Segregação de Voz Usando Mascaramento INM sobre o Banco de Filtros Gammatone

Christian Arcos Gordillo, Marley Vellasco e Abraham Alcaim

***Resumo*— Este artigo apresenta uma abordagem inovadora, que emprega uma máscara de vizinhança ideal (INM), que tem a capacidade de usar eficientemente os *Local Binary Pattern* (LBP) para indicar quais unidades Tempo-Frequência da voz corrompida são dominadas pelo ruído. Resultados experimentais obtidos com um reconhecedor de voz baseado em DNN em ambientes ruidosos demonstram que a técnica proposta alcança melhorias significativas em termos de taxa de erro de palavra. Confirmando a superioridade do esquema proposto em comparação com os algoritmos de mascaramento tradicionais IBM e IRM.**

***Palavras-Chave*— Realce, máscara, vizinhança, Tempo-Frequência, ruído.**

***Abstract*— This paper presents an innovative approach that employs an ideal neighbourhood mask (INM) that has the ability to efficiently use Local Binary Pattern (LBP) to indicate which Time-Frequency units of the corrupted voice are dominated by noise. Experimental results obtained with a DNN based voice recogniser in noisy environments demonstrate that the proposed technique achieves significant improvements in terms of word error rate corroborating the superiority of the proposed scheme in comparison with the traditional masking algorithms IBM and IRM.**

***Keywords*— Enhancement, mask, Neighborhood, Time-Frequency, noise.**

I. Introdução

No campo do reconhecimento automático de voz (RAV), os sistemas atuais têm conseguido passar do reconhecimento de palavras isoladas próprias de um vocabulário limitado a situações de reconhecimento de voz contínua onde o vocabulário é de grande dimensão, atingindo taxas de precisão de palavra até 95% para a língua inglesa em condições de laboratório (caso do sistema RAV do google) de acordo com o relatório anual de tendências da internet da Mary Meeker de março de 2017 [1]. No entanto, esses sistemas são sensíveis à alteração das condições acústicas, o que pode causar degradação significativa do desempenho, já que o sinal sofre distorções que não têm sido contempladas na etapa de treinamento.

Uma das principais causas da queda acentuada de desempenho é a existência de vários tipos de ruído de fundo, que fazem com que o desempenho dos sistemas em aplicações do mundo real tais como, serviços de comunicações de voz sem fio, dispositivos de aparelhos auditivos digitais, telefonia móvel com mãos-livres, transmissão de voz, entre outros, esteja longe de ser satisfatório, já que o ruído degrada os sistemas até níveis em que seu uso se torna definitivamente inaceitável.

A existência de ruído é inevitável em aplicações do mundo real. Segundo a teoria da comunicação, o ruído é um som inarticulado ou distúrbio anômalo que causa uma sensação de audição desagradável no sistema auditivo humano e que gera uma interferência não desejada no processo comunicativo, distorcendo a informação transmitida pela onda acústica portadora de informação, dificultando sua correta percepção e evitando que a informação chegue de forma clara.

Estas alterações acústicas limitam significativamente o funcionamento dos sistemas RAV em ambientes reais, já que se superpõem ao sinal de voz, mascarando e alterando as suas características. É por isto que na atualidade tem se acrescentado métodos de robustez nos sistemas RAV que não exigem uma carga computacional excessiva e que melhoram o desempenho dos sistemas evitando o descasamento entre as condições de treinamento e as de reconhecimento. Estes métodos constituem uma área de pesquisa fundamental no processamento de voz conhecida como robustez, e estão divididas em 3 categorias principais, segundo seu enfoque principal [2]; (i) realce de voz; (ii) compensação de atributos e (iii) adaptação de modelos.

Um conjunto de técnicas que têm mostrado bons resultados nos métodos de realce de voz em ambientes ruidosos nas últimas décadas são derivados a partir de estudos sobre a percepção auditiva humana. Esses métodos baseados na estimativa de máscara de segregação [3][4], utilizam as propriedades do sistema auditivo humano para calcular o limiar de mascaramento do ruído [5] a fim de selecionar da matriz de mascaramento quais unidades contêm energia predominante de voz considerando-as como úteis, e o resto como energia de voz dominada pelo ruído e portanto não contendo informação útil [6].

Wang *et al.* em [7] introduziram uma abordagem de segregação de voz, que é a tarefa de separar a fala do ruído de fundo, essa abordagem mostrou uma promessa considerável para melhorar os resultados do realce da fala. Ela considera que os sons que atingem o ouvido estão sujeitos a um processo chamado Análise da Cena Auditiva (ASA para o acrônimo em inglês) [8]. Com base nesse processo foi proposta a *Ideal Binary Mask* (IBM) [9] que foi sugerida como objetivo computacional dos algoritmos *Computational auditory scene analysis* (CASA) [10], cujo principal objetivo é fazer a separação das distintas fontes de som que compõem a entrada acústica. Continuando com a mesma linha de mascaramento discreto, Srinivasan *et al.* propuseram uma outra alternativa de mascaramento chamada Ideal Ratio Mask (IRM) [11]. Os

Christian Arcos Gordillo, Marley Vellasco e Abraham Alcaim, Centro de Estudos em Telecomunicações da PUC-Rio (CETUC), Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil, E-mails: christian@cetuc.puc-rio.br, marley@ele.puc-rio.br, aalcaim@gmail.com. Este trabalho foi parcialmente financiado pelo CNPq.

autores mostraram que em uma banda de frequência estreita, existe uma relação sistemática entre SNR *apriori* e valores de *Interaural Time Differences*(ITD) e *Interaural Intensity Differences*(IID), o que os motivou a estimar uma razão ideal de mascaramento usando estatísticas coletadas para ITD e IID em cada faixa de frequência individual. Matematicamente a IRM é estreitamente ligada ao ganho do filtro de Wiener [12].

Na literatura podem ser encontradas inúmeras variações referentes a essas técnicas de estimação de máscaras aplicadas tanto aos coeficientes de potência espectral em escala logarítmica quanto aos coeficientes cepstrais, proporcionando em ambos os casos melhorias similares nos resultados de reconhecimento [13] [14].

Recentemente, em [15], foi proposto um novo esquema inovador que emprega uma máscara de vizinhança ideal ou *Ideal Neighborhood Mask* (INM) que tem a capacidade de usar eficientemente os *Local Binary Pattern* (LBP) que indicarão quais unidades Tempo-Frequência (T-F) da voz corrompida são dominadas pelo ruído. Esta técnica aplicada sobre o sinal de voz melhorou o desempenho do sistema de reconhecimento minimizando o desajuste causado pelo ruído, preservando as componentes importantes da voz através da codificação das unidades T-F.

Neste artigo introduzimos uma nova abordagem baseada no realce de voz, que melhora a inteligibilidade e qualidade do sinal a partir da aplicação de uma estimativa de máscara INM sobre o banco de filtros gammatone. Será descrito o novo algoritmo INM e apresentada uma análise comparativa do esquema proposto com as máscaras tradicionais *Ideal Binary Mask* (IBM) e *Ideal Ratio Mask* (IRM).

Este artigo estrutura-se da seguinte forma. A Seção II inclui uma breve revisão dos métodos tradicionais de mascaramento IBM e IRM. A Seção III apresenta o algoritmo de mascaramento INM proposto neste trabalho. O procedimento experimental e discussão dos resultados são apresentados na Seção IV. Finalmente, as principais conclusões são resumidas na Seção V.

II. As Máscaras IBM e IRM

A. Ideal Binary Mask (IBM)

A *Ideal Binary Mask* (IBM) foi proposta como objetivo principal dos sistemas CASA [16] a fim de separar as diferentes fontes de som que compõem a entrada acústica. Tipicamente, nesses sistemas o sinal de entrada é transformado em uma série de segmentos chamados unidade tempo-frequência (T-F), onde cada unidade pertencente a um determinado tempo e a uma frequência específica. Na máscara IBM é atribuído o valor 1 se a energia da voz excede a energia do ruído e 0 caso contrário. Matematicamente, a IBM é definida como

$$IBM_{(t,f)} = \begin{cases} 1 & se \quad SNR_{(t,f)} > LC \\ 0 & c.c \end{cases} \tag{1}$$

onde LC é o critério local ou limiar determinado empiricamente para cada técnica usando conjuntos de validação (geralmente toma faixas de valores de $LC \in [-6, 6]$ [17][18]) e $SNR_{(t,f)}$ é a relação sinal-ruído instantânea para cada unidade T-F dada por

$$SNR_{(t,f)} = 10log\frac{X_{(t,f)}}{R_{(t,f)}} \tag{2}$$

onde $X_{(t,f)}$ e $R_{(t,f)}$ são a energia instantânea do sinal limpo e do ruído respectivamente em um tempo t e uma frequência f.

A IBM tem sido amplamente utilizada na literatura [17] [19], mostrando que sob certas restrições, ela é uma ótima máscara binária em termos de relação sinal ruído (SNR)[20].

B. Ideal Ratio Mask (IRM)

A aplicação de uma máscara binária aos espectros de voz corrompida pode afetar a qualidade da voz nesse processo de remoção de componentes espectrais, ou seja, quando às unidades T-F é atribuído 0, esse procedimento pode potencialmente produzir o *ruído musical*. A redução a zero das unidades T-F (ou remoção de componentes espectrais) pode criar picos pequenos e isolados no espectro que ocorrem em locais de frequência aleatórias em cada quadro. Convertidos ao domínio do tempo, esses picos são semelhantes a tons com frequências que mudam aleatoriamente de quadro para quadro e produzem o indesejável *ruído musical*.

Para resolver este problema, a *Ideal Ratio Mask* (IRM) foi proposta em [21] com o objetivo de suavizar as unidades T-F ao invés de removê-las. A IRM proporciona um melhor desempenho porque está intimamente relacionada com o filtro de Wiener[21], onde um valor de SNR alto indica baixa atenuação da energia das unidades T-F, enquanto um valor de SNR baixo indica alta atenuação, suavizando todas as unidades T-F em vez de removê-las como o caso da IBM. A IRM é definida por

$$IRM_{(t,f)} = \frac{10^{(SNR_{(t,f)}/10)}}{10^{(SNR_{(t,f)}/10)} + 1} \tag{3}$$

onde o $SNR_{(t,f)}$ é a relação sinal-ruído instantânea para cada unidade T-F (equação 2).

III. Máscara INM Sobre o Banco de Filtros e Estimador IMCRA

O mascaramento INM possui a capacidade de usar eficientemente os LBPs para estimar uma máscara ideal que identifica quais unidades T-F do sinal de voz corrompido são dominadas pelo ruído [15]. Nesta seção se faz uma estimativa da máscara sobre o banco de filtros auditivos com o objetivo de diminuir as taxas de erro de palavra e comparar seu desempenho com as máscaras tradicionais IBM e IRM. Esta nova proposta é representada em diagrama de blocos na Fig. 1.

As unidade T-F são extraídas dos bancos de filtros através de um procedimento de filtragem e janelamento chamado *cochleogram*. Através deste procedimento é possível conhecer a SNR ideal das unidades T-F quando o sinal limpo e o ruído são conhecidos. No entanto, o que se busca nesta seção é utilizar um algoritmo que estime a SNR de cada unidade T-F do sinal corrompido a fim de criar a máscara que possa agir em condições reais.

Após se obter a SNR das unidades T-F, é realizado o procedimento de mascaramento INM considerada ideal, já que

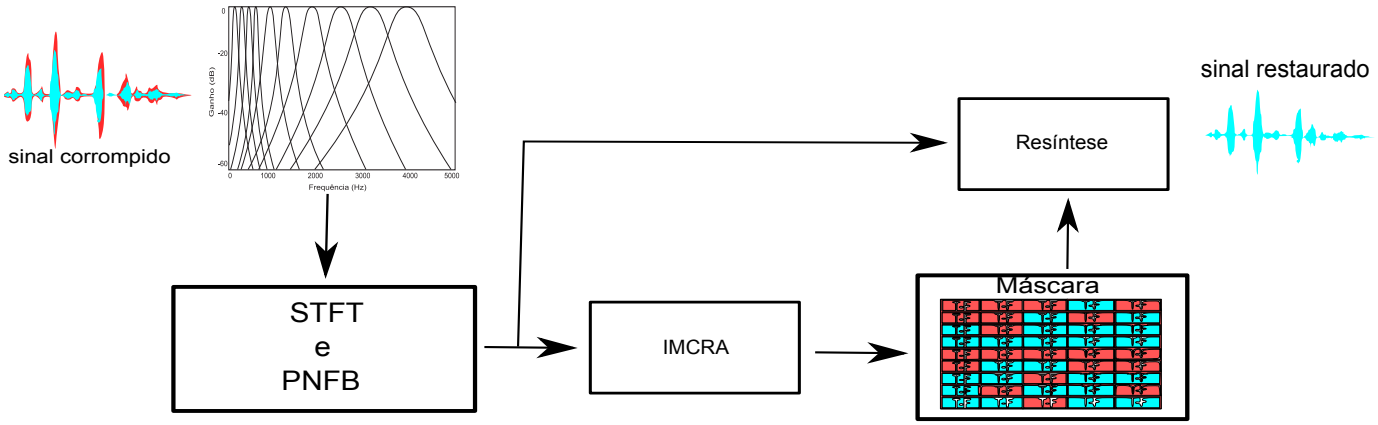

Fig. 1. Diagrama em blocos do sistema proposto.

se dá uma separação perfeita entre voz e ruído, devido ao fato de conhecer a *priori* os componentes do sinal. O INM é descrito em detalhes em [15]. Este tipo de mascaramento é útil para conhecer o nível máximo de confiabilidade que pode chegar a ter a técnica, e por que são utilizadas como *targets* (alvos) em processos de separação supervisionados.

Nesta nova abordagem usamos o algoritmo IMCRA [14] a fim de estimar a SNR *a priori* do sinal corrompido de cada unidade T-F. A estimação da SNR é feita a partir de cada banda de frequência dos filtros gammatone. O estimador IMCRA é dividido em duas etapas, 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. Detalhes sobre o algoritmo IMCRA podem ser encontrados em [14]. Neste caso, a máscara INM será denominada ENM e matematicamente, é representada pela seguinte equação

$$INM[k] = \begin{cases} \dfrac{\gamma[k-1] + 2\gamma[k] + \gamma[k+1]}{4} & para LBP[k] = 0 \\ \sqrt{\dfrac{\gamma[k]}{1+\gamma[k]}} & para LBP[k] = 1 \; or \; 2 \\ 1 & para LBP[k] = 3 \end{cases} \tag{4}$$

onde $\gamma[k]$ é a SNR estimada com o algoritmo IMCRA na *k-ésima* unidade T-F. O algoritmo 1 resume o procedimento realizado para mascarar o sinal através do método ENM no caso de $p = 2$ vizinhos.

Finalmente, é aplicada uma variante na reconstrução do sinal depois de aplicar o mascaramento. São aproveitadas as características dos atributos *Power normalized filter bank* (PNFB) (que são a etapa prévia aos coeficientes cepstrais PNCC [22], ou seja, a etapa anterior à aplicação da DCT) que simulam o sistema auditivo humano. Os PNFB surgiram como um conjunto de características para o reconhecimento de voz que são mais robustos em relação à variabilidade acústica, e não apresentam perda de desempenho quando o sinal de fala não é degradado. Os PNFB são obtidos a partir de uma matriz de unidades T-F com as mesmas configurações de decomposição do *cochleogram*, seguindo o seguinte procedimento:

Algoritmo 1 Cálculo da máscara ENM para p=2

Input: Sinal de entrada.
Output: $INM[k]$.

1: Aplicar filtro pré-enfase
2: Passar o sinal por um banco de filtros gammatone de 64 canais, com frequências igualmente espaçadas entre 50Hz e 800Hz.
3: Segmentar a voz em quadros de 20-ms com 10-ms de superposição entre quadros.
4: Calcular $SNR[k]$ *a priori*, ou seja, $\gamma[k]$ de cada segmento usando o algoritmo IMCRA [14] .
5: **while** $\gamma[k]$ True (para $k = [0 : N-1]$) **do**
6: Calcular códigos LBP na janela de análise deslisante W de comprimento p.
7: **if** $\gamma[i \pm 1] \geq \gamma[i]$ **then**
8: **return** $W_p = 1$; incrementar p
9: **else**
10: **return** $W_p = 0$; incrementar p
11: **end if**
12: **if** $p = 2$ **then**
13: **return** $\gamma[k] = 2^{W_p} + 2^{W_{p+1}}$
14: **end if**
15: **end while**
16: Separar todos os segmentos com diferentes valores LBP
17: Calcular INM[k] de acordo com (4)

- O sinal de entrada é passado por um filtro de pré-enfase a fim aumentar a magnitude de algumas frequências, para compensar os efeitos dos pulsos glotais e a ressonância dos lábios.
- O sinal filtrado é transformado ao domínio da frequência usando a *short-time Fourier transform* (STFT) com quadros de 20 ms superpostos com 10 ms.
- A potência espectral em N bandas de análise é obtida ponderando a magnitude quadrada das saídas da STFT pela resposta em frequência associada com as N bandas do banco de filtros gammatone.
- É realizada uma estimativa e remoção de ruído em cada banda através de uma série de operações não-lineares,

variáveis no tempo, que são executadas usando uma análise temporal de longo tempo que faz subtração de ruído, acrescentando um grau de robustez em relação ao coeficientes tradicionais.

- Finalmente, a energia de cada banda é calculada aplicando uma não-linearidade denominada *power-law* com expoente $1/15$

Informação detalhada do procedimento acima exposto encontra-se em [22]. Multiplicando a matriz que se obteve com a decomposição em PNFB pela máscara INM obtida na etapa anterior busca-se dar maior robustez ao sinal já que os coeficientes PNFB incluem uma etapa de remoção de ruído chamada integração temporal para a análise de ambiente. Esse procedimento também é feito para as máscaras IBM e IRM.

IV. Resultados e Discussões

Os resultados que serão mostrados a seguir referem-se ao mascaramento ENM sobre o banco de filtros gammatone tomando como métrica as taxas de erro de palavras (WER) que são fornecidas por um sistema de reconhecimento de voz híbrido, usando uma rede neural profunda e os modelos ocultos de markov (DNN-HMM). O sistema é implementado usando o kit de ferramentas de reconhecimento de voz *Kaldi* [23]. Os parâmetros experimentais, incluindo o tipo de atributos, são os mesmos que para experimentos padrão usando a fórmula de Kaldi $s5$. Informação detalhada dos parâmetros experimentais do classificador encontram-se em [24] .

O Modelo acústico foi treinado usando o conjunto de treinamento limpo si84 (7138 sentenças) da AURORA-4 que consiste em sinais gravados usando um microfone Sennheiser e processados usando um filtro P.314. Este conjunto foi dividido em 90% para treinamento e 10% para validação cruzada. Utiliza-se o conjunto de dados Nov92 (330 sentenças) como o conjunto de teste que consiste, na realidade, de 7 subconjuntos de teste: 1 subconjunto limpo e 6 subconjuntos corrompidos por seis diferentes tipos de ruído *babble, airport, restaurant, street, car, e train* variando de 0dB a 15dB. O modelo de linguagem utiliza os trigramas fornecidos nas tarefas da WSJ. Finalmente, os alinhamentos forçados foram gerados a partir da fórmula tri4b de Kaldi.

A Tabela I mostra os resultados das taxas de erro de palavra (WER) do algoritmo de mascaramento proposto INM em comparação com os métodos tradicionais IBM e IRM em condições ideais (oráculo), onde cada método tem seu desempenho calculado como a média sobre as SNR de 0, 5, 10 e 15 dB. Esta condição nos mostra o limiar máximo até onde as técnicas podem realçar o sinal, já que para estes testes assume-se que o sinal de voz e o ruído são conhecidos. Nos outros experimentos é empregada uma estimativa de ruído através do algoritmo IMCRA [14] a fim de estimar a SNR do sinal corrompido simulando, portanto, condições reais (neste caso identificaremos os métodos como EBM, ERM e ENM).

O ponto de partida dos experimentos é mostrar como os resultados obtidos sem qualquer técnica de mascaramento (linha Noisy) apresentam o pior desempenho, mostrando como o ruído afeta severamente o sinal de voz. Eles serão tomados como referência para os testes posteriores, a fim de verificar como as técnicas de mascaramento melhoram a qualidade do sinal de voz. Pode-se ver claramente a tendência do WER ser consideravelmente maior quanto mais problemático for o tipo de ruído adicionado.

Dos resultados de desempenho em termos de WER mostrados na Tabela I pode-se ver como o mascaramento ENM baseado em códigos LBP fornece melhores resultados em todos os ambientes, melhorando o desempenho dos sistemas RAV em comparação com os métodos tradicionais. Os resultados mostram a importância de utilizar esse mascaramento sobre os bancos de filtros que simulam o sistema auditivo humano. As taxas de erro médias do ENM em relação ao sistema mais robusto da literatura ERM são reduzidas em média de $45,87\%$ para $44,62\%$ no ruído *babble*, de $45,67\%$ para $45,02\%$ no ruido *airport*, de $52,07\%$ para $51,96\%$ no ruído *restaurant*, de $36,33\%$ para $35,55\%$ no ruído *street*, de $44,60\%$ para $40,73\%$ no ruído *car*, e de $25,44\%$ para $21,69\%$ no ruído *train*. Por outro lado, pode-se ver a melhora relativa (MR) do mascaramento óraculo (INM) com uma diferença de $3,46\%$ e $14,26\%$ para IRM e IBM respectivamente e para o caso real usando o estimador IMCRA (ENM) uma diferença de $10,75\%$ e $24,27\%$ para ERM e EBM respectivamente. Em resumo, o ENM atinge menores erros de palavra em comparação com a EBM e ERM.

A Fig. 2 apresenta os espectrogramas das diferentes configurações de máscaras : (a) sinal de voz limpa "440c020a" do banco de dados AURORA-4 (b) ruído *babble* (c) sinal de voz corrompido com ruído *babble* de 0dB, (pode se ver como o ruído gera uma grande incompatibilidade entre a fala limpa e a corrompida, levando a graves problemas de inteligibilidade da voz), (d) e (e) mostram a importância das técnicas de máscara IBM e IRM e como elas reduzem a degradação da voz causada pelo ruído. Finalmente, a Fig. 2(e) apresenta o espectrograma com a nossa proposta INM. A partir de uma comparação visual, pode-se observar que a INM preserva informações mais detalhadas do que a IBM e a IRM.

V. Conclusões

Neste artigo, foi apresentada uma nova estrutura de mascaramento para melhorar a qualidade e inteligibilidade da voz corrompida por ruído de alta intensidade. O algoritmo estima uma máscara de vizinhança ideal (INM), que baseia-se na técnica *Local Binary Patterns* (LBP), originalmente empregada em processamento de imagens. A máscara foi aplícada sobre os filtros gammatone. Comparou-se a máscara INM com as técnicas de mascaramento espectral tradicionais da literatura em um ambiente real onde a máscara não depende da condição verdadeira ou ideal de conhecer todos os sinais *a priori* (condição oráculo). Os experimentos realizados com o reconhecedor de voz baseado em DNN revela que, em termos de taxa de erro de palavra, o ENM (INM com estimador IMCRA) é mais eficaz na redução de ruído. O ENM oferece vantagens de desempenho de reconhecimento de voz em relação às outras técnicas de mascaramento da literatura. Finalmente, a principal motivação do método de mascaramento baseado em LBPs foi o fato de que, com os códigos LBP, a energia da unidade central T-F original é codificada com a energia das unidades T-F

TABELA I

RESULTADOS DE RECONHECIMENTO IDEAL E USANDO O ESTIMADOR IMCRA, OBTIDOS PARA O BANCO DE DADOS AURORA-4. TOMANDO-SE A MÉDIA SOBRE AS DIFERENTES CONDIÇÕES DE SNR. MR SIGNIFICA MELHORIA RELATIVA EM RELAÇÃO AO SISTEMA RUIDOSO

system	babble	airport	restaurant	street	car	train	MR.
Noisy	59,24	56,27	58,88	46,18	56,27	34,90	0
IBM	11,18	11,81	14,11	9,88	14,24	8,37	76,02%
IRM	4,85	4,60	5,22	4,36	5,03	4,60	86,82%
INM	**3,38**	**3,39**	**3,51**	**3,32**	**3,36**	**3,39**	**90,28%**
EBM	53,20	52,95	59,23	43,69	53,39	30,16	13,58%
ERM	45,87	45,67	52,07	36,33	44,60	25,44	27,10%
ENM	**44,62**	**45,02**	**51,96**	**35,55**	**40,73**	**21,69**	**37,85%**

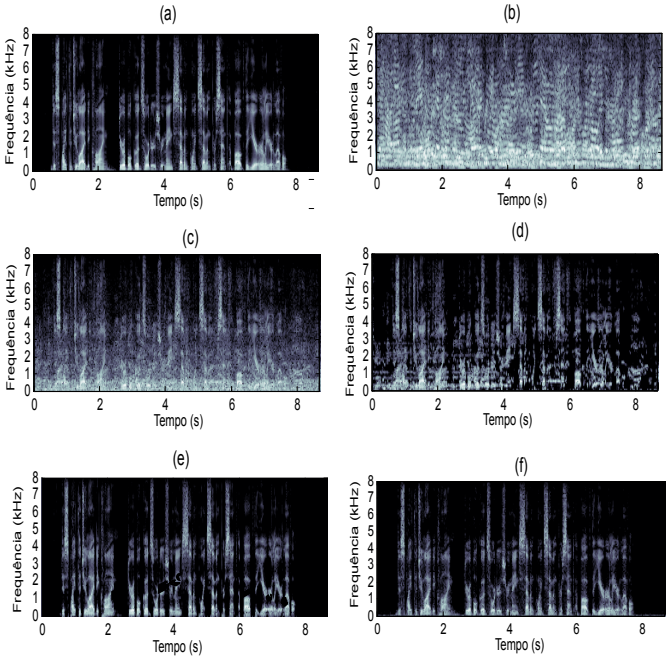

Fig. 2. Exemplo de INM, a figura representa o espectrograma da frase "440c020a" do banco de dados AURORA-4: (a) limpa (b) ruído bable (c) frase "440c020a" corrompida com ruído babble de 0dB, (d) IBM com $LC = 6$ (e) IRM e (f) INM

vizinhas, de modo que a informação da unidade T-F codificada será de um nível mais alto.

AGRADECIMENTOS

Os autores agradecem ao CNPq.

REFERÊNCIAS

[1] M. Meeker, "Kp internet trends 2017 code conference," pp. 48, 2017.

[2] J. Bellegarda, "Statistical techniques for robust asr: review and perspectives," *Fifth European Conference on Speech Communication and Technology*, pp. 33-36, 1997.

[3] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, v. 114, no. 4, pp. 2236-2252, 2003.

[4] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486-1494, 2009.

[5] J. Holmes and N. Sedgwick, "Noise compensation for speech recognition using probabilistic models," *International Conference on Acoustics Speech and Signal Processing, ICASSP*, vol. 11, pp. 741-744, IEEE, 1986.

[6] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech communication*, vol. 34, no. 3, pp. 267-285, 2001.

[7] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270-279, 2013.

[8] A. S. Bregman et al., "Auditory scene analysis" *Cambridge, ma: mit press*, vol. 10, 1990.

[9] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis" *Speech separation by humans and machines*, pp. 181-197, Springer, 2005.

[10] D. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications" *Springer*, 2006.

[11] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition" *Speech Communication,* vol. 48, no. 11, pp. 1486-1501, 2006

[12] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849-1858, 2014.

[13] B. Mellor and A. Varga, "Noise masking in a transform domain," *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, vol. 2, pp. 87-90, IEEE, 1993.

[14] P. C. Loizou, "Speech enhancement: theory and practice," *CRC press, Boca Raton*, London, 2013.

[15] Gordillo, C. A., Vellasco, M., and Alcaim, A. "Ideal neighbourhood mask for speech enhancement," *Electronics Letters*, v. 54, no. 5, pp. 317-318, 2018.

[16] C. Darwin, "Computational auditory scene analysis: Principles, algorithms and applications," *The Journal of the Acoustical Society of America*,vol. 124, no. 1, pp. 13-13, 2008

[17] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007-4018, 2006.

[18] S. Srinivasan and D. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," *Speech Communication*, vol. 52, no. 1, pp. 72-81, 2010.

[19] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 77-93, 2010.

[20] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230-239, 2009.

[21] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," *Blind Source Separation*, pp. 349-368, Springer, 2014.

[22] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," *International Conference on Acoustics Speech and Signal Processing ICASSP*, pp. 4101-4104, 2012.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584, 2011.

[24] C.D. Arcos., "Realce e Reconhecimento de Voz Contínua em Ambientes Adversos," *Tese de Doutorado*, Programa de Engenharía Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, 2018.