Classificação de Voz *versus* Silêncio via Dicionários Redundantes

Raffaello Claser e Ivandro Sanches

***Resumo*— O presente trabalho propõe uma técnica para a solução do problema de se classificar trechos de sinal entre voz e silêncio. Surpreendentemente, a técnica proposta não se baseia na variação de níveis de energia ao longo do sinal, mas nas características fundamentais que determinam a essência de cada uma dessas duas classes de sinais, nominalmente voz e silêncio. Para esse fim, constrói-se um dicionário redundante de funções básicas (átomos) e analisa-se o sinal via Matching Pursuit. Dessa análise, fase de treinamento, obtém-se a distribuição de probabilidade discreta *a priori* de ocorrência do conjunto de átomos para cada classe de interesse, permitindo a discriminação a posteriori entre as classes. O trabalho apresenta resultados promissores com sinais de voz com alta razão sinal-ruído.**

***Palavras-Chave*— *Matching Pursuit*, Histogramas, Detecção de Voz.**

***Abstract*— This paper presents a technique to solve the problem of classifying portions of signal between voice and silence. Surprisingly, the proposed technique is not based on the variation of energy levels over the signal, but on the fundamental characteristics that determine the essence of each of these two classes of signals, namely voice and silence. To this end, a redundant dictionary of basis functions (atoms) is built and the signal is analyzed via Matching Pursuit. From this analysis, the technique training phase, the *a priori* discrete probability distribution of occurrence of a set of atoms for each class of interest is computed, allowing subsequent discrimination between the classes. The paper presents promising results for signals with high signal to noise ratio.**

***Keywords*— *Matching Pursuit*, Histograms, Voice Detection.**

I. INTRODUÇÃO

Trabalhos recentes tais como [1], [2], [3] têm explorado como decompor sinais em termos de funções que estejam mais correlacionadas com o sinal de entrada, do que outras funções base, como por exemplo, seno e cosseno que são tipicamente utilizadas em Transformadas de Fourier. Um dos métodos capazes de realizar esta decomposição, é conhecido como *Matching Pursuit* (MP). Este, por sua vez, através de um dicionário redundante e sobre-completo, composto por funções limitadas no tempo (também conhecidas como átomos), encontra de forma iterativa quais sinais deste dicionário são mais correlacionados com o sinal de entrada $f(t)$ [4]. Esta seleção, visa maximizar o produto escalar entre $f(t)$ e os átomos do dicionário, o que por sua vez, minimiza o erro quadrático no processo de reconstrução. O sinal $f(t)$ pode então ser representado como uma combinação linear de N átomos acrescidos de um erro residual. Na equação (1) é apresentada a expressão de reconstrução de um sinal $f(t)$ baseando-se em átomos g_{γ_n} pertencentes a um dicionário D.

$$f(t) = \sum_{n=0}^{N-1} \langle R^n f, g_{\gamma_n} \rangle g_{\gamma_n} + R^N f \qquad (1)$$

Nesta expressão, o parâmetro $R^n f$ corresponde ao resíduo obtido conforme são acrescidos átomos, multiplicados por seus respectivos pesos, ao processo de decomposição do sinal de entrada; o parâmetro $R^N f$ corresponde ao erro atingido ao término do processo; e por fim, o parâmetro g_{γ_n} corresponde aos átomos de índice γ_n pertencentes ao dicionário D, os quais, além de possuirem norma unitária ($\|g_{\gamma_n}\| = 1$), devem ser limitados assintoticamente por $O(\frac{1}{t^2+1})$.

A cada iteração executada no algoritmo MP, os parâmetros dos átomos selecionados são armazenados e o processo se repete até que a energia do resíduo atinja um valor estipulado ou até que um número de átomos definido previamente, seja encontrado.

Cada átomo é caracterizado por três parâmetros: amplitude (s), frequência de modulação (ξ) e translação (u). Sendo assim, para cada átomo n do dicionário, denota-se $\gamma = (s, u, \xi)$ como sendo a tupla de parâmetros de cada átomo. Essas representações podem ser mais esparsas e flexíveis do que outros tipos de expansões (como por exemplo DCT, FFT, Wavelet, etc), mas a um custo de eficiência, convergência e processamento computacional.

Na Figura 1, é apresentado um exemplo de átomo obtido quando utiliza-se uma função de Gabor (cosseno janelado com uma gaussiana). Neste exemplo, utilizou-se um cosseno de frequência 500 Hz (adotada arbitrariamente).

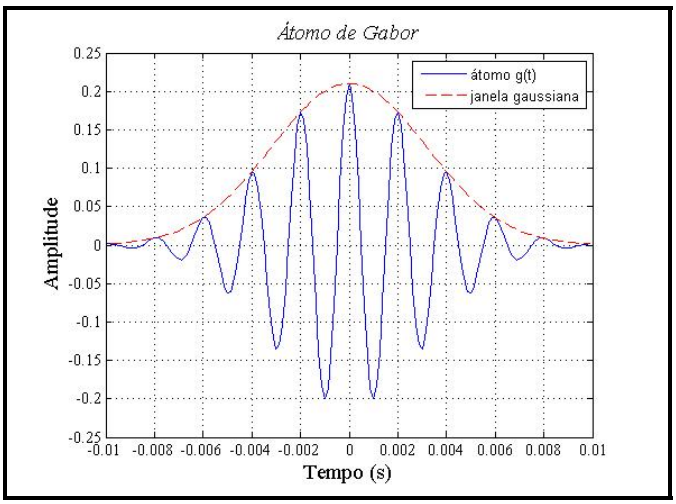

Fig. 1: Exemplo de átomo de Gabor.

Raffaello Claser e Ivandro Sanches. Departamento de Engenharia Elétrica, Centro Universitário da FEI, São Bernardo do Campo, SP, Brasil. E-mails: rclaser1@gmail.com, isanches@fei.edu.br. Este trabalho foi parcialmente financiado pela CAPES com uma bolsa de mestrado pelo período de 1 ano.

Variando-se o sinal da Figura 1 de acordo com a equação (2) e com os parâmetros γ mencionados anteriormente, obtém-se o dicionário de átomos *D*.

$$g_\gamma(t) = \frac{1}{\sqrt{s}}\ g(\frac{t-u}{s})\ e^{i\xi t}. \quad (2)$$

Dessa forma, a técnica MP têm sido largamente utilizada em codificação e reconstrução de sinais, extração de características, análise de sinais, classificação, transformação e visualização [5], [6], [7], [8]. Sendo assim, na seção II, será discutido como aplicar esta técnica junto ao uso de modelagem estatística de forma a permitir a detecção de sinais de voz e silêncio. Na seção III serão discutidos os resultados obtidos e por fim, na seção IV, a conclusão do trabalho.

II. IMPLEMENTAÇÃO

Uma vez que sinais de voz possuem um número elevado de amostras, se faz necessário segmentá-los de forma que as rotinas de processamento sejam feitas para cada trecho e não para o sinal inteiro. Uma vez feito isso, para cada janela de amostras, de forma simplificada para um sinal com alta razão sinal-ruído, podemos classificar cada segmento como sendo voz ou silêncio comparando a energia do trecho em análise com um limiar pré-determinado. Sendo assim, o trecho será caracterizado como voz, uma vez que a energia seja superior ao limiar, e silêncio no caso contrário.

Com base nisso, a seguir será discutida uma outra técnica de detecção que se baseia no uso de representação esparsa e histogramas construídos de forma a caracterizar as diferenças entre essas duas classes de sinais. Estas, por sua vez, são fundamentadas nas probabilidades de ocorrência de um determinado átomo quando o sinal em análise é caracterizado como voz ou como silêncio.

Antes de detalhar o conceito envolvido nesta nova técnica, destacamos a seguir o pseudo-código do algoritmo MP utilizado no algoritmo de detecção via histogramas.

Algoritmo 1 Matching Pursuit (x)

Inicialização das variáveis:

$n \leftarrow 0$
$R_f^n \leftarrow x$
$erro \leftarrow definido\ pelo\ projetista$

Rotina principal:

while $\|R_f^n\| \leq$ *erro* **do**
  for $m = 0$ to $\#D - 1$ **do**
    $a(m) \leftarrow \langle\ g_{\gamma_m}, R_f^n\ \rangle$
  end for
  $k \leftarrow indice\ em\ que\ ocorreu\ max\ do\ vetor\ \|a\|$
  $R_f^{n+1} \leftarrow R_f^n - a(k)g_{\gamma_k}$
  $n \leftarrow n + 1$
end while

Neste algoritmo 1, a variável de entrada x corresponde ao sinal de voz que se deseja decompor; o parâmetro R_f^n corresponde ao erro residual definido entre a variável x e os átomos dos dicionários multiplicados por seus respectivos pesos; o parâmetro a corresponde aos pesos que foram obtidos via produto escalar entre os átomos do dicionário (g_{γ_m}) e o resíduo R_f^n; por fim, o parâmetro $\#D$ corresponde ao número de átomos (ou cardinalidade) do dicionário.

Baseando-se no algoritmo anterior, elaborou-se uma rotina de processamento cuja saída são dois histogramas, um para voz e outro para silêncio. Esta rotina se utiliza de uma base de dados de treinamento de sinais de voz, sendo que, para cada trecho será analisado primeiramente se o mesmo pode ser considerado voz ou silêncio via método da energia.

Uma vez caracterizado como um desses dois, realiza-se a rotina de decomposição em átomos do MP e em paralelo um processo de contagem do número de ocorrências de cada átomo do dicionário em cada trecho de voz. Em outras palavras, se por exemplo um dado átomo de índice n ocorreu 10 vezes (no total) ao longo de todos os trechos que foram caracterizados como voz, será contabilizado no histograma de voz o valor 10 para o átomo de índice n (um dicionário típico pode conter 64000 átomos). Contudo, uma vez que o mesmo átomo pode aparecer tanto em trechos de voz quanto em silêncio, deve-se contabilizar também o número de ocorrências deste átomo em trechos de silêncio. No entanto, este novo valor deverá ser posicionado no histograma de silêncio e não no de voz. No algoritmo 2, é apresentado o pseudo-código de elaboração dos histogramas:

Algoritmo 2 Histogramas (x ∈ base de dados de treinamento)

Inicialização das variáveis:

$n \leftarrow 0$
$R_f^n \leftarrow x$
$erro \leftarrow definido\ pelo\ projetista$
$histo\_voz \leftarrow zeros(1, \#D)$
$histo\_sil \leftarrow zeros(1, \#D)$
$limiar\_energia \leftarrow definido\ pelo\ projetista$

Rotina principal:

if $\|R_f^n\| > limiar\_energia$ **then**
  while $\|R_f^n\| \leq$ *erro* **do**
    for $m = 0$ to $\#D - 1$ **do**
      $a(m) \leftarrow \langle\ g_{\gamma_m}, R_f^n\ \rangle$
    end for
    $k \leftarrow indice\ em\ que\ ocorreu\ max\ do\ vetor\ \|a\|$
    $R_f^{n+1} \leftarrow R_f^n - a(k)g_{\gamma_k}$
    $histo\_voz(k) \leftarrow histo\_voz(k) + 1$
    $n \leftarrow n + 1$
  end while
else
  while $\|R_f^n\| \leq$ *erro* **do**
    for $m = 0$ to $\#D - 1$ **do**
      $a(m) \leftarrow \langle\ g_{\gamma_m}, R_f^n\ \rangle$
    end for
    $k \leftarrow indice\ em\ que\ ocorreu\ max\ do\ vetor\ \|a\|$
    $R_f^{n+1} \leftarrow R_f^n - a(k)g_{\gamma_k}$
    $histo\_sil(k) \leftarrow histo\_sil(k) + 1$
    $n \leftarrow n + 1$
  end while
end if

Uma vez obtidos ambos os histogramas, precisamos normalizá-los de forma que se obtenha uma densidade de probabilidade cuja soma dos valores seja igual a 1. Para isso, basta dividirmos cada amostra do histograma de voz pela soma de todas amostras deste histograma. Em seguida, realiza-se a mesma operação para o histograma de silêncio.

A partir disso, podemos executar a rotina de detecção de voz via votação. Esta rotina por sua vez, utiliza-se de um sinal de voz de uma base de dados de teste, e o analisa segundo a probabilidade dos átomos utilizados pertencerem a classe voz ou silêncio. Ao final do processo, é estabelecido um processo de contagem do número de átomos pertencentes a classe voz e silêncio que foram necessários para sintetizar o sinal de voz em questão. Caso tenha sido necessário um número maior de átomos da classe voz do que da classe silêncio, o sinal em questao será classificado como voz. Na situação contrária, será classificado como silêncio. Em outras palavras, se um dado átomo de índice n possui uma probabilidade de 0,15 no histograma de voz e de 0,12 no histograma de silêncio, o contador de átomos de voz será acrescido de 1 unidade. Na situação contrária de probabilidade, acrescenta-se 1 no contador de átomos de silêncio. Ao final do processo, compara-se ambos os contadores e classifica-se o trecho como sendo voz ou silêncio, dependendo do contador que detenha o maior valor. Apresentamos a seguir o pseudo-código de detecção de voz via votação (algoritmo 3).

Algoritmo 3 Detecção via votação (x ∈ base de dados de teste)

Inicialização das variáveis:

$n \leftarrow 0$
$R_f^n \leftarrow x$
$erro \leftarrow definido\ pelo\ projetista$
$cont\_voz \leftarrow 0$
$cont\_sil \leftarrow 0$
$histo\_voz \leftarrow histo\_voz / \sum(histo\_voz)$
$histo\_sil \leftarrow histo\_sil / \sum(histo\_sil)$

Rotina principal:

while $\|R_f^n\| \leq erro$ **do**
  for $m = 0$ to $\#D - 1$ **do**
    $a(m) \leftarrow \langle\ g_{\gamma_m}, R_f^n\ \rangle$
  end for
  $k \leftarrow indice\ em\ que\ ocorreu\ max\ do\ vetor\ \|a\|$
  $R_f^{n+1} \leftarrow R_f^n - a(k) g_{\gamma_k}$
  if $histo\_voz(k) > histo\_sil(k)$ **then**
    $cont\_voz \leftarrow cont\_voz + 1$
  else
    $cont\_sil \leftarrow cont\_sil + 1$
  end if
  $n \leftarrow n + 1$
end while
if $cont\_voz > cont\_sil$ **then**
  $audio \leftarrow voz$
else
  $audio \leftarrow silencio$
end if

Além do método de detecção via votação, pode-se utilizar o conceito de perplexidade [9] para classificação entre voz e silêncio. Este método baseia-se no inverso da probabilidade de ocorrência de um determinado átomo para classificar o trecho como sendo voz ou silêncio. Na equação (3) é apresentada a expressão de perplexidade utilizada no processo de classificação.

$$PP = \prod_{n=1}^{N} \frac{1}{\sqrt[N]{P_n}} \tag{3}$$

Nesta expressão, o parâmetro PP corresponde a perplexidade, N corresponde ao número de átomos por janela e P_n corresponde a probabilidade de ocorrência de um determinado átomo em uma dada classe. Uma vez obtida a perplexidade para os átomos de voz e para os átomos de silêncio, compara-se ambos os valores e seleciona-se aquele que possuir menor valor para classificação, isto é, se a perplexidade referente a voz for menor do que a de silêncio, o trecho será classificado como voz. Na situação contrária, o trecho será classificado como silêncio. No algoritmo 4 é apresentado o pseudo-código de detecção via perplexidade.

Algoritmo 4 Detecção via perplexidade (x ∈ base de dados de teste)

Inicialização das variáveis:

$n \leftarrow 0$
$R_f^n \leftarrow x$
$n\_atomos \leftarrow definido\ pelo\ projetista$
$PP\_voz \leftarrow 1$
$PP\_sil \leftarrow 1$
$histo\_voz \leftarrow histo\_voz / \sum(histo\_voz)$
$histo\_sil \leftarrow histo\_sil / \sum(histo\_sil)$

Rotina principal:

while $n < n\_atomos$ **do**
  for $m = 0$ to $\#D - 1$ **do**
    $a(m) \leftarrow \langle\ g_{\gamma_m}, R_f^n\ \rangle$
  end for
  $k \leftarrow indice\ em\ que\ ocorreu\ max\ do\ vetor\ \|a\|$
  $R_f^{n+1} \leftarrow R_f^n - a(k) g_{\gamma_k}$
  $PP\_voz \leftarrow PP\_voz * hist\_voz(k)$
  $PP\_sil \leftarrow PP\_sil * hist\_sil(k)$
  $n \leftarrow n + 1$
end while
if $(PP\_voz)^{-1/N} < (PP\_sil)^{-1/N}$ **then**
  $audio \leftarrow voz$
else
  $audio \leftarrow silencio$
end if

Neste algoritmo, os valores de perplexidade são armazenados nas variáveis PP_voz e PP_sil e comparados no final da rotina principal. Vale ressaltar que a classificação utilizada não é baseada somente em qual átomo possui um valor maior nos histogramas de voz ou silêncio, como é o caso do algoritmo 3, mas utiliza todas as probabilidades dos histogramas envolvidas nos átomos selecionados (voz e silêncio) para obtenção das respectivas perplexidades e comparação das mesmas.

Na seção seguinte, apresentamos os resultados que foram obtidos executando-se as rotinas de processamento abordadas anteriormente.

III. RESULTADOS

Para a realização das simulações, utilizou-se o *software* MATLAB em uma plataforma *Windows 7* de 64 bits. Nestas simulações, utilizou-se uma base de dados de treinamento equivalente a 2 horas de sinais de voz (amostrados em 8 kHz) para elaboração dos histogramas.

Em todo processamento, padronizou-se a frequência de amostragem (f_s) em 8 kHz, tamanho do segmento de voz, para análise, em 20 ms ou 160 amostras e erro de $-35dB$. Com relação ao dicionário, utilizou-se como base a função Gabor e um tamanho equivalente a 64000 átomos. Na Figura 2a e 2b, são apresentados os histogramas de voz e de silêncio (sem normalização), que foram obtidos utilizando o algoritmo 2 abordado anteriormente.

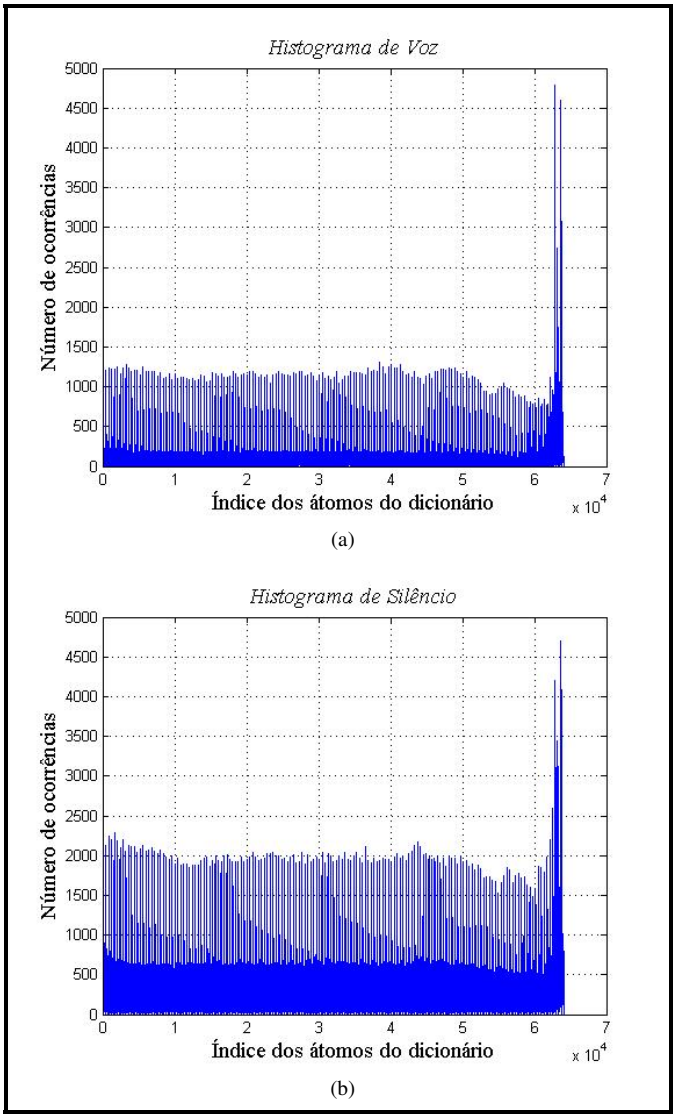

Fig. 2: (a) Histograma de Voz; (b) Histograma de Silêncio.

Com estes dois histogramas e com uma base de dados de teste com aproximadamente 15 min de sinais de voz (não se utilizou uma base de dados maior devido ao excessivo tempo de processamento do MATLAB), executou-se a rotina de processamento abordada no algoritmo 3 (a Figura 3 apresenta o resultado da classificação para um dos sinais de voz utilizados). No entanto, apesar dos sinais utilizados tanto na base de dados de treinamento quanto na de teste apresentarem uma alta razão sinal-ruído, ocorreram alguns erros de detecção ao longo do sinal de voz (principalmente em trechos de silêncio). Estes erros podem ser observados na Figura 3.

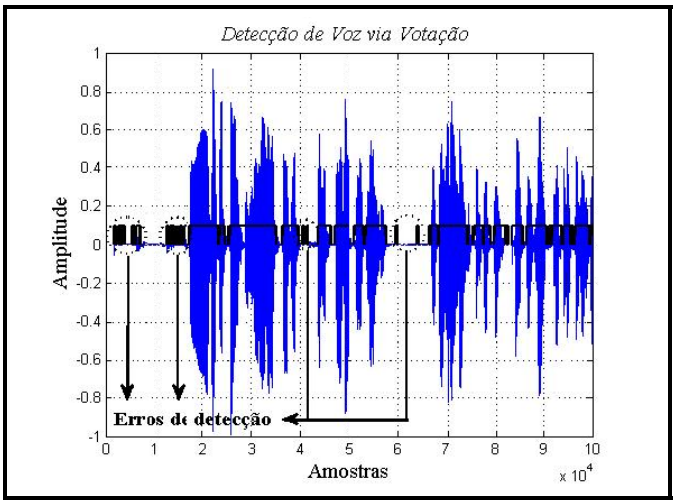

Fig. 3: Detecção de voz via votação.

Estes erros de detecção ocorreram em virtude do erro definido previamente ($-35dB$) não corresponder ao limiar ótimo de operação. Dessa forma, comparando-se a classificação via votação com a classificação pelo limiar de energia (sendo a última usada como referência), obtêm-se a taxa de acerto de detecção para o sinal de voz em questão. Contudo, uma vez que para diferentes valores de erro obtêm-se diferentes taxas de acerto, decidiu-se determinar qual o ponto ótimo de operação que maximiza esta taxa. Na Figura 4, é apresentada a curva que define o comportamento de variação entre a taxa de acerto e o erro definido.

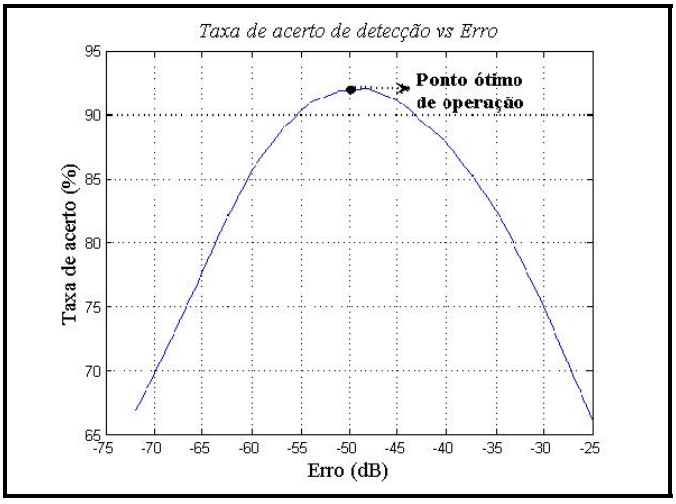

Fig. 4: Ponto ótimo de operação para o método de detecção via votação.

Percebe-se portanto, que o ponto ótimo de operação está definido para um erro de -50 dB, o qual resulta em uma taxa de acerto de aproximadamente 93%. Contudo, uma vez que ambos os parâmetros estão fortemente correlacionados com o tipo de dicionário que está sendo utilizado, observa-se que, quanto maior for o nível de redundância e completude do dicionário, maior será a taxa de acerto e por conseguinte menor será a necessidade de se utilizar limiares de conversão mais rigorosos/intensos (quanto mais rigoroso o limiar de conversão, maior será a quantidade de átomos necessários para sintetizar um sinal mais próximo do sinal original).

Contudo, observa-se também que para valores de erro abaixo de -50 dB, a taxa de acerto começa a decair. Este fato indica que o dicionário utilizado não é completo o suficiente (limitou-se o tamanho em 64k por motivos de processamento e memória), pois, além dos histogramas terem sido obtidos com erro $= -35$ dB (diferente do obtido no ponto ótimo), usou-se átomos de menor correlação (e por conseguinte menor probabilidade) no processo de classificação entre voz e silêncio, resultando em um aumento na taxa de erro de detecção.

No entanto, uma vez que a quantidade de átomos aumenta conforme o limiar de conversão se torna mais rigoroso, optou-se por utilizar o conceito da perplexidade no processo de classificação entre voz e silêncio. Para isso, limitou-se o número de átomos por janela e observou-se a taxa de acerto obtida. Na Figura 5 é apresentada a curva obtida para valores entre 10 e 80 átomos por janela.

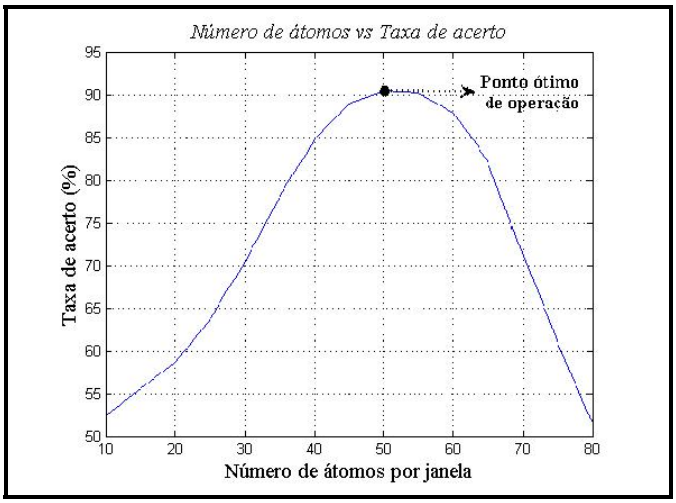

Fig. 5: Ponto ótimo de operação para o método de detecção via perplexidade.

Nesta figura, percebe-se que o ponto ótimo de operação é obtido para 50 átomos por janela o qual resulta em uma taxa de acerto de aproximadamente 91%. Esta taxa é inferior à obtida anteriormente (93%), no entanto, a quantidade de átomos utilizada é 47% menor pois, enquanto que para um limiar de -50 dB foram necessários aproximadamente 59000 átomos, pelo método da perplexidade foram necessários 31250 átomos, permitindo assim uma redução na quantidade de átomos necessários no processo de classificação entre voz e silêncio.

IV. CONCLUSÕES

O presente trabalho propõe uma nova forma de classificação voz/silêncio que se baseia nas densidades de probabilidade de átomos que caracterizam cada uma dessas classes de sinal, prescindindo da análise do comportamento da energia ao longo do sinal, como feito nos métodos mais tradicionais.

Conforme apresentado nos resultados, tanto o método de detecção via votação quanto o método via perplexidade, se mostraram muito promissores uma vez que conseguiram distinguir as duas classes existentes, no caso voz e silêncio com pelo menos 90% de acerto. No entanto, em virtude dos picos consecutivos de detecção que ocorreram principalmente nos trechos de silêncio, pretende-se em trabalhos futuros, utilizar a informação contida nos pesos dos átomos para refinar a detecção, uma vez que os mesmos contêm informação correlacionada com a energia.

Uma vez refinada a detecção, pretende-se verificar futuramente a precisão da detecção com sinais que possuam uma razão sinal-ruído mais baixa. Uma vez que o método apresentado não se utiliza da energia ao longo do sinal, ruídos com características bem determinadas poderão permitir boa classificação ainda sem se considerar o uso das energias envolvidas, pois haverá probabilidade significativa de um determinado conjunto de átomos formarem a base fundamental para a síntese de um dado ruído particular. Além disso, espera-se que ruídos que possuam um comportamento estacionário poderão ser convenientemente modelados pelos átomos do dicionário, dando chances para que a técnica também funcione bem para sinais ruidosos que possuam ruídos com características estacionárias.

Para trabalhos futuros, pretende-se não só utilizar dicionários de maior completude e redundância (o que permite melhorar o ponto ótimo de operação), mas também implementar os algoritmos que foram apresentados em uma linguagem mais eficiente, reduzindo assim o tempo de execução e o custo computacional.

REFERENCES

[1] P. Dymarski1, N. Moreau and G. Richard, *"Greedy sparse decompositions: a comparative study"*, *EURASIP Journal on Advances in Signal Processing*, 2011.

[2] D. Needell, R. Vershynin, *Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit*, *IEEE J. Sel. Topics Signal Process*, vol.4(2), pp. 310–316, 2010.

[3] B. Sturm, M. Christensen, *Cyclic matching pursuit with multiscale time-frequency dictionaries*, *Asilomar Conference on Signals Systems and Computers*, 2010.

[4] S. Mallat and Z. Zhang, *"Matching pursuit with time-frequency dictionaries"*, *IEEE Trans. Sig. Proc.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[5] P. Vera-Candeas, N. Ruiz-Reyes, M. Rosa-Zurera, D. Martinez-Munoz, and J. Curpián-Alonso, *"New matching pursuit based sinusoidal modelling method for audio coding"*, *IEEE Proc. Vis. Image Signal Process.*, vol. 151, no. 1, pp. 21–28, 2004.

[6] S. Jaggi, W. Carl, S. Mallat, and A. Willsky, *"High resolution pursuit for feature extraction"*, *Technical report, MIT, November*, 1995.

[7] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, *"Discrimination of pathological voices using a time-frequency approach"*, *IEEE Trans. Biomedical Eng.*, vol. 52, no. 3, pp. 421–430, 2005.

[8] G. Kling and C. Roads, *"Audio analysis, visualization, and transformation with matching pursuits"*, *in Proc. of the COST G-6 Conference on Digital Audio Effects (DAFX-04)*, 2004.

[9] D. Jurafsky and J. H. Martin, *"Speech and Language Processing"*, Pearson Prentice Hall, 2 edition, 2008.