Estudo de Métodos de Estimação de Frequência Fundamental em Sinais Reverberantes-Ruidosos

A. Queiroz e R. Coelho

***Resumo*— Este artigo apresenta um estudo dos efeitos de ruído e reverberação na acurácia da estimação da frequência fundamental em sinais de voz. Seis métodos de estimação (ACF, YIN, SWIPE, TAO, HHT-Amp e SFF) são considerados na análise para diferentes condições de reverberação e ruído incluindo o grau de não-estacionaridade destas interferências. Duas medidas de erro (GE e MAE) são aplicadas na avaliação da acurácia dos estimadores. Os resultados mostraram que o método HHT-Amp apresentou os menores erros médios de estimação para todos os cenários quando comparado com estimadores competitivos.**

***Palavras-Chave*— Estimação da frequência fundamental, avaliação da acurácia, não-estacionaridade.**

***Abstract*— This letter presents a study of the noise and reverberation effects in accuracy of fundamental frequency estimation of speech signals. Six estimation methods (ACF, YIN, SWIPE, TAO, HHT-Amp and SFF) are considered in the analysis for different reverberation and noise conditions, including the nonstationarity degree of these interferences. Two error measures (GE and MAE) are applied to accuracy evaluation of the estimators. The results shown by the HHT-Amp method presented the minor mean estimation errors for all scenarios when compared to competitive estimators.**

***Keywords*— Fundamental frequency estimation, accuracy evaluation, nonstationarity.**

I. INTRODUÇÃO

Ruídos e reverberações são efeitos reais comumente presentes em ambientes e cenários urbanos. Esses podem afetar severamente sinais de voz, alterando suas principais características temporais e espectrais. Tais alterações podem ser notadas em suas componentes harmônicas ou na frequência fundamental (F_0) desses sinais [1]. A F_0 ou *pitch*, consiste no menor componente periódico de segmentos sonoros da voz provenientes da vibração das pregas vocais. Sua estimação é importante em diversas áreas de processamento de sinais tais como o reconhecimento de locutor, síntese, detecção ou codificação de sinais de voz.

Diversas soluções para estimação de frequência fundamental de sinais sonoros ou harmônicos foram propostas na literatura, com atuação no domínio do tempo quanto no dimínio espectral. Os métodos ACF [2] e YIN [3] caracterizam-se por uma abordagem temporal baseda na função autocorrelação [2], enquanto que a solução SWIPE (*Sawtooth Waveform Inspired Pitch Estimator*) [4] atua no domínio da frequência. Além disso, alguns métodos como o SFF [6] e HHT-Amp [7] também analisaram o efeito dos ruídos em componentes harmônicos dos sinais de voz. Essas interferências mascaram os sinais diminuindo assim a acurácia da estimação da F_0.

A reverberação é outro efeito que também pode afetar o sinal de voz, e consequentemente a estimação da *pitch*. Tal efeito é causado pelas múltiplas reflexões de uma onda sonora em objetos e superfícies, geralmente observado com facilidade em ambientes fechados, como auditórios, igrejas, teatros ou salas de aula [8]. As primeiras reflexões chegam ao ouvinte entre 50 ms e 80 ms após a emissão do sinal. Por outro lado, a etapa das reverberações tardias é definida pelo decaimento e maior distorção do sinal de voz. Estes distorções têm consequências indesejáveis, como o comprometimento da inteligibilidade, principalmente em usuários de implantes cocleares [9]. A função de transferência que caracteriza o efeito da reverberação em uma sala é denominada RIR (*Room Impulse Response*). O tempo de reverberação (RT_{60}) consiste no tempo demandado para que a intensidade de um sinal decaia 60 dB. Este parâmetro é importante na avaliação da qualidade acústica do ambiente.

Este artigo apresenta um estudo detalhado dos efeitos da reverberação e dos ruídos na acurácia das técnicas de estimação da F_0. No trabalho, foram examinados seis métodos de estimação de F_0 (ACF [2], YIN [3], SWIPE [4], TAO [5], HHT-Amp [7] e SFF [6]) para diferentes ambientes reverberantes-ruidosos. A acurácia dos estimadores é analisada considerando-se as duas principais medidas de erro adotadas na literatura: GE (*Gross Error*) e MAE (*Mean Absolute Error*). A base de sinais de voz utilizada é a CSTR (*Centre of Speech Technology Research*) [10], pois a mesma disponibiliza os valores de F_0 que podem ser utilizados como referência na avaliação das soluções examinadas. Estes sinais são reverberados em três salas, sendo uma a *Stairway* da base AIR [11] e outras duas salas LASP1 e LASP2, ambas com diferentes valores de RT_{60}. Os sinais reverberados são corrompidos por quatro tipos diferentes de ruídos (*Babble*, *Traffic*, *Car*, *Helicopter*), considerando três valores de SNR: -5 dB, 0 dB e 5 dB. Para examinar o comportamento dos diferentes sinais reverberantes-ruidosos, são apresentadas as medidas do índice de não-estacionaridade (INS) [12]. Extensivos experimentos são realizados demonstrando o impacto da reverberação e do ruído na acurácia da estimação de F_0.

O restante do artigo está estruturado da seguinte maneira: A Seção II mostra o efeito da reverberação e dos ruídos nos sinais de voz, principalmente em suas medidas de não-estacionaridade. A Seção III descreve os métodos de estimação de *pitch* utilizadas no trabalho. Na Seção IV, as técnicas são avaliadas de acordo com a acurácia da estimação para diferentes cenários reverberantes-ruidosos. Finalmente, a Seção V conclui o trabalho.

A. Queiroz é mestrando do Programa de Pós-Graduação em Engenharia Elétrica do Instituto Militar de Engenharia (IME) e Bolsista da CAPES. O trabalho dos autores A. Queiroz e R. Coelho é desenvolvido no Laboratório de Processamento de Sinais Acústicos (LASP/IME) e parcialmente financiado pelo CNPq (308155/2019-0) e pela FAPERJ (203075/2016). E-mails: {anderson.queiroz,coelho}@ime.eb.br.

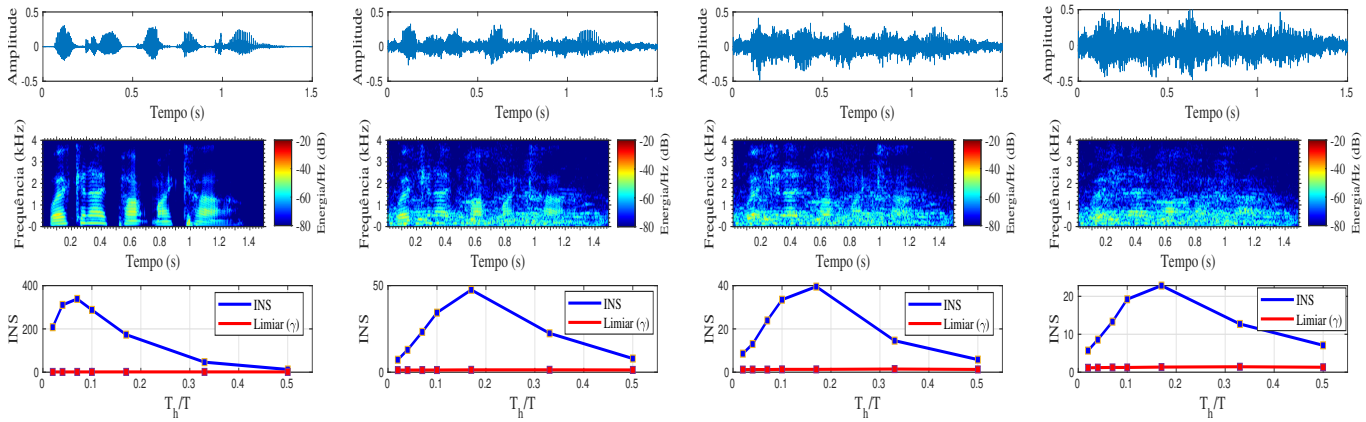

Fig. 1. Sinais de voz (a) limpo, (b) com ruído *Babble* (SNR=0 dB) e Reverberantes-Ruidosos com ruído *Babble* (SNR=0 dB) e RT_{60}: (c) 0,7 s e (d) 1,1 s, com seus respectivos espectrogramas e valores de INS.

II. Reverberação e a Não-Estacionaridade em Sinais Sonoros

Um sinal reverberante-ruidoso pode ser descrito por

$$x\left(t\right) = s\left(t\right) * h\left(t\right) + w\left(t\right) \tag{1}$$

onde $s\left(t\right)$ é o sinal de voz, $h\left(t\right)$ a resposta ao impulso da sala, e $w\left(t\right)$ o ruído acústico.

As características de não-estacionaridade dos sinais de voz podem ser mascaradas por ruídos e reverberações, pois estes efeitos são encontrados nas mesmas faixas de freqüência da voz. Consequentemente, há uma redução na acurácia da estimação da *pitch*. O INS (Índice de Não-Estacionaridade) [12] é adotado no trabalho para estudo em ambientes reverberantes-ruidosos. Esta medida é obtida a partir da comparação do sinal de voz com referenciais estacionários chamados *surrogates*. O INS é obtido de acordo com a escala de observação T_h/T, que consiste na razão entre o tamanho da janela utilizada na análise espectral (T_h), e a duração total do sinal (T). Em [12], um limiar γ é definido para cada valor da janela T_h, considerando uma precisão de 95%. Este limiar é comparado com o valor de INS para avaliação da hipótese de estacionaridade, ou seja

$$INS \begin{cases} \leq \gamma, & \text{sinal é estacionário;} \\ > \gamma, & \text{sinal é não-estacionário.} \end{cases} \tag{2}$$

A Figura 1 ilustra os valores de INS de uma locução diante de 4 cenários: sinal limpo (a), sinal com ruído *Babble* da base RSG-10 [24] (SNR = 0 dB) (b), e os demais reverberantes-ruidosos. Estes últimos são reverberados com as salas: (c) LASP1[1] e (d) *Stairway* (da base AIR [11]). Os valores de RT_{60} são 0,7 s e 1,1 s, respectivamente. Além da reverberação, os sinais também apresentam o ruído *Babble* com SNR 0 dB. Note que o ruído acústico e a reverberação mascaram as características temporais e espectrais do sinal de voz. O sinal limpo possui suas componentes de frequência bem definidas, possibilitando inclusive a distinção entre instantes com ou sem atividade vocal. Por outro lado, as locuções reverberantes-ruidosas apresentam distorções em todo o espectro. Grande parte da alteração se encontra nas baixas frequências, onde está concentrada a maior parte da energia dos ruídos acústicos. Estes efeitos resultam em uma queda da qualidade e inteligibilidade dos sinais [8][13][14].

Segundo os resultados de INS, os sinais nas quatro condições apresentam não-estacionaridade em todas as escalas temporais. Note que o sinal de voz limpo é aqui classificado como altamente não-estacionário pois atingiu o valor de INS máximo de 340. Já o sinal de voz corrompido pelo ruído *Babble* (Figura 1.b), obteve o valor de INS máximo de 45 sendo classificado como não-estacionário. Enquanto os sinais reverbertantes-ruidosos, considerando as duas salas ($RT_{60} = 0,7$ e $RT_{60} = 1,1$), alcançaram os valores máximos de INS de 40 e 25, respectivamente. Os resultados demonstram o impacto da reverberação e ruído na estrutura espectral e temporal do sinal de voz. Consequentemente, estes efeitos comprometem a acurácia dos estimadores.

III. Métodos de Estimação da F_0

Esta seção descreve brevemente os métodos de estimação da F_0 ACF, YIN, SWIPE, TAO, SFF e HHT-Amp investigadas neste trabalho. Estas soluções em geral, foram examinadas para a estimação da F_0 em sinais limpos e algumas também examinadas para sinais ruidosos [6][7]. No presente artigo, a acurácia da estimação de cada técnica é avaliada para sinais de voz sonoros em ambientes reverberantes-ruidosos.

A. ACF

Este estimador clássico baseia-se na aplicação da autocorrelação em sequências de amostras do sinal x_t, ou seja

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j \, x_{j+\tau} \tag{3}$$

onde $r_t(\tau)$ é a função autocorrelação do *lag* τ em um tempo t, e W o tamanho da janela de integração. A partir dos picos da curva de autocorrelação, é possível definir a periodicidade do quadro em análise [2]. Para o quadro de um sinal com taxa de amostragem f_s, com pico de autocorrelação no *lag* τ, seu período é definido por $T_0 = \tau / f_s$. Logo, a F_0 é obtida justamente pelo inverso deste período, ou seja, $F_0 = 1/T_0$ Hz.

[1]Disponível em: http://lasp.ime.eb.br/index.php?vPage=downloads.

B. YIN

Esta técnica baseia-se na AMDF (*Average Magnitude Difference Function*) [15], a qual consiste em uma variação da função autocorrelação. A AMDF é normalizada para que a periodicidade, que era observada nos picos da função, seja definida pelo *lag* (τ) para o qual o resultado é zero. Então, um limiar é aplicado para reduzir os erros de detecção de sub-harmônicos. A definição de limites de frequências superiores e inferiores $[F_{min}, F_{max}]$, contribuem na precisão das estimativas de F_0 [3].

C. SWIPE

Esta solução extrai a F_0 do sinal de voz, analisando suas características espectrais. A frequência fundamental é estimada pela obtenção de uma forma de onda com características espectrais similares ao espectro do sinal de entrada. Esta aproximação é obtida por uma função com um valor de frequência que maximiza a medida de distância do pico para o vale entre as suas harmônicas.

Ajustes são realizados no espectro aproximado para alinhar os componentes harmônicos com o sinal de entrada e ressaltar as características mascaradas pelas interferências do ambiente. Então a função aproximada é multiplicada por um envelope com decaimento de $1/f$, evitando a periodicidade que a autocorrelação apresenta em alguns sinais. Finalmente, a função é normalizada para que os lóbulos espectrais coincidam com os lóbulos positivos do cosseno [4].

D. TAO

A proposta deste método consiste em detectar a F_0 da fala, principalmente de idiomas com estruturas não-monotônicas, como é o caso do Mandarim [5]. Esta técnica baseia-se na aplicação da decomposição EEMD (*Ensamble Empirical Mode Decomposition*) [17]. O EEMD é um aprimoramento da versão EMD original [18], e consiste na aplicação de realizações de ruído branco gaussiano no sinal de entrada, antes da sua decomposição, de forma a minimizar ou evitar o efeito "mode mixing"[19]. Com isso, ocorre um aprimoramento na acurácia da decomposição. O resultado deste processo apresenta uma série de IMF's (*Intrinsic Mode Functions*), onde cada uma delas possui uma oscilação característica.

A decomposição EEMD é combinada com a transformada de Hilbert [21], resultando na transformada de Hilbert-Huang (HHT) [22]. Da HHT, derivam-se as amplitudes e frequências instantâneas das IMF's em função do tempo. Por fim, a estimação da F_0 é descrita por $F_0(t) = \{f_i(t), min|f_i(t) - F_r(t)|\}$, onde $f_i(t)$, $(i = 1, 2, 3)$ são as frequências instantâneas das três últimas IMF's para segmentos sonoros da voz, e $F_r(t)$ os valores da frequência de referência obtidos por [23]. Das IMF's analisadas, o valor mais próximo da referência obtida em certo quadro de tempo, é selecionada para a estimativa de *pitch*.

E. SFF

Nesta solução, a frequência fundamental é extraída do sinal, usando uma técnica denominada filtragem de frequência única (SFF), resultando em mútiplos envelopes do sinal de voz filtrado. Cada envelope é encontrado em função de valores de frequência f_k, de modo que 300 Hz $\leq f_k \leq$ 1200 Hz. Os instantes de silêncio do sinal de entrada são detectados a partir das menores amplitudes do envelope. Para decidir qual será adotado na extração da F_0, encontra-se o envelope com maior energia $E_k(v)$ para o *frame* de índice v. Desta forma, $F_D(v) = argmax(E_k(v))$, onde $F_D(v)$ consiste na frequência dominante. Esta frequência define o envelope que possui o maior valor de SNR para um quadro do sinal de voz. Para extração da F_0, a função autocorrelação é aplicada ao envelope. Os picos da função autocorrelação localizados fora do intervalo $\tau_{min} \leq \tau_0 \leq \tau_{max}$ são desconsiderados. Desta forma é estabelecida uma faixa de frequências $[F_{min}, F_{max}]$, onde encontram-se os valores de F_0 dos sinais de voz [6].

F. HHT-Amp

Neste método [7], a transformada de Hilbert-Huang [22] é aplicada nos segmentos sonoros dos sinais. Diferentemente das técnicas que estimam a F_0 por meio das frequências instantâneas [5], esta solução propõe a utilização das amplitudes instantâneas. Os passos a seguir descrevem o método:

1) Aplicação do EEMD [17] na decomposição de sequências de amostras $x_q(t)$ em IMF's e um residual $r_q(t)$, $x_q(t) = \sum_{k=1}^{K} \text{IMF}_{k,q}(t) + r_q(t)$.
2) Cálculo das amplitudes instantâneas $a_{k,q}(t) = |Z_{k,q}(t)|, k = 1, \ldots, K$, dos sinais analíticos, definidos como $Z_{k,q}(t) = \text{IMF}_{k,q}(t) + j\,H\{\text{IMF}_{k,q}(t)\}$, onde $H\{\text{IMF}_{k,q}(t)\}$ é a transformada de Hilbert da $\text{IMF}_{k,q}(t)$.
3) Cálculo da função autocorrelação $r_{k,q}(\tau) = \sum_t a_k(t)\, a_k(t + \tau)$ das amplitudes instantâneas $a_{k,q}(t), k = 1, \ldots, K$.
4) Para cada conjunto k, τ_0 é o menor valor de τ que corresponde a um pico da função autocorrelação, sendo que $\tau_{min} \leq \tau_0 \leq \tau_{max}$. Estes limites são definidos de acordo com a faixa $[F_{min}, F_{max}]$ de possíveis valores de F_0. Desta forma, os candidatos a estimativas de F_0 são definidos como τ_0/f_s, onde f_s é a taxa de amostragem.
5) Aplicação do critério de decisão definido em [7] para selecionar o melhor candidato a *pitch*. Finalmente, a F_0 estimada é obtida por $\hat{F}_0 = 1/\hat{T}_0$.

IV. Resultados e Discussão

Esta seção apresenta os resultados de acurácia dos estimadores de F0 referentes aos seis métodos avaliados neste trabalho. As medidas de erro adotadas no estudo comparativo, incluem a GE, utilizada pelos métodos competitivos [3][4][5][6][7], e também a medida MAE. A MAE permite uma maior percepção do erro, visto que é uma distância absoluta (em Hz) entre a F_0 de referência e a estimada. Os sinais de voz da base CSTR [10] foram utilizados pois a mesma disponibiliza os valores da F_0 de referência dos segmentos sonoros. Ela é composta por 100 locuções (50 masculinas e 50 femininas) com uma taxa de amostragem de 8 kHz. Os sinais foram reverberados por meio de uma convolução com a RIR das salas apresentadas na Tabela I. Ela descreve as condições dos sinais, adotadas neste trabalho para reverberar os sinais de voz. Por fim, foram adicionados quatro ruídos aos sinais reverberados: *Babble* da base RSG-10 [24]; e *Car*, *Helicopter* e *Traffic* da base FreeSfx com valores de SNR: -5 dB, 0 dB e 5 dB.

TABELA I

REVERBERAÇÕES SELECIONADAS E SUAS CARACTERÍSTICAS.

Reverberação	RT_{60} (s)	d_{fm} (m)	DRR (dB)
Stairway (AIR)	1,1	1,0	-9,10
LASP1	0,7	1,2	-1,74
LASP2	0,8	1,5	-2,77

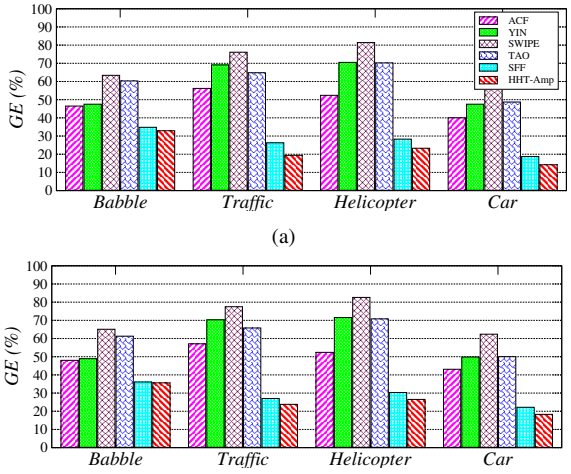

Fig. 2. Resultados de GE dos sinais reverberantes-ruidosos das salas LASP1 (a) e LASP2 (b).

TABELA II

GE (%) DOS SINAIS REVERBERANTES-RUIDOSOS PARA A SALA *Stairway* PARA DIFERENTES VALORES DE SNR.

Ruído	SNR	ACF	YIN	SWIPE	TAO	SFF	HHT-Amp
Babble $\overline{\text{INS}}$=16,6	-5 dB	92,8	76,0	93,2	84,5	68,9	**54,9**
	0 dB	74,6	49,2	73,1	76,0	50,4	**42,5**
	5 dB	**24,4**	29,4	43,3	63,9	32,5	31,2
Média		63,9	51,5	69,9	74,8	50,6	**42,9**
Traffic $\overline{\text{INS}}$=10,9	-5 dB	84,7	88,0	89,9	84,6	62,7	**33,1**
	0 dB	61,4	70,9	74,8	74,1	40,5	**25,6**
	5 dB	43,1	47,3	57,2	62,6	27,3	**22,5**
Média		63,1	68,7	74,0	73,8	43,5	**27,1**
Helicopter $\overline{\text{INS}}$=1,3	-5 dB	96,3	92,0	98,2	88,0	75,0	**41,1**
	0 dB	62,5	72,5	83,8	80,7	54,0	**31,3**
	5 dB	31,7	44,1	60,8	67,4	32,0	**23,7**
Média		63,5	69,5	80,9	78,7	53,7	**32,0**
Car $\overline{\text{INS}}$=1,1	-5 dB	60,1	61,9	74,8	72,4	39,5	**21,6**
	0 dB	39,0	33,8	56,2	54,6	27,7	**20,3**
	5 dB	21,0	21,2	40,1	44,2	21,7	**19,4**
Média		40,0	39,0	57,0	57,1	29,6	**20,4**
Média Total		57,6	57,2	70,5	71,1	44,4	**30,6**

Em todas as técnicas, as frequências de referência da base CSTR foram ajustadas, onde cada instante de tempo da base foi deslocado em +20 ms. Este passo é importante no aumento da precisão da estimação, pois constatou-se que há um atraso temporal imposto pelo efeito da reverberação.

A. Resultados de GE

A medida de erro GE é descrita como $(E_{F0}/N_{F0}) * 100$, onde N_{F0} é a quantidade total de estimações e E_{F0} a parcela de estimativas que satifazem a condição $|(\hat{F}_0/F_0) - 1| > 0,2$. Este resultado reflete a porcentagem dos quadros de um sinal cuja estimativa $\hat{F}_0$ possui uma diferença maior do que 20% do valor da F_0 de referência. A Tabela II apresenta os resultados de GE obtidos para sinais reverberantes-ruidosos da sala *Stairway* da Base AIR [11], de acordo com o INS médio ($\overline{\text{INS}}$) dos ruídos acústicos. Note que o método HHT-Amp apresenta os menores resultados de GE para a maioria dos casos (11 casos de 12). A média total mostra que o HHT-Amp [7] obteve um GE de 30,6 p.p. (pontos percentuais), 13,8 p.p. menor que o método SFF, que obteve a segunda melhor média. Considerando o valor $\overline{\text{INS}}$ dos ruídos, veja que o erro de estimação da F_0 diminui com o decéscimo destes valores. A Figura 2 (a) e (b) apresenta os valores do GE das salas LASP1 e LASP2, respectivamente. Para cada ruído, é apresentada uma média dos três valores de SNR. Em ambas as salas, o HHT-Amp obteve menores valores de erro para praticamente todos os ruídos, com resultado semelhante ao SFF [6] somente para o ruído *Babble*. Note nos valores de GE para a sala *Stairway* da Tabela II que o método TAO resultou no maior erro médio total, assim como apontado nos resultados para sinais ruidosos apresentados em [7]. Por outro lado, o GE das salas LASP aponta que o estimador SWIPE possui menor acurácia.

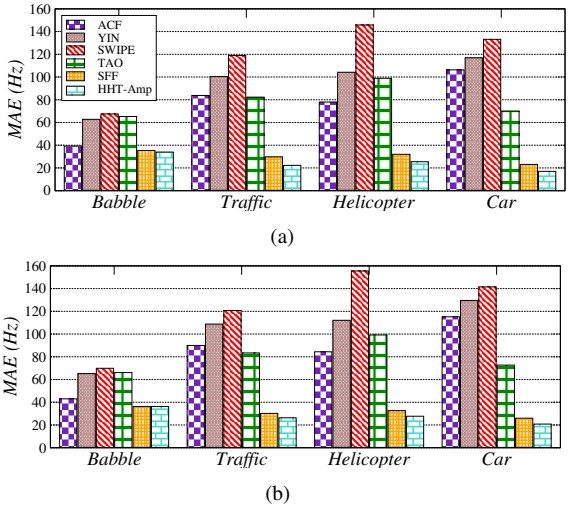

Fig. 3. Resultados de MAE dos sinais reverberantes-ruidosos das salas LASP1 (a) e LASP2 (b).

TABELA III

MAE (HZ) DOS SINAIS REVERBERANTES-RUIDOSOS PARA A SALA *Stairway* PARA DIFERENTES VALORES DE SNR.

Ruído	SNR	ACF	YIN	SWIPE	TAO	SFF	HHT-Amp
Babble $\overline{\text{INS}}$=16,6	-5 dB	60,7	95,3	60,3	99,3	74,4	**53,9**
	0 dB	**42,1**	59,4	42,1	87,8	50,5	42,5
	5 dB	**26,8**	37,4	29,6	75,3	34,1	32,5
Média		43,2	64,0	44,0	87,5	53,0	**43,0**
Traffic $\overline{\text{INS}}$=10,9	-5 dB	113,5	119,2	105,0	100,3	65,1	**34,5**
	0 dB	72,9	92,0	72,3	84,9	40,7	**27,7**
	5 dB	46,2	59,0	49,7	72,1	28,2	**25,5**
Média		77,5	90,1	75,7	85,8	44,7	**29,2**
Helicopter $\overline{\text{INS}}$=1,3	-5 dB	147,2	128,9	114,6	107,2	85,8	**40,7**
	0 dB	72,0	94,4	60,8	99,2	54,9	**32,2**
	5 dB	37,5	59,5	50,3	79,4	31,9	**25,9**
Média		85,6	94,3	75,2	95,3	57,5	**32,9**
Car $\overline{\text{INS}}$=1,1	-5 dB	143,0	93,5	63,6	82,1	42,4	**24,4**
	0 dB	85,7	54,4	56,1	61,8	29,8	**22,2**
	5 dB	33,7	34,7	36,6	51,9	24,3	**22,3**
Média		87,5	60,9	52,1	65,3	32,2	**32,0**
Média Total		73,4	77,3	61,8	83,4	46,8	**32,0**

B. Resultados de MAE

O erro médio absoluto (MAE) em Hz é definido pela seguinte equação:

$$MAE = \left(\sum_{i=1}^{n} \left| \hat{F}_0(i) - F_0(i) \right| \right) / n \qquad (4)$$

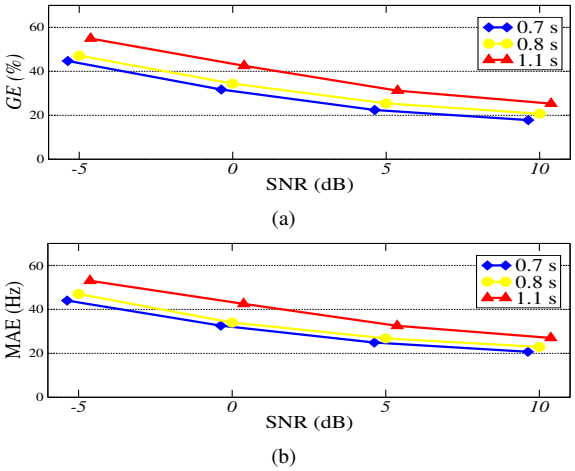

Fig. 4. Comportamento do (a) GE e (b) MAE para diferentes RT_{60} afetados pelo ruído *Babble*.

onde $\hat{F}_0(i)$ é a estimação e $F_0(i)$ a referência. A Figura 3 ilustra os resultados médios de MAE dos sinais reverberantes-ruidosos para as salas LASP. A Tabela III refere-se aos valores de MAE para sala *Stairway* ($RT_{60} = 1,1s$). Pode-se também verificar o efeito dos valores de $\overline{\text{INS}}$ no resultados de MAE. Note novamente que o método HHT-Amp apresentou os menores erros em praticamente todos os cenários, com erro médio total de 32,0 Hz, contra 46,8 Hz do estimador SFF.

A Figura 4 apresenta o comportamento das medidas de erro dos sinais com diferentes tempos de reverberação. As curvas representam a variação do SNR para GE (a) e MAE (b) com o ruído *Babble* para RT_{60} = 0,7 s, 0,8 s e 1,1 s. Note que o aumento dos valores de RT_{60} provocou um aumento dos valores de erro evidenciados pelas medidas GE e MAE. Isso ocorre, pois a elevação deste tempo é reflexo da amplificação das reflexões tardias da reverberação, responsáveis pela distorção de sinais de voz. Estas, por sua vez, mascaram os componentes temporais e espectrais dos sinas de voz [8] [9]. Considerando todos os cenários avaliados, o estimador HHT-Amp apontou os melhores resultados médios de acurácia para a estimação da F_0, quando comparado com os métodos competitivos.

V. Conclusões

Este artigo apresentou um estudo da acurácia de seis estimadores de frequência fundamental em sinais reverberantes-ruidosos. As medidas de erro de estimação GE e MAE foram adotadas na análise da acurácia dos métodos. Nos experimentos, foram utilizadas três salas com diferentes valores de RT_{60} e quatro ruídos com diferentes valores de INS e SNR. Os resultados mostraram que o aumento dos valores de RT_{60} e INS impactam o erro de estimação da F_0. As medidas de erro confirmaram que a extração do contorno da F_0 de sinais reverberantes-ruidosos com o método HHT-Amp alcançou a maior acurácia.

Referências

[1] J. Zhaozhang e D. L. Wang, " HMM-Based Multipitch Tracking for Noisy and Reverberant Speech," *IEEE Trans. Audio, Speech and Lang. Process.*, v. 19, no. 5, pp. 1091-1102, 2011.

[2] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech Signal Process.*, v. 25, pp. 24-33, Feb. 1977.

[3] A. de Cheveigné e H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, v. 111, no. 4, pp. 1917-1930, Apr. 2002.

[4] A. Camacho e J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, v. 124, no. 3, pp. 1638-1652, Sep. 2008.

[5] H. Hong, Z. Zhao, X. Wang, e Z. Tao, "Detection of dynamic structures of speech fundamental frequency in tonal languages," *IEEE Signal Process. Lett.*, v. 17, no. 10, pp. 843-846, Oct. 2010.

[6] G. Aneeja and B. Yegnanarayana, "Extraction of fundamental frequency from degraded speech using temporal envelopes at high SNR frequencies," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, v. 25, no. 4, pp. 829-838, Apr. 2017.

[7] L. Zão e R. Coelho, "On the estimation of fundamental frequency from nonstationary noisy speech signals based on the Hilbert-Huang Transform," *IEEE Signal Process. Lett.*, v. 25, no. 2 pp. 248-252, Feb. 2018.

[8] R. J. Bolt, e A. D MacDonald, "Theory of speech masking by reverberation," *The Journal of the Acoustical Society of America.*, v. 21, no. 6, pp. 577-580, 1949.

[9] P. Assmann e Q. Summerfield, " The perception of speech under adverse conditions," *Speech processing in the auditory system*, pp. 231-308, Springer, 2004.

[10] P. C. Bagshaw, S. M. Hiller e M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," *Proc. EUROSPEECH?93.*, pp. 1003-1006, Sep. 1993.

[11] M. Jeub, M. Schaefer, e P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *2009 16th International Conference on Digital Signal Processing.*, pp. 1-5, Jul. 2009.

[12] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, e J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, v. 58, no. 7, pp. 3459-3470, Jul. 2010.

[13] R. Tavares e R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Processing Letters*, v. 23, pp. 6-10, Jan. 2016.

[14] L. Wang, D. Zheng e F. Chen, "Understanding low-pass-filtered Mandarim sentences: Effects of fundamental frequency contour and single-channel noise suppression," *Acoustical Society of America*, v. 143 no. 3, pp. 141-145, Mar. 2018.

[15] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust. Speech and Signal Processing*, v. ASSP-22, pp. 353-362, Oct. 1974.

[16] D. D. O'Shaughnessy, "Speech communications - human and machine," *IEEE*, 2 ed., 2000.

[17] Z. Wu e N. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, v. 1, no. 1, pp. 1-41, 2009.

[18] N. Huang et al., "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A, Math. Phys. Eng. Sci.*, v. 454, no. 1971, pp. 903-995, 1998.

[19] P. Flandrin, G. Rilling e P. Gonçalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, v. 11, no. 2, pp. 112-114, Feb. 2004.

[20] M. Torres, M. Colominas, G. Schlotthauer e P. Flandrin, " A complete ensemble empirical mode decomposition with adaptive noise," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 4144-4147, May. 2011.

[21] Huang, et al., " The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond.*, pp. 903-995, 1998.

[22] H. Huang e J. Pan, " Speech pitch determination based on Hilbert-Huang transform," *Signal Process.*, v. 86, no. 4 pp. 792-803, 2006.

[23] S. A. Zahorian e H. Hu, " A spectral/temporal method for robust fundamental frequency tracking," *J. Acoust. Soc. Amer.*, v. 123, no. 6, pp. 4559-4571, 2008.

[24] H. J. Steeneken e F. W. Geurtsen, "Description of the RSG-10 noise-database," *report IZF*, v. 3, 1988.

[25] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, v. 5, no. 9/10, pp. 341-345, 2001.