INSTITUTO MILITAR DE ENGENHARIA

RODRIGO TAVARES DOS SANTOS

REALCE DE SINAIS DE VOZ COM ESTIMAÇÃO ROBUSTA DE RUÍDOS ACÚSTICOS NÃO-ESTACIONÁRIOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Engenharia Elétrica do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Ciências em Engenharia Elétrica.

Orientador: Prof. Rosângela Fernandes Coelho - Docteur ENST

Rio de Janeiro
2014

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

Santos, Rodrigo Tavares
Técnicas de Realce de Sinais de Voz com Uso de Métodos de Detecção e Estimação Aplicados à Identificação e Seleção de Bandas Corrompida / Rodrigo Tavares dos Santos, orientado por Rosângela Fernandes Coelho. - Rio de Janeiro : Instituto Militar de Engenharia, 2014.

Dissertação (mestrado) - Instituto Militar de Engenharia - Rio de Janeiro, 2014.

1. Engenharia elétrica - teses, dissertações. 2. Processamento de sinais. 3. Sinais de voz 4. Acústica I. Coelho, Rosângela Fernandes II. Título III. Instituto Militar de Engenharia.

INSTITUTO MILITAR DE ENGENHARIA

RODRIGO TAVARES DOS SANTOS

REALCE DE SINAIS DE VOZ COM ESTIMAÇÃO ROBUSTA DE RUÍDOS ACÚSTICOS NÃO-ESTACIONÁRIOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Engenharia Elétrica do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Ciências em Engenharia Elétrica.

Orientador: Prof. Rosângela Fernandes Coelho - Docteur ENST

Aprovada em 21 de maio de 2014 pela seguinte Banca Examinadora:

Prof. Rosângela Fernandes Coelho - Docteur ENST do IME - Presidente

---

Prof. Weiler Alves Finamore - Ph.D. da UFJF

Prof. Ernesto Pinto Leite - D.Sc. do IME

Prof. Paulo Fernando Ferreira Rosa - Ph.D. do IME

Rio de Janeiro
2014

AGRADECIMENTOS

À Prof. Rosângela Fernandes Coelho, minha orientadora, por toda a paciência e incentivo, que foram essenciais para o desenvolvimento desta Dissertação. Não posso esquecer-me de agradecer também por me ensinar a montar as peças deste imenso quebra-cabeça chamado ciência.

À minha esposa, Juliana, por todo o amor e cuidado durante a realização deste Mestrado,

Aos meus pais Murilo e Ligia, à minha irmã Lucelia, por todo carinho e amor fundamentais nesta etapa da minha jornada,

A todos os familiares que compreenderam e me perdoaram pelo afastamento para dedicação a este curso,

Aos colegas Zão, Dranka e Zucatelli do Laboratório de Processamento de Sinais Acústicos, do Instituto Militar de Engenharia, pela amizade que tornou a caminhada menos desgastante e ainda mais prazerosa,

Ao Instituto Militar de Engenharia, instituição que me proporcionou a realização deste curso de Mestrado,

A todos os professores e funcionários do IME, por contribuírem direta e indiretamente para minha formação,

Ao Banco do Brasil pela liberação parcial para realizar o curso de Mestrado,

A Deus, por estar presente na minha vida, na minha família e nos meus estudos, e por guiar sempre o meu caminho.

"Eu gosto do impossível porque lá a concorrência é menor."
Walt Disney

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

AI	*articulation index*
CSII	*Coherence and Speech intelligibility index*
DATE	*d-dimensional trimmed estimator*
EMD	*empirical mode decomposition*
EMD-DT	*EMD-based detrending*
EMDF	*EMD-based Hurst*
FAI	*fractional articulation index*
fGn	*fractional Gaussian noise*
fwSegSNR	*frequency-weighted segmental signal-to-noise ratio*
IMCRA	*improved minima controlled recursive averaging*
INS	*index of nonstationarity*
IS	distância de Itakura-Saito
LLR	*log-likelihood ratio*
LSA	*log-spectral amplitude*
MMSE	*minimum mean-square error*
MS	*minimum statistics*
OMLSA	*optimally-modified log-spectral amplitude*
OQCM	*overall quality composite measure*
PESQ	*Perceptual Evaluation Of Speech Quality*
SDR	*signal-to-distortion ratio*
SNR	*signal-to-noise ratio*

SS	*spectral subtraction*
STFT	*short-time Fourier transform*
STOI	*short-time objective intelligibility*
UnB-MMSE	*unbiased minimum mean-square error*
VAD	*voice activity detector*
WSS	*weighted spectral slope*

RESUMO

Nesta Dissertação, são estudadas soluções para reduzir o efeito de distorções acústicas em sinais de voz. Para tratar as distorções causadas por ruídos acústicos ambientais, é introduzido um método de realce de sinais de voz no domínio do tempo. Esta proposta adota o estimador robusto de desvio padrão como um critério para a seleção e estimação das componentes do ruído. Para avaliar o método de realce proposto, são utilizados ruídos coletados de diversas fontes acústicas com diferentes índices de não-estacionariedade. O método proposto aprimorou os resultados de seis medidas objetivas, selecionadas para avaliar a qualidade e a inteligibilidade dos sinais de voz. Cinco técnicas de realce existentes na literatura são adotadas como referência. A proposta alcançou os melhores resultados para a maioria dos experimentos realizados, principalmente para aqueles com ruídos altamente não-estacionários.

ABSTRACT

The main issue of this work is to reduce the effects of noise corruption in speech signals. A speech enhancement technique is proposed to reduce or suppress the signals distortion caused by acoustic noises. The proposed technique adopts a noise standard deviation estimator as a criterion to select and reckon noise components. Corrupted speech signal with different sources and nonstationarity indices are used to evaluate the proposed speech enhancement experiment. The proposed method improves the results of six objective measures, adopted to evaluate the speech signals in terms of both quality and intelligibility. For comparison, five other techniques are also considered in the experiments. The proposed technique leads to best results for most of the noise scenarios considered in the experiments, mainly for the highly nonstationary noises.

1 INTRODUÇÃO

Nas últimas décadas, o avanço da pesquisa no processamento de sinais de voz, impulsionou o desenvolvimento de importantes sistemas. Entre estes, destaca-se o reconhecimento de voz (ATAL, 1976; DODDINGTON, 1985), o reconhecimento de locutor ou individuo (REYNOLDS, 1995; MING, 2007) e a identificação acústica de emoções (KAISER, 1990; SCHULLER, 2009; ZÃO, 2014a). Uma das principais razões do uso da voz como sinal biométrico nestas soluções, se deve ao fato desta conter informações do indivíduo tais como idade, sexo, idioma e condições físico-emocionais. Além disso, a voz é o meio mais natural de comunicação do homem e também de fácil aquisição, não sendo necessário o uso de aparelhos sofisticados para a sua captação.

Um importante objetivo da área de pesquisa de processamento de voz é evitar a perda de qualidade destes sinais em ambientes com presença de ruídos acústicos. Por exemplo, sistemas de identificação de locutor podem ter a taxa de acertos reduzida em até 80% (MING, 2007; ZÃO, 2011). Desde a década de 1970 (BOLL, 1979), métodos de realce de sinais de voz têm sido propostos para atenuar as distorções causadas pelos ruídos.

As técnicas de realce de sinais de voz podem ser classificadas como espectrais e temporais. As principais soluções espectrais propostas na literatura são a subtração espectral (SS - *spectral subtraction*) (BOLL, 1979), a minimização do erro médio quadrático (MMSE - *minimum mean-square error*) (EPHRAIM, 1984) e o método de Cohen (COHEN, 2001, 2003). O desafio da área de pesquisa é atribuído às diferentes fontes acústicas, como por exemplo, pessoas conversando ao mesmo tempo, buzina de carro no engarrafamento, avião, toque de celular entre outros. Os ruídos também possuem diferentes distribuições de amplitude e estatísticas, e podem ser não-estacionários. Isto dificulta a obtenção e estimação das características dos ruídos e, consequentemente, reduz a precisão da informação fundamental para a eficiência das soluções de realce. O método de subtração espectral divide o sinal corrompido em quadros de curta duração para efetuar a análise no domínio da frequência de cada quadro com o uso da transformada de Fourier de tempo curto (STFT - *short-time Fourier transform*). Após a aplicação da STFT, emprega-se um método de estimação para determinar as componentes espectrais do ruído presentes no sinal de voz. Estas componentes são então subtraídas ou suprimidas do espectro do sinal corrompido,

e uma versão realçada do sinal de voz é reconstruída no domínio do tempo utilizando a transformada inversa de Fourier. O método de minimização do erro médio quadrático da magnitude dos coeficientes espectrais é utilizado para estimar o espectro do sinal de voz com o uso de modelos com distribuição Gaussiana. O método MMSE é empregado sobre o logaritmo da magnitude dos coeficientes espectrais (LSA - *log-spectral amplitude*).

Para lidar com a não-estacionariedade dos ruídos acústicos, foram propostos métodos que realizam a estimativa do ruído em longos segmentos, inclusive em regiões onde há presença da voz (MARTIN, 2001; COHEN, 2003). Ainda assim, mesmo os mais recentes métodos se mostraram incapazes de estimar fielmente as oscilações de ruídos altamente não-estacionários (MANOHAR, 2006).

Recentemente, surgiram métodos de realce no domínio do tempo baseados na teoria tempo-frequência (TF) (COHEN, 1995), tais como a decomposição *wavelets* (DONOHO, 1994) e a decomposição empírica de modos (EMD - *empirical mode decomposition*) (HUANG, 1998). A decomposição EMD foi proposta como uma forma não-linear e adaptativa para análise de sinais não-estacionários. A principal diferença entre as decomposições *wavelets* e a empirica de modos é que o EMD resulta em um conjunto de funções intrínsecas de modo (IMF - *intrinsic mode functions*), que são totalmente dependentes do próprio sinal, ou seja, as bases não são fixas. Dentre estes métodos baseados na TF, destacam-se o EMD-DT (EMD-*based detrending*) (FLANDRIN, 2004a), o EMDF (EMD-*based filtering*) (CHATLANI, 2012) e o EMDH (ZÃO, 2014b). Para estes, é necessário um critério de decisão para identificar quais componentes são mais afetadas pelo ruído. Em seguida, faz-se a exclusão e finalmente, a reconstrução do sinal com as componentes remanescentes. A primeira proposta apresentada na literatura de realce de sinais com o uso do EMD foi realizada por (FLANDRIN, 2004a). O método EMD-DT tem por objetivo eliminar o ruído de sinais de naturezas diversas. Neste, após a análise do sinal ruidoso com o uso do método EMD, as médias das IMFs resultantes são obtidas para selecionar quais modos são mais afetados por ruído. A reconstrução do sinal aprimorado é obtida a partir da soma dos demais modos. Para lidar com ruídos não-estacionários, a proposta de pós-realce EMDF foi empregada sobre sinais previamente aprimorados por técnicas espectrais. O método EMDH utiliza o expoente de Hurst (HURST, 1951) para selecionar os modos mais afetados por ruídos não-estacionários que apresentam altas concentrações de energia nas baixas frequências. Para avaliar os métodos de realce, a maioria das propostas apresentadas na literatura consideram apenas medidas de qualidade da voz. A

despeito dos testes subjetivos perceptuais serem a forma mais precisa para julgamento da qualidade de um sinal de voz, estes são frequentemente substituidos por medidas objetivas devido ao seu alto custo operacional (QUACKENBUSH, 1988; RIX, 2001; HU, 2008; BISPO, 2010).

Para uma medida objetiva ser considerada satisfatória, ela deve demonstrar uma alta correlação com os resultados perceptuais de qualidade obtidos por meio de testes subjetivos (HU, 2008). Todavia, não é possível julgar o grau inteligibilidade através destas medidas (LOIZOU, 2007b). Isso porque, embora os métodos de realce promovam melhora da qualidade dos sinais de voz, o seu uso pode degradar, por exemplo, a inteligibilidade de palavras (LOIZOU, 2007b). Desta forma, para avaliar os algoritmos de realce com relação à inteligibilidade, outras medidas objetivas devem ser empregadas. A busca e definição por medidas de inteligibilidade com tais características ainda é um dos principais objetivos da área de processamento de sinais acústicos.

Nesta Dissertação, é apresentada uma proposta de realce de sinais de voz que utiliza um estimador robusto de desvio padrão do ruído (PASTOR, 2012) como critério para a seleção e estimação das componentes de ruídos acústicos ambientais com características não-estacionárias. O método proposto é avaliado em termos de qualidade e inteligibilidade da voz utilizando seis medidas objetivas. Os ruídos acústicos considerados nos experimentos possuem diferentes índices de não-estacionariedade (INS - *index of nonstationarity*) (BORGNAT, 2010). Cinco métodos de realce são utilizados como referência na avaliação da proposta de realce de sinais de voz. Três destes métodos são espectrais: a subtração espectral, o método de Cohen e o método baseado na filtragem de Wiener. Os outros dois são baseados no método EMD: EMDF e EMDH.

Para os testes de avaliação, foram utilizadas medidas objetivas de qualidade e inteligibilidade. Nos experimentos, foram utilizados 24 locutores selecionados aleatoriamente da base de voz TIMIT, sendo 8 mulheres e 16 homens. Cada locutor gerou 10 gravações com duração média de 3s e amostradas à taxa de 16 kHz, totalizando 240 sinais de voz utilizados nos testes. Os sinais de voz foram corrompidos com seis ruídos ambientais provenientes de diferentes fontes de ruídos acústicos: balbúrdia, britadeira, fábrica, helicóptero, serra elétrica e trem. A escolha destes ruídos se deu em função dos diferentes valores de INS e dos espectrogramas possuírem formas distintas. Para os testes, os ruídos foram adicionados aos sinais de voz limpo para a obtenção de cinco diferentes valores de SNR: 10 dB, 5 dB, 0 dB, -5 dB e -10 dB.

1.1 OBJETIVOS

Os principais objetivos deste trabalho são:

- propor um método para realce, no domínio do tempo, de sinais de voz corrompidos por ruídos acústicos não-estacionários. Nesta proposta, denominada PRO, o sinal de voz corrompido é inicialmente dividido em quadros de mesmo tamanho. Em seguida, são obtidas as estimativas das componentes ruidosas com o uso de um estimador robusto de desvio padrão do ruído. A partir desta estimação, é realizado um teste para extrair as amplitudes constituídas predominantemente por ruídos. Finalmente, as demais amplitudes são atenuadas baseadas na estimação do desvio padrão do ruído, e o sinal de voz é reconstruído.

- investigar o uso do estimador robusto de desvio padrão (DATE - *d-dimensional trimmed estimator*) (PASTOR, 2012) como critério de identificação e estimação das componentes de ruído. Para isto, o desempenho do estimador adotado na presente proposta é comparado com o estimador de desvio médio absoluto (MAD - *median absolute desviation*) (HUBER, 2009).

- avaliar o método de realce proposto para sinais de voz distorcidos por ruídos de distintas fontes acústicas reais. Avaliar os ruídos acústicos segundo os seus índices de não-estacionariedade. Adotar seis medidas objetivas que apresentam alta correlação com a qualidade e a inteligibilidade da voz para examinar o método proposto;

1.2 RESULTADOS OBTIDOS

Os principais resultados e contribuições obtidos no desenvolvimento desta Dissertação são:

- proposta de um método de realce para sinais de voz corrompidos por ruídos acústicos reais não-estacionários. Os resultados obtidos nos experimentos de realce demonstraram que o método proposto aprimorou seis medidas objetivas utilizadas para avaliar a qualidade e a inteligibilidade dos sinais de voz. Pode-se destacar que os ganhos de razão sinal-ruído segmental (SegSNR - *segmental signal-to-noise ratio*) da proposta para valores de SNR maiores que zero foram duas vezes maior que as técnicas espectrais (SS, Cohen e Wiener), e uma vez das técnicas temporais (EMDF e

EMDH). A proposta obteve ainda incremento de aproximadamente 1 dB em relação aos demais algoritmos para SegSNR com ponderação em frequência (fwSegSNR - *frequency weighted SegSNR*) nos ruídos altamente não-estacionários.

- avaliação da proposta de realce em termos de inteligibilidade do sinal de voz. A proposta conseguiu aumentar a taxa de acertos de sentenças em 12% quando avaliada pela medida objetiva de inteligibilidade em tempo curto (STOI - *short-time objective intelligibility*) (TAAL, 2011). Este resultado foi bem acima dos obtidos pelos demais métodos. Estes ganhos de inteligibilidade também foram vistos na medida de coerência e inteligibilidade de voz (CSII - *coherence and speech intelligibility index*) (KATES, 2005), que avalia além da inteligibilidade as distorções causadas pelo método de realce. Nesta, a proposta obteve incrementos de 6% em relação às soluções temporais e 12% em relação às espectrais.

1.3 ORGANIZAÇÃO DA TESE

O restante deste trabalho está organizado da seguinte forma:

- **Capítulo 2**: neste Capítulo, são primeiramente introduzidos três métodos espectrais de realce de sinais de voz: a subtração espectral, a proposta de Cohen e o método baseado na filtragem de Wiener. Ainda neste Capítulo, são introduzidos os principais conceitos sobre o método de decomposição EMD, seguido da apresentação dos métodos EMDF (CHATLANI, 2012) e EMDH (ZÃO, 2014b). Ao final do Capítulo, são apresentadas duas medidas objetivas de qualidade: a razão sinal-ruído segmental e uma medida composta de qualidade de voz (OQCM - *overall quality composite measure*) (LOIZOU, 2007b), e quatro medidas de inteligibilidade: fwSegSNR, CSII, STOI e a medida de articulação fracionária (FAI -*fractional articulation index*) (LOIZOU, 2011a).

- **Capítulo 3**: são descritas as três etapas do realce de sinais de voz proposto neste trabalho. Na primeira etapa, de identificação e estimação das componentes de ruído, é apresentado o método de estimação robusta do desvio padrão do ruído (DATE) utilizado como critério de seleção e estimação das componentes ruidosas. Para validar a sua escolha, ele é comparado com o método de estimação MAD, considerado na literatura como o mais robusto, por meio de testes de estimação para diferentes

ruídos não-estacionários em diferentes razões sinal-ruído. Os resultados mostram que o estimador DATE consegue obter um alto grau de precisão. Finalmente, são descritas as alterações realizadas no DATE para a sua utilização nos ruídos não-estacionários. Na etapa seguinte é definida a forma de extração dos componentes do ruído a partir da estimação realizada pelo DATE.

- **Capítulo 4**: os experimentos para avaliação da algoritmo de realce PRO são apresentados neste Capítulo. Os resultados são obtidos utilizando sinais de voz da base TIMIT (GAROFOLO, 1993) corrompidos por seis ruídos coletados em diferentes fontes acústicas reais. Inicialmente, apresenta-se a definição e os resultados de INS (BORGNAT, 2010) para os ruídos selecionados. Em seguida, os métodos de realce são avaliados por seis medidas objetivas de qualidade e inteligibilidade: SegSNR, OQCM, fwSegSNR, CSII, STOI e FAI.

- **Capítulo 5**: Finalmente, este Capítulo expõe as principais conclusões e contribuições desta Dissertação. Também são destacadas sugestões para trabalhos futuros.

2 MÉTODOS DE REALCE DE SINAIS DE VOZ E MEDIDAS DE QUALIDADE E INTELIGIBILIDADE

As distorções causadas pelos ruídos acústicos ambientais no sinal de voz, representam um grande desafio para área de processamento de sinais. A captação da voz em ambientes com presença de ruídos acústicos reduz o desempenho das soluções, como por exemplo, o reconhecimento automático de locutor. A literatura apresenta como forma de lidar com estes tipos de distorções métodos de realce de sinais de voz, que têm por objetivo remover ou atenuar os efeitos causados pelos ruídos acústicos.

A maior parte das soluções de realce utilizam a transformada rápida de Fourier para estimar o espectro do ruído. Para isso, é necessária a localização de trechos do sinal onde não ocorra a atividade de voz, por este motivo, geralmente, são utilizados detectores de atividade de voz (VAD - *voice activity detector*). O grande desafio enfrentado por estes métodos é a obtenção de estimativas precisas das estatísticas do ruído. Esta dificuldade ocorre porque os ruídos são proveniente de diversas fontes acústicas, apresentam diferentes tipos de distribuições de amplitude (Gaussianas e não-Gaussianas) e são não-estacionários.

Com objetivo de melhorar o desempenho das aplicações de realce de sinais de voz foram propostos alguns métodos de estimação de ruídos não-estacionários, dentre os quais podem-se destacar estatísticas mínimas (MS - *minimum statistics*) (MARTIN, 2001) e o método IMCRA (*improved minima controlled recursive averaging*) (COHEN, 2003). Nestas propostas, as estimativas são realizadas por meio de observações de uma dada quantidade de quadros passados. Segundo (MANOHAR, 2006), estes algoritmos tornam-se lentos no acompanhamento das variações espectrais de ruídos não-estacionários, devido à necessidade de observação de momentos anteriores. Uma solução a este problema pode ser encontrado no método de estimação UnB-MMSE (*unbiased minimum mean-square error*) (GERKMANN, 2012), obtido a partir da minimização de erro médio quadrático. Esta solução foi desenvolvida para capturar com menor tempo de resposta as variações espectrais dos ruídos não-estacionários. Todavia, em (GERKMANN, 2012) é demonstrado que nenhum destes estimadores consegue acompanhar precisamente estas oscilações.

Outra abordagem, introduzida na literatura recentemente, é baseada na análise tempo-frequência para o realce de sinais de voz utilizando como ferramenta a decomposição

empírica de modos (HUANG, 1998). Ao contrário dos métodos espectrais, o realce baseado no EMD não necessita da estimação explícita das estatísticas dos ruídos acústicos, nem de que os sinais analisados sejam estacionários.

Para avaliar o aprimoramento do sinal gerado pelos métodos de realce de sinais de voz, a literatura apresenta duas formas:

- Avaliação perceptual subjetiva;
- Avaliação objetiva de qualidade de voz.

A avaliação perceptual subjetiva utiliza ouvintes para julgar a qualidade do sinal após a aplicação do método de realce de sinais de voz. Esta forma de avaliação é considerada a mais apropriada para examinar as soluções de realce de sinais de voz. No entanto, é necessário despender muito tempo e recursos financeiros. As avaliações objetivas de qualidade de voz, utilizam medidas que comparam o sinal de voz limpo com o sinal de voz aprimorado pelo método de realce. As medidas existentes na literatura geralmente apresentam alto coeficiente de correlação com os resultados alcançados por testes subjetivos. Todavia, a melhora na qualidade não implica necessariamente em aumento na inteligibilidade dos sinais de voz, pois como apresentado por (LOIZOU, 2007b) na avaliação do aprimoramento gerado por treze métodos de realce, a melhora na qualidade do sinal de voz provocou redução na taxa de acerto de sentenças. Deste modo, o ideal é que todo método de realce de sinais de voz leve em consideração tanto o aspecto da qualidade de voz quanto da inteligibilidade.

Neste Capítulo, são apresentados alguns dos principais métodos de realce de sinais de voz em situações de ruídos acústicos não-estacionários. Primeiramente, são apresentados três algoritmos espectrais: a proposta de Cohen (COHEN, 2001, 2003), a UnB-MMSE com o filtro de Wiener (SCALART, 1996; GERKMANN, 2012) e o método clássico de subtração espectral (BOLL, 1979). Em seguida, são introduzidos os algoritmos tempo-frequência EMDF (CHATLANI, 2012) e EMDH (ZÃO, 2014b). São apresentadas ainda medidas objetivas relacionadas à qualidade e à inteligibilidade do sinal de voz. Primeiramente duas medidas de qualidade do sinal de voz, a razão sinal ruído segmental (SegSNR) e a medida OQCM (HU, 2006). E depois serão mostradas medidas objetivas de inteligibilidade: a razão sinal-ruído com ponderação em frequência (fwSegSNR) (HU, 2008), a STOI (TAAL, 2011), a CSII (KATES, 2005) e a FAI (LOIZOU, 2011a).

2.1 MÉTODOS DE REALCE DE SINAIS DE VOZ

Esta Seção contém os principais métodos de realce de sinais de voz. Primeiramente são apresentados os algoritmos que utilizam a transformada de Fourier de tempo curto para avaliar o sinal ruidoso no domínio da frequência, SS, Cohen e Wiener. Depois são mostradas as propostas de realce no domínio do tempo, EMDF e EMDH. Antes da apresentação destas, é exposto um breve resumo do método de decomposição empírica de modos.

2.1.1 SUBTRAÇÃO ESPECTRAL

Para a realização deste método de realce de sinais de voz é necessário utilizar a transformada de Fourier de tempo curto para analisar o sinal ruidoso no domínio da frequência. Seja $y(t)$ o sinal resultante de um sinal de voz limpo $x(t)$ distorcido por um ruído aditivo $\eta(t)$. Então, pode-se escrever $y(t) = x(t)+\eta(t)$. Se $Y(\kappa,\tau)$, $X(\kappa,\tau)$ e $\mathcal{N}(\kappa,\tau)$ representam a STFT de $y(t)$, $x(t)$ e $\eta(t)$, respectivamente, então

$$Y(\kappa,\tau) = X(\kappa,\tau) + \mathcal{N}(\kappa,\tau)\,, \tag{2.1}$$

onde τ e κ são, respectivamente, os índices de quadro e frequência (LOIZOU, 2007a).

A subtração espectral considera o ruído aditivo ao sinal de voz, e estabelece que a estimação do espectro do sinal limpo é obtida através da subtração da estimativa do espectro do ruído do espectro do sinal de voz corrompido. Originalmente em (BOLL, 1979), o ruído foi considerado como estacionário e a estimação e a atualização de suas componentes deveria ocorrer apenas nos momentos em que não houvesse presença de voz. Para reconstruir o sinal de voz foi utilizado o espectro estimado para o sinal de voz limpo juntamente com a informação de fase do sinal corrompido.

Primeiro deve-se considerar a separação em magnitude e fase obtida pela forma polar da STFT do sinal corrompido,

$$Y(\kappa,\tau) = |Y(\kappa,\tau)|\, e^{j\phi_y(\kappa,\tau)}\,. \tag{2.2}$$

No algoritmo SS (BOLL, 1979), a magnitude do sinal limpo é obtida por

$$|\hat{X}(\kappa,\tau)| = \begin{cases} |Y(\kappa,\tau)| - |\hat{\mathcal{N}}(\kappa,\tau)| & \text{, se } |Y(\kappa,\tau)| > |\hat{\mathcal{N}}(\kappa,\tau)|, \\ 0 & \text{, caso contrário.} \end{cases} \tag{2.3}$$

No momento seguinte, cada quadro τ do sinal realçado $\hat{x}(t)$ é reconstruído a partir da transformada inversa de Fourier aplicada no espectro estimado de $\hat{X}(\kappa,\tau)$. Para obter $\hat{X}(\kappa,\tau)$ é utilizada a informação de fase do sinal corrompido, ou seja,

$$\hat{X}(\kappa,\tau) = |\hat{X}(\kappa,\tau)|\, e^{j\phi_y(\kappa,\tau)}\,. \tag{2.4}$$

2.1.2 O MÉTODO DE COHEN

O método de Cohen foi proposto em (COHEN, 2001) e (COHEN, 2003) para realçar sinais de voz corrompidos por ruídos não-estacionários. Isto é possível pois este emprega o método IMCRA para a atualização das estimativas do espectro de potência dos ruídos. Com a estimativa obtida com o método IMCRA, o sinal de voz é reconstruído utilizando-se o algoritmo OMLSA (*optimally-modified log spectral amplitude*) (COHEN, 2001), que minimiza o erro médio quadrático do logaritmo da magnitude espectral.

O estimador IMCRA é dividido em duas iterações, onde cada uma possui duas fases, uma de suavização do espectro de potência do sinal ruidoso e outra de localização por estatísticas mínimas, que tem o objetivo de estimar o espectro de potência do ruído acústico presente no sinal de voz.

A primeira iteração começa com o uso da STFT sobre o sinal de voz corrompido. Logo após, uma versão suavizada de $|Y(\kappa,\tau)|^2$ na frequência ($S_f(\kappa,\tau)$) e no tempo ($S(\kappa,\tau)$) é obtida por

$$\begin{cases} S_f(\kappa,\tau) = \sum_{i=-w}^{w} W(i)|Y(\kappa-i,\tau)|^2\,, \\ S(\kappa,\tau) = \delta_s\, S(\kappa,\tau-1) + (1-\delta_s) S_f(\kappa,\tau)\,, \end{cases} \tag{2.5}$$

onde $W(i)$ é uma janela normalizada ($\sum_{i=-w}^{w} W(i) = 1$) para calcular a média entre valores vizinhos em frequência de $|Y(\kappa,\tau)|^2$, e $\delta_s \in [0,1]$ é o parâmetro de suavização no tempo utilizado para atualizar os valores de $S(\kappa,\tau)$ a partir de $S_f(\kappa,\tau)$. Adotando o mesmo princípio do método MS, uma estimativa para o espectro de potência do ruído pode ser adquirida pelos valores mínimos de $S(\kappa,\tau)$ em um conjunto de Q quadros passados,

$$S_{\min}(\kappa,\tau) = \min\left\{S(\kappa,\tau') \,|\, \tau - Q + 1 \leq \tau' \leq \tau\right\}\,. \tag{2.6}$$

Deste modo, é considerado que em pelo menos um dentre estes Q quadros anteriores, a voz estará ausente, e

$$E\left\{S_{\min}(\kappa,\tau)\right\} = B_{\min}^{-1}\, E\left\{|\mathcal{N}(\kappa,\tau)|^2\right\}\,, \tag{2.7}$$

onde $B_{\min}$ é um fator de correção de tendência (*bias*) que pode ser obtido de maneira empírica. Como indicado em (COHEN, 2003), o valor adotado para o fator de correção de *bias* é $B_{\min} = 1,66$.

São definidas as seguintes grandezas para determinar o VAD na primeira iteração,

$$\begin{aligned} \gamma_{\min}(\kappa,\tau) &\triangleq \frac{|Y(\kappa,\tau)|^2}{B_{\min}\, S_{\min}(\kappa,\tau)} \quad ; \\ \zeta(\kappa,\tau) &\triangleq \frac{S(\kappa,\tau)}{B_{\min}\, S_{\min}(\kappa,\tau)}\,. \end{aligned} \tag{2.8}$$

Em cada quadro e índice de frequência, a decisão sobre a ausência ou presença de voz é dada por

$$I(\kappa,\tau) = \begin{cases} 1\,, & \text{se } \begin{matrix}\gamma_{\min}(\kappa,\tau) < \gamma_0 \\ \text{e } \zeta(\kappa,\tau) < \zeta_0\end{matrix} \quad \text{(voz está ausente)} \\ 0\,, & \text{caso contrário} \qquad\quad \text{(voz está presente)} \end{cases} \tag{2.9}$$

Na segunda iteração, um novo espectro suavizado $\tilde{S}_f(\kappa,\tau)$ é definido usando apenas as regiões do sinal corrompido onde o algoritmo não detectou atividade de voz, isto é, $I(\kappa,\tau) = 1$. A partir de $\tilde{S}_f(\kappa,\tau)$, as grandezas $\tilde{S}_{min}(\kappa,\tau)$, $\tilde{\gamma}_{\min}(\kappa,\tau)$ e $\tilde{\zeta}(\kappa,\tau)$ são definidas de forma análoga às EQS. 2.6 e 2.8.

Considere as hipóteses de ausência $\mathcal{H}_0(\kappa,\tau)$ e presença de voz $\mathcal{H}_1(\kappa,\tau)$ no quadro τ e no índice de frequência κ. A probabilidade condicional de presença de voz $p(\kappa,\tau) \triangleq P\left(\mathcal{H}_0(\kappa,\tau)|\gamma(\kappa,\tau)\right)$ foi deduzida em (COHEN, 2003) como

$$p(\kappa,\tau) = \left(1 + \frac{q(\kappa,\tau)}{1-q(\kappa,\tau)}(1+\xi(\kappa,\tau))\exp\left\{v(\kappa,\tau)\right\}\right)^{-1}, \tag{2.10}$$

onde $v \triangleq \gamma\xi/(\xi+1)$ e a probabilidade *a priori* de ausência de voz, $q(\kappa,\tau) = P(\mathcal{H}_0(\kappa,\tau))$, pode ser estimada por

$$\hat{q}(\kappa,\tau) = \begin{cases} 1\,, & \text{se } \begin{matrix}\hat{\gamma}_{\min}(\kappa,\tau) \le 1 \\ \text{e } \hat{\zeta}(\kappa,\tau) < \zeta_0\end{matrix} \quad ; \\ \dfrac{\gamma_1 - \tilde{\gamma}_{\min}(\kappa,\tau)}{\gamma_1 - 1}\,, & \text{se } \begin{matrix}1 < \hat{\gamma}_{\min}(\kappa,\tau) \le \gamma_1 \\ \text{e } \hat{\zeta}(\kappa,\tau) < \zeta_0\end{matrix} \quad ; \\ 0\,, & \text{em outros casos.} \end{cases} \tag{2.11}$$

Através da probabilidade $p(\kappa,\tau)$, o espectro de potência do ruído do quadro seguinte ($|\bar{\mathcal{N}}(\kappa,\tau+1)|^2$) é recursivamente estimado por

$$|\bar{\mathcal{N}}(\kappa,\tau+1)|^2 = \tilde{\delta}_\eta(\kappa,\tau)|\bar{\mathcal{N}}(\kappa,\tau)|^2 + [1-\tilde{\delta}_\eta(\kappa,\tau)]|Y(\kappa,\tau)|^2\,, \tag{2.12}$$

onde $\tilde{\delta}_\eta(\kappa,\tau)$ é um parâmetro de suavização variável que depende de $p(\kappa,\tau)$.

É empregado, para estimar a versão final para o espectro do ruído, um fator de compensação multiplicativo,

$$|\hat{\mathcal{N}}(\kappa,\tau)|^2 = B\,|\bar{\mathcal{N}}(\kappa,\tau)|^2\,. \tag{2.13}$$

Novamente faz-se necessário a utilização de um fator de correção, pois o espectro do ruído $|\bar{\mathcal{N}}(\kappa,\tau)|^2$ é subestimado pelo estimador IMCRA, uma vez que este é derivado do método de estatísticas mínimas.

Depois da aplicação do estimador IMCRA (EQS. 2.5 a 2.13), o algoritmo OMLSA é utilizado para obter o espectro do sinal de voz $|\hat{X}(\kappa,\tau)|$. O OMLSA é uma versão modificada do estimador LSA (EPHRAIM, 1985), cuja finalidade é minimizar o erro médio quadrático entre os logaritmos das magnitudes espectrais dos sinais de voz limpo e realçado,

$$E_{\min}\left\{\Big(\log|X(\kappa,\tau)| - \log|\hat{X}(\kappa,\tau)|\Big)^2\right\}\,. \tag{2.14}$$

O ganho $G_{\mathrm{OMLSA}}(\kappa,\tau)$ a ser aplicado sobre o espectro do sinal de entrada é dado por (COHEN, 2001)

$$G_{\mathrm{OMLSA}}(\kappa,\tau) = \{G_{\mathrm{LSA}}(\kappa,\tau)\}^{p(\kappa,\tau)} G_{\min}^{1-p(\kappa,\tau)}\,, \tag{2.15}$$

onde a probabilidade condicional de presença de voz é calculada pela EQ. 2.10 e o limiar mínimo $G_{\min}$ para o ganho corresponde a -25 dB. Já o ganho do estimador LSA foi deduzido em (EPHRAIM, 1985) como

$$G_{\mathrm{LSA}}(\kappa,\tau) = \frac{\xi(\kappa,\tau)}{1+\xi(\kappa,\tau)} \exp\left\{\frac{1}{2}\int_{v(\kappa,\tau)}^{\infty} \frac{e^{-t}}{t}\,\mathrm{d}t\right\}\,, \tag{2.16}$$

onde o valor da SNR *a priori* é recursivamente estimado por

$$\hat{\xi}(\kappa,\tau) = \delta_{\mathrm{LSA}}\, G_{\mathrm{LSA}}^2(\kappa,\tau-1)\gamma(\kappa,\tau-1) + (1-\delta_{\mathrm{LSA}})\max\{\gamma(\kappa,\tau)-1,0\}\,. \tag{2.17}$$

Foram estipulados em (COHEN, 2003) os valores recomendados para os diversos parâmetros utilizados no estimador IMCRA e no método OMLSA. Estes valores foram definidos considerando taxa de amostragem de 16 kHz. Para os limiares das EQS. 2.9 e 2.11, foram sugeridos $\gamma_0 = 4,6$, $\zeta_0 = 1,67$ e $\gamma_1 = 3$. O valor de γ_1 está relacionado com o fator de compensação da EQ. 2.13 por

$$B = \frac{\gamma_1 - 1 - e^{-1} + e^{-\gamma_1}}{\gamma_1 - 1 - 3e^{-1} + (\gamma_1+2)e^{-\gamma_1}}\,, \tag{2.18}$$

resultando em $B = 1,47$. Já o coeficiente de suavização da EQ.2.17 foi determinado como $\delta_{\mathrm{LSA}} = 0,92$.

2.1.3 FILTRAGEM DE WIENER COM ESTIMADOR UNB-MMSE

O estimador UnB-MMSE (GERKMANN, 2012) é um método utilizado para obter as componentes espectrais do ruído, que são então suprimidas do espectro do sinal de voz através da filtragem de Wiener (WIENER, 1949), baseada na estimação da SNR *a priori* estabelecida em (SCALART, 1996).

O estimador UnB-MMSE é derivado da minimização de erros médios quadráticos definida em (HENDRIKS, 2010). O autor considera a hipótese de que os coeficientes espectrais tanto do ruído quanto do sinal de voz apresentam distribuição Gaussiana (HENDRIKS, 2010). Dessa forma, foi deduzido o estimador MMSE para o valor do peridograma do ruído $|\mathcal{N}(\kappa,\tau)|^2$,

$$E\left(|\mathcal{N}(\kappa,\tau)|^2 \,|\, Y(\kappa,\tau)\right) = \left(\frac{1}{1+\hat{\xi}(\kappa,\tau)}\right)^2 |Y(\kappa,\tau)|^2 + \frac{\hat{\xi}(\kappa,\tau)}{1+\hat{\xi}(\kappa,\tau)} |\hat{\mathcal{N}}(\kappa,\tau-1)|^2 . \quad (2.19)$$

Partindo do pressuposto que em quadros consecutivos, o espectro do ruído possui variação menor que o da voz, foram estimados o valor da SNR *a posteriori* $\hat{\gamma}(\kappa,\tau)$ adotando o espectro de potência do ruído obtido no quadro anterior,

$$\hat{\gamma}(\kappa,\tau) = \frac{|Y(\kappa,\tau)|^2}{|\hat{\mathcal{N}}(\kappa,\tau-1)|^2} , \quad (2.20)$$

e a SNR *a priori* é estimada por

$$\hat{\xi}(\kappa,\tau) = \max\left\{\hat{\gamma}(\kappa,\tau) - 1, 0\right\} . \quad (2.21)$$

Assim sendo, a estimação do espectro de potência do ruído pode ser atualizada de um quadro para outro pela relação recursiva

$$|\hat{\mathcal{N}}(\kappa,\tau)|^2 = \alpha_p |\hat{\mathcal{N}}(\kappa,\tau-1)|^2 + (1-\alpha_p) E\left(|\mathcal{N}(\kappa,\tau)|^2 | Y(\kappa,\tau)\right) . \quad (2.22)$$

Em (GERKMANN, 2012), foi proposta uma alteração do estimador MMSE apresentado em (HENDRIKS, 2010). Para isto, a estimação do periodograma da EQ. 2.19 foi reformulada utilizando as probabilidades de ausência e presença de voz:

$$E\left(|\mathcal{N}|^2|Y\right) = P(\mathcal{H}_0|Y)\,|Y|^2 + P(\mathcal{H}_1|Y)\,|\hat{\mathcal{N}}|^2 . \quad (2.23)$$

Com o objetivo de resolver a EQ. 2.23, as probabilidades condicionais são definidas como

$$P(\mathcal{H}_1|Y(\kappa,\tau)) = \left(1 + (1+\xi_{\text{opt}}) e^{-\hat{\gamma}(\kappa,\tau)\frac{\xi_{\text{opt}}}{1+\xi_{\text{opt}}}}, \right)^{-1} \quad (2.24)$$

e $P(\mathcal{H}_0|Y(\kappa,\tau)) = 1-P(\mathcal{H}_1|Y(\kappa,\tau))$. O valor considerado ótimo para a SNR *a priori*, ξ_{opt} na EQ. 2.24, foi definido como 15 dB (GERKMANN, 2012). Existe ainda outra vantagem do UnB-MMSE sobre os estimadores IMCRA e de outros algoritmos baseados no MS, o estimador UnB-MMSE não precisa captar informações de uma grande quantidade de quadros anteriores para a estimação do espectro do ruído. Isto dá ao UnB-MMSE um atraso menor na captação das variações espectrais dos ruídos não-estacionários.

Após a estimação das componentes espectrais do ruído, o espectro do sinal de voz é obtido pelo método baseado no filtro de Wiener exibida em (SCALART, 1996). O filtro de Wiener foi escolhido por ser um estimador ótimo, que adota as mesmas hipóteses do estimador UnB-MMSE. Isto é, os coeficientes espectrais do ruído e do sinal de voz obedecem a distribuições Gaussianas. Nesta abordagem, o ganho de Wiener $G_W(\kappa,\tau)$, aplicado sobre o espectro do sinal corrompido, é dado por (SCALART, 1996)

$$G_W(\kappa,\tau) = \frac{\xi(\kappa,\tau)}{1+\xi(\kappa,\tau)}\,. \tag{2.25}$$

Para a estimação da SNR *a priori*, $\hat{\xi}(\kappa,\tau)$, é empregada a decisão direta demonstrada em (EPHRAIM, 1984),

$$\hat{\xi}(\kappa,\tau) = \alpha_W\, G_W^2(\kappa,\tau-1)\,\gamma(\kappa,\tau-1) + (1-\alpha_W)\,\max\{\gamma(\kappa,\tau)-1,0\}\,. \tag{2.26}$$

Os valores utilizados, em (GERKMANN, 2012), para as constantes de suavização das EQS. 2.22 e 2.26 foram $\alpha_p = 0,8$ (HENDRIKS, 2010) e $\alpha_W = 0,98$ (SCALART, 1996).

2.1.4 O MÉTODO EMD

No trabalho (HUANG, 1998) a decomposição empírica de modos foi apresentada como uma forma não-linear para análise de sinais não-estacionários. Este método gera um conjunto de funções intrínsecas de modo e um resíduo. As IMFs são inteiramente dependentes do sinal analisado. Seja um sinal $y(t)$ contendo dois máximos locais consecutivos nos pontos t_- e t_+. Para valores de t no intervalo $t_- \leq t \leq t_+$, pode-se definir uma componente de altas frequências do sinal que passa por estes máximos e pelo mínimo local que existe entre eles. A partir desta componente, chamada de detalhe $d(t)$, localiza-se uma componente de tendência local ou resíduo $r(t)$, tal que $y(t) = d(t) + r(t)\,,\ t_- \leq t \leq t_+\,.$

Quando a decomposição é aplicada sobre todo o sinal $y(t)$, a IMF será definida pelo conjunto das componentes de detalhes. O sinal residual é definido pelo conjunto de todas

as componentes de tendência local. Aplicando-se repetidamente o procedimento sobre o sinal residual, chega-se a um conjunto de IMFs e a um resíduo de baixas frequências.

O algoritmo para o método EMD aplicado sobre um sinal $y(t)$ pode ser dividido nas seguintes etapas (HUANG, 1998) (FLANDRIN, 2004b):

a) Identificação de todos os pontos de máximo $y_{max}(t)$ e mínimo $y_{min}(t)$ locais;

b) Obtenção das envoltórias $e_{max}(t)$ e $e_{min}(t)$, a partir da interpolação dos pontos de máximo e de mínimo, respectivamente. Para isso, adota-se nesta etapa o uso de interpolação polinomial de terceiro grau utilizando o método de *splines*.;

c) Cálculo do resíduo como a média entre as envoltórias: $r(t) = (e_{min}(t) + e_{max}(t))/2$;

d) Extração das componentes de detalhes: $d(t) = y(t) - r(t)$;

e) Repetição da iteração sobre o sinal residual $r(t)$.

De acordo com (HUANG, 1998), por definição, toda IMF deve obedecer às seguintes propriedades:

- O número de extremos e de cruzamentos em zero devem ser iguais ou se diferenciar em uma unidade;
- O valor médio definido pelas envoltórias dos seus máximos e mínimos deve ser nulo.

Caso a componente de detalhes $d(t)$, extraída no passo (d) do algoritmo EMD, não obedeça às propriedades acima, as etapas (a-d) serão novamente efetuadas, com $d(t)$ no lugar de $y(t)$. Este processo, denominado *sifting*, deve ser repetido até garantir que a nova função $d(t)$ seja considerada uma IMF. Ao final de um número finito (M) de iterações, o sinal pode ser escrito como

$$y(t) = \sum_{m=1}^{M} \mathrm{IMF}_m(t) + r(t)\,, \tag{2.27}$$

onde $\mathrm{IMF}_m(t), 1 \leq m \leq M$, são as funções de detalhes $d(t)$ obtidas no passo (d) de cada iteração, e $r(t)$ é o sinal residual obtido na última iteração.

É possível verificar, a partir do algoritmo da decomposição, que o número de extremos (máximos e mínimos locais) diminui de uma IMF para a próxima. Ou seja, localmente, as

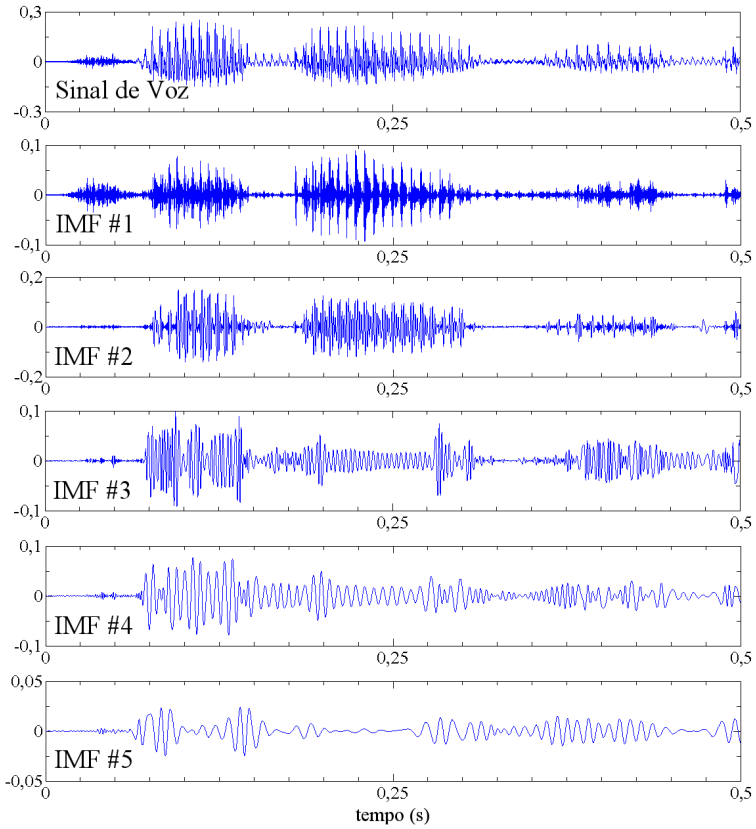

FIG. 2.1: Forma de onda das cinco primeiras IMFs extraídas da decomposição de um segmento de um sinal de voz limpo de 0,5 s da base de voz TIMIT.

primeiras IMFs possuem oscilações mais rápidas (altas frequências) que as IMFs de maior índice. A Fig. 2.1 ilustra este fenômeno, mostra a forma de onda das cinco primeiras IMFs extraídas de um trecho de 0,5 s de uma locução limpa da base de voz TIMIT (GAROFOLO, 1993). Em (FLANDRIN, 2004b) foi exposto que, quando aplicado sobre sinais representados por um processo estocástico fGn (*fractional Gaussian noise*), o método EMD decompõe o sinal em IMFs cujas componentes espectrais são equivalentes às saídas de um banco de filtros diádicos com sobreposição de bandas passantes. Ou seja, o primeiro filtro é passa-altas com banda passante igual à metade da banda do sinal. Os demais são filtros passa-faixas, com banda passante correspondente à metade superior da banda rejeitada pelo filtro anterior.

Em cada uma dos algoritmos de realce baseadas na análise tempo-frequência apresentados nesta Dissertação, o método EMD é primeiramente utilizado para decompor o sinal de voz em um número finito de IMFs. Em seguida, um critério de seleção é utilizado para identificar quais IMFs são predominantemente compostas por ruídos. A reconstrução do

sinal de voz é então realizada utilizando as N IMFs de menor índice,

$$\tilde{y}(t) = \sum_{m=1}^{N} \mathrm{IMF}_m(t), \text{ com } N < M\,. \tag{2.28}$$

Isso corresponde à remoção dos modos que, quando seus espectros são analisados, ocupam as mais baixas frequências do sinal de voz $y(t)$. Segundo o que foi apresentado em (CHATLANI, 2012), as quatro primeiras IMFs concentram a maior parte da energia do sinal de voz. Assim, de forma a evitar distorções no sinal de voz reconstruído, pelo menos as quatro primeiras IMFs devem ser consideradas na reconstrução da EQ. 2.28. Isto é, os valores de N devem ser restritos a $N \geq 4$.

2.1.5 EMDF

O método EMDF (CHATLANI, 2012) foi proposto como um algoritmo de pós-realce para atenuar o ruído residual de baixas frequências. Para isto, a decomposição EMD foi utilizada sobre sinais de voz previamente realçados pelo algoritmo de Cohen. Logo após, as IMFs utilizadas na reconstrução do sinal de voz foram selecionadas por um critério baseado nos valores de variância amostral estimados das amostras das próprias IMFs.

Em (CHATLANI, 2012), é apresentado que, para um sinal de voz limpo, a variância amostral estimada da $\mathrm{IMF}_m(t)$ decai à medida que o índice m aumenta. Na FIG. 2.2 este padrão pode ser verificado na linha contínua, que mostra os valores das variâncias $\mathrm{Var}(m) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IMF}_m^2(t)$ obtidas de um sinal de voz extraído da base TIMIT. Observe que a variância só não decai da primeira para a segunda IMF. De outro modo, quando corrompidos por ruídos acústicos de baixas frequências, as IMFs com índices mais altos apresentam um acréscimo nos valores das variâncias. Na linha tracejada da FIG. 2.2, são exibidos os valores das variâncias obtidas do mesmo sinal de voz, mas agora corrompido pelo ruído acústico fábrica, coletado da base NOISEX-92 (VARGA, 1993), para SNR de 0 dB. Como pode-se notar, a presença do ruído fábrica leva a um ápice de variância na sétima IMF. Deste modo, o objetivo do método EMDF é selecionar qual é o índice (N) de IMFs mais indicado para a reconstrução do sinal de voz (EQ. 2.28).

No trabalho (CHATLANI, 2012), o algoritmo adotado para a seleção deste índice N foi apresentado com as seguintes etapas:

a) Decomposição do sinal de voz $y(t)$ em M modos ($\mathrm{IMF}_m(t), m = 1, \ldots, M$), conforme a EQ. 2.27;

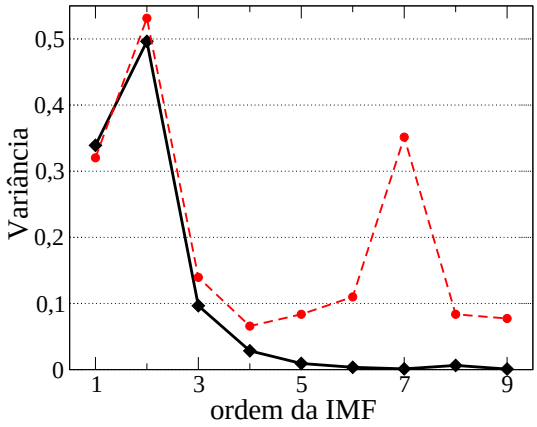

FIG. 2.2: A linha contínua indica os valores de variância amostral estimados das amostras das IMFs de um sinal de voz limpo coletado da base TIMIT. Na linha tracejada, são apresentados os valores referentes ao mesmo sinal de voz corrompido pelo ruído fábrica com SNR de 0 dB. (ZÃO, 2014)

b) Estimação da variância empírica de cada modo utilizando todas as suas T amostras, $\mathrm{Var}(m) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IMF}_m^2(t)$;

c) Identificação, se houver, do índice do primeiro pico (m_p) tal que $\mathrm{Var}(m_p) > \mathrm{Var}(m_p - 1)$ e $\mathrm{Var}(m_p) > \mathrm{Var}(m_p + 1)$, tal que $m_p > 4$;

d) Determinação do índice (m_v) do vale imediatamente anterior ao pico m_p, isto é, $\mathrm{Var}(m_v) < \mathrm{Var}(m_v - 1)$ e $\mathrm{Var}(m_v) < \mathrm{Var}(m_v + 1)$, para $m_v < m_p$;

e) Reconstrução do sinal de voz de acordo com a EQ. 2.28, onde $N = \max\{m_v\,,\,4\}$.

Note que o índice N selecionado pelo algoritmo do método EMDF refere-se ao último vale anterior ao primeiro pico. Todavia, conforme apresentado na Seção anterior, ao menos quatro IMFs devem ser empregadas na reconstrução, de modo a não suprimir os componentes do próprio sinal de voz. Em (CHATLANI, 2012), os resultados de medidas objetivas para sinais de voz, corrompidos por três ruídos acústicos reais, foram aprimorados pelo EMDF. Entretanto, o realce foi significativamente inferior aos outros dois ruídos quando em presença do ruído não-estacionário balbúrdia. Nesta Dissertação, o algoritmo EMDF é avaliado não apenas como pós-realce, mas também é aplicado diretamente sobre os sinais de voz corrompidos por ruídos.

2.1.6 EMDH

Na proposta EMDH (ZÃO, 2014b) de realce de sinais de voz, o expoente de Hurst (H) (HURST, 1951) é utilizado como critério de seleção para a identificação das IMFs a serem removidas do sinal corrompido por ruído. Além disso, tanto a seleção quanto a reconstrução do sinal de voz são realizadas quadro a quadro, de forma a identificar as variações nas características do ruído ao longo do tempo.

O expoente de Hurst ($0 \leq H \leq 1$) de um processo estocástico $y(t)$ é definido pela taxa de decaimento da sua função de autocorrelação normalizada $\rho(k)$. O valor de H está relacionado com as características espectrais de $y(t)$. Isto significa que a densidade espectral de potência de $y(t)$, $S_y(f)$, é predominantemente composto por altas frequências para valores $H < 1/2$. Para o caso $H = 1/2$, $S_y(f)$ é aproximadamente constante ao longo de todo o espectro de frequências, correspondendo ao ruído branco. Já para os valores de $H \in (1/2\,,\,1]$, a maior parte da energia de $y(t)$ está concentrada nas baixas frequências. Devido a esta característica, o expoente de Hurst foi proposto em (SANT'ANA, 2006) como um vetor de atributos de voz, sua aplicação em reconhecimento de locutor foi bem sucedida. A FIG. 2.3 ilustra os valores médios do expoente de Hurst calculados das IMFs obtidas das locuções limpa e das corrompidas da FIG. 2.2. Note que as primeiras IMFs, que englobam as componentes de mais altas frequências do sinal de voz, possuem valores de $\hat{H}$ no intervalo $(0, 1/2)$. Já os modos de maior índice (IMFs de 7 a 9) possuem $H \approx 1$, o que corresponde às componentes onde os ruídos acústicos (baixas frequências) estão geralmente concentrados. Note que a presença do ruído fábrica leva a um aumento nos valores de $\hat{H}$ para as IMFs de 4 a 6. Isto é uma indicação de que o expoente de Hurst é capaz de identificar as IMFs que possuem a maior parte de sua energia devido à presença de ruídos de baixas frequências.

Na proposta de realce EMDH, o sinal de voz ruidoso $y(t)$ é primeiramente decomposto em M modos, conforme a EQ. 2.27. Em seguida, cada uma das IMFs é dividida em quadros, não sobrepostos, de curta duração,

$$\text{w-IMF}_{m,q}(t) = \begin{cases} \text{IMF}_m(t + qT_d) & , t \in [0, T_d]\,, \\ 0 & , \text{caso contrário}, \end{cases} \tag{2.29}$$

onde $q \in \{0, \ldots, Q-1\}$ representa o índice dos quadros e T_d a duração (fixa) de cada quadro. Para cada quadro q, estima-se o valor do expoente de Hurst, H_m, da m-ésima IMF janelada, w-IMF$_{m,q}(t)$. Isso leva à construção de um vetor $\mathbf{H}_q$ com M componentes

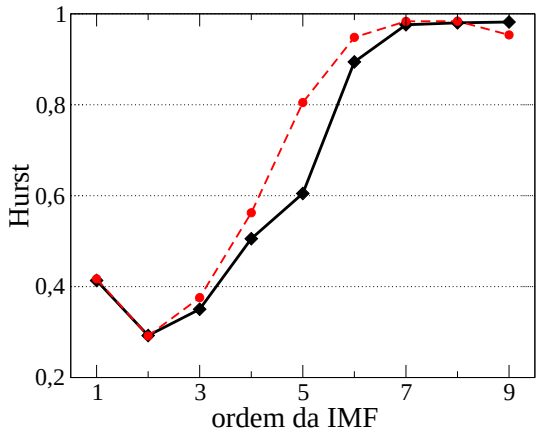

FIG. 2.3: A linha contínua indica os valores de H estimados das IMFs do mesmo sinal de voz limpo da FIG. 2.2. Na linha tracejada, são apresentados os valores referentes ao mesmo sinal de voz corrompido pelo ruído fábrica com SNR de 0 dB. (ZÃO, 2014)

$(m = 1, \ldots, M)$. Em seguida, determina-se a última IMF janelada cujo valor estimado de H está abaixo do limiar $H_{\text{lim}} = 0,9$, determinado de maneira empírica. Se N_q representa este índice desta IMF janelada, pode-se escrever que $\mathbf{H}_q(N_q) < H_{\text{lim}}$.

Cada quadro do sinal de voz realçado $\hat{x}_q(t)$ é então reconstruído como

$$\hat{x}_q(t) = \sum_{m=1}^{N_q} \text{w-IMF}_{m,q}(t),\ q = 0, \ldots, Q-1\,, \tag{2.30}$$

e o sinal de voz $\hat{x}(t)$ é finalmente dado por

$$\hat{x}(t) = \sum_{q=0}^{Q-1} \hat{x}_q(t - qT_d)\,. \tag{2.31}$$

2.2 MEDIDAS DE QUALIDADE E INTELIGIBILIDADE

Nesta Seção são apresentadas medidas de qualidade e inteligibilidade de voz para avaliar o desempenho dos métodos de realce, principalmente, do algoritmo proposto (PRO) nesta Dissertação. As medidas de qualidade têm por objetivo medir o nível de atenuação do ruído gerado pelo método de realce. Já as de inteligibilidade avaliam o número de acertos de sentenças obtidas a partir de um sinal de voz aprimorado por um algoritmo de realce de sinais de voz.

2.2.1 RAZÃO SINAL-RUÍDO SEGMENTAL

A razão sinal-ruído segmental é a primeira medida objetiva utilizada para estudo dos métodos de realce em termos de qualidade. O valor de SegSNR é obtido através da média entre os valores de SNR, em dB, calculados em quadros de curta duração do sinal de voz. Seja $x(t)$ um sinal de voz limpo, e $\hat{x}(t)$ uma versão corrompida ou distorcida deste mesmo sinal, a SegSNR de $\hat{x}(t)$ é estimada por (HANSEN, 1998):

$$\text{SegSNR} = \frac{10}{Q}\sum_{\tau=0}^{Q-1}\log\frac{\sum_{t=\tau T_{\text{sh}}}^{\tau T_{\text{sh}}+T_d-1} x^2(t)}{\sum_{t=\tau T_{\text{sh}}}^{\tau T_{\text{sh}}+T_d-1}\left[x(t)-\hat{x}(t)\right]^2}\,, \tag{2.32}$$

onde T_d é a quantidade de amostras de cada quadro, T_{sh} é o deslocamento (em amostras) entre quadros consecutivos e Q é o total de quadros. Os valores de cada parcela do somatório forem limitados ao intervalo $[-10\text{dB}, 35\text{dB}]$ (HANSEN, 1998). Desse modo, evita-se a necessidade de um detector de atividade de voz.

2.2.2 MEDIDA OQCM DE QUALIDADE DE SINAIS DE VOZ

A medida de qualidade OQCM (*overall quality composite measure*), apresentada em (HU, 2006), foi motivada por estudos da correlação entre cinco medidas objetivas e os resultados de testes subjetivos na avaliação de algoritmos de realce de voz. No trabalho citado, foram escolhidas as medidas SegSNR, PESQ (*perceptual evaluation of speech quality*), WSS (*weighted spectral slope*) (KLATT, 1982), LLR (*log-likelihood ratio*), e IS (*Itakura-Saito distance*) (QUACKENBUSH, 1988) por serem frequentemente adotadas na avaliação de algoritmos para supressão de ruídos. Ainda no trabalho citado, a correlação entre as medidas objetivas e os testes subjetivos foi abordada com sinais de voz realçados por treze algoritmos de realce de voz distintos, incluindo o SS, o de Cohen e o baseado no filtro de Wiener, todos utilizados nesta Dissertação. Em (HU, 2006), estes métodos de realce foram aplicados em 16 sinais de voz da base NOIZEUS (HU, 2007) corrompidos por quatro ruídos acústicos ambientais (balbúrdia, carro, rua e trem) e dois valores de SNR (5 dB e 10 dB). Três medidas subjetivas foram avaliadas nos experimentos: distorção do sinal de voz, distorção do ruído e qualidade total do sinal. Com relação à qualidade total, os estudos demonstraram que as medidas PESQ, LLR e WSS foram as que apresentaram maior coeficiente de correlação com os testes subjetivos. De modo que, foi proposto uma combinação de medidas para obter maior correlação com os resultados subjetivos de qualidade total dos sinais de voz. Desta forma, a medida OQCM é

representada pela combinação linear entre PESQ, LLR e WSS,

$$\text{OQCM} = 1.594 + 0.805\ \text{PESQ} - 0.512\ \text{LLR} - 0.007\ \text{WSS}\,. \tag{2.33}$$

Os resultados apresentaram que, em relação às cinco medidas examinadas separadamente, a medida de qualidade OQCM obteve maior correlação com os testes subjetivos. Por este motivo, ela é também adotada no presente trabalho para avaliação da proposta PRO e dos demais algoritmos de realce em termos de qualidade do sinal de voz.

A medida PESQ foi calculada a partir da recomendação ITU-T P.862.2. Esta versão foi proposta para substituir a PESQ definida em ITU-T P.862, que considerava apenas sinais de banda estreita (3,2 kHz), correspondente à largura de banda de um canal telefônico.

Para o cálculo de OQCM definida na EQ. 2.33, a medida LLR é calculada como (QUACKENBUSH, 1988)

$$\text{LLR}(\vec{a}_p, \vec{a}_c) = \log\left(\frac{\vec{a}_p \mathbf{R}_c \vec{a}_p^T}{\vec{a}_c \mathbf{R}_c \vec{a}_c^T}\right)\,, \tag{2.34}$$

onde $\vec{a}_c$ e $\vec{a}_p$ são os vetores formados pelos coeficientes de predição linear do sinal de voz limpo e do sinal realçado, respectivamente, e $\mathbf{R}_c$ é a matriz de autocorrelação do sinal limpo.

Para o cálculo da medida WSS, os sinais de voz limpo e realçado são inicialmente divididos em Q quadros de curta duração. A magnitude do espectro de cada quadro τ do sinal limpo ($|X(j,\tau)|$) e realçado ($|\hat{X}(j,\tau)|$) é calculada a partir da divisão da sua banda de frequências em $K = 25$ sub-bandas, utilizando filtros Gaussianos, sendo j o índice das sub-bandas ($j = 1, \ldots, 25$). A medida WSS é obtida em cada quadro por uma soma ponderada entre as diferenças das magnitudes do espectro (em dB) do sinal calculadas em bandas adjacentes. Ou seja, se

$$\begin{cases} S_x(j,\tau) = |X(j+1,\tau)|_{(\text{dB})} - |X(j,\tau)|_{(\text{dB})}\,; \text{ e} \\ S_{\hat{x}}(j,\tau) = |\hat{X}(j+1,\tau)|_{(\text{dB})} - |\hat{X}(j,\tau)|_{(\text{dB})}\,, \end{cases} \tag{2.35}$$

a medida WSS é definida por (KLATT, 1982)

$$\text{WSS} = \frac{1}{Q}\sum_{\tau=0}^{Q-1} \frac{\sum_{j=1}^{K-1} W_{\text{WSS}}(j,\tau)\left(S_x(j,\tau) - S_{\hat{x}}(j,\tau)\right)^2}{\sum_{j=1}^{K} W_{\text{WSS}}(j,\tau)}\,, \tag{2.36}$$

onde os pesos $W_{\text{WSS}}(j,\tau)$ foram determinados em (KLATT, 1982).

2.2.3 SNR COM PONDERAÇÃO EM FREQUÊNCIA PARA INTELIGIBILIDADE

A utilização da razão sinal-ruído com ponderação em frequência (fwSegSNR) é motivada pelos resultados de inteligibilidade de voz descritos em (MA, 2009), onde foi demonstrado que os resultados de fwSegSNR apresentam alta correlação com as taxas de acertos de palavras obtidos em testes subjetivos. Ainda neste trabalho, foi comprovado que medidas como SegSNR, WSS e LLR apresentam baixo coeficiente de correlação com os resultados de inteligibilidade apesar de conseguirem representar a qualidade. Os resultados alcançados em (MA, 2009) corroboram com a conclusão apresentada em (LOIZOU, 2007b), que demonstrou a degradação da inteligibilidade gerada por diversos algoritmos propostos para melhorar a qualidade dos sinais de voz.

A medida fwSegSNR pode ser considerada com uma versão no domínio da frequência da razão sinal-ruído segmental (EQ. 2.32), sendo definida como

$$\text{fwSegSNR} = \frac{10}{Q}\sum_{\tau=0}^{Q-1}\frac{\sum_{j=1}^{K} W_f(j,\tau)\log\frac{|X(j,\tau)|^2}{\left(|X(j,\tau)|-|\hat{X}(j,\tau)|\right)^2}}{\sum_{j=1}^{K} W_f(j,\tau)}\,, \tag{2.37}$$

onde τ e j são os índices de quadro e de sub-banda, respectivamente, Q é o número total de quadros e as magnitudes das sub-bandas dos sinais de voz ($|X(j,\tau)|$ e $|\hat{X}(j,\tau)|$) são obtidas com filtros Gaussianos, conforme descrito na Seção 2.2.2. Em (LOIZOU, 2007b), a função de ponderação $W_f(j,\tau)$ que acarretou na maior correlação com os resultados de inteligibilidade foi dada por

$$W_f(j,\tau) = |X(j,\tau)|^{\gamma}\,, \tag{2.38}$$

com $\gamma = 0,2$. Por este motivo, esta definição também é adotada nos experimentos elaborados nesta Dissertação. Assim como na avaliação de SegSNR, os valores de SNR calculados em cada quadro e em cada sub-banda, são limitados ao intervalo $[-10\text{dB}, 35\text{dB}]$.

2.2.4 FAI

Uma das medidas mais utilizadas na avaliação da inteligibilidade é o índice de articulação (AI - *articulation index*) (KRYTER, 1962). Esta é baseada na ideia de que a resposta de um sistema de comunicação de voz pode ser dividida em vinte faixas de frequências, onde cada uma exerce uma contribuição independente para a inteligibilidade do sistema. A Razão sinal-ruído é calculada para cada faixa individual, depois estes valores são ponderados e combinados para produzir um índice de inteligibilidade.

Todavia, a medida AI tem uma série de restrições. A primeira é que ela foi desenvolvida para avaliar casos em que a voz está adicionada a ruídos estacionários. Isto significa que em presença de sinais ou ruídos não-estacionários gera resultados distorcidos, pois usa a média de longo prazo do sinal corrompido e do sinal aprimorado para obter o SNR, e a média destes variam com o tempo. Outra limitação é que a medida AI não pode avaliar sinais aprimorados por métodos que utilizem a subtração espectral, devido às alterações não-lineares provocadas por estes algoritmos. Para lidar com estes efeitos não-lineares gerados pelo processamento da voz (realce de sinal de voz) e a não-estacionariedade do ruído, em (LOIZOU, 2011a) foi proposto o índice de articulação fracionária (FAI). A ideia desta medida é considerar que o valor de SNR de cada banda do sinal aprimorado não pode exceder o valor de SNR do sinal sem tratamento. Para a realização desta medida é necessário entender o novo SNR de saída $\overline{\text{SNR}}_j$.

$$\overline{\text{SNR}}_j = \frac{\hat{x}_j^2}{\eta_j^2}, \tag{2.39}$$

onde $\hat{x}_j$ é o sinal aprimorado pela técnica de realce e j é o índice banda. Após localizar todas estas bandas é feito o cálculo para apurar a proporção ou fração do SNR transmitido do sinal corrompido para o sinal aprimorado pelo método de realce pela seguinte equação:

$$\text{fSNR}_j = \begin{cases} \frac{(\min(\overline{\text{SNR}}_j, \text{SNR}_j)}{\text{SNR}_j} & \text{SNR}_j \geq \text{SNR}_l, \\ 0 & \text{, caso contrário,} \end{cases} \tag{2.40}$$

onde SNR_l representa a menor valor de SNR permitido para cada banda. O valor de fSNR_j é limitada $0 \leq \text{fSNR}_j \leq 1$ e os valores próximos de 1 são obtidos quando $\hat{x} \approx x$, ou seja, o algoritmo de realce conseguiu produzir uma estimativa da voz muito precisa para a banda j. Para calcular o FAI:

$$f_{AI} = \frac{1}{\sum_{k=1}^{M} W_k} W_k \, \text{fSNR}_k \tag{2.41}$$

onde W_k representa a função de ponderação ou funções de importância de banda aplicadas à banda k, M é o número total de bandas usadas e fSNR_k indica a fração de SNR de entrada transmitido pelo algoritmo de realce de sinais de voz. Para avaliar a inteligibilidade com o uso do FAI em (LOIZOU, 2011a), foram realizados testes subjetivos com 72 ruídos, e foi possível o desenvolvimento de uma função logística usada para a predição de inteligibilidade que é dada por:

$$I = (1 - 10^{-f*P/Q})^2, \tag{2.42}$$

onde f é o valor obtido em FAI, $P = 27,5$ e $Q = 8,4$. Com estes valores em (LOIZOU, 2011a), a função logística obteve uma correlação de 0,9 com testes subjetivos de inteligibilidade.

2.2.5 STOI

A medida STOI foi introduzida por (TAAL, 2011) para estimar a degradação na inteligibilidade de sinais de voz causada por algoritmos de supressão de ruídos. A diferença desta medida em relação ao índice de articulação (KRYTER, 1962) e medidas derivadas (STEENEKEN, 1980; RHEBERGEN, 2005; LOIZOU, 2011b), se dá pela não utilização do cálculo de SNR para avaliar a inteligibilidade dos sinais de voz. De outro modo, é adotado o coeficiente de correlação entre os espectros dos sinais de voz limpo e realçado, evitando assim a necessidade de estimação explícita da distorção presente no sinal de voz.

Na obtenção da medida STOI, o sinal de voz limpo $x(t)$ é inicialmente re-amostrado a taxa de 10 kHz e segmentado em quadros de 256 amostras utilizando janelas de Hanning com 50% de sobreposição. A taxa de amostragem é aqui fixada em 10 kHz de forma a manter a mesma resolução em frequência da análise realizada em (TAAL, 2011). Na sequência, cada quadro é transformado para o domínio da frequência utilizando-se a DFT com 512 pontos. Seja $X(\kappa, \tau)$ o κ-ésimo ponto resultante da aplicação da DFT sobre o quadro τ. Os pontos $X(\kappa, \tau)$ são agrupados em 15 bandas cujas frequências centrais variam de 150 Hz a 4300 Hz, com três bandas por oitava. A norma da j-ésima banda ($j = 1, 2, \ldots, 15$) é definida por:

$$\bar{X}_j(\tau) = \sqrt{\sum_{\kappa=\kappa_l(j)}^{\kappa_u(j)-1} |X(\kappa, \tau)|^2}, \tag{2.43}$$

onde $\kappa_l(j)$ e $\kappa_u(j)$ são os seus limites inferior e superior, respectivamente. Em cada região de tempo e frequência, a envoltória temporal de cada banda do sinal limpo é representada pelo vetor

$$\mathbf{x}_{(j,\tau)} = \left[\bar{X}_j(\tau - 29), \bar{X}_j(\tau - 28), \ldots, \bar{X}_j(\tau)\right]^T. \tag{2.44}$$

O uso de 30 coeficientes para o vetor $\mathbf{x}_{(j,\tau)}$ foi definido em (TAAL, 2011) a partir de resultados experimentais. A análise temporal com 30 quadros consecutivos corresponde a 384 ms, ou seja, um quadro a cada 12,8 ms.

De maneira análoga à estimação de $\mathbf{x}_{(j,\tau)}$, obtém-se o vetor $\mathbf{y}_{(j,\tau)}$ a partir do sinal de voz corrompido $y(t)$. Em seguida, $\mathbf{y}_{(j,\tau)}$ é normalizado para compensar eventuais diferenças de energia em relação a $\mathbf{x}_{(j,\tau)}$. Seja $\mathbf{y}_{(j,\tau)}(n)$ o n-ésimo coeficiente do vetor $\mathbf{y}_{(j,\tau)}$, a versão normalizada de $\mathbf{y}_{(j,\tau)}$ é obtida por

$$\bar{\mathbf{y}}_{(j,\tau)}(n) = \min\left(\frac{\|\mathbf{x}_{(j,\tau)}\|}{\|\mathbf{y}_{(j,\tau)}\|}\mathbf{y}_{(j,\tau)}(n)\,,\,(1+10^{-\beta/20})\mathbf{x}_{(j,\tau)}(n)\right)\,, \tag{2.45}$$

onde $\|\cdot\|$ representa a norma ℓ^2 e $\beta_{SDR} = -15$ dB indica o valor máximo para a grandeza SDR (*signal-to-distortion ratio*) definida em (TAAL, 2011). A medida intermediária $\text{STOI}_{(j,\tau)}$ é definida como o coeficiente de correlação entre os vetores $\bar{\mathbf{y}}_{(j,\tau)}$ e $\mathbf{x}_{(j,\tau)}$. Ou seja,

$$\text{STOI}_{(j,\tau)} = \frac{(\mathbf{x}_{(j,\tau)} - \mu_{\mathbf{x}_{(j,\tau)}})^T(\bar{\mathbf{y}}_{(j,\tau)} - \mu_{\bar{\mathbf{y}}_{(j,\tau)}})}{\|\mathbf{x}_{(j,\tau)} - \mu_{\mathbf{x}_{(j,\tau)}}\|\|\bar{\mathbf{y}}_{(j,\tau)} - \mu_{\bar{\mathbf{y}}_{(j,\tau)}}\|}\,, \tag{2.46}$$

onde $\mu_{(\cdot)}$ indica a média amostral do vetor correspondente. Finalmente, a medida STOI é dada pela média de todos os valores intermediários calculados de cada quadro τ e de cada banda j,

$$\text{STOI} = \frac{1}{15\,Q}\sum_{j=1}^{15}\sum_{\tau=1}^{Q}\text{STOI}_{(j,\tau)}\,, \tag{2.47}$$

onde Q é o número total de quadros.

Além da proposta da medida STOI, os autores aplicaram uma função monótona não-linear para mapear os resultados de STOI na predição de taxas de acertos de palavras em experimentos subjetivos de inteligibilidade. A função de mapeamento foi dada por

$$f(\text{STOI}) = \frac{100}{1+\exp(a\,\text{STOI}+b)}\,, \tag{2.48}$$

com a e b são constantes. A conclusão dos testes demonstrou boa precisão para sinais provenientes de duas bases de voz, uma delas em língua inglesa. Apesar da utilização da base de voz TIMIT, gravada neste mesmo idioma, para a realização dos experimentos de realce de voz nesta Dissertação foi necessária a alteração da função de mapeamento. Na EQ. 2.48, novos valores de a e b, foram definidos, isto é, $a = -13,45$ e $b = 9,361$.

2.2.6 CSII

A medida CSII elaborada por (KATES, 2005), é um aprimoramento da SII *(speech intelligibility index)*. A principal diferença entre as duas é que a CSII utiliza medida

de coerência quadrática (MSC - *magnitude-squared coherence*) no cálculo da razão sinal-ruído para a computação dos índices. Assim como no AI e na SII, a CSII resulta num número entre zero e um, onde valores maiores indicam maior inteligibilidade. O principal ganho do uso desta técnica é que ela leva em consideração as distorções causadas pelos métodos de realce, principalmente quando estes provocam redução da amplitude a zero em regiões com atividade de voz ("*center-clipping*") e ganho de amplitude acima do limite de saturação ("*peak-clipping*"). Para calcular a CSII, um sinal de referência $x(t)$ é utilizado para medir o ganho ou a perda de inteligibilidade causado pelo processamento, tendo como sinal resultante $y(t)$. Os espectros dos sinais são obtidos por meio da aplicação da transformada discreta de Fourier (DFT) em versões segmentadas dos sinais obtidas via janelamento. A MSC é estimada usando

$$\mathrm{MSC}(f) = \frac{|\sum_{j=0}^{J-1} X_j(f) Y_j^*(f)|^2}{\sum_{j=0}^{J-1} |X_j(f)|^2 \times \sum_{j=0}^{J-1} |Y_j(f)|^2} \tag{2.49}$$

onde $X_j(f)$ e $Y_j(f)$ são, respectivamente o espectro do segmento j dos sinais $x(t)$ e $y(t)$ e f é o índice da DFT. Sendo $x(t)$ e $y(t)$ respectivamente a entrada e a saída de um sistema, a MSC representa o quanto da potência do sinal de saída é linearmente dependente da entrada (KATES, 2005), assim como $1 - \mathrm{MSC(f)}$ representa a presença de distorção e de ruído. Sendo $S_y(f)$ a amostra f da densidade espectral de potência do sinal de saída, estimada via DFT, a razão sinal-ruído e interferência (*SDR - Signal-to-noise and Distortion Ratio*) pode ser estimada utilizando:

$$\mathrm{SDR(b)} = \frac{\sum_{f=0}^{F} I_b(\tau) MSC(f) S_y(f)}{\sum_{f=0}^{\Upsilon}(f)[1 - MSC(f))] S_y(f)} \tag{2.50}$$

onde $I_b(f)$ é um filtro que implementa o peso da banda de frequências b na inteligibilidade da fala, assim como efeitos como o mascaramento de frequências.

Além disso, ao avaliar o processo de obtenção da CSII (KATES, 2005) foi verificado que quando calculada em três níveis de amplitudes diferentes, apresentava um alto coeficiente de correlação com testes perceptuais subjetivos. Para o cálculo dos três níveis de CSII, o sinal de entrada de voz é dividido em três regiões de amplitude. O cálculo usa um tamanho de bloco de 16 ms com janelas de Hamming de 50% de sobreposição entre os segmentos. A magnitude do sinal em cada segmento é calculada e armazenada ao longo da duração da sequência, de modo que seja obtido o valor médio quadrático (RMS - *root mean square*) de cada nível dos segmentos. O valor de $\mathrm{CSII_{alto}}$ é obtido a partir dos

segmentos que apresentam valores superiores ao RMS. O $\text{CSII}_{\text{medio}}$ é adquirido com os segmentos que apresentam valores entre 0 e 10 dB abaixo do nível RMS, e o $\text{CSII}_{\text{baixo}}$ é calculado com segmentos entre 10 e 30 dB abaixo do nível RMS. Para a predição de inteligibilidade é utilizada uma função de mapeamento com os três níveis de CSII. Nesta Dissertação, a equação foi modificada para ser ajustada a base de voz TIMIT.

$$\begin{aligned} c &= -3,47 + 1,84\text{CSII}_{\text{baixo}} + 9.99\text{CSII}_{\text{médio}} + 0.00\text{CSII}_{\text{alto}} \\ I_3 &= \frac{100}{1+\exp(ac+b)} \end{aligned} \tag{2.51}$$

onde $a = -10,09$ e $b = 4,65$. A predição de inteligibilidade (I_3) é determinada pelo $\text{CSII}_{\text{médio}}$, com alguma entrada do $\text{CSII}_{\text{baixo}}$. O peso para o $\text{CSII}_{\text{alto}}$ é zero, de modo que, este termo não tem nenhum efeito aparente sobre a inteligibilidade no contexto do modelo. O grau de correlação entre (I_3) e testes subjetivos de inteligibilidade é superior a 90%.

2.3 RESUMO

Neste Capítulo foi apresentado um grupo de métodos extraídos da literatura para realce de sinais de voz corrompidos por ruídos acústicos. O algoritmo de subtração espectral clássica é o único que assume a estacionariedade do ruído, dentre as soluções que utilizam a transformada de Fourier para estimação e supressão das componentes do ruído no domínio da frequência. As demais (Cohen e Wiener) empregam métodos de estimação que conseguem atualizar o espectro de potência do ruído mesmo durante a atividade da voz. Foram introduzidos também dois métodos baseados na análise tempo-frequência que utilizam a decomposição empírica de modos, EMDF e EMDH. Estes algoritmos, por sua vez, não necessitam assumir hipóteses sobre as características do sinal de voz, nem estimar previamente e de maneira explícita as componentes do ruído. Ainda neste Capítulo foram apresentadas duas medidas objetivas de qualidade de voz (SegSNR e OQCM) e quatro medidas objetivas de inteligibilidade, fwSegSnR, FAI, STOI, CSII.

3 REALCE DE SINAIS DE VOZ NO DOMÍNIO DO TEMPO: PROPOSTA

O emprego de soluções de realce de sinais é fundamental para amenizar ou atenuar o efeitos de distorções provocadas por ruídos acústicos. O estado da arte das principais técnicas de realce de sinais espectrais e temporais propostas na literatura, está descrito no Capítulo 2. Neste Capítulo, é introduzida uma nova proposta de realce de sinais de voz, no domínio do tempo, para aprimorar sinais de voz corrompidos por ruídos acústicos não-estacionários. Para identificação das componentes de ruído, o seu desvio padrão é estimado do sinal corrompido. A estimação robusta do desvio padrão é aplicada considerando qualquer distribuição de amplitude do sinal de voz e do ruído acústico. O uso deste critério no domínio do tempo, deve permite o aprimoramento do sinal de voz corrompido por ruídos acústicos. Para a obtenção da estimação robusta é adotado o algoritmo DATE proposto em (PASTOR, 2012). Este algoritmo foi inicialmente definido para estimar o desvio padrão de um ruído aditivo Gaussiano, de espectro branco, e considerando sinais de voz com distribuições de amplitudes desconhecidas.

O método de realce de sinais de voz apresentados nesta Dissertação é realizado em três etapas: identificação e estimação das componentes de ruído, extração destas componentes do sinal corrompido e reconstrução do sinal.

3.1 PRIMEIRA ETAPA: IDENTIFICAÇÃO E ESTIMAÇÃO DAS COMPONENTES DE RUÍDO

Considere $y(t)$ um sinal de voz corrompido por um ruído acústico aditivo $\eta(t)$. Logo, pode-se escrever $y(t) = x(t) + \eta(t)$, onde $x(t)$ é o sinal de voz limpa. Para a estimação do desvio padrão (σ_i) das componentes do ruído, o sinal corrompido é dividido em i quadros de tamanho j.

3.1.1 ESTIMADOR ROBUSTO DE CORTE d-DIMENSIONAL - DATE

A estimação robusta data do século passado e até os dias de hoje é um grande desafio (STIGLER, 1973; KAY, 1993). Os principais estimadores robustos podem ser classificados em três famílias (ZOUBIR, 2012). Os de máxima verossimilhança (HUBER, 2009), os

de combinação linear de ordens estatísticas (HAMPEL, 2005) e os derivados do teste de posto (*rank tests*) (DONOHO, 1983). A categoria de estimadores lineares é amplamente adotada por apresentar baixo custo computacional. Entre os mais populares destacam-se o desvio médio absoluto (MAD - *median absolute desviation*) e o estimador de corte (T-*trimmed estimator*). A popularidade do MAD deve-se à obtenção de estimações precisas mesmo quando a quantidade de valores discrepantes (*outliers*) correspondem a 50% do total de amostras utilizadas no cálculo do desvio padrão. Assim, o estimador MAD é apresentado pela literatura como o principal estimador robusto. O estimador de corte T é considerado muito preciso quando o número de valores discrepantes é menor do que 25%. No entanto, a precisão deste grupo de estimadores cai significativamente quando a proporção de valores discrepantes excede este valor. Como alternativa, em (PASTOR, 2012) foi introduzido o estimador de corte DATE. Neste estimador o número de valores discrepantes mesmo sendo muito grandes e não conhecidos, não interfere na precisão da estimação, o que o torna um bom candidato. Além disso, no DATE, não é necessário o conhecimento prévio da distribuição das amostras do sinal para obter a estimativa do desvio padrão do ruído. Este adota duas hipóteses: a norma das amplitudes do sinal deve estar acima de um limite inferior conhecido e a probabilidadede de ocorrência do sinal de voz deve ser menor que 0,5. Neste trabalho este algoritmo foi modificado para estimar o desvio padrão das componentes de ruídos acústicos não-estacionários. Além disso, é considerado que o sinal de voz e o ruído possuem qualquer tipo distribuição de amplitude, Gaussiana ou não-Gaussiana.

3.1.2 ALGORITMO DE ESTIMAÇÃO DATE

A estimação robusta do desvio padrão é realizada em duas etapas preparo da sequência amostral e estimação do desvio padrão do ruído acústico

Etapa 1: preparo da sequência amostral

- definição e busca de uma sequência amostral $\{y(1), y(2), \ldots, y(K)\}$; $1 \leq k \leq K$ que deve satisfazer as seguintes premissas:
  - para todo $k \in \mathbb{N}$, $y(k)$, $x(k)$ e $\varepsilon(k)$[1] são independentes;
  - $x(k)$ não está sempre presente em $y(k)$, ou seja, existe uma probabilidade p de ausência de sinal de voz;

[1] $\varepsilon(k)$ é uma variável aleatória $[0, 1]$ que indica a presença do sinal de voz $x(k)$.

– $\mathbb{E}(y(k)^{\nu}) < \infty$, em (PASTOR, 2012) foi utilizado o valor de $\nu = 2$ para a estimação do desvio padrão de ruídos Gaussianos;

- Inicialização do limiar de estimação
  $\xi(\rho) = \frac{1}{2}\rho + \frac{1}{\rho}\log\left(1 + \sqrt{1 - \exp^{-\rho^2}}\right)$, onde ρ representa a razão entre a média de todos os valores de amplitude do sinal corrompido e o desvio padrão dos valores dos seus valores mínimos. Para um ruído Gaussiano, tem-se que $\rho = 4$, e $\xi(\rho) = 3,4742$.

- definição do grau de confiança (Q);
  $Q \leq 1 - \frac{K}{4(\frac{K}{2}-1)^2}$

- rearranjar a sequência amostral de $\{y(1), y(2), \ldots, y(k)\}$ em ordem crescente de valor de amplitude $Y_1, \leq Y_2, \ldots \leq Y_k$

Etapa 2: estimação do desvio padrão do ruído acústico

- busca do intervalo inicial de estimação

  – cálculo de k_{min}, que indica a quantidade de amostras na qual os k primeiros valores de Y_k (sinal corrompido) são constituídos apenas por ruídos, para um dado grau de confiança. Segundo a desigualdade de Bienaymé-Chebyshev (ROUSSEEUW, 1981) o valor de k_{min} pode ser obtido por:

  $$k_{min} = K/2 - hK \tag{3.1}$$

  onde $h = \frac{1}{\sqrt{4K(1-Q)}}$, K é o tamanho total de $y_{(k)}$ e Q é o grau de confiança. Os resultados de experimentos realizados em (PASTOR, 2012) indicam que para ruídos Gaussianos o valor de Q deve ser igual a 95%.

- verificar se existe um valor inteiro mínimo b em $\{Y_{(kmin)}, \ldots, Y_{(k)}\}$ tal que:

  $$\|Y_{(k-1)}\| \leq R < \|Y_{(k+1)}\| \tag{3.2}$$

  onde $\| \bullet \|$ é a norma Euclidiana, $R = \frac{[\sum_{i=1}^{k} \|Y\| \xi(\rho)]}{\lambda k}$ onde λ é o fator de ajuste do limiar de estimação em função da dimensão da sequência amostral.

  Se positivo então, $b = k$; caso contrário: $b = k_{min}$.

- cálculo de σ
  $\sigma = \frac{[\sum_{i=1}^{b} \|Y\|] \xi(\rho)}{\lambda b}$

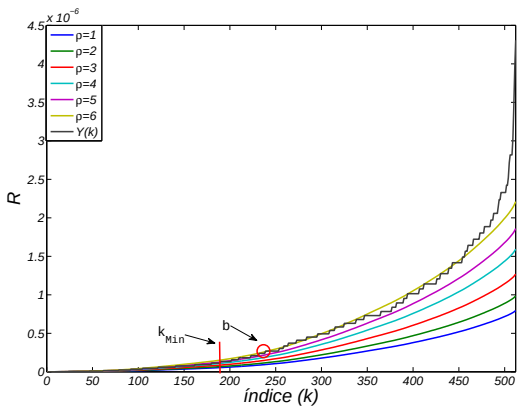

FIG. 3.1: Estimação do desvio padrão do ruído, a partir de um quadro com 600 amostras, de um sinal de voz corrompido por ruído britadeira a razão sinal ruído de 10 dB.

Para ser aplicado na estimação de desvio padrão de ruídos não-estacionários e com qualquer distribuição de amplitude, foram feitas algumas motificações no algoritmo DATE.

- aplicação do algoritmo em segmentos de curta duração e não em todo sinal;
- definir um limiar variável. Ou seja, um limiar para cada quadro i definindo um novo valor de ρ para garantir a existência de um valor de b_{min} para cada quadro.

A FIG. 3.1 ilustra a alteração da curva R quando o limiar $\xi(\rho)$ é alterado para a estimação do desvio padrão de um ruído britadeira[2] adicionado a um sinal de voz para razão sinal ruído de 10 dB. Note que para $\rho = 4$,valor do algoritmo DATE original, a curva da função R se distancia da sequência original $|Y(k)|$. Isso se deve ao fato de que R é menor $\|Y_{(k-1)}\|$ para todo o k violando a desigualdade da EQ. 3.2.

Para avaliar a estimação do desvio padrão após as modificações foram realizados testes com três ruídos acústicos reais: fábrica, serra elétrica e trem coletados da base NOISEX-92[2] (VARGA, 1993), Freesfx.co.uk[3] e Freesound.org[4] respectivamente. Estes ruídos foram adicionados a um sinal de voz extraído da base TIMIT (GAROFOLO, 1993) com valor de SNR de 10 dB. A taxa de amostragem é de 16 kHz com duração 1,5 s, ou 600 amostras por quadro.

[2]Disponível em http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

[3]Disponível em http://www.freesfx.co.uk.

[4]Disponível em http://www.freesound.org.

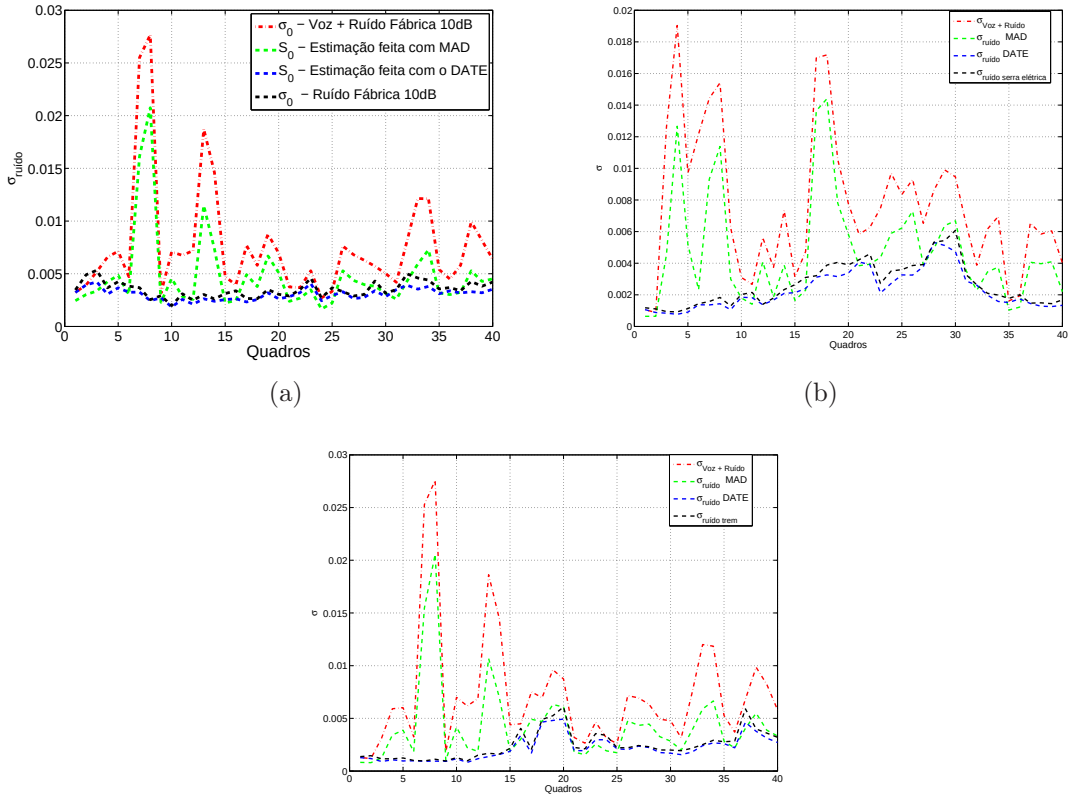

FIG. 3.2: Uso do DATE e do MAD para estimar o desvio padrão dos ruídos (a) fábrica, (b) serra elétrica e (c) trem

A FIG. 3.2 apresenta os resultados da estimação do desvio padrão a partir de quadros de curta duração dos ruídos utilizando os estimadores DATE e MAD, com SNR 10 dB. A linha em vermelho representa o desvio padrão do sinal corrompido, a linha preta o desvio padrão original do ruído, a linha verde é a estimativa obtida a partir do MAD e a azul a estimada pelo DATE. Pode-se notar que a estimação do desvio padrão dos ruídos pelo DATE é próxima dos valores reais de desvio padrão dos ruídos.

Para examinar o DATE para a estimação do desvio padrão de ruídos não-estacionários, em diferentes valores de SNR foi realizado um outro teste para tanto, Para isso, foram utilizados outros ruídos como: balbúrdia, britadeira e helicóptero.

Na TAB. 3.1 podem ser observados os resultados da estimação do desvio padrão de seis ruídos acústicos (balbúrdia, britadeira, fábrica, helicóptero, serra elétrica e trem) para diferentes valores de SNR (-10 dB, -5 dB, 0 dB, 5 dB, 10 dB).

A boa precisão na estimação do desvio padrão de ruídos não-estacionários obtidas após

TAB. 3.1: Comparação entre a estimação de σ_{ruido} com o uso do DATE e MAD.

Ruído	SNR	σ_{real}	$\sigma_{(DATE)}(10^{-3})$	$\sigma_{(MAD)}(10^{-3})$
balbúrdia	10 dB	2,6	2,6	4,6
	5 dB	4,6	4,6	5,5
	0 dB	8,1	7,8	12,4
	-5 dB	14,4	12,2	18,1
	-10 dB	25,6	23,0	29,2
britadeira	10 dB	1,0	1,0	4,6
	5 dB	5,1	5,1	7,7
	0 dB	9,0	8,8	11,0
	-5 dB	16,1	19,8	23,2
	-10 dB	28,6	24,9	30,2
fábrica	10 dB	3,1	3,1	4,8
	5 dB	5,4	5,5	6,1
	0 dB	9,6	9,4	10,6
	-5 dB	17,2	14,6	18,4
	-10 dB	30,5	27,4	32,4
helicóptero	10 dB	2,7	2,8	4,6
	5 dB	4,9	4,9	5,6
	0 dB	8,7	8,3	9,7
	-5 dB	9,4	7,6	11,9
	-10 dB	27,4	22,2	32,6
serra elétrica	10 dB	2,3	2,3	4,2
	5 dB	4,1	4,1	6,2
	0 dB	7,3	7,1	8,9
	-5 dB	12,9	10,6	14,3
	-10 dB	23,0	19,9	23,5
trem	10 dB	2,2	2,2	4,3
	5 dB	3,8	3,8	6,1
	0 dB	6,8	6,6	7,7
	-5 dB	12,1	9,8	15,8
	-10 dB	21,5	18,4	23,5

as modificações no DATE, mostram que o mesmo pode ser usado para a realização da primeira etapa do realce de sinais proposto nesta Dissertação.

3.2 SEGUNDA ETAPA: EXTRAÇÃO DAS COMPONENTES RUÍDOS

Para a remoção das componentes ruidosas em cada quadro do sinal $y(t)$ aplica-se o seguinte teste:

- se valor da amplitude de $y(t) \geq y(b)$ então o valor $y(t)$ é selecionado e deste é subtraído o valor do desvio padrão estimado do seu respectivo quadro.
- se o valor da amplitude de $y(t) < y(b)$ então $y(t) = 0$, ou seja, esta amplitude é

removida por ser considerada a hipótese que esta é formada somente por ruído.

Note que $y(b)$ é o último valor utilizado para o cálculo do desvio padrão e como os valores da sequências estão em ordem crescente de amplitude então $y(b)$ contém a maior amplitude do ruído. Logo, todos os valores acima $y(b)$ tem maior presença de sinal de voz.

3.3 TERCEIRA ETAPA: RECONSTRUÇÃO DO SINAL DE VOZ

A reconstrução sinal é a última nesta etapa do processo de realce e é realizado procedimento: Nesta o sinal é composto pelas amplitudes remanescentes, selecionadas na etapa anterior.

3.4 RESUMO

Neste Capítulo, apresentou-se uma proposta de realce de sinais de voz que possui duas características importantes:

- a sua realização ocorre no domínio do tempo;
- aprimora sinais de voz corrompidos por ruídos não-estacionários.

Para a execução desta proposta é utilizado como critério um estimador robusto de desvio padrão (DATE). Este estimador foi proposto inicialmente para obter o desvio padrão de um ruído branco Gaussiano, por isso algumas alterações foram realizadas de forma a permitir a sua utilização para o cálculo de qualquer tipo de ruído (não-estacionário e não-Gaussiano).

As modificações empregadas permitiram uma estimação mais robusta, que foi comprovada a partir de dois testes: o primeiro comparou a obtenção do desvio padrão do ruído pelos métodos MAD e DATE modificado. E este último apresentou maior exatidão para todos os ruídos testados, apesar da literatura apontar o MAD como um dos estimadores mais robustos. O segundo teste avaliou a influência da razão sinal ruído na obtenção de uma estimativa superior do desvio padrão. Os resultados mostraram que o DATE apresenta respostas mais precisas sob condições onde a razão sinal ruído seja maior que zero dB. Além disso, vale ressaltar que mesmo nas outras razões sinais ruído o estimador consegue obter resultados mais acurados quando comparado com o MAD. Portanto, fato o

DATE foi adotado na primeira etapa do processo de realce de sinais realizado no domínio do tempo.

4 RESULTADOS DE QUALIDADE E INTELIGIBILIDADE

As técnicas de realce de sinais de voz podem ser examinadas de forma subjetiva ou objetiva. Na primeira abordagem, utilizam-se ouvintes para o exame e o julgamento da qualidade do sinal de voz através de testes perceptuais. No entanto, esta forma consome muito tempo e é altamente custosa. A segunda forma de análise, emprega medidas objetivas de qualidade para avaliação do sinal de voz (KATES, 2005; HU, 2008). Uma das limitações do uso destas medidas de qualidade reside no fato de que a maioria foi originalmente desenvolvida para julgar codificadores de voz ou canais de comunicações, e não necessariamente métodos de realce de sinais.

Em (HU, 2008), é descrito um estudo de diversas medidas objetivas na avaliação da qualidade de sinais de voz considerando 13 diferentes métodos de realce de sinais. Os resultados mostraram que a razão sinal-ruído segmental (SegSNR) e a distância de Itakura-Saito (IS), apresentam baixa correlação com a qualidade do sinal indicada por testes subjetivos. Outro aspecto desta avaliação é que o aumento da qualidade não é suficiente para confirmar o aprimoramento do sinal de voz, sendo portanto necessário o exame do ganho de inteligibilidade. Este último, determina as taxas de acertos das palavras e sentenças transmitidas pela voz. Os testes realizados em (HU, 2008), também demonstraram que os algoritmos de realce, apesar de melhorarem a qualidade do sinal de voz, reduzem o grau de inteligibilidade, ou seja, degradam a taxa de acertos de palavras ou sentenças. Por este motivo, medidas objetivas de inteligibilidade são usadas neste trabalho para analisar o método proposto. As medidas de inteligibilidade são: fwSegSNR, CSII, STOI e FAI. Para avaliação da qualidade dos sinais de voz são utilizadas as medidas SegSNR e OQCM.

O método de realce proposto é comparado com três algoritmos espectrais (SS, Cohen e Wiener), e dois temporais (EMDF e EMDH). Para os testes, foram utilizados seis ruídos acústicos ambientais (balbúrdia, britadeira, fábrica, helicóptero, serra elétrica e trem), com diferentes índices de não-estacionariedade (BORGNAT, 2010). Nos experimentos de realce, os métodos são aplicados de forma direta nos sinais de voz corrompidos pelos ruídos acústicos.

4.1 DESCRIÇÃO DOS EXPERIMENTOS DE REALCE DE VOZ

Para analisar o método de realce de sinais de voz proposto (PRO), foram realizados testes com 24 locutores selecionados aleatoriamente da base de voz TIMIT (GAROFOLO, 1993), sendo 8 mulheres e 16 homens. Cada locutor gerou 10 gravações com duração média de 3 s e amostradas à taxa de 16 kHz, totalizando 240 sinais de voz. Para os testes, os ruídos foram adicionados aos sinais de voz limpos para a obtenção de cinco diferentes valores de SNR: 10 dB, 5 dB, 0 dB, -5 dB, e -10 dB. A escolha destes ruídos se deu em função dos diferentes valores de INS e dos espectrogramas possuírem formas distintas. Da base de ruídos NOISEX-92[5] (VARGA, 1993) foram coletados os ruídos fábrica e balbúrdia e da base Freesfx.com.uk[6], helicóptero e trem, os últimos dois britadeira e serra elétrica, da base Freesound.org[7].

A FIG.4.1 apresenta os espectrogramas de segmentos de 3 s dos ruídos utilizados nos testes. É possível notar que os ruídos balbúrdia, fábrica e serra elétrica possuem componentes espectrais em toda a faixa de frequência 0-4 kHz. Já os ruídos helicóptero e trem estão concentrados principalmente na faixa de 0-2,5 kHz. Cabe ressaltar que no ruído trem há concentração de energia nas altas frequências, no intervalo de 2 s a 3 s. O ruído britadeira apresenta componentes nas frequências 0-4 kHz, até aproximadamente 1 segundo. Depois, existe um corte brusco e as suas componentes espectrais ficam na região de 0-2 kHz. Isto se deve à redução quase instantânea das taxas de rotação motor. Destaca-se ainda a presença de harmônicos no ruído serra elétrica, fruto da rotação do motor do próprio equipamento.

4.1.1 ÍNDICE DE NÃO-ESTACIONARIEDADE

O índice de não-estacionariedade (INS) é um método tempo-frequência, proposto em (BORGNAT, 2010), para determinar de forma objetiva o grau de não-estacionariedade de sinais e ruídos. O INS é obtido para um sinal $x(t)$ em três etapas. Na primeira, são construídos referenciais estacionários (*surrogates*) de $x(t)$. Para isto, é aplicada a transformada discreta de Fourier (DFT - *discrete Fourier transform*) sobre $x(t)$. Em seguida,

[5]Disponível em http://www.speech.cs.cmu.edu/comp.speech/section1/data/noisex.html.
[6]Disponível em http//www.freesound.org.
[7]Disponível em http://www.freesfx.co.uk

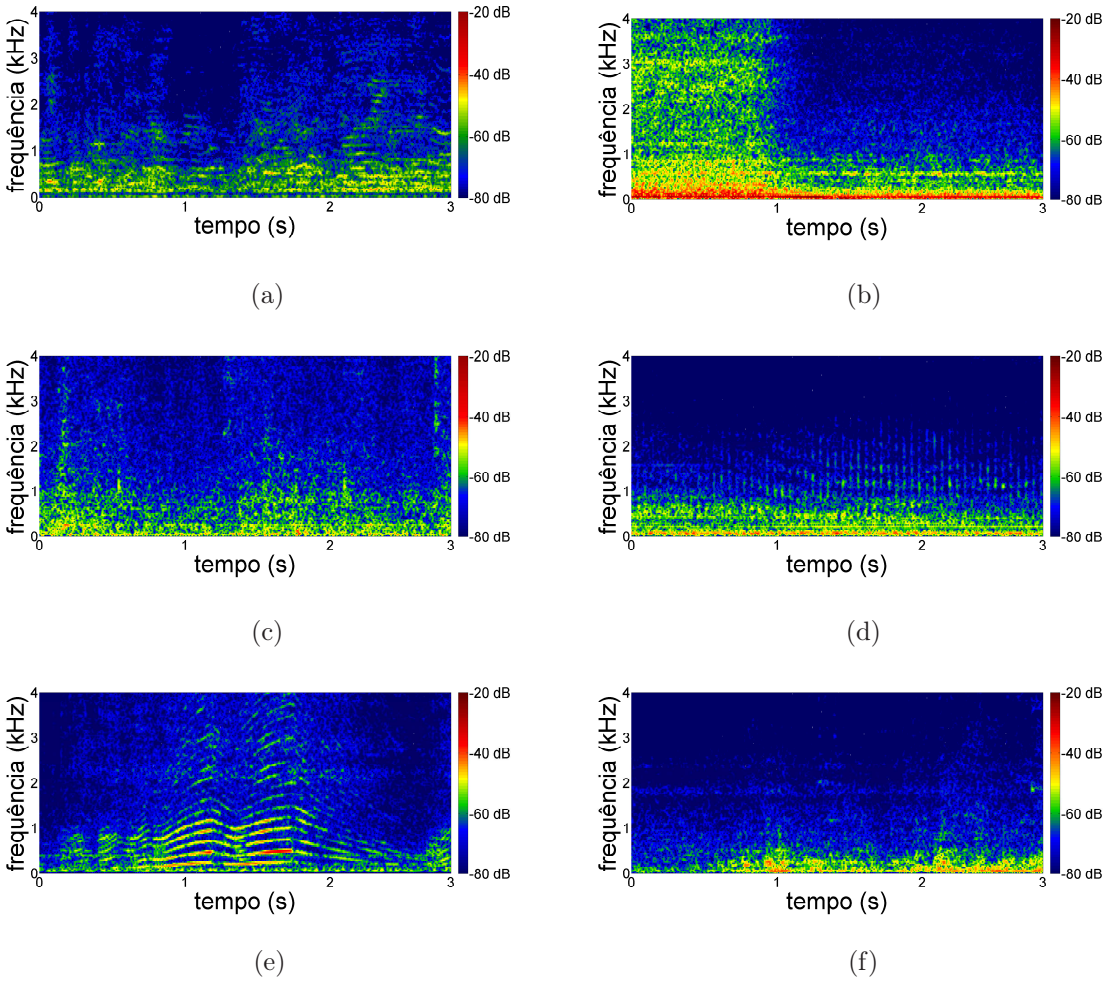

FIG. 4.1: Espectrogramas de segmentos de 3 segundos de duração dos ruídos (a) balbúrdia, (b) britadeira, (c) fábrica, (d) helicóptero (e) serra elétrica, e (f) trem.

a fase do sinal original é substituída por uma sequência aleatória com amostras independentes e uniformemente distribuídas em $[-\pi, \pi]$. Com o uso da transformada inversa de Fourier da sequência obtida, é criada uma versão "estacionária" de $x(t)$. A segunda etapa consiste em comparar o sinal original com seus referenciais estacionários. Esta avaliação é realizada a partir da distância de Kullback-Leibler (D_{kl}) simétrica (BASSEVILLE, 1989), por meio de comparação entre os espectrogramas do sinal $x(t)$ com os referenciais substitutos. Na terceira etapa, o índice de não-estacionariedade é calculado pela razão entre a variância das distâncias observadas ($\Theta_0(j)$) do sinal em análise e a média das variâncias obtidas por meio dos sinais referenciais Θ_1. INS $:= \sqrt{\frac{\Theta_1}{\langle\Theta(j)\rangle_j}}$. Para o teste de não-estacionariedade do sinal em análise, é utilizado um limiar de estacionariedade (γ) que define com uma precisão de 95%, que os valores abaixo deste são estacionários. Ou

seja,

$$\text{INS} \begin{cases} \leq \gamma & , \textit{x(t)} \text{ é estacionário;} \\ > \gamma & , \textit{x(t)} \text{ não é estacionário;.} \end{cases} \tag{4.1}$$

A FIG.4.2 mostra os valores de INS obtidos de segmentos de 3 s dos seis ruídos acústicos. As linhas tracejadas (em verde) representam o limiar γ de estacionariedade. A escala temporal T_h/T corresponde à razão entre o tamanho da janela de análise de tempo curto (T_h) e a duração total do segmento do ruído (T= 3 segundos). Os valores de INS foram obtidos com $J = 50$ referenciais estacionários. Neste trabalho é adotado o seguinte critério para classificar os ruídos não-estacionários segundo o valor de INS:

- Critério 1: INS $> 10\gamma$ − o sinal é considerado altamente não-estacionário;
- Critério 2: $\gamma <$ INS $\leq 10\gamma$ − o sinal é considerado como moderadamente não-estacionário.

Os resultados observados na FIG. 4.2 mostram que, com exceção do ruído helicóptero, todos os demais ruídos acústicos são não-estacionários para todas as janelas de tempo. Cabe ressaltar que, os ruídos balbúrdia, britadeira, serra elétrica e trem, por apresentarem valores de INS superior ao critério 1, neste trabalho são considerados altamente não-estacionários. Já o ruído fábrica por apresentar INS menor que o estabelecido no critério 2 é moderadamente não-estacionário.

4.2 RESULTADOS DE QUALIDADE PARA REALCE

Nesta Seção, a qualidade do sinal de voz do método de realce proposto é verificada utilizando duas medidas, SegSNR e OQCM. Os resultados obtidos para o método PRO são comparados com três algoritmos espectrais de realce de sinais de voz, SS, Cohen e Wiener, e dois temporais EMDF e EMDH.

4.2.1 SegSNR

A FIG. 4.3 mostra os incrementos de SegSNR obtidos pelo método PRO e demais métodos de realce de sinais de voz. Os resultados estão organizados em ordem decrescente de INS dos ruídos acústicos. O valor de incremento de SegSNR obtido em cada experimento é calculado pela diferença entre o SegSNR do sinal realçado e o do sinal ruidoso. Nota-se que o método PRO (linha preta) alcançou melhores resultados de SegSNR em relação aos

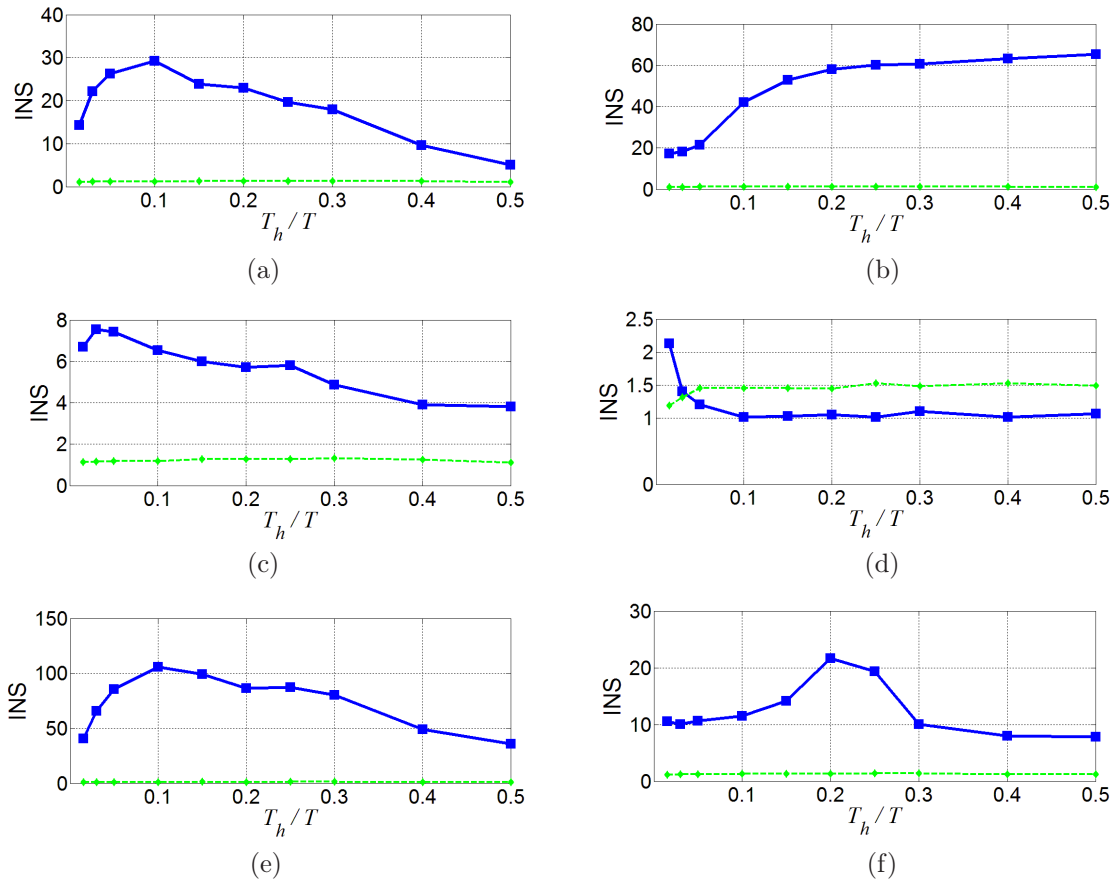

FIG. 4.2: Os valores de INS obtidos de segmentos de 3 s de duração dos ruídos acústicos (a) balbúrdia, (b) britadeira, (c) fábrica, (d) helicóptero, (e) serra elétrica, e (f) trem. As linhas tracejadas indicam os valores correspondentes do limiar γ para os testes de estacionariedade.

algoritmos temporais. Quando a proposta é comparada com os métodos espectrais (SS, Cohen e Wiener), nos ruídos classificados como altamente não-estacionários, o incremento médio de SegSNR é de aproximadamente 1 dB. Esta diferença aumenta para cerca de 2 dB para SNR $>$ 0 dB. É interessante notar que, nos ruídos fábrica e helicóptero, que possuem os menores valores de INS, o método de Cohen aprimorou o sinal em mais de 5 dB quando SNR = -5 dB. Observa-se que, nesta mesma razão sinal-ruído, em serra elétrica e balbúrdia, este valor é inferior a 2 dB e 3 dB, respectivamente. Essa diferença pode ser explicada pelo atraso na atualização do espectro de potência do ruído, característica do estimador IMCRA em presença de ruídos altamente não-estacionários. A justificativa para o bom desempenho do método Cohen em britadeira, se deve ao fato deste ruído apresentar bruscas variações apenas no primeiro segundo. Estes fatores permitiram que o método de Wiener obtivesse o melhor resultado, dentre os espectrais, em ruídos altamente

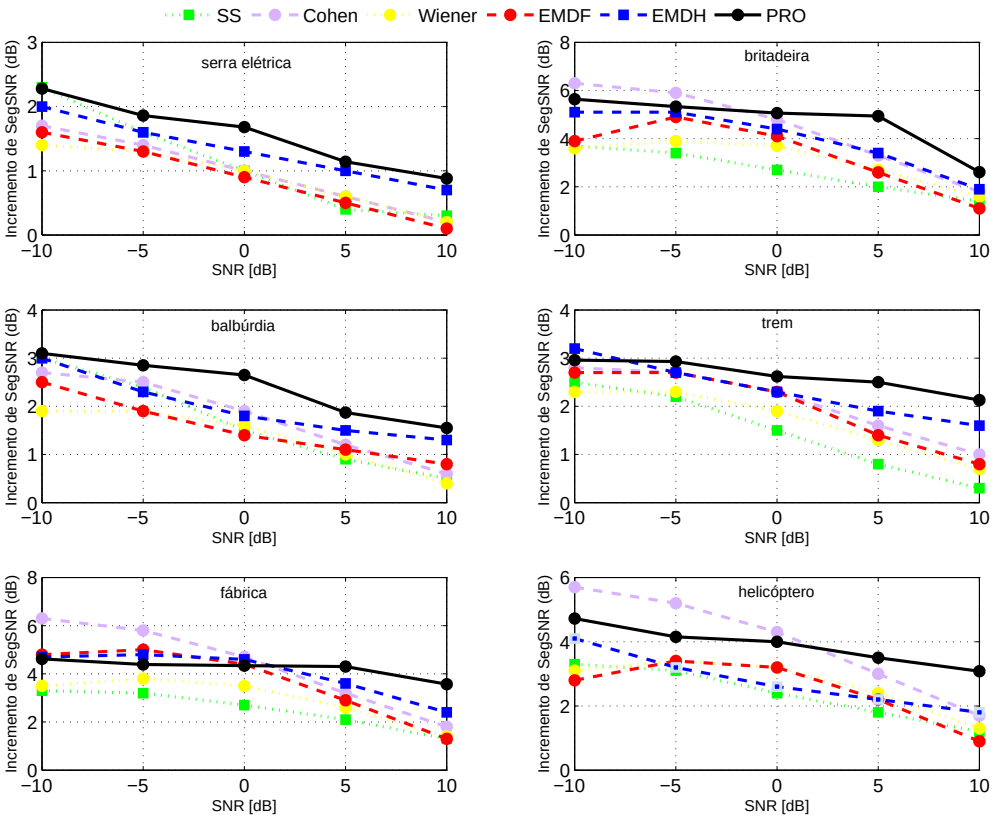

FIG. 4.3: Incrementos de SegSNR (dB) obtidos com as métodos de realce de voz SS, Cohen, Wiener, EMDF, EMDH e a proposta PRO.

não-estacionários.

4.2.2 OQCM

A FIG. 4.4 ilustra os resultados de OQCM obtidos com os métodos PRO, SS, Cohen, Wiener, EMDF e EMDH. O valor de incremento de OQCM obtido em cada experimento é calculado pela diferença entre o OQCM do sinal após a aplicação do método e do ruidoso. Dos métodos temporais de realce, PRO é o que mostra maior ganho de OQCM. Quando comparado com os algoritmos espectrais, assim como nos resultados de SegSNR, PRO apresenta maiores incrementos em ruídos altamente não-estacionários.

Os valores de OQCM para os ruídos serra elétrica, balbúrdia e trem evidenciam que os métodos espectrais de Cohen e de Wiener degradaram a qualidade do sinal de voz para SNR $<$ -5 dB. Para o ruído estacionário helicóptero, Wiener é o que obteve maior incremento de OQCM.

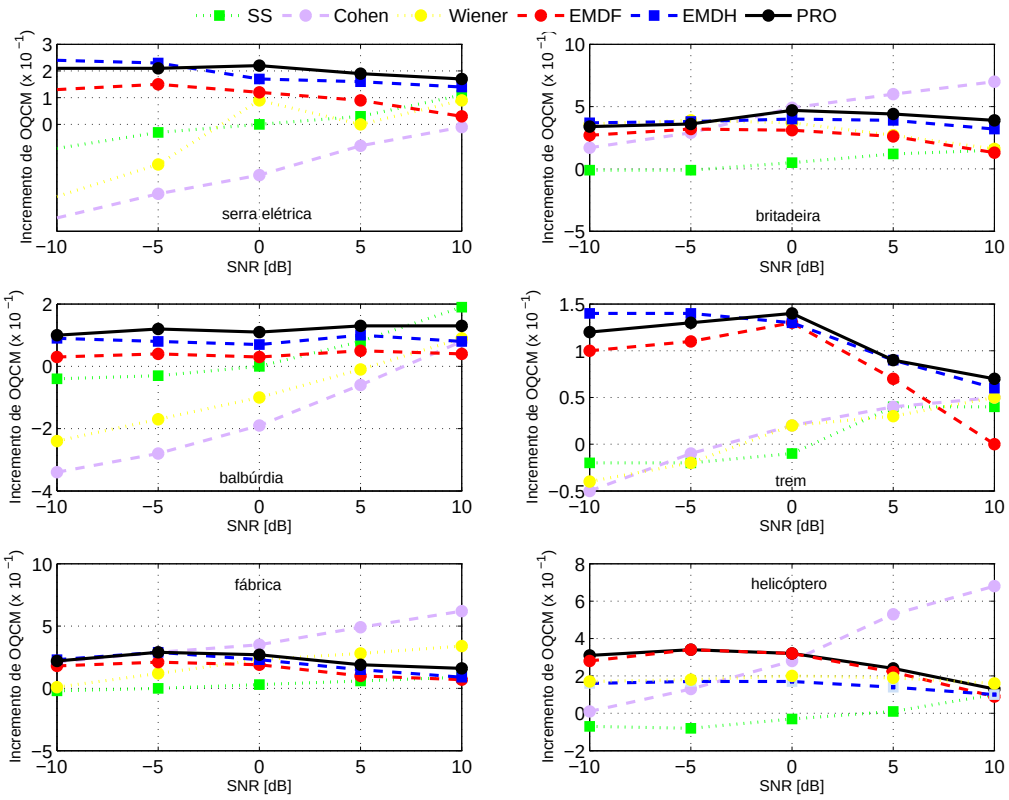

FIG. 4.4: Incrementos na medida OQCM obtidos com as métodos de realce de voz SS, Wiener, EMDF, EMDH e a proposta PRO.

4.3 RESULTADOS DE INTELIGIBILIDADE

Quanto à inteligibilidade, os resultados são analisados de forma objetiva, com o uso de quatro medidas fwSegSNR, CSII, STOI e fAI.

4.3.1 fwSegSNR

A FIG. 4.5 apresenta os resultados para a medida de inteligibilidade fwSegSNR. O valor de incremento de fwSegSNR, em dB, obtido em cada experimento é calculado pela diferença entre o fwSegSNR do sinal após a aplicação do método e do sinal ruidoso. Dos métodos temporais, PRO alcança o maior ganho, superior a 1 dB nos ruídos britadeira e balburdia em valores de SNR $>$ 0 dB. No ruído serra elétrica, a proposta PRO também obtém os melhores resultados, com incrementos próximos de 1 dB. Para SNR $<$ 0 dB, o EMDH apresenta maiores ganhos em britadeira e fábrica. Em relação aos espectrais, Cohen obteve, em média, fwSegSNR de 1 dB acima das demais nos ruídos helicóptero e trem. E o algoritmo de Wiener com ruído trem a 0 dB mostra resultado melhor que os

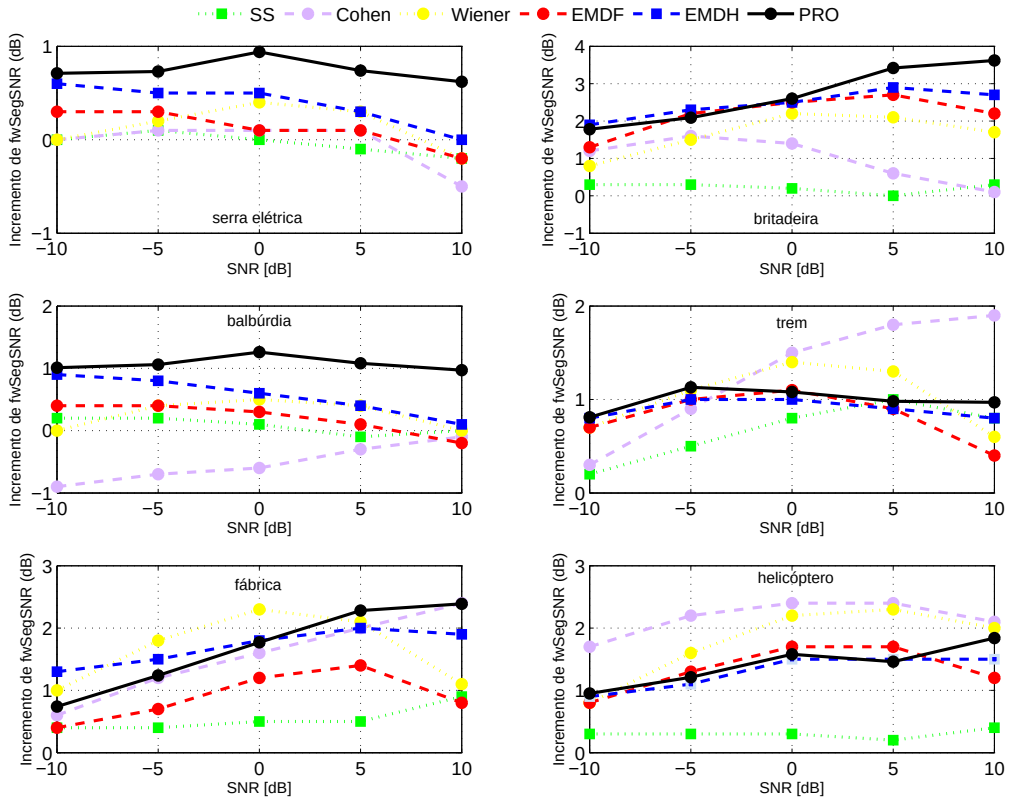

FIG. 4.5: Incrementos de fwSegSNR (dB) obtidos com os métodos de realce de voz SS, Cohen, Wiener, EMDF, EMDH e a proposta PRO.

demais algoritmos (espectrais e temporais) com ganho de 2,32 dB. Considerando todos os métodos, PRO consegue alcançar o maior ganho de inteligibilidade, de 3,72 dB em britadeira em SNR de 10 dB.

4.3.2 CSII

A TAB. 4.1 expõe os resultados de predição das taxas de acertos em sentenças obtidos com a medida de inteligibilidade CSII para o método proposto e os demais algoritmos. Esta medida apresenta um coeficiente de correlação maior que 90% com testes subjetivos de inteligibilidade. E a principal vantagem do seu uso, é que ela leva em consideração além da ação do ruído, a distorção causada pela uso da solução de realce na coerência das sentenças.

Em relação aos demais métodos de realce, a proposta PRO obtém o melhor resultado para todos os ruídos altamente não-estacionários. O ganho médio de inteligibilidade foi de 5,6%. O incremento médio alcançado pelos métodos espectrais, considerando todos os ruídos, é de $-0,3\%$, enquanto o dos temporais foi de $1,2\%$. Parte desta diferença

TAB. 4.1: Predição das taxas de acertos (%) de inteligibilidade obtidos com o resultado do CSII do mapeamento determinado pela EQ. 2.3.

Ruído	SNR	Sem Realce	SS	Cohen	Wiener	EMDF	EMDH	PRO
serra elétrica	10 dB	83,0	82,5	80,9	80,6	80,5	82,3	93,1
	5 dB	48,8	44,0	45,4	45,6	46,5	49,1	54,9
	0 dB	18,7	14,7	13,7	15,3	17,9	18,8	20,8
	-5 dB	5,7	4,3	3,4	4,2	5,4	5,8	6,45
	-10 dB	2,0	1,7	1,4	1,6	2,0	2,1	2,29
	Média	31,6	29,4	29,0	29,5	30,5	31,6	35,5
britadeira	10 dB	99,0	96,1	97,6	97,5	96,9	97,4	99,2
	5 dB	91,2	76,6	90,2	91,6	88,3	90,5	94,6
	0 dB	66,7	50,4	67,2	74,3	64,0	66,9	75,5
	-5 dB	33,9	23,5	30,6	40,3	29,2	30,1	32,3
	-10 dB	12,6	9,1	12,3	14,5	10,6	10,4	11,1
	Média	54,9	51,1	59,6	63,7	57,8	59,1	66,1
balbúrdia	10 dB	94,0	93,0	93,5	92,2	92,1	92,2	96,7
	5 dB	72,8	68,1	73,6	71,4	71,2	71,5	82,4
	0 dB	37,1	29,1	34,1	34,6	36,3	36,4	41,1
	-5 dB	12,7	9,7	8,3	10,0	12,6	12,6	14,2
	-10 dB	4,0	3,0	2,2	2,8	4,0	4,0	4,5
	Média	45,9	40,6	42,3	42,2	43,2	43,4	47,8
trem	10 dB	99,5	97,9	98,0	97,7	97,5	97,6	98,8
	5 dB	92,5	92,3	93,0	91,9	90,6	90,8	94,3
	0 dB	69,2	69,0	74,4	72,0	67,9	67,8	76,0
	-5 dB	33,3	28,8	37,9	36,9	33,1	32,8	38,6
	-10 dB	11,5	8,6	10,8	11,8	11,4	11,4	12,3
	Média	58,6	59,3	62,8	62,1	60,1	60,1	64,2
fábrica	10 dB	99,6	97,5	98,5	98,0	97,4	97,9	98,8
	5 dB	93,3	86,4	94,6	93,3	91,0	92,0	96,3
	0 dB	71,0	55,6	78,7	76,6	69,0	70,3	76,0
	-5 dB	34,6	26,7	43,7	42,4	34,1	34,5	38,6
	-10 dB	11,7	9,7	13,3	14,2	11,6	11,6	11,0
	Média	61,5	55,2	65,8	64,9	60,6	61,3	61,9
helicóptero	10 dB	98,6	95,7	97,7	97,2	96,4	96,9	97,4
	5 dB	89,8	76,7	91,8	90,7	87,7	88,4	89,7
	0 dB	63,8	45,9	70,5	70,4	62,1	62,9	67,3
	-5 dB	28,5	19,4	33,6	35,7	28,0	28,4	27,5
	-10 dB	9,4	6,7	9,3	11,3	9,2	9,3	9,0
	Média	58,6	48,9	60,6	61,0	56,7	57,2	58,2

pode ser atribuída à baixa distorção causada pelo uso de PRO. Isto porque, diferente dos outros métodos, o método proposto não realiza nenhum tipo de transformação no sinal corrompido, como os espectrais que utilizam a transformada de Fourier. O maior ganho de CSII para o ruído britadeira foi alcançado pela proposta PRO: $11,2\%$. Neste mesmo ruído, o algoritmo de Wiener alcançou $8,8\%$, sendo este o melhor resultado obtido pelo

realce com o uso dos métodos espectrais.

A maior redução nos resultados de inteligibilidade foi apresentado pelo método de subtração espectral, que obteve média de $-4,4\%$. Os demais métodos espectrais, Cohen e Wiener, atingiram ganhos médios de $1,5\%$ e $2,1\%$, respectivamente. Já o método proposto teve um incremento médio de $3,7\%$.

Os melhores resultados com o PRO são particularmente interessantes nos ruídos mais não estacionários (serra elétrica, britadeira e balbúrdia): média de $5,7\%$ contra $1,0\%$ do método de Wiener. Note que os algoritmos espectrais Cohen e Wiener nos ruídos com menores valores de INS (helicóptero e fábrica) alcançaram ganho médio de $3,2\%$ e $2,9\%$, respectivamente. O método de Cohen obteve melhores resultados porque, devido aos menores valores de INS, o estimador IMCRA consegue melhor precisão na estimação do espectro de potência destes ruídos, em relação ao estimador UnB-MMSE. No entanto, é interessante ressaltar que, mesmo para estes ruídos, o método SS reduz as taxas de inteligibilidade em 8%.

4.3.3 STOI

Na FIG. 4.6, são exibidos os resultados de predição das taxas de acertos com a medida de inteligibilidade STOI. PRO atinge os melhores resultados, tanto em relação aos métodos temporais quanto aos espectrais, com aumento de cerca de 12% na taxa de acertos de sentenças. O incremento alcançado pelos algoritmos espectrais e temporais foram, respectivamente, 9% e 5%. Estes ganhos se dão, sobretudo, nos valores de SNR acima de 0 dB. O resultado de Wiener é, em média, 8% superior em relação aos demais algoritmos espectrais.

No ruído serra elétrica a 0 dB, os métodos temporais apresentam uma diferença no grau de acertos da ordem de 18% em relação aos algoritmos espectrais. PRO obteve os melhores resultados médios, onde destacam-se os $66,7\%$ alcançados na razão sinal-ruído de 5 dB.

Em balbúrdia, os métodos temporais mostram resultados superiores aos espectrais, cerca de 10% em média. Quanto aos algoritmos espectrais, Cohen e Wiener apresentam valores de ganho aproximados, cerca de 11%. A diferença, no entanto, está na composição da média. Os incrementos de Cohen são de 13%, nos SNR maiores que 0 dB, e os de Wiener são de 12%, nos SNR menores que 0 dB.

A proposta PRO foi superior em 13% na taxa de acertos de inteligibilidade em relação

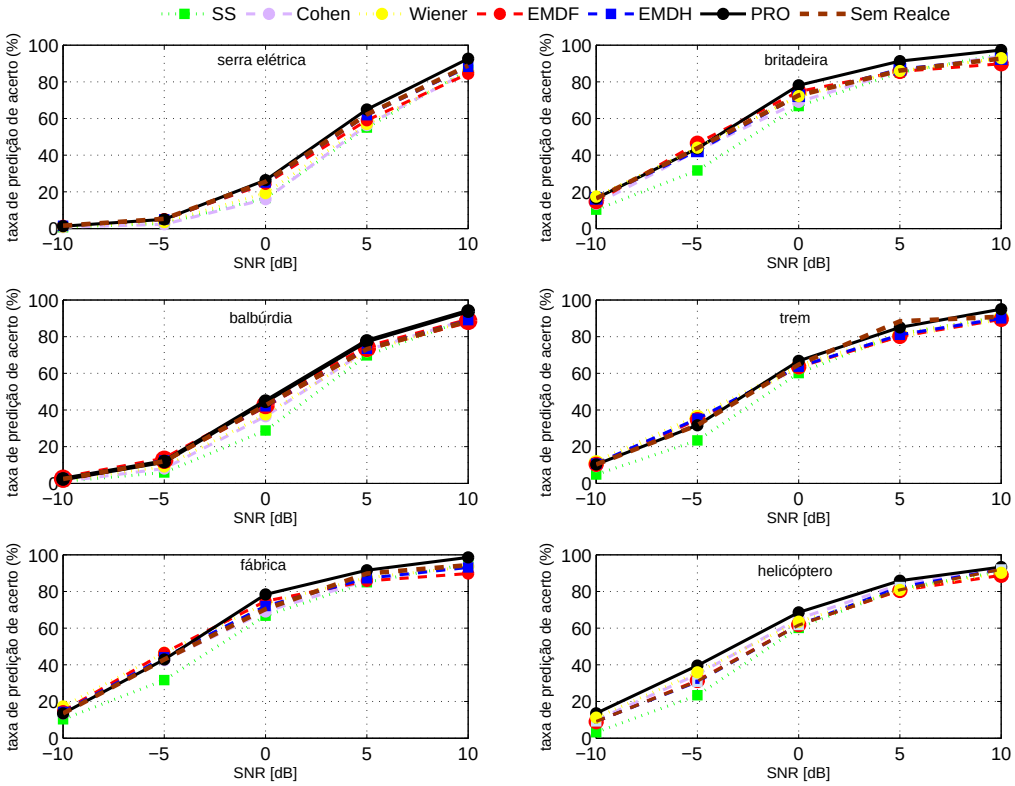

FIG. 4.6: Predição de inteligibilidade com STOI das métodos de realce de voz SS, Wiener, EMDF, EMDH e a proposta PRO.

a todos os outros métodos no ruído fábrica. Os melhores resultados alcançados pelos algoritmos espectrais foram atingidos pelo método Wiener, especificamente no ruído trem, com aumento de 9% na taxa de acertos.

4.3.4 FAI

A TAB. 4.2 mostra os resultados de predição das taxas de acertos de sentenças dos algoritmos examinados com a medida de inteligibilidade FAI. Os métodos temporais foram melhores que os espectrais em, aproximadamente, 12%. Em relação à voz sem tratamento, eles obtêm 1% e o espectrais reduzem a inteligibilidade em 11%. PRO obteve o maior incremento médio geral, $2,3\%$ quando ele é comparado com o sinal sem realce, e este desempenho advém dos ganhos em SNR > 0 dB. Em serra elétrica e britadeira, PRO alcança uma diferença de, aproximadamente, 5% em relação aos outros métodos temporais, e 9% dos espectrais.

Nos ruídos trem e balbúrdia, o método PRO é o único que consegue aumentar a taxa média de acertos. Com relação aos algoritmos espectrais, aquele que consegue as maiores

TAB. 4.2: Predição das taxas de acertos (%) de inteligibilidade obtidos com o resultado do FAI do mapeamento determinado pela EQ.2.5

Ruído	SNR	Sem Realce	SS	Cohen	Wiener	EMDF	EMDH	PRO
serra elétrica	10 dB	99,1	96,1	96,1	96,4	97,0	97,4	99,0
	5 dB	97,6	88,5	91,7	92,3	93,6	94,1	97,7
	0 dB	87,7	64,3	77,1	80,5	85,6	86,3	89,8
	-5 dB	71,6	36,3	52,1	61,8	68,5	70,6	72,8
	-10 dB	40,5	6,3	15,3	27,6	38,3	40,3	41,2
	Média	79,3	58,3	66,5	71,7	76,6	76,6	80,3
britadeira	10 dB	99,1	96,9	97,5	97,8	98,2	98,5	99,2
	5 dB	97,5	87,3	94,9	96,1	96,7	97,4	98,7
	0 dB	96,1	57,1	88,3	92,2	93,7	94,8	97,3
	-5 dB	87,7	18,6	67,7	81,0	84,9	87,8	88,6
	-10 dB	65,0	4,2	23,7	53,2	63,2	65,6	69,0
	Média	89,1	52,8	74,4	84,1	87,3	88,8	90,7
balbúrdia	10 dB	97,2	94,5	94,1	94,3	95,3	95,3	98,7
	5 dB	90,5	83,3	85,1	86,0	88,7	88,7	91,8
	0 dB	66,5	44,0	58,8	60,6	65,8	65,9	71,4
	-5 dB	28,6	9,6	20,4	23,4	28,3	28,3	28,8
	-10 dB	5,9	0,6	2,6	3,9	5,8	5,8	6,1
	Média	57,7	46,4	52,2	53,6	56,8	56,8	59,4
trem	10 dB	99,4	97,8	97,6	97,5	97,8	97,9	99,3
	5 dB	97,1	95,9	95,7	95,4	96,1	96,2	98,4
	0 dB	92,5	89,6	90,3	89,8	91,8	92,0	94,2
	-5 dB	80,3	62,0	72,9	73,0	78,3	78,6	80,8
	-10 dB	47,8	20,1	37,0	39,5	47,5	47,7	51,0
	Média	83,4	73,1	78,7	79,1	82,3	82,5	90,7
fábrica	10 dB	99,1	97,2	97,6	97,6	97,8	98,1	99,0
	5 dB	97,4	92,4	95,3	95,5	96,3	96,6	98,0
	0 dB	94,4	66,9	88,9	90,1	92,4	92,7	95,5
	-5 dB	79,9	21,5	68,1	73,1	78,9	79,2	80,8
	-10 dB	43,8	3,6	24,9	34,7	43,2	43,6	45,8
	Média	82,9	56,3	75,0	78,2	81,7	82,0	83,8
helicóptero	10 dB	98,7	96,4	96,9	97,1	97,6	97,9	99,2
	5 dB	96,5	87,0	93,5	94,3	95,9	96,1	98,0
	0 dB	90,7	53,9	82,7	86,1	90,4	90,8	93,4
	-5 dB	76,6	15,3	56,5	66,0	75,7	76,0	78,6
	-10 dB	35,5	2,5	14,9	24,5	34,9	35,0	36,0
	Média	79,6	51,0	68,9	73,6	78,9	79,1	81,0

taxas médias para todos os ruídos é o de Wiener. Todavia, é importante ressaltar que nenhum deles (SS, Cohen ou Wiener) consegue aumentar as taxas de acertos de sentenças em relação aos sinais de voz ruidosos.

4.3.5 AVALIAÇÃO GERAL DE INTELIGIBILIDADE

A aplicação das quatro medidas de avaliação de inteligibilidade mostra que, apesar de serem calculados de formas distintas, elas apresentam interpretações muito similares. Em geral, os métodos temporais superam os ganhos de inteligibilidade dos espectrais nos ruídos considerados altamente não-estacionários. A proposta PRO obtém, em média, melhores resultados que os demais algoritmos. O principal diferencial nos resultados está nos incrementos de inteligibilidade, que ocorrem acima da razão sinal-ruído de 0 dB. Um dos motivos para este desempenho é que este método não emprega nenhuma transformação no sinal corrompido, como é feito pelas soluções espectrais que usam a transformada de Fourier ou os métodos tempo-frequência que, utilizam técnicas de decomposição para realizar o realce.

Os maiores ganhos do método proposto ocorrem em SNR > 0 porque fica mais fácil para o algoritmo de PRO separar os sinais dos ruídos. E dentre os métodos espectrais, Wiener é o que apresenta os melhores resultados, e isto se deve ao menor tempo de atraso da estimativa do espectro.

4.4 RESUMO

Este Capítulo apresentou experimentos para a avaliação da proposta PRO sob o aspecto da qualidade e da inteligibilidade de voz. Foram usadas duas medidas para julgar a qualidade da voz, SegSNR e OQCM. Para examinar a inteligibilidade, foram utilizadas quatro medidas, fWSegSNR, CSII, STOI e FAI. Nos testes experimentais, foram empregados seis métodos de realce de sinais, divididos em dois grupos, os espectrais (SS, Cohen e Wiener) e os temporais (EMDF, EMDH e PRO). Os métodos foram aplicados em sinais de voz corrompidos por seis ruídos acústicos coletados de fontes reais. Segundo os seus valores de INS, os ruídos foram divididos da seguinte forma: quatro ruídos altamente não-estacionários, um moderadamente não-estacionário, e o último estacionário. Na avaliação de qualidade de voz, realizada com o uso do SegSNR e OQCM, verifica-se que PRO apresenta o melhor resultado na maioria dos testes. Diferentemente dos demais métodos, os ganhos em qualidade se deram de forma mais robusta, em valores de SNR maiores que 0 dB. Isto se deve à característica de PRO, que entende o sinal de voz como um valor discrepante em relação ao ruído. E, nestes valores de SNR fica mais fácil para o algoritmo discriminar o sinal de voz do ruído. Os quatro diferentes testes de inteligibi-

lidade mostram que PRO apresenta, em média, resultado melhor que os outros métodos. As quatro medidas fwSegSNR, CSII, STOI e FAI, apesar de serem calculadas de formas distintas, mostram resultados muito similares. Este fato reforça o ganho de inteligibilidade de PRO, tanto em relação aos espectrais como aos temporais, nos ruídos altamente não-estacionários.

5 CONCLUSÃO E TRABALHOS FUTUROS

Nesta Dissertação, foi proposta uma solução para o realce de sinais de voz no domínio do tempo. Nesta abordagem, o sinal de voz corrompido é dividido em janelas e, para cada uma destas janelas, é obtido o desvio padrão do ruído através do estimador robusto DATE. A fim de extrair as componentes do ruído, são excluídas primeiramente as componentes mais afetadas com o uso de uma regra de decisão. Os valores restantes são atenuados com o desvio padrão de seu respectivo quadro, calculado com o DATE. O sinal é reconstruído com as componentes remanescentes.

Para a avaliação do método proposto, os sinais de voz foram corrompidos por seis ruídos acústicos com diferentes índices de não-estacionariedade. Este método de realce foi ainda comparado com outros cinco algoritmos de supressão de ruídos.

Os resultados comparativos confirmam o bom desempenho do método proposto, principalmente em ruídos altamente não-estacionários, onde foram obtidos incrementos de SegSNR acima de 1 dB, e não houve distorção da qualidade na medida OQCM. Adicionalmente, este método apresentou ganhos acima de 1 dB para SNR maiores que 0 dB em fwSegSNR. A proposta ainda aumentou a predição das taxas de acertos de sentenças: 7% na medida CSII , 12% na STOI, e 14% na FAI. Por outro lado, os métodos espectrais reduziram as taxas de acertos.

As principais contribuições apresentadas nesta Dissertação podem ser resumidas da seguinte forma:

- proposta de um método de realce de sinais de voz corrompidos por ruídos não-estacionários, que utiliza como critério de seleção das componentes mais afetadas pelo ruído um limitante obtido a partir da estimação robusta do desvio padrão do ruído. A proposta aprimorou as medidas objetivas utilizadas para avaliar a qualidade e a inteligibilidade dos sinais de voz. Em comparação às soluções utilizadas como referência, os resultados da proposta foram particularmente interessantes para os ruídos com maiores valores de INS.

5.1 SUGESTÕES PARA TRABALHOS FUTUROS

Nesta Seção serão destacadas algumas sugestões para trabalhos futuros:

- estudar a utilização de outros estimadores robustos como critério para identificação e estimação de componentes ruidosas para a realização do realce no domínio do tempo;
- investigar o uso do índice de não-estacionariedade como critério para melhorar a estimação das componentes do ruído;
- utilizar a proposta de realce de sinais de voz como pós-realce, aplicada aos sinais de voz previamente tratados pelo algoritmos SS, Wiener, EMDF e EMDH;
- avaliar as taxas de acertos em reconhecimento automático de locutor de sinais de voz previamente realçados pela proposta.

5.2 COMENTÁRIOS FINAIS

Nesta Dissertação foi apresentada uma proposta de realce para o problema de distorções acústicas nos sinais de voz. Para sinais corrompidos por ruídos acústicos, o método proposto é realizado no domínio do tempo e utiliza um estimador de desvio padrão do ruído para obter um critério de seleção das amplitudes altamente distorcidas. Os experimentos de realce mostraram que o método proposto apresentou resultados promissores, principalmente para ruídos altamente não-estacionários.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ATAL, B. Automatic recognition of speakers from their voices. **Proceedings of the IEEE**, 64(4):460–475, April 1976.

BASSEVILLE, M. Distance measures for signal processing and pattern recognition. **Signal Processing**, 18(4):349–369, December 1989.

BISPO, B., ESQUEF, P., BISCAINHO, L., LIMA, A., FREELAND, F., JESUS, R., SAID, A., LEE, B., SCHAFER, R. e KALKER, T. EW-PESQ: A quality assessment method for speech signals sampled at 48 khz. **Journal of the Audio Engineering Society**, 58(4):251–268, April 2010.

BOLL, S. Suppression of acoustic noise in speech using spectral subtraction. **IEEE Transactions on Acoustics, Speech and Signal Processing**, 27(2):113–120, April 1979.

BORGNAT, P., FLANDRIN, P., HONEINE, P., RICHARD, C. e XIAO, J. Testing stationarity with surrogates: A time-frequency approach. **IEEE Transactions on Signal Processing**, 58(7):3459–3470, July 2010.

CHATLANI, N. e SORAGHAN, J. EMD-based filtering (EMDF) of low-frequency noise for speech enhancement. **IEEE Transactions on Audio, Speech, and Language Processing**, 20(4):1158–1166, May 2012.

COHEN, I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. **IEEE Transactions on Speech and Audio Processing**, 11(5):466–475, September 2003.

COHEN, I. e BERDUGO, B. Speech enhancement for non-stationary noise environments. **Signal Processing**, 81(11):2403–2418, 2001.

COHEN, L. **Time Frequency Analysis**. Prentice-Hall, New York, USA, 1995.

DODDINGTON, G. Speaker verification - identifying people by their voices. **Proceedings of the IEEE**, 73(11):1651–1664, November 1985.

DONOHO, D. e JOHNSTONE, I. Threshold selection for wavelet shrinkage of noisy data. **Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'94)**, 1:A24–A25, November 1994.

DONOHO, D. e HUBER, P. The notion of breakdown point. **A Festschrift for Erich Lehmanng**, (157), january 1983.

EPHRAIM, Y. e MALAH, D. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, 32(6):1109–1121, December 1984.

EPHRAIM, Y. e MALAH, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. **IEEE Transactions on Acoustics, Speech and Signal Processing**, 33(2):443–445, April 1985.

FLANDRIN, P., GONÇALVES, P. e RILLING, G. Detrending and denoising with empirical mode decompositions. **Proceedings of the European Signal Processing Conference (EUSIPCO'04)**, págs. 1581–1584, September 2004a.

FLANDRIN, P., RILLING, G. e GONCALVES, P. Empirical mode decomposition as a filter bank. **IEEE Signal Processing Letters**, 11(2):112–114, February 2004b.

GAROFOLO, J., LAMEL, L., FISHER, W., FISCUS, J., PALLETT, D., DAHLGREN, N. e ZUE, V. TIMIT acoustic-phonetic continuous speech corpus. **Linguistic Data Consortium**, 1993.

GERKMANN, T. e HENDRIKS, R. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. **IEEE Transactions on Audio, Speech, and Language Processing**, 20(4):1383–1393, 2012.

HAMPEL, F., RONCHETTI, E., ROUSSEEUW, P. e STAHEL, W. **Robust Statistics: The Approach Based on Influence Functions**. Wiley, New York, USA, abril 2005.

HANSEN, J. e PELLOM, B. An effective quality evaluation protocol for speech enhancement algorithms. **Proceedings of the International Conference on Speech and Language Processing (ICSLP'98)**, págs. 2819–2822, December 1998.

HENDRIKS, R., HEUSDENS, R. e JENSEN, J. MMSE based noise psd tracking with low complexity. **Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'10)**, págs. 4266–4269, 2010.

HU, Y. e LOIZOU, P. Evaluation of objective measures for speech enhancement. **Proceedings of INTERSPEECH**, págs. 1–4, September 2006.

HU, Y. e LOIZOU, P. Subjective evaluation and comparison of speech enhancement algorithms. **Speech Communication**, 49(7):588–601, July 2007.

HU, Y. e LOIZOU, P. Evaluation of objective quality measures for speech enhancement. **IEEE Transactions on Audio, Speech and Language Processing**, 16(1):229–238, January 2008.

HUANG, N., SHEN, Z., LONG, S., WU, M., SHIH, H., ZHENG, Q., YEN, N., TUNG, C. e LIU, H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. **Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences**, 454(1971):903–995, March 1998.

HUBER, P. e RONCHETTI, E. **Robust statistics**. Wiley, New York, USA, 2009.

HURST, E. Long-term storage capacity of reservoirs. **Transaction of the American Society of Civil Engineers**, 116(11):770–799, April 1951.

KAISER, J. On a simple algorithm to calculate the 'energy' of a signal. **Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'90)**, págs. 381–384, April 1990.

KATES, J. Coherence and the speech intelligibility index. **The Journal of the Acoustical Society of America.**, 4(1):2224–2237, April 2005.

KAY, S. M. **Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory**. Prentice-Hall Inc, New Jersey, 1993.

KLATT, D. Prediction of perceived phonetic distance from critical-band spectra: A first step. **Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'82)**, 7:1278–1281, May 1982.

KRYTER, K. Methods for the calculation and use of the articulation index. **The Journal of the Acoustical Society of America**, 34(11):1689–1697, November 1962.

LOIZOU, P. **Speech Enhancement: theory and practice**. CRC Press, 2007a.

LOIZOU, P. e HU, Y. A comparative intelligibility study of single-microphone noise reduction algorithms. **The Journal of the Acoustical Society of America**, 22(3): 1777–1786, 2007b.

LOIZOU, P. e MA, J. Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms. **The Journal of the Acoustical Society of America**, 130(2):986–995, 2011a.

LOIZOU, P. e MA, J. Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms. **The Journal of the Acoustical Society of America**, 130(2):986–995, August 2011b.

MA, J., HU, Y. e LOIZOU, P. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. **The Journal of the Acoustical Society of America**, 125(5):3387–3405, 2009.

MANOHAR, K. e RAO, P. Speech enhancement in nonstationary noise environments using noise properties. **Speech Communication**, 48:96–109, January 2006.

MARTIN, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. **IEEE Transactions on Speech and Audio Processing**, 9(5): 504–512, July 2001.

MING, J., HAZEN, T., GLASS, J. e REYNOLDS, D. Robust speaker recognition in noisy conditions. **IEEE Transactions on Audio, Speech, and Language Processing**, 15(5):1711–1723, July 2007.

PASTOR, D. e SOCHELEAU, F. Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurences. **IEEE Transactions on Signal Processing**, 60(4):1545–1555, April 2012.

QUACKENBUSH, S., BARNWELL, T. e CLEMENTS, M. **Objective Measures Of Speech Quality**. Prentice-Hall, Inc., 1988.

REYNOLDS, D. e ROSE, R. Robust text independent speaker identification using gaussian mixture speaker models. **IEEE Transactions on Speech and Audio Processing**, 3:72–82, 1995.

RHEBERGEN, K. e VERSFELD, N. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. **The Journal of the Acoustical Society of America**, 117(4): 2181–2192, April 2005.

RIX, A., BEERENDS, J., HOLLIER, M. e HEKSTRA, A. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. **Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)**, 2:749–752, May 2001.

ROUSSEEUW, P. J. e RONCHETTI, E. Influence curves of general statistics. **Journal of Computational and Applied Mathematics**, 7(3):161 – 166, 1981.

SANT'ANA, R., COELHO, R. e ALCAIM, A. Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model. **IEEE Transactions on Audio, Speech, and Language Processing**, 14 (3):931–940, May 2006.

SCALART, P. e FILHO, J. Speech enhancement based on a priori signal to noise estimation. **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)**, 32(6):629–632, December 1996.

SCHULLER, B., VLASENKO, B., EYBEN, F., RIGOLL, G. e WENDEMUTH, A. Acoustic emotion recognition: A benchmark comparison of performances. **IEEE Workshop on Automatic Speech Recognition Understanding**, págs. 552–557, 2009.

STEENEKEN, H. e HOUTGAST, T. A physical method for measuring speech transmission quality. **The Journal of the Acoustical Society of America**, 67(1):318–326, January 1980.

STIGLER, S. Simon newcomb, percy daniell and the history of robust estimation 1885-1920. **Journal American Statistical Association**, 68(344):872–879, 1973.

TAAL, C., HENDRIKS, R., HEUSDENS, R. e JENSEN, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. **IEEE Transactions on Audio, Speech and Language Processing**, 19(7):2125–2136, September 2011.

VARGA, A. e STEENEKEN, H. Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. **Speech Communication**, 12(3):247–251, 1993.

WIENER, N. **Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications**. MIT Press, Cambridge, MA, 1949.

ZÃO, L., CAVALCANTE, D. e COELHO, R. Time-frequency feature and AMS-GMM mask for acoustic emotion classification. **IEEE Signal Processing Letters**, 21(5): 620–624, May 2014a.

ZÃO, L. e COELHO, R. Colored noise based multicondition training technique for robust speaker identification. **IEEE Signal Processing Letters**, 18(11):675–678, November 2011.

ZÃO, L., COELHO, R. e FLANDRIN, P. Speech enhancement with emd and hurst-based mode selection. **IEEE/ACM Audio, Transactions on Speech, and Language Processing**, 21(99):899–911, 10 2014b.

ZOUBIR, A., KOIVUNEN, V., CHAKHCHOUKH, Y. e MUMA, M. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. **IEEE Signal Processing Magazine**, 29(4):61–80, July 2012.