabela 5.2 e com a Figura 5.8, observa-se que o SPHINS alcança uma EER significativamente inferior às obtidas em ambos os trabalhos em Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) em condições de degradação por ruído aditivo com SNR menor ou igual a 25 dB.

Para uma avaliação mais realista do desempenho do método em condições desfavoráveis de relação sinal ruído, são realizados testes de validação cruzada em estratégia *10-fold*. Para tal, a base de dados Carioca 1 original é dividida em 10 subgrupos. O classifica-

Tabela 5.2: Taxas EER para a Base de Dados Carioca 1 Corrompida por Ruído Branco Gaussiano. SPHINS Comparado com EAB-2014 (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014) e EAB-2015 (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2015)

SNR(dB)	EER-SPHINS (%)	EER EAB-2014 (%)	EER EAB-2015 (%)
30	4,00	4,00	2,30
25	4,60	5,40	4,70
20	6,45	12,30	12,20
15	11,25	24,70	22,90
10	21,05	38,70	36,30
5	22,50	45,00	43,90

dor é treinado utilizando-se áudios da base Carioca 1 original provenientes de 9 dos 10 subgrupos formados, tendo como ponto de operação o limiar $t = 0$ correspondente à máxima margem de separação. O procedimento de teste é realizado classificando-se os áudios correspondentes ao subgrupo não utilizado na fase de treinamento, e provenientes de uma das versões ruidosas da base de dados Carioca 1 para uma determinada SNR. Esse processo é repetido variando-se os subgrupos na estratégia *10-fold*. Além disso, o procedimento é repetido para classificar os áudios de todas as 10 versões ruidosas de Carioca 1 para uma determinada SNR, de tal forma que, para cada SNR prescrita, são realizadas 2000 mil classificações (uma vez que a base de dados Carioca 1 possui originalmente 200 arquivos de áudio, 100 editados e 100 não editados). As classificações incorretas que resultem em falsos positivos e falsos negativos são contadas obtendo-se os valores de FPR, FNR e OER. A Tabela 5.3 e a Figura 5.9 resumem os desempenhos obtidos. Os mesmos testes de validação cruzada não foram encontrados nos trabalhos de Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) que permitissem uma comparação apropriada, porém verifica-se que os valores de OER obtidos nos testes de validação cruzada são inferiores as EER correspondentes[14].

Também foi alcançado um desempenho melhor do que o obtido em Fuentes et al. (2016), onde os autores reportam uma OER de 50 % para uma degradação correspondente a SNR de 10 dB. Neste trabalho, os autores reportam uma taxa de erro de 50 % obtida a partir da classificação de todos os sinais de áudio da base de dados Carioca 1 degradada para uma SNR de 10 dB, considerando o ponto de operação correspondente a EER na base de dados Carioca 1 original utilizando também todos os sinais.

[14]Os valores de EER correspondentes aos trabalhos de Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) constam da Tabela 5.2.

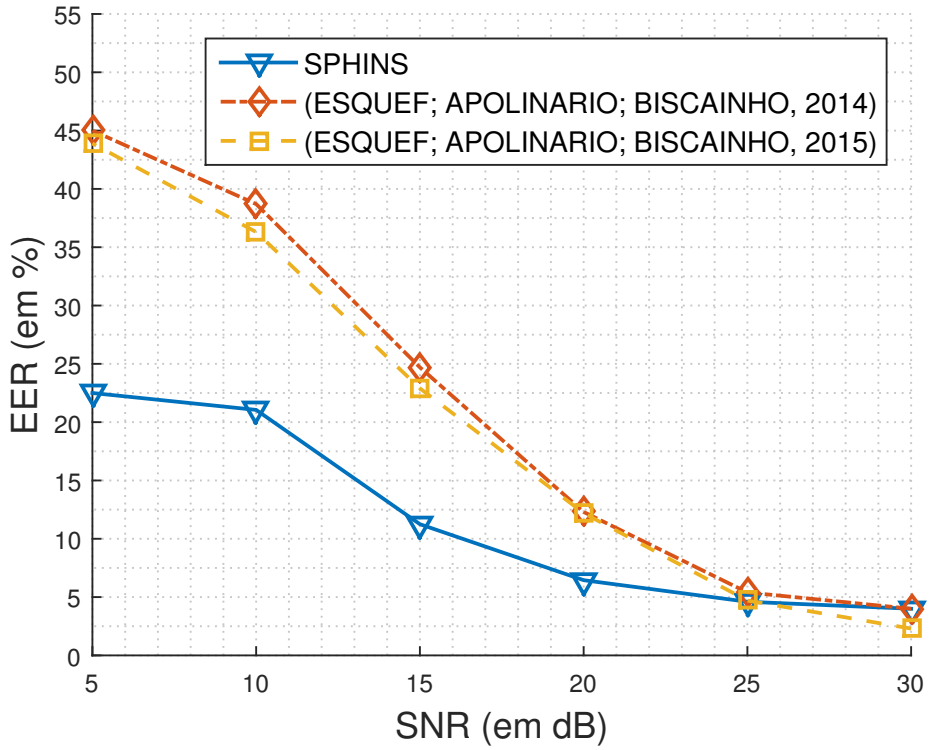


Figura 5.8: Valores de EER obtidos nas bases de dados corrompidas por ruído branco Gaussiano comparados com os resultados obtidos em (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014)

O desempenho do SPHINS também foi avaliado variando-se o valor da largura de banda $\mathrm{BW_{ENF}}$ entre valores de 0,4 Hz a 1,4 Hz, na classificação de sinais de áudio contidos nas dez versões ruidosas da base de dados Carioca 1 correspondentes a uma SNR de 10 dB. Como pode ser visto na Figura 5.10, há uma degradação considerável do desempenho à medida que aumenta-se a largura de banda $\mathrm{BW_{ENF}}$. O melhor valor de desempenho obtido considerando esse conjunto de larguras de banda corresponde à largura de banda $\mathrm{BW_{ENF}} = 0,6$ Hz. Entretanto, o uso de uma largura de banda muito estreita, embora reduza a influência deletéria do ruído no processo de classificação, pode comprometer o desempenho nos casos em que há considerável flutuação dos valores da ENF, especialmente nos casos em que a SNR é suficientemente alta para não ser um fator crítico na determinação do desempenho do sistema. Visando prevenir a ocorrência de desempenhos insatisfatórios nessas situações, sugere-se a utilização de uma largura de banda ligeiramente superior de 0,8 Hz.

Tabela 5.3: Taxas de Erro para o Método SPHINS Obtidas por Validação Cruzada na Base de Dados Carioca 1 sujeita a Degradação por Ruído Branco Gaussiano

SNR(dB)	OER(%)	FPR(%)	FNR(%)
30	4,40	1,00	7,8
25	4,40	1,00	7,8
20	5,60	1,00	10,20
15	11,40	5,40	17,40
10	24,20	30,20	18,20
5	33,15	51,80	14,50

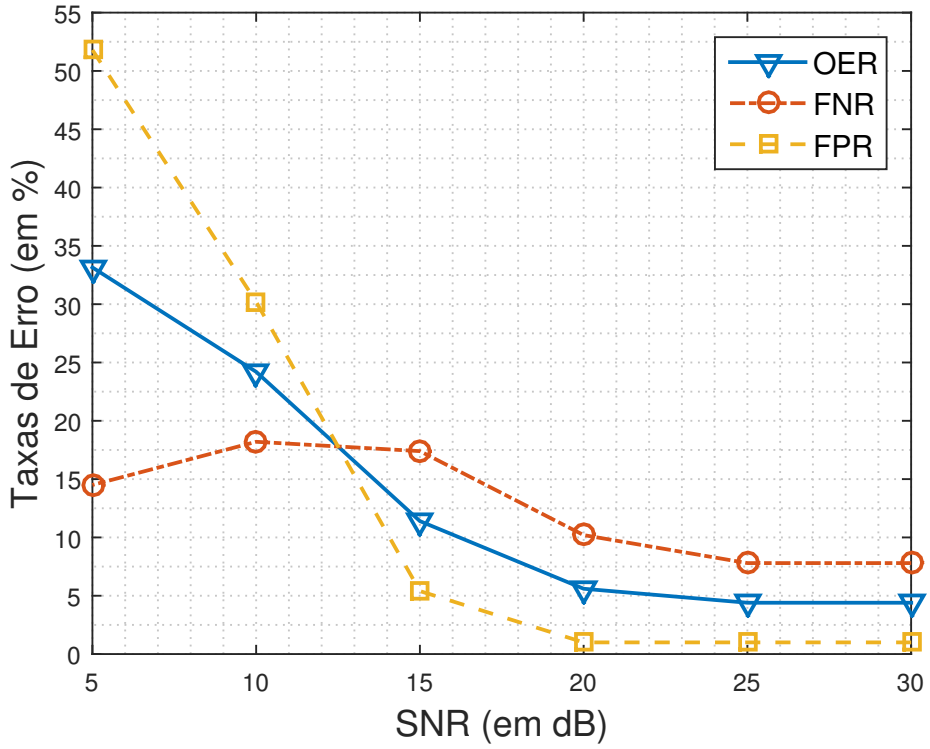


Figura 5.9: Taxas de erro da técnica SPHINS obtidas por validação cruzada a partir da base de dados Carioca 1 degradada por ruído branco Gaussiano.

5.5 RESULTADOS EM DIFERENTES NÍVEIS DE SATURAÇÃO

Uma outra fonte de deterioração com interesse prático consiste em não linearidades causadas pela saturação do sinal de áudio. Para avaliar o efeito deste tipo de degradação no desempenho da técnica de detecção, os sinais de áudio em Carioca 1 são submetidos a diferentes níveis de saturação (SL). Usando a mesma abordagem e o mesmo algoritmo VAD empregado em Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015), os instantes de atividade de voz em cada áudio da base de dados Carioca 1 são determinados, e um

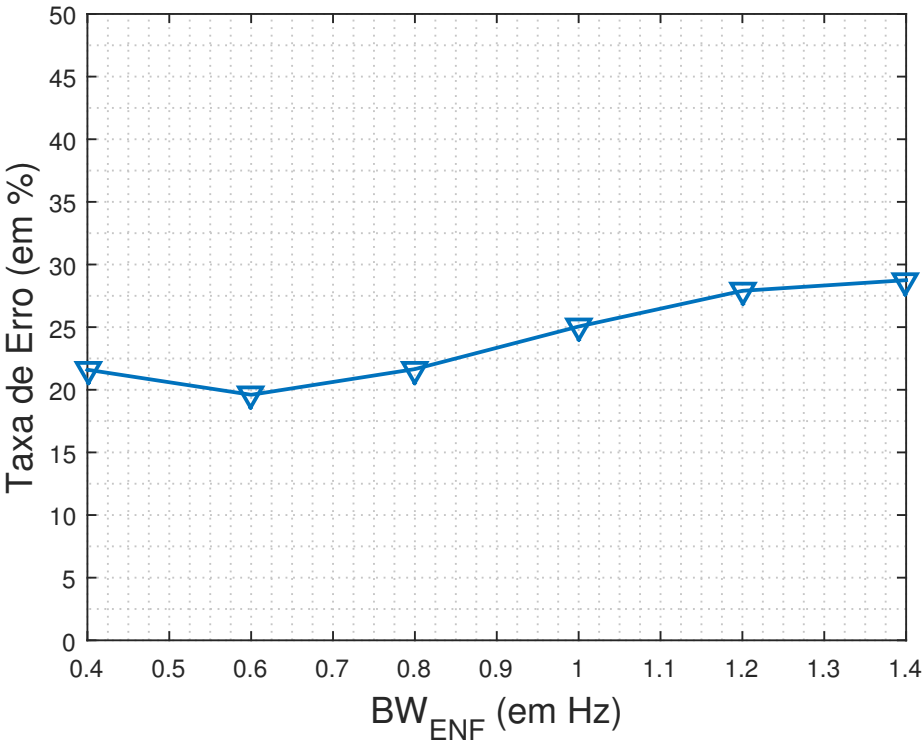


Figura 5.10: Taxa de erro global do método proposto para a base de dados Carioca 1 degradada por ruído branco Gaussiano, em uma SNR=10 dB, para diferentes valor de largura de banda $\mathrm{BW_{ENF}}$

percentual das amostras de áudio correspondente ao nível de saturação SL prescrito é ceifado para um valor absoluto máximo. O conjunto de níveis de saturação prescritos nos teste realizados é igual a SL = [0 0,2 0,5 1 2 4] %. A Figura 5.11 mostra os limiares de ceifamento para cada uma dos valores de SL = [0 0,2 0,5 1 2 4] %, e a Figura 5.12 ilustra a forma de onda do áudio após uma saturação com SL=4 %.

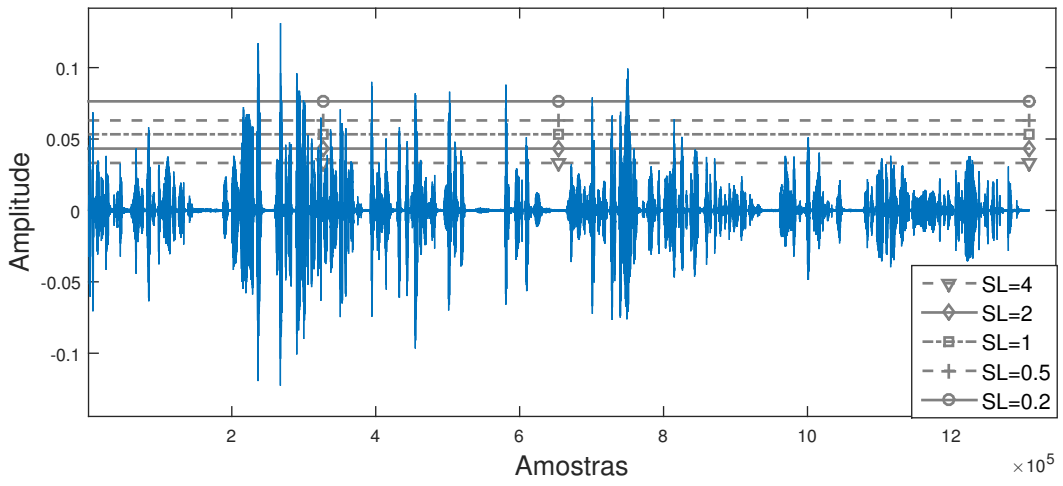


Figura 5.11: Limiares de ceifamento para SL = [0 0.2 0.5 1 2 4] %

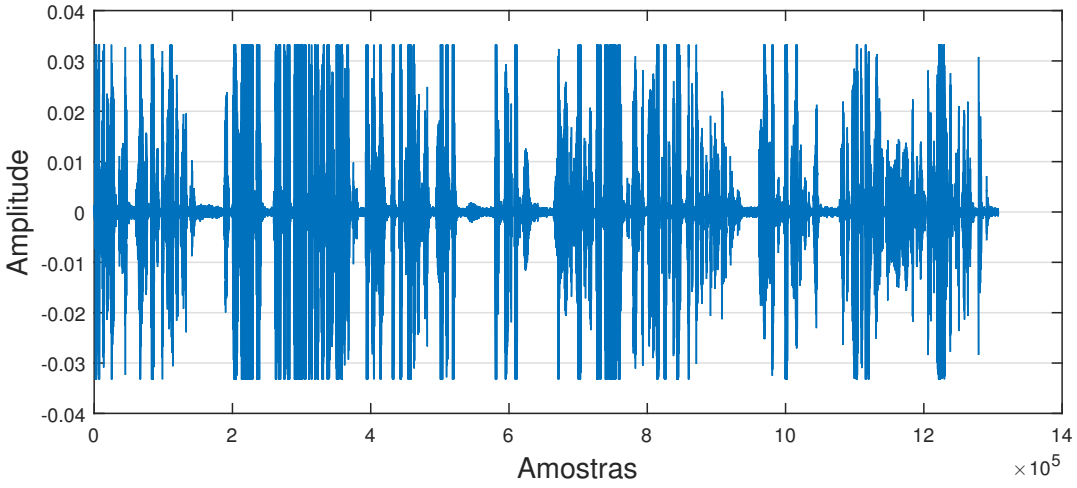


Figura 5.12: Forma de onda de um sinal de áudio após uma saturação com SL=4 %

O método proposto foi aplicado para cada nível de saturação SL prescrito, computando-se a EER. De acordo com a Tabela 5.4 e a Figura 5.13, o método SPHINS apresenta um desempenho superior ao obtido por Esquef, APOLINÁRIO JR. & Biscainho (2014) em cenários de saturação não linear. No entanto, o método com melhorias descrito em Esquef, APOLINÁRIO JR. & Biscainho (2015) apresenta melhor desempenho para valores de saturação de baixo a moderado, sendo superado pelo SPHINS para valores de SL elevados, superiores a 1 %.

Tabela 5.4: Taxas EER Para a Base de Dados Carioca 1 em Diferentes Níveis de Saturação. SPHINS Comparado com EAB-2014 (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2014) e EAB-2015 (ESQUEF; APOLINÁRIO JR.; BISCAINHO, 2015)

SL(%)	EER-SPHINS(%)	EER EAB-2014	EER EAB-2015(%)
0	4,00	4,00	2,00
0,2	8,00	12,00	6,00
0,5	9,00	12,00	8,00
1	9,00	13,00	10,00
2	10,50	14,00	13,00
4	13,00	18,00	15,00

Além disso, também é realizada classificação por validação cruzada em estratégia *10-fold*, observando-se que mesmo nos testes de validação cruzada, em que as taxas de erro são mais realistas, o SPHINS alcança resultados similares de taxa de erro. A Tabela 5.5 e a Figura 5.14 mostram o desempenho do método para essas condições de saturação. Os mesmos testes de validação cruzada não foram encontrados nos trabalhos de Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) que permitissem uma comparação apropriada.

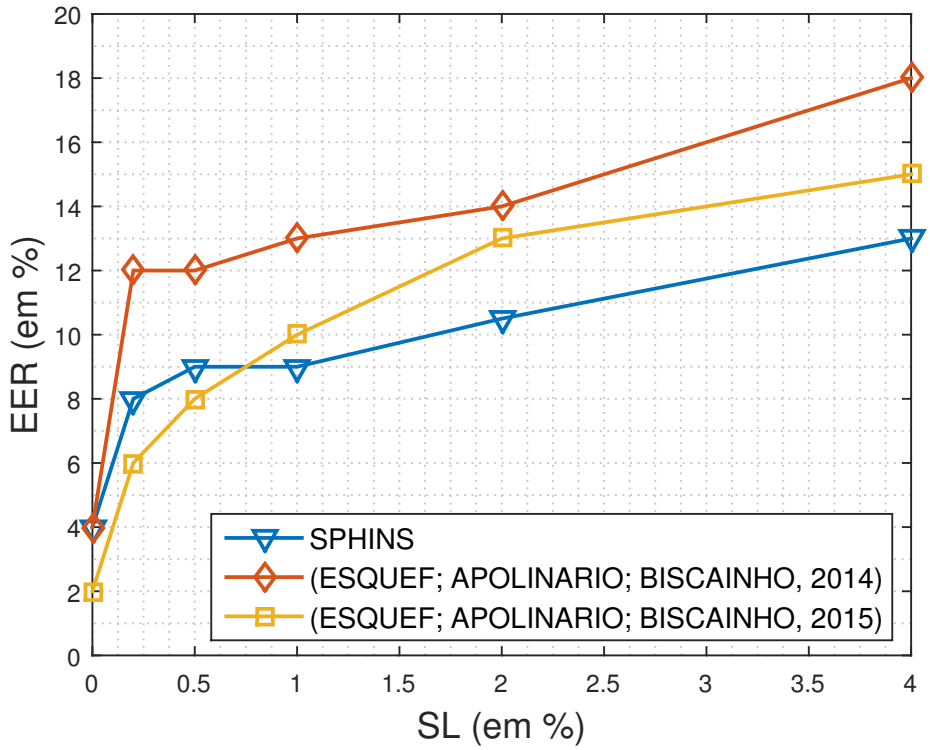


Figura 5.13: Taxas EER para o método proposto em comparação com os taxas EER obtidas em Esquef, APOLINÁRIO JR. & Biscainho (2014)

Também obtém-se um desempenho superior ao obtido em Fuentes et al. (2016), onde os autores reportam uma OER de 34 % obtida a partir da classificação de todos os áudios da base de dados Carioca 1 degrada para uma SL de 0,5 %, considerando o ponto de operação correspondente a EER na base de dados Carioca 1 original, sem empregar a mesma estratégia 10-fold de validação cruzada.

Tabela 5.5: Taxas de Erro por Validação Cruzada do Método SPHINS para a Base de Dados Carioca 1 em Diferentes Condições de Níveis de Saturação.

SL(%)	OER(%)	FPR(%)	FNR(%)
0	4,50	1,00	8,00
0,2	7,35	6,40	8,30
0,5	10,70	12,30	9,10
1	13,45	16,90	10,00
2	11,80	15,70	7,90
4	16,50	22,00	11,00

Por último, para mensurar o grau de influência do parâmetro σ (Equação 4.26) no desempenho do classificador, experimentos são realizados com a base de dados Carioca 1 em condições de degradação por saturação leve. Para isso testes de validação cruzada

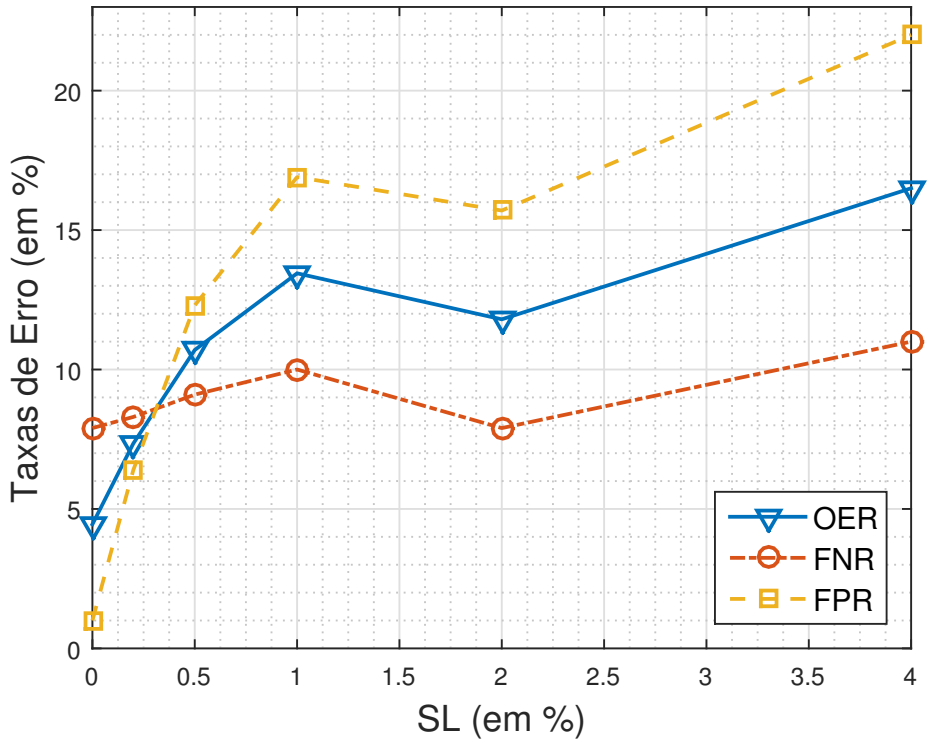


Figura 5.14: Taxas de erro por validação cruzada para o método SPHINS para a base de dados Carioca 1 degradadas por diferentes de níveis de saturação.

são realizados para diferentes valores de σ entre 1 e 6 para um nível de saturação SL $= 0,2$ %. A Figura 5.15 mostra a variação do desempenho do classificador com as mudanças no valor σ. Valores muito baixos (*overfitting*) ou muito altos (*underfitting*), tendem a piorar o desempenho do classificador. Os melhores valores encontrados nessa condição correspondem a σ entre 2,5 e 3,75, resultando numa OER entre 7,25 e 7,5 %. Em todos os demais testes realizados neste trabalho foi utilizado $\sigma = 3$.

5.6 RESULTADOS AVALIANDO-SE O 3º HARMÔNICO DA ENF

O método proposto também pode ser aplicado para identificar edições a partir de perturbações no terceiro harmônico da ENF, causados por descontinuidade de fase produzidas por inserções e supressões de segmentos de áudio. Para avaliar o desempenho do SPHINS nessa condição, o método foi aplicado aos áudios da base de dados Carioca 1. Dessa forma , considerou-se $f_{\text{nom}} = 180$ Hz, $f_s = 3600$ Hz, e $\text{BW}_{\text{ENF}} = 2,4$ Hz. Os demais parâmetros forma mantidos idênticos aos utilizados com a ENF em 60 Hz.

Para permitir uma comparação com trabalhos anteriores, foi calculada a curva DET

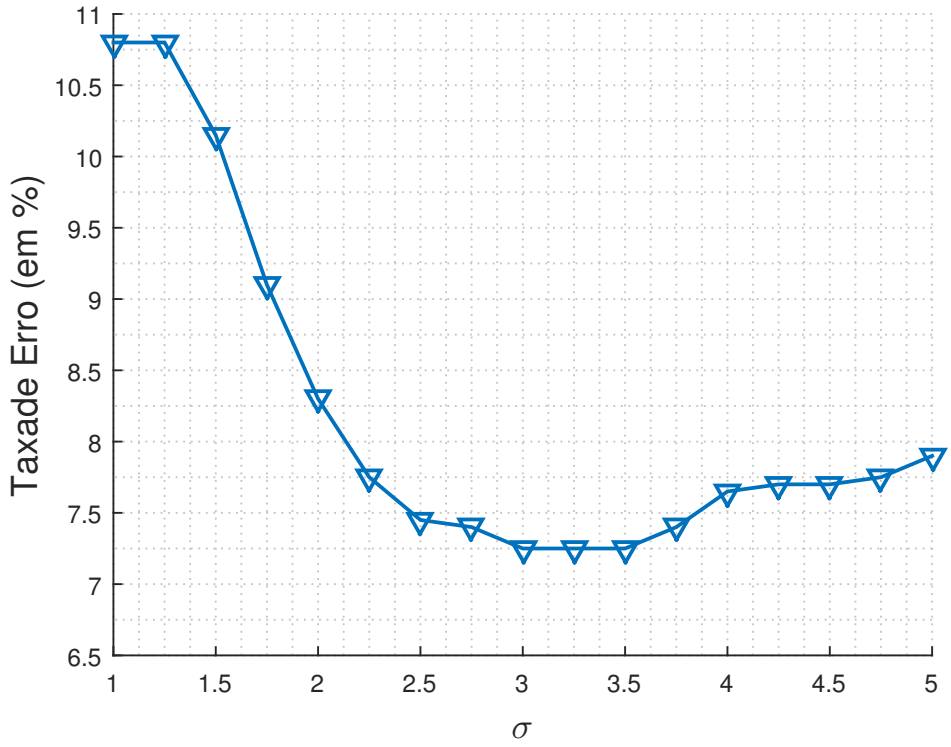


Figura 5.15: Taxa de erro global para base de dados Carioca 1 em diferentes valores de σ, para uma SL $= 0,2$ %.

do método utilizando-se a base de dados Carioca 1, como mostrado na Figura 5.4. Verifica-se que o método SPHINS apresenta uma EER de 21%, superando a EER de 24 % do método apresentado em Rodríguez, APOLINÁRIO JR. & Biscainho (2013), de 28 % reportado em Esquef, APOLINÁRIO JR. & Biscainho (2014), e de 22 % em Esquef, APOLINÁRIO JR. & Biscainho (2015). Apesar da sutil melhora nos valores de EER observados, a taxa de erro na aplicação do método nessas condições ainda é por demasiado elevada, demonstrando pouca aplicação prática nesses casos. A Figura 5.16 ilustra a curva DET obtida para aplicações do método no terceiro harmônico da ENF.

5.7 SUMÁRIO

Neste capítulo foram apresentados os resultados de experimentos onde o método SPHINS foi aplicado a uma base de dados conhecida, visando avaliar o desempenho na classificação de áudios editados e não editados. A base de dados utilizada para realizar os experimentos, denominada Carioca 1 (RODRÍGUEZ; APOLINÁRIO JR.; BISCAINHO, 2010), contém áudios provenientes de ligações telefônicas, editados e não editados, de locutores masculinos e femininos. O desempenho do método foi estimado, usando-se a EER como parâmetro para aferição de sua acurácia. Em seguida o SPHINS teve

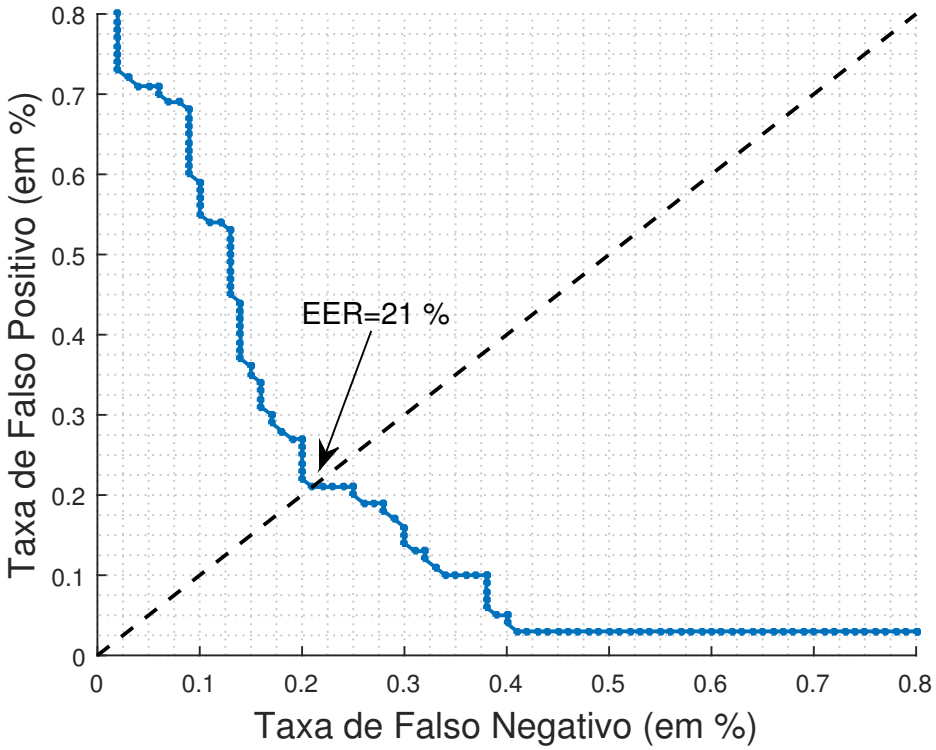


Figura 5.16: Curva DET para a aplicação do método no terceiro harmônico da ENF, mostrando as taxas FPR vs. FNR para a base de dados Carioca 1. A EER de 21 % está marcada.

seu desempenho comparado com os valores de EER obtidos em trabalhos anteriores. Visando a uma estimativa menos otimista do desempenho de classificação, realizou-se também a aferição das taxas de erro global (OER), e das taxas de falso positivo (FPR) e falso negativo (FNR) por meio de testes de validação cruzada em estratégia *10-fold*.

Por fim o desempenho do SPHINS foi avaliado em condições desfavoráveis onde algumas das premissas para uma correta aplicação do método foram relaxadas. Inicialmente avaliou-se o desempenho do SPHINS em diversas condições de degradação por ruído aditivo branco Gaussiano. Foi computada a EER nas diversas SNR de interesse, comparando-se os resultados com aqueles obtidos nos trabalhos correlatos. Também foram realizados testes de validação cruzada em estratégia *10-fold*, computando-se as taxas de OER, FPR e FNR. Os mesmos testes foram realizados para uma variedade de condições adversas de saturação, computando-se a EER e comparando-se com trabalhos anteriores, bem como verificando-se o desempenho em testes de validação cruzada. Por último foi avaliado a aplicação do método na detecção de edições a partir do 3º harmônico da ENF.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propõe uma nova técnica para detectar adulterações em registros de áudio digital por meio da análise de distúrbios de sinais interferentes da rede elétrica (ENF), com bons resultados. Trabalhos anteriores, como os descritos em Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015), abordam o problema de detectar adulterações em áudio a partir de variações anormais no sinal ENF causadas por descontinuidades bruscas de fase devido a supressões ou inserções de segmentos de áudio.

A técnica aqui proposta, batizada SPHINS, baseia-se na mesma ideia, porém por meio de um método distinto. Para isso estima conjuntamente a ENF por técnicas baseadas na frequência instantânea da aproximação analítica de Hilbert (HEE) e na técnica ESPRIT (3E). A estimativa ENF por meio da técnica de Hilbert, embora seja mais sensível a presença de descontinuidades de fase do que a técnica ESPRIT, também gera estimativas sensíveis a degradação por ruído e por saturação do sinal de áudio. Já a técnica ESPRIT apresenta-se mais robusta a tais degradações.

Como diferencial, o SPHINS utiliza a curtose amostral como uma medida de anomalia na distribuição de valores da ENF (PEÑA; PRIETO, 2012), onde, a partir dos valores das curtoses computadas, extrai-se um vetor de características que agrega informação sobre a ocorrência de valores caudais na distribuição da ENF estimada pelos dois métodos HEE e 3E. Valores elevados de curtose são indicativos da ocorrência de anomalias na ENF e, portanto, indicativos de alguma edição por inserção ou supressão de segmentos de áudio. As curtoses estimadas são então empregadas para formar um vetor de características que é aplicado a um classificador SVM devidamente treinado para indicar a presença de adulterações.

A vantagem do SPHINS é utilizar um vetor de características que, obtido a partir das curtoses amostrais das estimativas ENF HEE e 3E, consegue discriminar áudios editados e não editados, por inserção ou supressão, com um melhor desempenho que as técnicas correlatas em condições severas de degradação por ruído e por saturação não linear.

Para avaliar o SPHINS, o mesmo banco de dados usado em Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) e Fuentes et al. (2016) é empregado. São realizados testes semelhantes a tais trabalhos, simulando-se degradações devidas à adição de ruído e a um processo não linear de saturação, visando obter comparações relevantes.

Os resultados mostram que o método proposto, quando aplicado ao banco de dados em sua forma original, alcança um desempenho de 4 % EER, superando a EER de 7 % do método apresentado em Rodríguez, APOLINÁRIO JR. & Biscainho (2010) e com o mesmo desempenho de Esquef, APOLINÁRIO JR. & Biscainho (2014) e Fuentes et al. (2016). Entretanto, em Esquef, APOLINÁRIO JR. & Biscainho (2015) os autores reportam uma EER de 2 % para a base de dados Carioca 1 original, inferior à obtida no método proposto.

Nos testes de validação cruzada, o método apresenta 95,5 % de precisão (4,5 % OER), com uma FPR = 1 % e uma FNR = 8 %. Há, portanto, nos teste de validação cruzada da base de dados Carioca 1 em sua forma original uma maior incidência de erros por falso negativo, correlacionados a ocorrência de valores de curtose reduzidos em alguns áudios editados. Associa-se esses valores reduzidos a áudios com pequenos desvios de fase provocado pelas edições, bem como a flutuações mais elevadas na ENF. Todos os falsos negativos correspondem a supressões de segmento de áudio, uma vez que inserções de segmentos produzem dois pontos de descontinuidade de fase, aumentando a probabilidade de se inserir uma descontinuidade de valor suficientemente elevado para acusar uma detecção. Os mesmos testes de validação cruzada não foram encontrados nos trabalhos de Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014, 2015) que permitissem uma comparação apropriada. Em Fuentes et al. (2016) os autores testaram os áudios da base de dados Carioca 1 a partir de um esquema de detecção treinado com uma base de áudios distinta, configurando um esquema de validação cruzada diferente que reportou uma OER de 4 %.

Tal qual os trabalhos de Rodríguez, APOLINÁRIO JR. & Biscainho (2010), Esquef, APOLINÁRIO JR. & Biscainho (2014) e Fuentes et al. (2016), o método proposto apresenta degradação de desempenho na presença de ruído e saturação não linear, porém alcança resultados melhores do que os trabalhos anteriores para cenários de degradação da SNR e com nível de saturação superior a 1 %.

6.1 SUGESTÕES PARA PESQUISAS FUTURAS

A aplicação do SPHINS como técnica forense necessita de certos cuidados uma vez que uma classificação assertiva acerca da autenticidade de um determinado arquivo de áudio não é determinada por somente uma característica, e sim fruto de um conjunto de análises de diversos parâmetros. As condições de SNR e de SL são em primeira análise fundamentais para o desempenho do esquema uma vez que inserem variações de fase espúrias e que podem comprometer o desempenho do método. Como trabalhos futuros, considera-se a aplicação da abordagem proposta a bases de dados maiores tomadas em localidades distintas, em condições variadas de sinal ruído e de saturação. Além disso, espera-se que o desempenho global possa ser melhorado ainda mais combinando-se a técnica com abordagens de estimação de perturbação da ENF a partir de outros domínios, como por exemplo métodos baseados na taxa de cruzamento de zero, e de nível. Vislumbra-se ainda estudar a adaptação do método para esquemas de extração da ENF a partir de sinais de imagem de vídeo (GARG; VARNA; WU, 2011). A utilização de técnicas sofisticadas de aprendizado de máquina, como por exemplo metaclassificadores, combinando informação vinda de diferentes domínios pode resultar em uma melhora no processo de classificação.

Sugere-se a avaliação do método frente a diferentes técnicas contra-forenses que podem ser aplicadas para induzir um resultado errôneo na classificação. Entre as possíveis técnicas contra-forenses visando a mitigar a aplicação do método proposto existe a possibilidade de se realizar edições com o cuidado para que segmentos inseridos e/ou suprimidos tenham duração correspondente ou próxima a um múltiplo inteiro do período nominal da ENF. Isto porque, dessa forma, a transição de fase é pequena o suficiente para não gerar perturbações detectáveis na análise por meio da curtose das estimativas ENF. Outra possibilidade é a inserção de ruído pós-edição para mascarar a ENF e não permitir a detecção das eventuais edições. Por outro lado, técnicas contra-forenses também são possíveis por meio da inserção de um sinal moderada ou fortemente intenso e sem descontinuidade de fase, simulando o sinal ENF em áudios editados que não apresentem a ENF, ou se possuírem que seja em baixa intensidade, com o objetivo de deixá-lo com características de autêntico (CHUANG; GARG; WU, 2013).

Uma área a ser pesquisada é a utilização de tensores visando obter ganhos de desempenho fruto de uma melhor caracterização das mudanças abruptas de de fase por meio de um representação multidimensional dos dados nos domínio temporal ou frequencial. Também pode se considerar a aplicação de técnicas de regressão robustas a

anomalias como forma de melhor estimar a ENF e caracterizar distúrbios na evolução linear da fase, provocados por edições, tais como o RANSAC (FISCHLER; BOLLES, 1981), ou modelos baseados em processos Gaussianos (PILON et al., 2015). Uma outra abordagem sugerida é explorar técnicas de arranjos de microfones visando a agregar informação adicional oriunda de dispositivos de captação de áudio que utilizem mais de um microfone de captação (SILVEIRA et al., 2013; DENK; COSTA; SILVEIRA, 2014). Modelos baseados em subespaço visando a estimação parametrizada de múltiplos harmônicos da ENF também podem ser investigados, visando a explorar o fato de que desvios de fase em harmônicos superiores são múltiplos inteiros do desvio na frequência fundamental. Também vislumbra-se a aplicação de técnicas de seleção de ordem de modelo e de análise de perfil de autovalores da matriz de correlação dos dados de sinal, visando a identificar perturbações na ENF associadas a tarefas de inserção ou supressão de segmentos de áudio.

Pode-se também avaliar a eficácia do método na detecção de adulteração de áudio por meio de outros sinais em banda estreita eventualmente incorporados ao registro questionado como, por exemplo, os harmônicos estáveis derivados de outros fenômenos, como, por exemplo, interferências geradas por pulsos de sincronismo horizontal e vertical em sinais de vídeo.

REFERÊNCIAS BIBLIOGRÁFICAS

ANEEL. *Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional-PRODIST, Módulo 8 - Qualidade da Energia Elétrica*. 2013. Disponível em: ⟨http://www.aneel.gov.br/arquivos/PDF/Modulo3_Revisao_5.pdf⟩. Acesso em: 20/05/2016.

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738.

BÖHME, R. et al. Multimedia forensics is not computer forensics. In: *Third International Workshop on Computational Forensics*. [S.l.]: Springer, 2009. p. 90–103.

CHUANG, W.-H.; GARG, R.; WU, M. Anti-forensics and countermeasures of electrical network frequency analysis. *IEEE Transactions on Information Forensics and Security*, IEEE, v. 8, n. 12, p. 2073–2088, 2013.

COOPER, A. J. The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings–an automated approach. In: AUDIO ENGINEERING SOCIETY. *Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice*. [S.l.], 2008.

DENK, F.; COSTA, J. P. C. da; SILVEIRA, M. A. Enhanced forensic multiple speaker recognition in the presence of coloured noise. In: IEEE. *Signal Processing and Communication Systems (ICSPCS), 2014 8th International Conference on*. [S.l.], 2014. p. 1–7.

ESQUEF, P. A.; APOLINÁRIO JR., J. A.; BISCAINHO, L. W. Improved edit detection in speech via ENF patterns. In: IEEE. *Information Forensics and Security (WIFS), 2015 IEEE International Workshop on*. [S.l.], 2015. p. 1–6.

ESQUEF, P. A. A.; APOLINÁRIO JR., J. A.; BISCAINHO, L. W. Edit detection in speech recordings via instantaneous electric network frequency variations. *Information Forensics and Security, IEEE Transactions on*, IEEE, v. 9, n. 12, p. 2314–2326, 2014.

FANTINI, D. F. Interceptação telefônica e linguagem. *Revista Brasileira de Ciências Policiais*, v. 3, n. 1, p. 11–25, 2013.

FARID, H. *Detecting Digital Forgeries Using Bispectral Analysis*. Relatório Técnico. AI Lab, Massachusetts Institute of Technology. Cambridge, MA, USA, 1999. Disponível em: ⟨ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-1657.pdf⟩. Acesso em: 20/05/2016.

FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, ACM, v. 24, n. 6, p. 381–395, 1981.

FUENTES, M. et al. Detection of ENF discontinuities using PLL for audio authenticity. In: IEEE. *2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS)*. [S.l.], 2016. p. 79–82.

GARG, R.; VARNA, A. L.; WU, M. Seeing ENF: natural time stamp for digital video via optical sensing and signal processing. In: ACM. *Proceedings of the 19th ACM international conference on Multimedia*. [S.l.], 2011. p. 23–32.

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

HAJJ-AHMAD, A.; GARG, R.; WU, M. Instantaneous frequency estimation and localization for ENF signals. In: IEEE. *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. [S.l.], 2012. p. 1–10.

HAWLEY, A. H. Ecology and human ecology. *Social Forces*, JSTOR, p. 398–405, 1944.

IKRAM, S.; MALIK, H. Microphone identification using higher-order statistics. In: AUDIO ENGINEERING SOCIETY. *Audio Engineering Society Conference: 46th International Conference: Audio Forensics*. [S.l.], 2012.

INMAN, K.; RUDIN, N. The origin of evidence. *Forensic Science International*, Elsevier, v. 126, n. 1, p. 11–16, 2002.

LIU, Q.; SUNG, A. H.; QIAO, M. Detection of double MP3 compression. *Cognitive Computation*, Springer, v. 2, n. 4, p. 291–296, 2010.

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

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

MANOLAKIS, D. G.; INGLE, V. K.; KOGON, S. M. *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*. [S.l.]: Artech House Norwood, 2005. v. 46.

MARPLE, S. L. Computing the discrete-time "analytic" signal via FFT. *IEEE Transactions on Signal Processing*, IEEE, v. 47, n. 9, p. 2600–2603, 1999.

MARTIN, A. et al. The DET curve in assessment of detection task performance. In: . [S.l.: s.n.], 1997. p. 1895–1898.

MCKENZIE, R. D. The ecological approach to the study of the human community. *American Journal of Sociology*, JSTOR, p. 287–301, 1924.

MELOAN, C. E. *Criminalistics: An Introduction to Forensic Science*. [S.l.]: Pearson College Div, 2000.

MORISSON, A. L. d. C.; MACHADO, C. E. P.; REIS, P. M. G. I. Exames de registro de áudio e imagens. In: _____. *Criminalística: Procedimentos e Metodologias*. Campinas/SP: Millennium Editora, 2015. cap. 12, p. 321–357.

PAN, X.; ZHANG, X.; LYU, S. Detecting splicing in digital audios using local noise level estimation. In: IEEE. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. [S.l.], 2012. p. 1841–1844.

PEÑA, D.; PRIETO, F. J. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, Taylor & Francis, 2012.

PILON, B. H. et al. Gaussian process for regression in business intelligence: A fraud detection application. Citeseer, 2015.

REIS, P. M. G. I. et al. Audio authentication using the kurtosis of ESPRIT based ENF estimates. In: *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS) (ICSPCS'2016)*. Surfers Paradise, Australia: [s.n.], 2016. Submetido.

REIS, P. M. G. I. et al. ESPRIT-Hilbert based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency. *Information Forensics and Security , IEEE Transactions on*, IEEE, 2016. Submetido.

RODRIGUES, A. R.; BERNACCHI, P. E. E.; WIESELTHALER, S. C. M. A valoração das escutas telefônicas como meio de obtenção de prova. Uma relação entre direitos fundamentais. *Temiminós Revista Científica*, v. 6, n. 1, p. 43–54, 2016.

RODRÍGUEZ, D. P. N. *Autenticação de Áudio Digital baseada na Presença da Frequência da Rede Elétrica*. Dissertação (Mestrado) — Instituto Militar de Engenharia, Rio de Janeiro, 2010.

RODRÍGUEZ, D. P. N.; APOLINÁRIO JR., J. A. Evaluating digital audio authenticity with spectral distances and ENF phase change. In: IEEE. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. [S.l.], 2009. p. 1417–1420.

RODRÍGUEZ, D. P. N.; APOLINÁRIO JR., J. A.; BISCAINHO, L. W. Audio authenticity based on the discontinuity of ENF higher harmonics. In: IEEE. *21st European Signal Processing Conference (EUSIPCO 2013)*. [S.l.], 2013. p. 1–5.

RODRÍGUEZ, D. P. N.; APOLINÁRIO JR., J. A.; BISCAINHO, L. W. P. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *Information Forensics and Security, IEEE Transactions on*, IEEE, v. 5, n. 3, p. 534–543, 2010.

ROY, R.; KAILATH, T. ESPRIT-estimation of signal parameters via rotational invariance techniques. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, IEEE, v. 37, n. 7, p. 984–995, 1989.

SCHMIDT, R. O. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, IEEE, v. 34, n. 3, p. 276–280, 1986.

SHORT, J. A.; INFIELD, D. G.; FRERIS, L. L. Stabilization of grid frequency through dynamic demand control. *Power Systems, IEEE Transactions on*, IEEE, v. 22, n. 3, p. 1284–1293, 2007.

SILVEIRA, M. A. et al. Convolutive ica-based forensic speaker identification using mel frequency cepstral coefficients and Gaussian mixture models. *The International Journal of Forensic Computer Science–IJoFCS*, v. 1, p. 27–34, 2013.

TÁVORA, R. G.; NASCIMENTO, F. A. Detecting replicas within audio evidence using an adaptive audio fingerprinting scheme. *Journal of the Audio Engineering Society*, Audio Engineering Society, v. 63, n. 6, p. 451–462, 2015.

WELCH, P. D. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, v. 15, n. 2, p. 70–73, 1967.

YANG, R.; QU, Z.; HUANG, J. Exposing MP3 audio forgeries using frame offsets. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, ACM, v. 8, n. 2S, p. 35, 2012.