DETECÇÃO DE ORADOR E PALAVRAS EM TELEVIGILÂNCIA MÉDICA COM TREINAMENTO MÍNIMO: UMA AMOSTRA POR PALAVRA

Jugurta Montalvão*, Márcia V. P. Montalvão†, Christiane Raulino*

* *Universidade Federal de Sergipe (UFS)*
São Cristóvão, Sergipe, Brazil

† *Tribunal Regional do Trabalho, 20a Região (TRT)*
Aracaju, Sergipe, Brazil

Emails: jmontalvao@ufs.br, marcia.montalvao@trt20.jus.br, chrisraulino@gprufs.org

Abstract— A method for both speaker (Biometrics) and speech recognition is proposed. It was adjusted to obtain high performance under minimum training requirement: only one sample of each target-word, uttered by the target-subject. This minimum training requirement fits with remote health-care convenience, where long and tedious enrolment sessions, with patients or elderly people, are not welcome. In extract a maximum number of relevant cues from single samples, we improve two fragile steps in usual approaches based on Dynamic Time-Warping (DTW). First, we use a resampling strategy based on energy profile, associated to a greedy local search. Moreover, we replace usual scoring based on distances with an evidence accumulation score, in bits of information. Experimental results with two databases (more than 1000 samples) show that both strategies indeed provide improved performances.

Keywords— Remote health-care, Speaker recognition, Speech recognition, Robust DTW, Evidence.

Resumo— Um método para reconhecimento de orador (biometria) e de comandos pronunciados é proposto neste artigo. Sua principal característica é sua adequação a uma fase de treinamento mínimo: apenas um exemplo de cada comando a ser reconhecido, pronunciado pelo orador alvo. Essa característica torna o método particularmente útil em aplicações de tele-vigilância médica baseada em sinais sonoros, por evitar a imposição de longas fases de aquisição de dados a pacientes potencialmente debilitados. Para tanto, dois pontos frágeis dos métodos tradicionais foram reforçados. Primeiramente, uma sub-amostragem baseada no perfil de energia acústica é usada para aumentar a robustez do alinhamento temporal dinâmico (*Dynamic Time-Warping* (DTW)), associada a uma busca local "gulosa". Além disso, a métrica para tomada de decisão, tipicamente baseada em distâncias, é substituída pelo acúmulo de evidências, em bits de informação. Resultados de comparações deste método com os tradicionais ilustram as vantagem da proposta.

Palavras-chave— Tele-vigilância médica, Reconhecimento de orador, Biometria, DTW robusto, Evidência.

1 Introdução

A tele-vigilância médica é um campo de aplicação estimulante para muitas das técnicas de reconhecimento automático de padrões. Através da análise e classificação automática de dados coletados por dispositivos de captura de sinais vitais, imagens, sons ou movimento, apenas para citar alguns, é possível alimentar um sistema de monitoramento do paciente em seu próprio domicílio. Um tal sistema, por sua vez, pode decidir, de forma automática, quando disparar um alarme ou solicitar ajuda através da linha telefônica ou da Internet.

Em particular, o reconhecimento de situações de risco através da análise dos sons no ambiente de um paciente com mobilidade (total ou parcial) apresenta as vantagens do baixo custo e da comodidade para o paciente, pois não depende de dispositivos presos ao corpo. De fato, esse foi o foco escolhido no projeto AtenSom (MCT/CNPq N º 14/2010), no âmbito do qual este trabalho vem sendo desenvolvido. Dentre as situações de risco mais importantes, a queda se destaca, pois lesões decorrentes de quedas são uma causa importante de morte entre os idosos (McClure et al., 2008) e se estima que esse tipo de acidente se tornará cada vez mais freqüente com o fenômeno mundial de envelhecimento da população.

As quedas ocorrem devido à perda de equilíbrio postural e podem ser decorrentes de problemas primários do sistema ósteo-articular e/ou neurológico, ou, secundariamente, de uma condição clínica que afete os mecanismos do equilíbrio, sendo considerada um evento sentinela, sinalizador do início do declínio da capacidade funcional, ou o sintoma de uma nova doença (Buksman et al., 2008). A estimativa da incidência de quedas por faixa etária é de 28% a 35%, nos idosos com mais de 65 anos, e de 32% a 42%, naqueles com mais de 75 anos(Buksman et al., 2008). Estudos prospectivos têm mostrado que, em metade dos idosos que já sofreram quedas em domicílio, o problema é recorrente. Isto é, os que já sofreram uma queda apresentam risco maior de novas quedas (60 a 70% no ano subsequente). Sabe-se também que entre 40% e 60% desses episódios levam a algum tipo de lesão, sendo que 30% a 50% de menor gravidade, 5% de fraturas e 5% a 6% de injúrias mais graves (Buksman et al., 2008).

O desenvolvimento de métodos para a detecção de quedas tem se multiplicado nos últimos anos, a maioria deles assentados na premissa de que idosos que vivem sozinhos têm um risco aumentado de morte devida à inabilidade de obter

ajuda após uma queda (Popescu et al., 2008; Li et al., 2011). Considerando que, nas quedas de maior gravidade, o indivíduo pode permanecer deitado, inconsciente ou impedido de se locomover e que a demora no atendimento médico e hospitalização podem aumentar o risco de mortalidade em algumas condições clínicas, a funcionalidade mais desejável em um método de detecção de quedas é a possibilidade de acionamento automático do cuidador ou do serviço de saúde. Os sistemas atualmente disponíveis são baseados em 2 tipos de dispositivos: com ou sem contato com o indivíduo (Popescu et al., 2008; Li et al., 2011). São exemplos de dispositivos com contato os acelerômetros e giroscópios. Entre os dispositivos sem contato, podemos destacar os sensores de vibração do piso, vídeo câmeras, câmaras infravermelhas e tapetes "inteligentes". A maioria desses sistemas requer investimentos em *hardware*, *software*, instalação e treinamento (Sposaro and Tyson, 2009). Não obstante suas diferenças de implementação, todos compartilham os mesmos requerimentos: confiabilidade, facilidade de uso e restrição de falsos positivos, possivelmente o maior desafio enfrentado por todos os sistemas de detecção de quedas (Sposaro and Tyson, 2009; Popescu et al., 2008).

Este artigo apresenta um sistema de detecção de pedido de socorro após um evento agudo, como as quedas, visando as situações não raras em que a vítima não consegue se locomover, mas ainda consegue falar. Para simplificar seu uso e baratear o sistema, ele tem sido concebido para ser embarcado em um telefone celular com plataforma Android, de custo acessível e amplamente disponível no mercado. Em operação, o dispositivo deve permanecer ligado, no ambiente, em estado de alerta, monitorando sinais acústicos continuamente. Na fase de desenvolvimento atual do protótipo, parcialmente relatada neste artigo, busca-se um sistema capaz de detectar e reconhecer palavras-chaves, tais como comandos e expressões de desconforto ou dor, bem como verificar a identidade do orador (verificação Biométrica). A verificação biométrica tem por finalidade evitar que sinais vindos de outras fontes sonoras (de uma televisão ligada, por exemplo) disparem falsos alarmes com uma freqüência intolerável para uma aplicação real. Nas fases subsequentes do projeto, ainda serão incluídas as probabilidades *a priori* mencionadas nesta introdução, em estruturas Bayesianas de tomada de decisão que também buscam reduzir as taxas de falsos alarmes.

Por hora, nos restringimos às tarefas de reconhecer palavras e oradores. Para isso, tanto um sistema de reconhecimento de orador quanto o de comandos vocálicos necessitam de fases de treinamento, ou ajuste dos modelos. Sabe-se que o segundo pode ser adaptado a partir de um modelo geral, um *Universal Backgroung Model* (UBM) (Reynolds et al., 2000), previamente treinado com amostras de vozes diversas. Mas, o reconhecimento de orador, em contraste, necessariamente envolve a coleta de amostras de vozes do orador-alvo, pronunciando repetidas vezes, sejam os comandos-alvos, sejam falas livres.

Embora as sessões de coletas de amostras — também chamadas no jargão da biometria computacional de *enrolment* — sejam viáveis em muitas das aplicações conhecidas, solicitar a um paciente e/ou idoso que repita dezenas de vezes cada comando a ser reconhecido pode inviabilizar o uso de biometria. De fato, vencer essa limitação é a meta principal deste trabalho (como parte essencial do projeto no qual se insere), concentrando esforços no sentido de maximizar o aproveitamento de uma única amostra de voz por comando alvo.

Os resultados desses esforços são apresentados neste texto como seguem: na Seção 2, são modeladas as tarefas de reconhecimento de orador e de palavra, enquanto que, na Seção 3, são detalhados os procedimentos de extração de características e subamostragem temporal dos vetores de MFCCs. Na Seção 4, o critério do acúmulo de evidências é delineado. A Seção 5 apresenta as bases que foram usadas no ajuste e teste dos novos métodos propostos. Finalmente, a Seção 6 apresenta os resultados experimentais deste trabalho.

2 Modelagem do problema e de suas restrições

Na exposição de conceitos deste artigo, consideramos um conjunto fechado de C palavras pronunciadas por K oradores. Isso claramente não modela uma situação de vigilância real onde, embora o conjunto de palavras a serem detectadas possa ser finito, o mesmo não se aplica ao conjunto de oradores, que é potencialmente ilimitado. Não obstante, os conjuntos finitos usados neste trabalho permitem uma avaliação objetiva dos ganhos comparativos de desempenho dos novos métodos propostos. Por essa razão, os experimentos relatados na Seção 6 tratam apenas de conjuntos finitos de oradores.

Cada comando c, pronunciado por cada orador k, é associado a uma fonte estocástica de sinais acústicos, de tal forma que cada sinal coletado, $x_{c,k}(n)$, é visto como uma instância de um processo estocástico distinto, indexado por c e k, onde $n = 1, 2, 3, \ldots, N_x$ é um contador de amostras digitalizadas de som limitado em tempo, $c = 1, 2, \ldots, C$ indica um comando em um conjunto fechado de C comandos e $k = 1, 2, \ldots, K$ indica o orador em um conjunto fechado de K indivíduos. Por conveniência de notação, os índices c e k só serão indicados quando necessários.

Assumindo que há uma medida de similaridade conveniente, $J(x_{c,k}, x_0)$, entre dois sinais, onde x_0 corresponde a um sinal cuja fonte é desconhecida (vale notar que dois sinais podem possuir

ISBN: 978-85-8001-069-5

números de amostras diferentes), se restringirmos o número de sinais de referência a apenas 1 por orador, por comando, podemos definir duas tarefas de detecção:

- do comando, se $\max_c(J(x_0, x_{c,k})) > \lambda_C$
- do orador, se $\max_k(J(x_0, x_{c,k})) > \lambda_K$

onde, respectivamente, λ_C e λ_K são limiares de detecção de comando e orador.

Uma estratégia usual para extração de características de sinais de voz é a segmentação do sinal em janelas curtas e superpostas de poucos milisegundos, e o posterior mapeamento dos segmentos de sinais em vetores de coeficientes mel-cepstrais (Mel-frequency Cepstral Coefficients - MFCC) (Bridle and Brown, 1974; Mermelstein, 1976). Dessa forma, de cada sinal obtemos uma seqüência de vetores MFCC, que denotaremos aqui como matrizes, $X_{c,k}$, onde a j-ésima coluna representa o vetor MFCC para a j-ésima janela curta de sinal. Neste trabalho, usamos o algoritmo de extração de MFCC proposto por Malcolm Slaney (disponível publicamente desde 1993[1])

Vale notar que, assim como os sinais no tempo, essas matrizes podem não ter o mesmo número de colunas. De fato, mesmo duas instâncias da mesma palavra, pronunciadas pelo mesmo orador, dificilmente terão a mesma duração em segundos. Mais grave ainda, dificilmente os fonemas equivalentes nas duas instâncias ocorrerão de forma sincronizada, o que gera um problema bem conhecido em reconhecimento de fala: o problema do alinhamento temporal.

Até a década de 80, a solução mais usada para alinhar sinais (ou vetores de características) foi o Dynamic Time Warping (DTW) (Itakura, 1975; Sakoe and Chiba, 1978), que foi posteriormente substituído com vantagens por métodos baseados em modelos de Markov. Notadamente, o *Hidden Markov Model* (HMM) (Rabiner, 1989) tem sido, desde então, o modelo preferido nas tarefas de reconhecimento de fala, enquanto modelos que não levam em conta a estrutura temporal do sinal, tais como as *Gassian Mixture Models* (GMM) (Reynolds and Rose, 1995; Reynolds et al., 2000) são preferidas no reconhecimento de orador.

Na contramão dessa tendência, a restrição extrema de termos apenas uma amostra de sinal por comando/orador nos remete de volta ao DTW. Ao contrário do HMM, que usa modelos com muitos parâmetros livres, e que, por isso, demanda grandes volumes de dados de treinamento (o mesmo valendo para os GMM usuais), o DTW é mais adequado à comparações entre sinais dois-a-dois, o que é precisamente a situação na qual a restrição do problema nos coloca, quando há apenas dois sinais a serem comparados: um de referência, e outro desconhecido.

Em termos de notação, o alinhamento entre duas matrizes de coeficientes MFCC, X_a e X_b, pelo método DTW, é representado por $w_{DTW}(n) \in \{1, 2, 3, \ldots, N_b\}$, $n = 1, 2, 3, \ldots, N_a$, tal que:

$$J(w) = \left(\frac{\sum_{n=1}^{N_a} dist\left[\mathbf{x}_a(n) - \mathbf{x}_b\left(w_{DTW}(n)\right)\right]}{N_a} \right)^{-1} \quad (1)$$

é o critério a ser maximizado, dado pelo inverso da distância acumulada ao longo do *caminho de alinhamento*, w_{DTW}. A medida de distância, $dist(\cdot)$, usada neste trabalho é a Euclideana, e os vetores $\mathbf{x}_a(n)$ e $\mathbf{x}_b(m)$ representam, respectivamente, a n-ésima coluna de X_a e a m-ésima coluna de X_b.

3 Subamostragem baseada em perfil de energia

Como método de referência, usamos aqui o DTW com as restrições propostas por Itakura, em 1975 (Itakura, 1975). Alternativamente, propomos alterações que se mostraram particularmente robustas a situações em que as dinâmicas dos sinais de teste e referência diferem muito (e.g. pronúncias muito rápidas comparadas a pronúncias com alongamento anormal de algumas vogais). Essas alterações podem ser sintetizadas no seguinte algoritmo, aplicável a cada sinal:

- obter os perfis de energias, $\rho_a(n)$ e $\rho_b(m)$, correspondentes às primeiras linhas de X_a e X_b, respectivamente (i.e. coeficiente MFCC de ordem 0);
- obter as variações positivas de energias correspondente, $\delta_a(n)$ e $\delta_b(m)$, de acordo com:

  $$\delta(i) = \begin{cases} \rho(i+\Delta_i) - \rho(i), ; \rho(i+\Delta_i) > \rho(i) \\ 0, \qquad\qquad\qquad\quad \rho(i+\Delta_i) \leq \rho(i) \end{cases}$$

  onde $\Delta_i = 5$ para uma taxa de amostragem de 8000 Hz e $\rho(i) = 0$ para todo i maior que o número de colunas da matriz X.

  O sinal δ tende a indicar, com valores mais altos, o início de sílabas (embora sílabas iniciadas com vogais não possuam necessariamente essa característica).
- combinar os perfis de energia às variações de perfis: $\rho \leftarrow \rho + 2\delta$
- acumular os perfis de energia, de acordo com:

  $$\alpha(i) = \frac{1}{\sum_{\forall j} \rho(j)} \sum_{j=1}^{i} \rho(j)$$
- sub-amostrar cada seqüência de MFCC de acordo com:

  $$\tilde{X}(k) = X(i_k)$$

[1] https://engineering.purdue.edu/∼malcolm/interval/1998-010/

ISBN: 978-85-8001-069-5

onde $i_k = \arg\min_i(k\Delta_\alpha - \alpha(i))$, $0 < \Delta_\alpha < 1$ e $k = 1, 2, 3 \ldots, \frac{1}{\Delta_\alpha}$.

Dessa forma, com uma escolha apropriada do parâmetro (passo arbitrário) Δ_α, é possível se subamostrar as colunas das matrizes de MFCC, X_a e X_b, convertendo-as em $\tilde{X}_a$ e $\tilde{X}_b$, com o mesmo número fixo de colunas, igual a $N_0 = 1/\Delta_\alpha$. Como essa subamostragem é guiada pelo perfil de energia, ela produz um pré-alinhamento de energias acumuladas, de tal forma que podemos usar o vetor $w_0 = [12 \ldots N_0]$ como um candidato inicial ao alinhamento entre $\tilde{X}_a$ e $\tilde{X}_b$. A figura 1 ilustra como o perfil de energia original é deformado, fazendo com que regiões com perfis intensos sejam alongadas, ao passo que regiões com baixa energia sejam abreviadas.

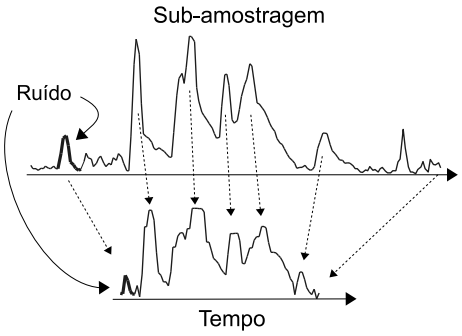

Figura 1: Ilustração da subamostragem não-linear através do perfil de energia. Detalhamento do pulso de energia de um ruído de batida do microfone.

Em seguida, para cada posição de alinhamento em w_0, podemos testar, de forma sistemática (*greedy search*, no sentido usado em algoritmos de otimização), alguns caminhos de alinhamento vizinhos. Neste trabalho, optamos arbitrariamente por permitir uma busca sobre os 3 caminhos mais próximos de cada lado do caminho atual, retendo aquele que maximiza localmente o critério $J(\tilde{X}_a, \tilde{X}_b)$. Repetindo essa busca "greedy" um número finito de vezes, obtém-se um ajuste iterativo do caminho original w_0, gerando assim um caminho localmente otimizado que denotamos como w_{SGTW}, onde o 'S' faz referência ao processo de subamostragem, e o G, à busca "greedy". Neste trabalho, optamos arbitrariamente por iterar a busca "greedy" apenas 10 vezes para cada par de sinais comparados.

Alternativamente, também podemos aplicar o DTW tradicional tanto às matrizes X_a e X_b (método DTW usual), quanto às matrizes subamostradas $\tilde{X}_a$ e $\tilde{X}_b$ (método SDTW). A figura 2 ilustra os resultados obtidos com os três métodos, para dois sinais correspondentes às mesmas 3 sílabas, pronunciadas pelo mesmo orador, mas de maneiras propositalmente discrepantes. Isto é, no sinal associado ao eixo horizotal, as vogais foram alongadas arbitrariamente, enquanto que o sinal associado ao eixo vertical teve suas vogais abreviadas. Claramente, o método DTW, com as

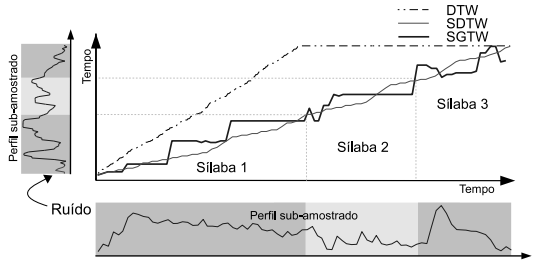

Figura 2: Ilustração dos desempenhos dos DTW, SDTW e SGTW, no alinhamento de um sinal muito longo a outro, muito curto.

restrições de Itakura, não foi capaz de encontrar uma solução de alinhamento aceitável, enquanto que os dois métodos baseados na subamostragem encontraram alinhamentos semelhantes e visualmente aceitáveis.

4 Acúmulo de evidências

Após o alinhamento de duas seqüências de vetores, ou duas matrizes de características, representando dois sinais, é usual se medir a "distância" entre os dois sinais como a média de todas as distâncias pareadas entre vetores, o que nos remete ao critério definido na Eq. 1. Isto é, o critério é baseado na média das distâncias pontuais entre vetores pareados, ou resíduos, representadas como:

$$r(n) = ||\mathbf{x}_a(n) - \mathbf{x}_b(w(n))||$$

No entanto, neste trabalho, estudamos uma medida alternativa ao acúmulo de distâncias. Se assumirmos que o alinhamento entre duas instâncias do mesmo processo estocástico é satisfatório, então cada instância $r(n)$ é uma *peça de evidência* independente das demais, na tomada de decisão. Isto é equivalente a assumir que a dependência temporal entre vetores de características já foi, por hipótese, completamente capturada e usada no processo de alinhamento temporal, de forma análoga a um processo de "branqueamento" estatístico. Assim, estudamos e modelamos probabilisticamente o comportamento de r para os casos em que X_a e X_b são extraídos de quatro situações de interesse, a saber, distribuição de r quando:

A orador e palavra são os mesmos,

B os oradores são distintos mas a palavra pronunciada é a mesma,

C o orador é o mesmo mas as palavra são distintas,

D oradores e palavras pronunciadas são distintos.

Nos quatro casos, o resíduo apresentou um comportamento aproximadamente Log-Normal[2]. (Keeping, 1999), como ilustrado na figura 3.

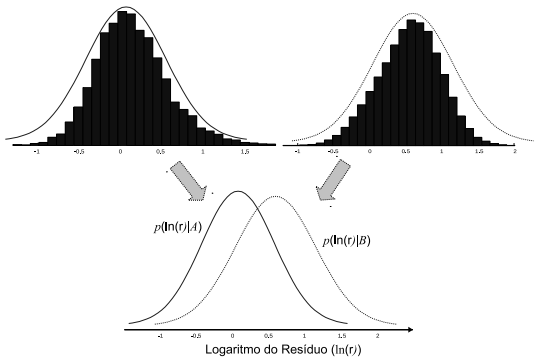

Figura 3: Ilustração das distribuições de $log_e(r)$, condicionados à mudança de orador para uma mesma palavra (situações A e B, respectivamente).

Experimentalmente, com os mesmos dados descritos na seção 6, podemos estimar os parâmetros das quatro densidades (i.e. $p(ln(r)|A)$, $p(ln(r)|B)$, $p(ln(r)|C)$ e $p(ln(r)|D)$), aproximadas como Normais, respectivamente μ_A, σ_A, μ_B, σ_B, μ_C, σ_C, μ_D e σ_D. Em seguida, dada a assumida independência de cada *peça de evidência*, $r(n)$, podemos, nas tarefas de detecção, seja de orador ou de comando, considerar duas hipóteses:

H_0 : O sinal analisado é um ruído, ou um sinal qualquer, não pertencente à classe do sinal de referência.

H_1 : O sinal analisado é um sinal alvo, da mesma classe do sinal de referência.

Assim, dada uma seqüência de K instâncias de r, resultantes da comparação de um sinal qualquer ao *único* sinal disponível como referência para a classe, temos que cada $r(n)$ fornece uma peça de evidência em favor de H_1, em *bits de informação*, dada por (MacKay, 2003): $b(n) = \log_2\left(\frac{p(r(n)|H_1)}{p(r(n)|H_0)}\right)$, e o critério para comparação de sinais resultante é finalmente dado por:

$$J_{ev}(\tilde{X}_a, \tilde{X}_b) = \sum_{n=1}^{N_0} b(n) \qquad (2)$$

[2]Muitos dados empíricos seguem aproximadamente leis Log-Normais, tais como as medidas de pressão sanguínea de seres humanos, o tempo de sobrevida de uma bactéria em presença de desinfetantes, e até o número de palavras em sentenças escritas por George Bernard Shaw (Keeping, 1999)

5 Base de dados, pré-processamento e extração de características

Os teste relatados aqui foram realizados com duas bases, A e B, de 5 palavras curtas, pronunciadas 10 vezes por cada um dos 8 oradores da base A (6 homens e 2 mulheres), bem como pelos 20 oradores da base B (14 homens e 6 mulheres) com idades entre 19 e 60 anos, num total de 28 oradores e 50 registros por orador (1400 registros). As amostras foram coletadas em ambientes não controlados, como domicílios e salas de aulas, numa taxa de 8000 amostras por segundo, e quantização de 16 bits por amostra. A aquisição das amostras foi feita com dispositivos móveis (como smartphones), usando seus respectivos microfones embutidos.

As palavras pronunciadas por cada orador são os comandos, em português: *avance*, *direita*, *esquerda*, *pare* e *recue*. Infelizmente, este conjunto de palavras não corresponde a expressões de pedidos de ajuda, o que seria preferível. De fato, uma nova base com simulações de tais pedidos de ajuda está em preparação, no âmbito do projeto AtenSom. No entanto, para o teste do método proposto neste trabalho, os comandos curtos e a forma não controlada como os sinais foram adquiridos servem à comparação com métodos tradicionais.

A principal diferença entre a base A e a base B é que, na aquisição da base A, os voluntários não receberam nenhuma instrução de como deveriam pronunciar as palavras, ao passo que, antes da aquisição das amostras da base B, cada voluntário foi instruído a repetir as palavras com ritmos e entonações variadas.

Cada sinal foi sistematicamente pré-enfatizado (filtro: $H(z) = 1 - 0.97z^{-}1$) e segmentado em blocos de 256 amostras, com avanço de 80 entre blocos consecutivos (superposição de $\approx 70\%$). Cada bloco foi então atenuado através de uma janela de Hamming e, finalmente, mapeado em 13 coeficientes cepstrais (13 MFCC). Isto é, cada sinal x, com N_x amostras, foi mapeado em uma matriz com 13 linhas e, aproximadamente, $(N_x - 256)/80$ colunas.

O vetor correspondente à primeira linha da matriz de coeficientes é o coeficiente de ordem zero, que expressa a energia acústica de cada bloco de sinal, que é usada no processo de subamostragem (ver seção 3) e descartada da montagem da matriz de características, $\tilde{X}$, com 12 linhas e um número de colunas fixo $N_0 = 83$. São essas matrizes, 12×83, que são entregues, duas-a-duas, como entradas ao processo de alinhamento. Mais precisamente, uma das matrizes sempre corresponde a um sinal de exemplo, uma referência (da classe palavra=c, ou da classe orador=k, dependendo da tarefa), enquanto que a outra corresponde a um sinal desconhecido, que deve ser classificado como sendo ou não da mesma classe do sinal de referên-

ISBN: 978-85-8001-069-5

cia.

6 Resultados Experimentais

A base B foi usada no ajuste dos parâmetros e a base A foi usada para testes. Os parâmetros obtidos com a base B estão apresentados na tabela 1.

Tabela 1: Médias e desvios estimados na base B.

Situação (ver seção 4)	μ	σ
A	0.09	0.43
B	0.54	0.43
C	0.72	0.57
D	0.81	0.52

Para simular uma situação em que um paciente fornece apenas uma amostra de cada comando (restrição extrema que guia este trabalho), tanto na fase de ajuste dos parâmetros quanto na etapa de testes, uma única amostra de voz foi selecionada aleatoriamente por vez, e separada como sendo 'a referência de treinamento', juntamente com os rótulos representando o comando pronunciado e o orador. Em seguida, as demais amostras da base foram amostradas aleatoriamente e comparadas, uma-a-uma, à referência, gerando medidas de similaridades que foram registradas e testadas contra os limiares de decisão.

Para uma apresentação sucinta dos desempenhos comparados, optamos por variar os limiares de decisão até que as taxas de falsos positivos e falsas rejeições se igualassem, em cada sessão de testes. Essa taxa de erros iguais, ou *Equal Error Rate* EER é uma medida simples que indica de forma compacta o desempenho aproximado de cada detector – quanto menor o EER, melhor o detector.

Também, para abreviar a apresentação de resultados, destacando apenas as tarefas de detecção mais difíceis, nas comparações entre palavras iguais proferidas por oradores distintos (reconhecimento biométrico), apenas oradores do sexo masculino foram amostrados, pois a diferença de timbres entre sexos poderia facilitar a detecção do orador-alvo. Além disso, na tarefa de detecção de palavras, apenas amostras do usuário de referência foram sorteadas em cada simulação. Dessa forma, também dificultamos a tarefa do detector, forçando-o a trabalhar sempre com o mesmo timbre de voz.

Para cada medida de EER, foram realizadas 3 baterias de simulação, com aproximadamente 1500 sorteios de pares de amostras. Para cada par de amostra, os métodos usados foram o SDTW (DTW com restrições de Itakura, aplicado às matrizes subamostradas), sob o critério descrito na equação 1, e o SGTW (busca *Greedy* a partir do alinhamento w_0 das matrizes subamostradas), sob o critério do acúmulo de evidências. Vale notar que ambos usam a subamostragem pelo perfil de energia, pois o DTW convencional, sem a subamostragem, produz resultados aceitáveis apenas para a base A, mas leva a resultados muito ruins para a base B.

A figura 4 ilustra, através de resultados de um dos experimentos com a base B (mais difícil), as tendências antagônicas das taxas de falsas detecções e falsas rejeições, em função do limiar de detecção, em bits de evidência.

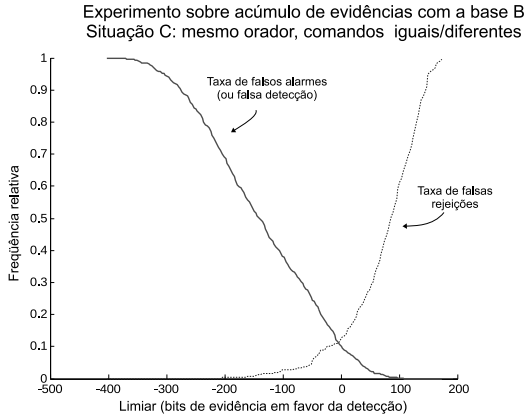

Figura 4: Ilustração da relação de compromisso entre falso alarme e falsa rejeição de comandos (situação C). Vale notar que os valores negativos de evidências em favor de H_1, no eixo horizontal, correspondem a evidências positivas em favor de H_0.

A tabela 2 apresenta os resultados médios obtidos para os 5 comandos pronunciados apenas por oradores masculinos e femininos das duas bases, e comparados através dos dois métodos concorrentes.

Tabela 2: Resultados (EER) no reconhecimento de comandos.

Método	Base A	Base B
SDTW	0.10 ± 0.01	0.13 ± 0.03
SGTW	0.04 ± 0.01	0.09 ± 0.02

A tabela 3 apresenta os resultados médios obtidos na detecção biométrica dos oradores masculinos nas duas bases, e comparados através dos dois métodos concorrentes. Juntamente com os EER

Tabela 3: Resultados (EER) no reconhecimento de oradores.

Método	Base A	Base B
SDTW	0.16 ± 0.02	0.19 ± 0.03
SGTW	0.06 ± 0.01	0.11 ± 0.04

ISBN: 978-85-8001-069-5

estimados em cada caso, também são apresentados, nas tabelas 2 e 3, os respectivos intervalos de 95% de confiança.

7 Conclusões

Um método para detecção de orador e comando, com base em apenas uma amostra de voz por comando, foi proposto neste artigo. Os resultados, em termos de EER, evidenciam o melhor desempenho do método proposto. Seus principais melhoramentos, em relação ao método baseado em DTW, são, em ordem de importância, a subamostragem dos vetores de características pelo perfil de energia, a busca *greedy* da solução de alinhamento, a partir do alinhamento inicial resultante da subamostragem, e o acúmulo de evidências (em bits de informação) em lugar do acúmulo de distâncias Euclideanas.

O principal resultado, em termos de robustez, foi decorrente da subamostragem pelo perfil de energia, sem o que, a base B (na qual os oradores foram induzidos a alongar e abreviar sílabas de forma não natural) não seria tratável. Em seguida, a busca *greedy* superou, em termos de EER, o desempenho do DTW com as restrições de Itakura.

Em contraste, o critério baseado em acúmulo de evidência propiciou um ganho mais sutil, em termos de EER, se comparado aos efeitos conjugados da subamostragem e da busca *greedy*. No entanto, o acúmulo de evidências fornece uma formulação adequada à restrição do efeito, por exemplo, de ruídos como o bater de portar e passos, registrados simultaneamente aos pedidos de ajuda. Isto é, no caso de ruídos cuja energia se concentra em intervalos curtos de tempo, algumas peças de evidências podem assumir valores fortemente negativos (evidências em favor de H_0). Como continuação deste trabalho, pretendemos estabelecer limites (valores de saturação) às peças de evidências, de forma a limitar o efeito dos ruídos curtos e fortes sobre a detecção de comandos e orador, aumentando assim a robustez do sistema final.

Agradecimentos

Este trabalho contou com o apoio financeiro do CNPq. Também agradecemos aos voluntários que forneceram suas amostras de vozes às bases de dados usadas nos experimentos.

BIBLIOGRAFIA

Referências

Bridle, J. S. and Brown, M. D. (1974). An experimental automatic word-recognition system, *Technical report*, JSRU, Joint Speech Research Unit.

Buksman, S., Vilela, A., Pereira, S., Lino, V. and Santos, V. (2008). Quedas em idosos: Prevenção, *Technical report*, Sociedade Brasileira de Geriatria e Gerontologia, AMBCFM.

Itakura, F. (1975). Minimum prediction residual applied to speech recognition, *IEEE Trans. Acoustics, Speech, Signal Proc.* .

Keeping, E. (1999). *Introduction to Statistical Inference*, Dover Publication, Inc., New York.

Li, Y., Popescu, M., Ho, K. C. and Nabelek, D. P. (2011). Improving acoustic fall recognition by adaptive signal windowing, *Proceedings of 33st Annual International Conference of the IEEE EMBS*.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

McClure, R., Turner, C., Peel, N., Spinks, A., Eakin, E. and Hughes, K. (2008). Population-based interventions for the prevention of fall-related injuries in older people (review), *Technical report*, The Cochrane Collaboration, Wiley Pub.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental, *Pattern Recognition and Artificial Intelligence* .

Popescu, M., Li, Y., Skubic, M. and Rantz, M. (2008). An acoustic fall detector system that uses sound height information to reduce the false alarme rate, *Proceedings of 30st Annual International Conference of the IEEE EMBS*.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, pp. 257–286.

Reynolds, D., Quatieri, T. and Dunn, R. (2000). Speaker verification using adapted gaussian mixture models, *IDigital Signal Processing* .

Reynolds, D. and Rose, R. (1995). Robust text-independent speaker identification using gaussian mixture speaker models, *IEEE Trans. on Speech and Audio Processing* .

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing* .

Sposaro, F. and Tyson, G. (2009). ifall: An android application for fall monitoring and response, *Proceedings of 31st Annual International Conference of the IEEE EMBS*.