Detecção de Atividade Vocal com o uso de Máquinas de Boltzmann Restritas Discriminativas

Rogério G. Borin e Magno T. M. Silva

***Resumo*— A detecção de atividade vocal (VAD – *Voice Activity Detection*) tem apreciável impacto no desempenho de aplicações tecnológicas atuais, tais como comunicações sem fio e reconhecimento de fala. Neste trabalho, aborda-se a tarefa de VAD por meio da aprendizagem de máquina usando-se uma estrutura denominada Máquina de Boltzmann Restrita Discriminativa (DRBM – *Discriminative Restricted Boltzmann Machine*). Com o emprego de vetores de características baseados tanto em coeficientes mel-cepstrais quanto em energias em bandas do espectro, chega-se a desempenhos levemente superiores aos do VAD denominado *Long-Term Spectral Divergence* (LTSD), usado frequentemente como base de comparação para outros detectores. Os resultados indicam ainda que a DRBM é capaz de lidar com vetores de características fortemente correlacionados.**

***Palavras-Chave*— Máquina de Boltzmann Restrita Discriminativa, Detecção de Atividade Vocal, Coeficientes Mel-Cepstrais, Aprendizagem de Máquina para Processamento de Sinais.**

***Abstract*— Voice Activity Detection (VAD) has substantial impact on the performance of present technological applications, such as wireless communications and speech recognition. In this work we approach the VAD task through machine learning by using a structure known as Discriminative Restricted Boltzmann Machine (DRBM). Employing feature vectors based either on mel-frequency cepstral coefficients or on filterbank energies, the resulting detector slightly outperforms the VAD known as Long-Term Spectral Divergence (LTSD), frequently used as benchmark for detector comparison. Results also indicate that the DRBM is able to deal with strongly correlated feature vectors.**

***Keywords*— Discriminative Restricted Boltzmann Machine, Voice Activity Detection, Mel-Frequency Cesptral Coefficients, Machine Learning for Signal Processing.**

I. Introdução

A detecção de atividade vocal (VAD – *Voice Activity Detection*) é um assunto que tem recebido a atenção da comunidade científica por muitos anos [1–15]. Em sistemas de telefonia, detectores de atividade vocal permitem uma significativa redução na largura de banda utilizada para comunicações de voz. VADs são também utilizados em sistemas de redução de ruído, nos quais se faz a estimação do espectro do ruído durante períodos de ausência de voz [1].

Técnicas tradicionais de detecção de atividade vocal incluem medidas baseadas em energia, parâmetros de codificação preditiva linear (LPC – *Linear Predictive Coding*) [2], taxa de cruzamento por zero [3], medidas de periodicidade [4], características cepstrais [5], configuração de formantes [6] e entropia espectral [7]. Destacam-se também as técnicas envolvendo estatísticas de ordem superior [8, 9] e modelos estatísticos avançados [10–12]. Essas técnicas buscam estabelecer medidas que salientem as diferenças entre as condições de presença e de ausência de voz. Tais medidas são então usadas em regras de decisão definidas empiricamente ou segundo algum critério objetivo como, por exemplo, satisfazer uma condição definida estatisticamente. Dois VADs que se utilizam de técnicas tradicionais ou estatísticas são aqui destacados. O primeiro deles, conhecido como G.729-B [16], foi adotado pela indústria como parte de um *codec* de voz. Esse VAD faz uso de medidas de energia, taxa de cruzamento por zero e parâmetros de codificação preditiva. O segundo, chamado de LTSD (*Long-Term Spectral Divergence*) [1], opera comparando a envoltória de longo prazo do espectro do sinal com uma estimativa do espectro do ruído. Esse VAD é frequentemente usado como base de comparação para outros detectores, dado o seu bom desempenho em uma larga faixa de relações sinal-ruído (SNR – *signal-to-noise ratio*). Além disso, o mesmo mostrou ter um desempenho superior a VADs adotados em *codecs* de voz de padrões europeus (e.g., AMR – *Adaptive Multirate* e AFE – *Advanced Front-End*) [1].

Recentemente, técnicas envolvendo aprendizagem de máquina vem se popularizando em muitas tarefas, inclusive na detecção de atividade vocal. Em uma das abordagens, utilizam-se medidas obtidas pelas já citadas técnicas tradicionais como entradas do classificador, o qual é treinado de modo a se obter a classificação desejada. Nesse caso, apenas o encargo da tomada de decisão do detector é passado para o classificador. Enqing et al. [13] apresentaram um dos primeiros VADs de que se tem notícia segundo essa abordagem. Nesse trabalho, uma Máquina de Vetor de Suporte (SVM – *Support Vector Machine*) recebe as medidas usadas pelo VAD G.729-B como entradas, obtendo resultados superiores aos do G.729-B convencional. Em outra abordagem, o mecanismo de aprendizagem é alimentado com medidas que buscam representar o som propriamente dito, como seria o caso dos coeficientes mel-cepstrais (MFCCs – *Mel-frequency cepstral coefficients*) [14]. Assim, o classificador torna-se responsável não apenas pela decisão final do detector, mas também pela descoberta das características importantes para discriminar a presença ou ausência de voz. Independentemente da abordagem utilizada, é interessante notar que o uso de mecanismos de aprendizagem torna direta a adição de novas informações a um detector. Em [15], por exemplo, vetores de características contendo uma vasta gama de medidas são utilizados como entrada para diferentes modelos usados em aprendizagem de máquina. Dentre esses modelos, destaca-se o das redes neuronais baseadas em redes de crença profunda (DBN-DNN – *Deep Belief Network-Deep Neural Network*). Os resultados obtidos mostram a capacidade dos mecanismos de aprendizagem, em especial da DBN-DNN, de fundir diferentes tipos de informações e assim obter detectores mais poderosos.

Rogério G. Borin e Magno T. M. Silva, Departamento de Engenharia Elétrica, Escola Politécnica da USP, São Paulo-SP, Brasil, E-mails: rborin@lps.usp.br, magno@lps.usp.br.

No presente trabalho, utiliza-se uma estrutura pouco explorada na área de aprendizagem de máquina: a Máquina de Boltzmann Restrita Discriminativa (DRBM – *Discriminative Restricted Boltzmann Machine*). Até onde os autores deste artigo conhecem, não existem trabalhos anteriores na literatura que façam uso da DRBM para VAD. A DRBM na sua forma original lida com dados de entrada binários [19]. Como contribuição deste trabalho, apresenta-se uma variante do modelo, aqui denominada DRBM Gauss-Bernoulli, que é capaz de lidar com dados de entrada contínuos. Essa adaptação do modelo é então utilizada na tarefa de VAD. Por meio de simulações, verifica-se que uma DRBM Gauss-Bernoulli com um pequeno número de unidades ocultas é capaz de obter um desempenho levemente superior ao do LTSD em termos de área sob a curva de característica de operação do receptor (ROC – *Receiver Operating Characteristic*), da taxa de acertos de classificação (acurácia) e de custo computacional. Adicionalmente, o detector proposto é comparado com o já citado G.729-B e com uma melhoria desse, aqui indicada como G.729-II [17].

O artigo está organizado da seguinte forma: na Seção II é fornecida uma breve introdução ao modelo usado como classificador; na Seção III são apresentados os detalhes da configuração experimental bem como os resultados comparativos entre os detectores. Por fim, na Seção IV tem-se a conclusão do trabalho.

II. RBM E DRBM COMO CLASSIFICADORES

Uma RBM (*Restricted Boltzmann Machine*) é uma rede neuronal estocástica capaz de produzir dados segundo uma distribuição de probabilidade [18]. RBMs foram aplicadas de forma bem sucedida como blocos construtivos de modelos com múltiplas camadas, tais como DBNs e DNNs. Na Figura 1, tem-se uma representação desse modelo em que os círculos representam as unidades. No retângulo superior estão as unidades ocultas formando a camada oculta do modelo e na parte inferior as unidades visíveis divididas em dois grupos que conjuntamente formam a camada visível da RBM: à direita, um grupo que representa a entrada e à esquerda, a classe (rótulo) a ela associada [19]. As unidades são formalmente modeladas como variáveis aleatórias cuja distribuição de probabilidade conjunta é dada por

$$P(y, \mathbf{x}, \mathbf{h}) = \frac{\exp(-E(y, \mathbf{x}, \mathbf{h}))}{Z}, \tag{1}$$

em que $\mathbf{x} = [x_1, \ldots, x_{n_d}]^T$ e $\mathbf{h} = [h_1, \ldots, h_{n_h}]^T$ são os vetores de estado das variáveis de entrada e ocultas, respectivamente, $y \in \{1, \ldots, n_c\}$ é a classe associada ao vetor de entrada, $E(y, \mathbf{x}, \mathbf{h})$ é a chamada função de energia global e Z, denominada função de partição, é uma constante que garante que o somatório de $P(y, \mathbf{x}, \mathbf{h})$ sobre seu domínio seja unitário.

Diferentes definições de $E(y, \mathbf{x}, \mathbf{h})$ levam a diferentes variantes do modelo. Neste trabalho, a definição original em [19] foi adaptada de modo a resultar em variáveis visíveis condicionalmente Gaussianas, produzindo

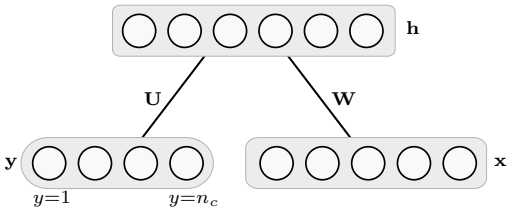

Fig. 1. Representação sucinta de uma RBM com camada de classificação.

$$E(y, \mathbf{x}, \mathbf{h}) = -\sum_{i=1}^{n_d}\sum_{j=1}^{n_h} h_j w_{ji} \frac{x_i}{\sigma_i^2} - \sum_{j=1}^{n_h} b_j h_j - \sum_{k=1}^{n_c} d_k \delta_{k,y} - \sum_{k=1}^{n_c}\sum_{j=1}^{n_h} h_j u_{jk} \delta_{k,y} + \sum_{i=1}^{n_d} \frac{(x_i - c_i)^2}{2\sigma_i^2}. \tag{2}$$

Essa forma para $E(y, \mathbf{x}, \mathbf{h})$ foi inspirada na adaptação proposta em [20], então direcionada a RBMs sem uma camada de classificação. Na Equação (2), a notação $\delta_{r,s}$ representa o delta de Kronecker e w_{ji}, b_j, c_i, σ_i^2, d_k, u_{jk}, com $i{=}1, \ldots, n_d$, $j{=}1, \ldots, n_h$, $k{=}1, \ldots, n_c$, constituem os parâmetros do modelo. Considera-se aqui que as variáveis ocultas possam assumir individualmente valores no conjunto $\{0, 1\}$. Pode-se demonstrar que, dados y e $\mathbf{x}$, o vetor de variáveis ocultas é conjuntamente independente e suas componentes têm distribuição de Bernoulli com probabilidade de sucesso

$$P(h_j{=}1|y, \mathbf{x}) = \varphi\left(b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\right), \tag{3}$$

em que $\varphi(z) = 1/(1 + e^{-z})$ é a chamada função sigmóide.

O treinamento de RBMs é comumente realizado por meio do algoritmo CD (*Contrastive Divergence*) [21], o qual busca determinar o conjunto de parâmetros do modelo que minimizem a função perda generativa

$$\mathcal{L}_{gen} = -\sum_{t=1}^{n_t} \log P(y^{(t)}, \mathbf{x}^{(t)}), \tag{4}$$

em que n_t é a quantidade de amostras de treinamento e o par $(\mathbf{x}^{(t)}, y^{(t)})$ representa a t-ésima amostra de treinamento, composta por uma entrada, $\mathbf{x}^{(t)}$, e sua respectiva classe, $y^{(t)}$. A busca eficiente pelos pontos de mínimo da função perda envolveria o cálculo do seu gradiente com relação aos parâmetros do modelo. Infelizmente, a determinação exata de tal gradiente é intratável, fato com o qual o algoritmo CD lida fazendo uso de certas aproximações [21].

A estrutura de uma DRBM é idêntica àquela mostrada na Figura 1 para a RBM. Na DRBM, entretanto, tem-se como objetivo de treinamento minimizar a função perda discriminativa

$$\mathcal{L}_{disc} = -\sum_{t=1}^{n_t} \log P(y^{(t)}|\mathbf{x}^{(t)}). \tag{5}$$

Com a definição dada na Equação (2), demonstra-se que

$$P(y|\mathbf{x}) = \frac{\exp\left(d_y + \sum_{j=1}^{n_h} \zeta\left(b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\right)\right)}{\sum_{y^*=1}^{n_c} \exp\left(d_{y^*} + \sum_{j=1}^{n_h} \zeta\left(b_j + u_{jy^*} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\right)\right)}, \tag{6}$$

com $\zeta(z) = \ln(1 + e^z)$. A Equação (6) tem a mesma forma que a obtida quando as variáveis de entrada são binárias. Como

observado em [19], pode-se computar $P(y|\mathbf{x})$ em um tempo $\mathcal{O}(n_h n_d + n_h n_c)$. Além disso, o gradiente de $P(y|\mathbf{x})$ pode ser avaliado de forma exata e eficiente. Como consequência, o gradiente de $\mathcal{L}_{disc}$ também pode ser obtido de forma exata. Com o emprego dessa função perda no algoritmo do gradiente descendente estocástico (SGD – *Stochastic Gradient Descent*), chega-se às regras de atualização de parâmetros sumarizadas na Tabela I. Nessa tabela, não são fornecidas as regras de atualização para os parâmetros c_i nem os σ_i^2, visto que os primeiros não são relevantes para o treinamento discriminativo enquanto que os últimos podem ter sua aprendizagem evitada fazendo-se uma normalização na variância dos dados de entrada.

TABELA I

REGRAS DE ATUALIZAÇÃO DE PARÂMETROS PARA A DRBM.

Definições:

$$\Delta b_j = P(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) - \sum_{y^*=1}^{n_c} P(y^*|\mathbf{x}^{(t)}) P(h_j{=}1|y^*, \mathbf{x}^{(t)})$$

$$\Delta d_k = \delta_{k,y^{(t)}} - P(y{=}k|\mathbf{x}^{(t)})$$

$$\Delta w_{ji} = \Delta b_j \left(\frac{x_i^{(t)}}{\sigma_i^2} \right)$$

$$\Delta u_{jk} = \Delta d_k P(h_j{=}1|y{=}k, \mathbf{x}^{(t)})$$

Nota: As funções $P(h_j{=}1|y, \mathbf{x})$ e $P(y|\mathbf{x})$ empregadas nas definições acima são aquelas apresentadas nas equações (3) e (6), respectivamente.

Regras de atualização: De acordo com o SGD, a atualização de um parâmetro ϕ qualquer do modelo é feita segundo a regra:

$$\phi \leftarrow \phi + \lambda \Delta\phi,$$

sendo λ a taxa de aprendizagem do algoritmo.

Em termos práticos, uma vantagem das DRBMs é que normalmente são obtidos bons resultados de classificação com modelos de menores dimensões em relação às RBMs. Além disso, como o gradiente da função perda discriminativa é exato, normalmente consegue-se usar taxas de aprendizagem maiores durante o treinamento de DRBMs sem que ocorra a divergência do algoritmo de treinamento.

III. ANÁLISE EXPERIMENTAL

Nesta seção, são fornecidos os detalhes dos experimentos conduzidos. Todos eles foram feitos usando o MATLAB 7.11 executado sobre o *Windows 7* em uma máquina com processador *Intel Xeon* com 6 núcleos físicos operando a 2,4 GHz e um total de 32 GB de memória (RAM).

A. Corpus de teste

Nos experimentos, utilizou-se uma versão modificada do *corpus* de teste (conjunto de arquivos de áudio) denominado NOIZEUS [22], o qual foi escolhido por três motivos: primeiramente, como o *corpus* é constituído de 30 frases com duração total de cerca de 80 segundos, a rotulação manual do áudio é viável; em segundo lugar, as frases foram concebidas de modo a conter todos os fonemas da língua inglesa; e, por fim, esse corpus é disponibilizado gratuitamente.

Vale aqui justificar as mudanças feitas no corpus de teste. O áudio de cada um dos arquivos que compõem o corpus foi manualmente rotulado de modo a indicar os trechos contendo voz. Segundo essa rotulação, a qual considerou como voz tanto sons vocalizados como não vocalizados, obteve-se um percentual de atividade vocal variando entre 64,9% e 91,2% entre os arquivos, e uma média de 83,4%. Tendo em vista esse percentual, pode-se dizer que essa base de informações é apreciavelmente desbalanceada, pois possui uma quantidade consideravelmente maior de exemplos positivos (presença de voz) do que negativos (ausência de voz). Para os detectores de voz baseados em processamento de sinais (G.729-B/II e LTSD), esse desbalanceamento significa uma quantidade relativamente pequena de dados para avaliação dos seus desempenhos na detecção de ausência de voz. No caso dos detectores baseados em aprendizagem de máquina, o treinamento nessa situação tenderia a produzir detectores enviesados no sentido de detectar voz. Por esses motivos, os arquivos de áudio foram modificados da seguinte forma: aos arquivos sem ruído adicionou-se 0,8 s de silêncio antes e após o áudio original e, em seguida, somou-se o ruído (obtido da gravação *car noise* da base AURORA-2) de modo a se obter as relações sinal-ruído desejadas (de -5 dB a 20 dB, em passos de 5 dB). Nessa operação, foi usado o mesmo procedimento empregado para geração dos arquivos ruidosos do corpus original [22]. O balanceamento por esse método garante ainda que os detectores G.729-B/II e LTSD operem apropriadamente, pois eles utilizam o início do áudio para estimar as características do ruído.

Os arquivos modificados foram separados aleatoriamente com 70% deles constituindo o conjunto de treinamento e os 30% restantes, o conjunto de teste.

B. Extração de características

Duas configurações de vetores de características se destacaram na tarefa de detecção de atividade vocal empregando DRBM. A primeira delas é baseada nos coeficientes mel-cepstrais (MFCCs) e a segunda, nas energias do banco de filtros (FBEs – *Filter Bank Energies*), um subproduto do cálculo dos MFCCs. Para obtenção dos MFCCs e FBEs, o áudio de cada arquivo passou por um filtro de pré-ênfase (com coeficiente 0,97) e o sinal resultante foi então segmentado em quadros sobrepostos com duração de 25 ms e deslocamento de 10 ms entre quadros. De cada quadro foram extraídos os 13 primeiros MFCCs e as energias (correspondentes aos FBEs) dos 23 canais usados no banco de filtros. A média quadrática das amostras de um quadro (FE – *Frame Energy*) foi também agregada ao vetor de características. A Tabela II detalha as configurações utilizadas e as dimensões resultantes dos vetores.

TABELA II

DETALHES DOS VETORES DE CARACTERÍSTICAS.

Config.	Conteúdo do vetor	Dimensão
C1	13 MFCCs normalizados (média nula e variância unitária) + log(FE), juntamente com as derivadas temporais de primeira e segunda ordem dos mesmos.	42
C2	23 FBEs + log(FE) normalizados para magnitude máxima unitária, juntamente com as derivadas temporais de primeira e segunda ordem dos mesmos.	72

C. Treinamento

Em razão de testes preliminares varrendo diversas configurações de número de unidades ocultas (n_h) e taxas de aprendizagem (λ), escolheu-se utilizar $n_h = 30$ e $\lambda = 0{,}005$ no treinamento da DRBM em todas as relações sinal-ruído. As amostras de treinamento, compostas pelos vetores de características e respectivos rótulos manuais, foram divididas em mini-lotes de 70 amostras para o cálculo do gradiente no algoritmo baseado em SGD.

D. Avaliação de desempenho

Para avaliação de desempenho geral dos detectores, usou-se a medida de área sob a curva ROC (sensibilidade *versus* (1−especificidade)), a qual tem emprego comum na área de telecomunicações. Os VADs G.729-B/II não possuem um limiar de detecção configurável e, assim, somente um ponto da curva ROC pode ser obtido, o qual, juntamente com os pontos teóricos $(0, 0)$ e $(1, 1)$ permitem o cálculo da área. Devido à possibilidade de se subestimar o desempenho desses VADs com esse método, empregou-se também, na comparação entre detectores, uma medida de uso difundido na área de aprendizagem de máquina: a acurácia balanceada (média da sensibilidade e especificidade). O desempenho do LTSD foi determinado executando-se esse VAD com diferentes limiares de detecção sobre todos os arquivos do *corpus*. No caso do detector baseado em DRBM, o limiar é aplicado à sua saída, que fornece a probabilidade (Equação (6)) de que uma dada amostra seja voz. Em ambos os casos (LTSD e DRBM), as acurácias apresentadas mais à frente são as melhores que se obteve variando-se o limiar de detecção. Além da área ROC e da acurácia, foram também realizadas medidas visando comparar aproximadamente o custo computacional dos detectores.

E. Resultados

Na Figura 2, são mostradas as medidas de área sob a curva ROC dos detectores avaliados para diferentes relações sinal-ruído. As DRBMs usando as configurações C1 e C2 (conforme Tabela II) estão identificadas como DRBM-C1 e DRBM-C2, respectivamente. Primeiramente, nota-se que o VAD G.729-B sofre uma forte degradação de desempenho com a diminuição de SNR. O G.729-II melhora sensivelmente essa situação, mas é ainda inferior ao LTSD, o qual, por sua vez, é levemente inferior aos VADs baseados em DRBM na maioria das relações sinal-ruído.

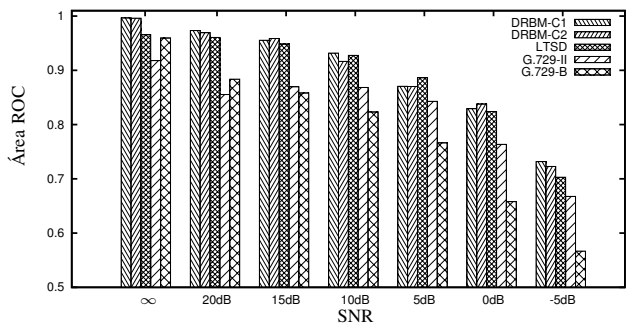

Fig. 2. Área sob a curva ROC para diferentes relações sinal-ruído.

Na Tabela III são apresentadas as medidas de acurácia dos VADs em que se percebe resultados qualitativamente semelhantes aos anteriores. Na média ao longo das relações sinal-ruído, observa-se uma vantagem dos detectores baseados em DRBM.

A similaridade dos resultados obtidos para as DRBMs com configurações de vetores de características envolvendo MFCCs e FBEs é algo a ser destacado. Os MFCCs são obtidos aplicando-se a transformada discreta do cosseno ao logaritmo dos FBEs. Essa transformada produz um vetor descorrelacionado, o que é geralmente considerado benéfico para blocos de processamento posteriores. Assim, a citada similaridade de desempenhos entre DRBM-C1 e DRBM-C2 é um indicativo de que as DRBMs têm a capacidade de lidar adequadamente com vetores de características fortemente correlacionados. A mesma característica foi apontada em [23] para o caso de DBN-DNNs.

TABELA III

ACURÁCIA DOS DETECTORES PARA DIFERENTES RELAÇÕES SINAL-RUÍDO.

SNR(dB)	DRBM-C1	DRBM-C2	LTSD	G729-II	G729-B
∞	97,76%	97,72%	93,73%	91,80%	95,95%
20	93,72%	91,66%	91,07%	85,55%	88,38%
15	90,93%	91,40%	89,70%	86,98%	85,82%
10	88,27%	86,81%	86,96%	86,85%	82,32%
5	82,60%	83,14%	82,83%	84,29%	76,63%
0	77,31%	78,52%	76,88%	76,35%	65,80%
-5	67,94%	67,37%	66,62%	66,76%	56,68%
Média	85,50%	85,23%	83,97%	82,65%	78,80%

A título de ilustração, as saídas dos detectores em diferentes relações sinal-ruído são mostradas na Figura 3. Nota-se claramente a queda de desempenho de todos eles com a piora de SNR.

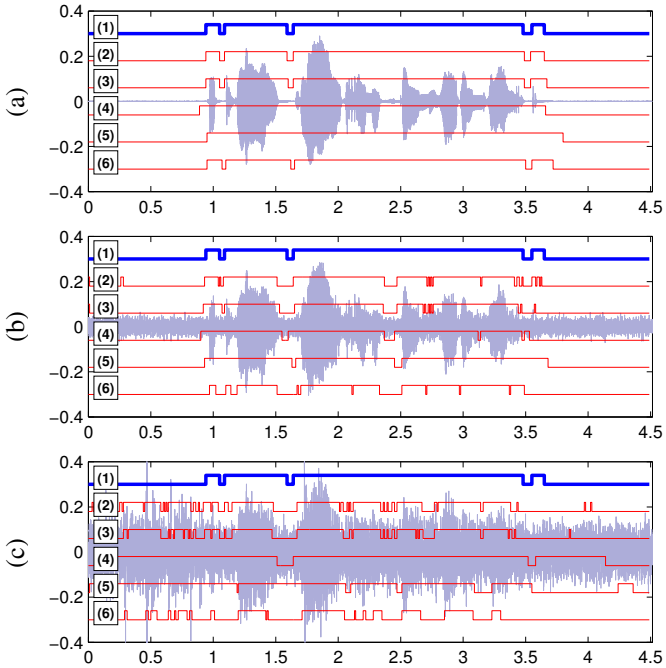

Fig. 3. Exemplo de operação dos detectores em diferentes SNRs: (a) Sinal sem ruído, (b) 10dB, (c) 0dB. Para cada SNR, são mostradas, sobrepostas à imagem do áudio, a (1) rotulação manual e as saídas dos detectores sendo: (2) DRBM-C1, (3) DRBM-C2, (4) LTSD, (5) G.729-II e (6) G.729-B.

Por fim, os custos computacionais dos detectores foram obtidos empiricamente. A medida usada para comparação entre eles é aqui chamada de taxa de trabalho e consiste na

TABELA IV
TAXA DE TRABALHO DOS DETECTORES.

DRBM-C1	DRBM-C2	LTSD	G729-II	G729-B
346,4	243,2	174,8	34,7	35,2

razão entre a quantidade de segundos de áudio processado e o tempo necessário para fazê-lo. Em outras palavras, essa medida indica quantos segundos de áudio são processados em 1 segundo de uso do processador. Os resultados são mostrados na Tabela IV. É importante mencionar que medidas de custo computacional são consideravelmente dependentes dos detalhes de implementação dos algoritmos. Portanto, esses valores permitem apenas a comparação das ordens de grandeza dos custos. Com essa ressalva, nota-se que os detectores baseados em DRBM têm custos computacionais comparáveis ao do LTSD. Os VADs G.729-B/II ficam em desvantagem nessas medidas porque eles baseiam suas operações em informações produzidas por um *codec* de voz, ou seja, parte do tempo gasto por esses VADs corresponde à codificação de voz.

IV. CONCLUSÕES

No presente trabalho, foi proposta a aplicação de uma DRBM à tarefa de detecção de atividade vocal. Em testes realizados em uma ampla faixa de relações sinal-ruído com o emprego de vetores de características baseados tanto em MFCCs quanto em FBEs, verificou-se que, na aplicação proposta, os detectores baseados em DRBM atingem desempenhos, em termos de área sob a curva ROC e de acurácia, bastante razoáveis. Com ambas as configurações de vetores de características, conseguiram-se resultados levemente superiores aos do VAD denominado LTSD, comumente usado como base de comparação para detectores, e consideravelmente superiores aos do G.729-B e de uma melhoria do mesmo, aqui indicada como G.729-II, utilizados pela indústria. Mais ainda, os bons desempenhos foram obtidos com uma DRBM de pequenas dimensões (30 unidades ocultas) e com um custo computacional comparável ao do LTSD.

Adicionalmente, a similaridade dos desempenhos observados para a DRBM tanto com o uso de vetores de características não correlacionados (MFCCs) quanto com vetores fortemente correlacionados (FBEs) assinalam a capacidade dessas estruturas de lidar apropriadamente com entradas correlacionadas, uma qualidade já apontada para o modelo mais complexo das DBN-DNNs [23], o qual é construído empilhando-se RBMs.

Num trabalho futuro, pretende-se comparar o VAD baseado em DRBM com um VAD empregando os mesmos vetores de características, mas tendo como classificador outro mecanismo de aprendizagem de máquina: uma Máquina de Vetor de Suporte (SVM).

AGRADECIMENTOS

Os autores agradecem as sugestões do Prof. Miguel Arjona Ramírez, o suporte recebido da FAPESP (processos 2012/24789-0 e 2015/25512-0), que possibilitou a utilização da base de dados AURORA-2 e também o suporte recebido do CNPq (processo 304275/2014-0).

REFERÊNCIAS

[1] Javier Ramırez et al. "Efficient voice activity detection algorithms using long-term speech information". Em: *Speech communication* 42.3 (2004), pp. 271–287.

[2] L Rabiner e MR Sambur. "Voiced-unvoiced-silence detection using the Itakura LPC distance measure". Em: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'77*. Vol. 2. IEEE. 1977, pp. 323–326.

[3] Jean-Claude Junqua, Ben Reaves e Brian Mak. "A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer". Em: *Second European Conference on Speech Communication and Technology*. 1991.

[4] R Tucker. "Voice activity detection using a periodicity measure". Em: *IEE Proceedings on Communications, Speech and Vision*. Vol. 139. 4. IET. 1992, pp. 377–380.

[5] JA Haigh e JS Mason. "Robust voice activity detection using cepstral features". Em: *TENCON'93. Proceedings. IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering*. Vol. 3. IEEE. 1993, pp. 321–324.

[6] John D Hoyt e Harry Wechsler. "Detection of human speech in structured noise". Em: *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP-94*. Vol. 2. IEEE. 1994, pp. II–237.

[7] Philippe Renevey e Andrzej Drygajlo. "Entropy based voice activity detection in very noisy conditions". Em: *threshold* 5.5.5 (2001), p. 6.

[8] Elias Nemer, Rafik Goubran e Samy Mahmoud. "Robust voice activity detection using higher-order statistics in the LPC residual domain". Em: *IEEE Transactions on Speech and Audio Processing* 9.3 (2001), pp. 217–231.

[9] Ke Li, MNS Swamy e M Omair Ahmad. "An improved voice activity detection using higher order statistics". Em: *IEEE Transactions on Speech and Audio Processing* 13.5 (2005), pp. 965–974.

[10] Jongseo Sohn, Nam Soo Kim e Wonyong Sung. "A statistical model-based voice activity detection". Em: *IEEE Signal Processing Letters* 6.1 (1999), pp. 1–3.

[11] Saeed Gazor e Wei Zhang. "A soft voice activity detector based on a Laplacian-Gaussian model". Em: *IEEE Transactions on Speech and Audio Processing* 11.5 (2003), pp. 498–505.

[12] Joon-Hyuk Chang, Nam Soo Kim e Sanjit K Mitra. "Voice activity detection based on multiple statistical models". Em: *IEEE Transactions on Signal Processing* 54.6 (2006), pp. 1965–1976.

[13] Dong Enqing et al. "Applying support vector machines to voice activity detection". Em: *6th International Conference on Signal Processing*. Vol. 2. IEEE. 2002, pp. 1124–1127.

[14] YX Zou et al. "Improved voice activity detection based on support vector machine with high separable speech feature vectors". Em: *19th International Conference on Digital Signal Processing (DSP)*. IEEE. 2014, pp. 763–767.

[15] Xiao-Lei Zhang e Ji Wu. "Deep belief networks based voice activity detection". Em: *IEEE Transactions on Audio, Speech, and Language Processing* 21.4 (2013), pp. 697–710.

[16] Adit Benyassine et al. "ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications". Em: *IEEE Communications Magazine* 35.9 (1997), pp. 64–73.

[17] *Appendix II – ITU-T G.729 Annex B enhancements in voice-over-IP applications – Option 1*. ITU, ago. de 2005.

[18] Paul Smolensky. "Parallel Distributed Processing: Explorations in the Microstructure of Cognition". Em: *Information Processing in Dynamical Systems: Foundations of Harmony Theory*. Ed. por David E. Rumelhart, James L. McClelland e Corp. PDP Research Group. Vol. 1. Cambridge, MA, USA: MIT Press, 1986. Cap. 6, pp. 194–281. ISBN: 026268053X.

[19] Hugo Larochelle e Yoshua Bengio. "Classification using discriminative restricted Boltzmann machines". Em: *Proceedings of the 25th international conf. on Machine learning*. ACM. 2008, pp. 536–543.

[20] KyungHyun Cho, Alexander Ilin e Tapani Raiko. "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines". Em: *Artificial Neural Networks and Machine Learning–ICANN 2011*. Springer, 2011, pp. 10–17.

[21] Geoffrey E. Hinton. "Training products of experts by minimizing contrastive divergence". Em: *Neural computation* 14.8 (2002), pp. 1771–1800.

[22] Yi Hu e Philipos C Loizou. "Evaluation of objective quality measures for speech enhancement". Em: *IEEE Transactions on Audio, Speech, and Language Processing* 16.1 (2008), pp. 229–238.

[23] Geoffrey Hinton et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". Em: *Signal Processing Magazine, IEEE* 29.6 (2012), pp. 82–97.