Uma Ferramenta para a Autenticação de Áudio com Aplicação em Fonética Forense

Daniel Patricio Nicolalde Rodríguez, José Antonio Apolinário Junior (Pós-graduação em Engenharia Elétrica, Instituto Militar de Engenharia, Rio de Janeiro, RJ, Brasil, E-mails: danielnicolalde@hotmail.com, apolin@ime.eb.br) e Luiz Wagner Pereira Biscainho (PEE/COPPE, Universidade Federal de Rio de Janeiro, Rio de Janeiro, RJ, Brasil, E-mail: wagner@lps.ufrj.br).

XX CONGRESSO NACIONAL DE CRIMINALÍSTICA
III CONGRESSO INTERNACIONAL DE PERÍCIA CRIMINAL

Resumo

Este trabalho propõe uma ferramenta de autenticação de áudio com aplicação em fonética forense. Esta ferramenta visa a determinar se um sinal de áudio foi ou não digitalmente editado (na forma de cortes e/ou inserções no conteúdo). O método se baseia em verificar o comportamento da frequência da rede elétrica, quase sempre embutida nas gravações efetuadas por equipamentos conectados à rede. São fornecidos um mecanismo visual para a detecção de edição e um mecanismo automático para discriminar sinais originais de editados. A técnica proposta foi avaliada em dois bancos de dados digitalmente editados: o primeiro deles apresentando condições favoráveis e o outro condições desfavoráveis. Apresentam-se os fundamentos teóricos e algumas importantes considerações práticas.

Palavras-Chave

Fonética forense, autenticação de áudio digital, frequência da rede elétrica.

I. Introdução

A fonética forense é utilizada no campo da criminalística sempre que existam sinais de voz que possam servir como evidência para a reconstituição de fatos ocorridos em um determinado caso. Ela pode então auxiliar na definição da inocência ou culpa dos implicados. As provas periciais, neste caso, serão conversações gravadas diretamente de um microfone ou oriundas de ligações telefônicas. Hoje em dia, com as facilidades da tecnologia digital, alterar, editando de alguma maneira, o conteúdo de sinais de

áudio pode ser considerada uma atividade muito simples. Neste quadro, uma das tarefas da perícia fonética é avaliar a autenticidade de gravações de áudio para aceitá-las ou não como evidências em procedimentos legais [1], [2].

Os sistemas de geração, transmissão e distribuição de energia elétrica utilizam como padrão para o valor nominal da frequência da rede elétrica (ENF – *Electric Network Frequency*) 50 Hz ou 60 Hz. O Brasil adota a frequência de 60 Hz como padrão. Para o correto funcionamento da rede, são projetados sistemas que buscam o sincronismo das diversas geradoras de energia com o objetivo de manter a voltagem e a frequência dentro de limites aceitáveis. Por isto, considera-se o comportamento da ENF estável em torno do seu valor nominal.

Considerando que um campo eletromagnético é irradiado por todo tipo de equipamento elétrico, a ENF está embutida na maioria de gravações efetuadas por equipamentos conectados à rede elétrica. [3] e [4] usam a ENF para os seus métodos de autenticação de autenticação de áudio. Estes métodos determinam o lugar e o instante onde a gravação foi realizada, com base na comparação do comportamento da ENF no sinal de áudio com o comportamento do sinal proveniente de uma tomada. Estes métodos precisam de armazenamento de dados de sinais elétricos de diferentes regiões.

Este trabalho apresenta os fundamentos e o desenvolvimento de uma ferramenta específica para a autenticação de áudio baseada na presença da ENF. Embora exista no mercado internacional um (complicado e bastante custoso) sistema de autenticação de áudio [5], justifica-se um desenvolvimento nacional desta tecnologia pela necessidade de adaptação da ferramenta ao ambiente local, onde a ENF possivelmente varia de maneira menos uniforme.

A organização deste trabalho é a seguinte. A Seção II explica o método usado para a autenticação de áudio digital e a Seção III apresenta algumas considerações práticas do método. As conclusões são apresentadas na Seção IV.

II. Método Proposto

Como anteriormente mencionado, a ENF encontra-se embutida na maioria das gravações. Considerando que em edição de áudio digital podem ser feitos recortes e/ou inserções no conteúdo do sinal, os mesmos são também feitos na ENF embutida. A idéia fundamental é determinar, da maneira mais precisa possível, se existem mudanças abruptas de fase na ENF devidas às edições feitas no sinal de áudio.

O método proposto visa a mostrar uma técnica visual, assim como um mecanismo para discriminação automática de sinais originais e editados.

A. *Mecanismo Visual*

Este processo é explicado passo a passo [6]:

1) Filtrar o sinal de áudio para considerar somente as componentes do sinal que contêm a ENF (50 ou 60 Hz).
2) Segmentar o sinal filtrado em blocos de duração de 3 ciclos de ENF nominal com sobreposição de 2 ciclos com respeito ao bloco seguinte (ver Fig. 1).

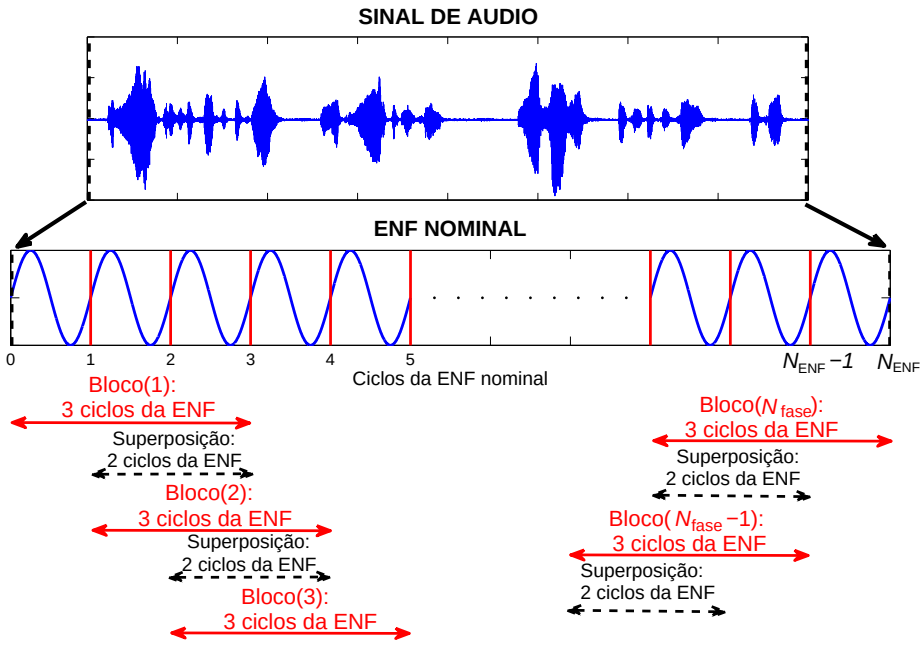

Fig. 1. Esquema de segmentação em blocos do sinal de áudio. N_{ENF} representa o número total de ciclos da ENF nominal contidos no sinal de áudio e N_{fase} o número total de blocos segmentados.

3) Obter a fase (ângulo em graus) da componente da frequência (achada via Transformada Discreta de Fourier de curta duração) no valor correspondente à ENF nominal (50 ou 60 Hz) em cada bloco segmentado do sinal. A este valor de fase no n-ésimo bloco denominamos $\phi(n)$. N_{fase} é o número total de blocos segmentados.
4) Calcular a variação da fase $\phi(n)$ entre duas amostras consecutivas (uma aproximação da derivada da fase no tempo contínuo), $\phi'(n)$. Os picos de $\phi'(n)$ ajudam na localização dos pontos de edição quando um sinal foi editado.

A Fig. 2 apresenta dois exemplos de edição de áudio: no primeiro (à esquerda) eliminou-se um fragmento do sinal e no segundo (à direita) inseriu-se um fragmento do sinal. Na eliminação de um fragmento, o ponto desse recorte possui uma mudança de fase significativa. A curva de estimação de fase ao longo do tempo, apresentada Fig. 2(c), ilustra este fenômeno. A inserção, por outro lado, provocou duas mudanças

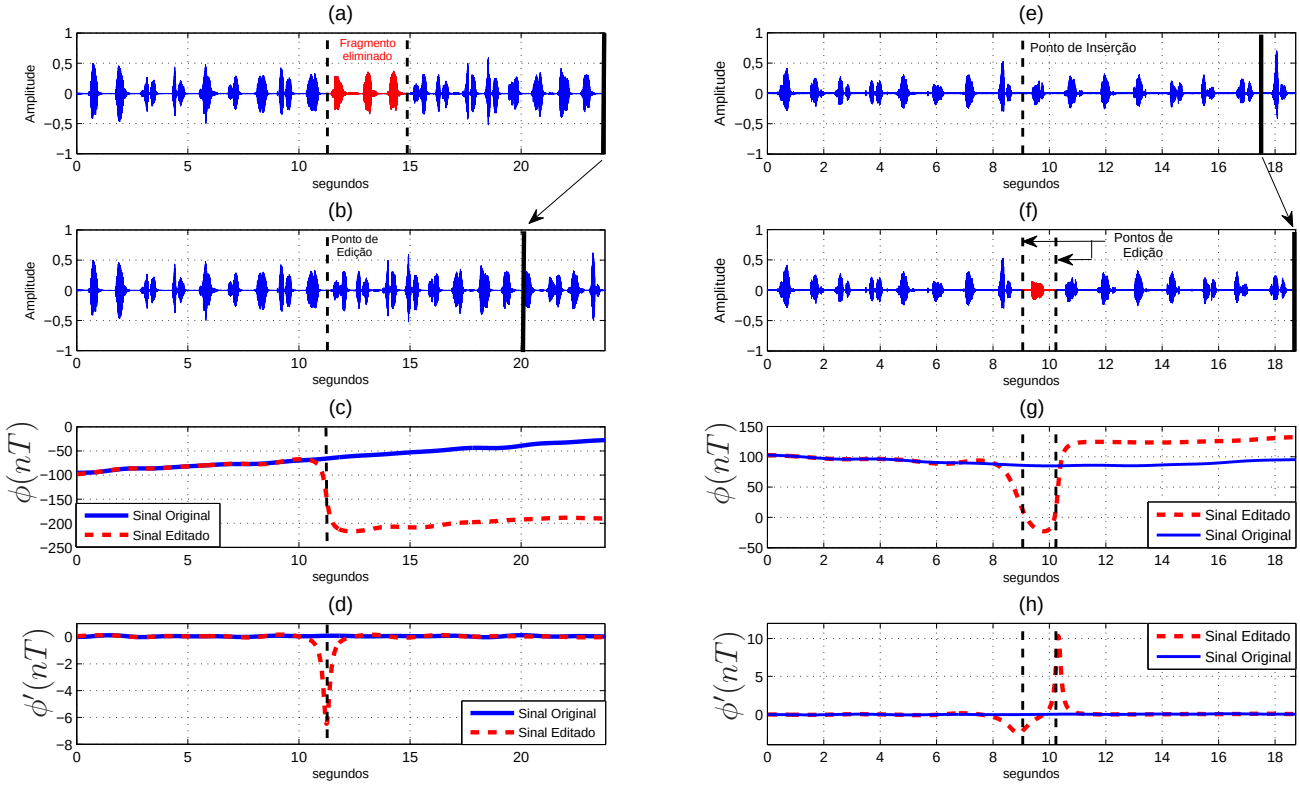

Fig. 2. Edição de áudio. Eliminação (à esquerda) e inserção (à direita) de um fragmento do sinal. O valor nominal da ENF é de 50 Hz, ou seja, $T = \frac{1}{50}$ segundo (1 ciclo da ENF nominal). (a) Sinal original (sinal em que será eliminado um fragmento); (b) Sinal editado (com um fragmento eliminado); (c) Curva de estimação de fase na ENF, $\phi(n)$; (d) Curva da variação de fase entre duas amostras consecutivas, $\phi'(n)$; (e) Sinal original (sinal em que será inserido um fragmento); (f) Sinal editado (com um fragmento inserido); (g) Curva de estimação de fase na ENF, $\phi(n)$; (h) Curva da variação de fase entre duas amostras consecutivas, $\phi'(n)$.

de fase. O resultado destas mudanças é apresentado na Fig. 2(g). Adicionalmente, nos casos de edição apresentados, os picos da curva $\phi'(n)$ nas Figs. 2(d) e 2(h) claramente ajudam na localização dos pontos de edição.

B. Mecanismo automático

Adicionalmente às ferramentas visuais apresentadas, é importante obter uma medida característica, M (conhecida pelo termo em inglês *feature*), para podermos automatizar o processo da autenticação de áudio mediante uma razão de verossimilhança, dada por:

$$M \underset{H_O}{\overset{H_E}{\gtrless}} \gamma, \tag{1}$$

onde H_O e H_E representam as hipóteses do que o sinal de áudio digital seja original e editado, respectivamente. Isto significa que com valores de M superiores a γ decide-se que o sinal analisado foi editado.

M proposta em [7] é obtida mediante a seguinte expressão:

$$M = 100 \log \left\{ \frac{1}{N_{fase} - 1} \sum_{n=2}^{N_{fase}} |\phi'(n) - m_{\phi'}| \right\}, \quad (2)$$

onde $m_{\phi'}$ representa o valor médio de $\phi'(n)$.

O valor do limiar, γ, é estabelecido mediante a preparação de um banco de dados com sinais editados e sua comparação com os sinais originais. Posteriormente, obtém-se o valor de M para todos os sinais da base (originais e editados). O limiar pode ser definido variando-se o valor de γ até um valor tal que a quantidade de sinais identificados como editados sendo na verdade autênticos ("falso alarme" de edição) seja igual à de sinais identificados como autênticos sendo na verdade editados ("perda" de edição). Isto corresponde a forçar erros iguais na decisão de sinais originais e editados. Somando o número de sinais que resultaram em decisão errada e dividindo o resultado pelo número total de sinais da base, obtém-se o erro percentual do sistema de autenticação.

Uma primeira avaliação do método foi realizada em sinais gravados sob condições controladas [7], em particular com pouca variabilidade da ENF, ausência de saturações e baixo ruído de fundo. Para tal avaliação, procedeu-se a editar um conjunto de sinais de voz provenientes de duas bases públicas em castelhano, AHUMADA e GAUDI, obtidas via http://atvs.ii.uam.es/databases.jsp [8]. A ENF característica é de 50 Hz, por ser uma base proveniente de Espanha. Os locutores usados foram 25 homens e 25 mulheres, produzindo cada um 2 sinais, o que significa um total de 100 sinais originais. Na sua edição, de 50 sinais eliminou-se um fragmento e nos restantes inseriu-se um fragmento do próprio sinal. Todas as edições foram feitas sem que se tomasse cuidado com as mudanças de fase da ENF nos sinais, como faria uma pessoa que desconhecesse este assunto específico. É importante destacar que quanto menor é a mudança de fase, mais complicada se torna a detecção. A distribuição das mudanças de fase dos sinais editados é uniforme entre $-180°$ e $+180°$.

Na Fig. 3 são apresentados os histogramas normalizados das distribuições da medida característica (Eq. 2) M para os sinais originais e editados deste banco de dados proveniente da Espanha. Nestes histogramas pode-se verificar uma boa separação entre as distribuições de M para o caso dos sinais originais e editados. O valor do limiar, γ, necessário para a decisão final é de $-113,81$. O erro obtido para o mecanismo de discriminação nesta base é de 7%.

III. Considerações Práticas

Os resultados apresentados na seção anterior foram obtidos a partir de um conjunto de sinais gravados sob condições bastante ideais. O pequeno erro de 7% possivelmente corresponde à probabilidade da pessoa que editou o sinal ter realizado (de modo não intencional) tal edição num ponto onde a mudança de fase

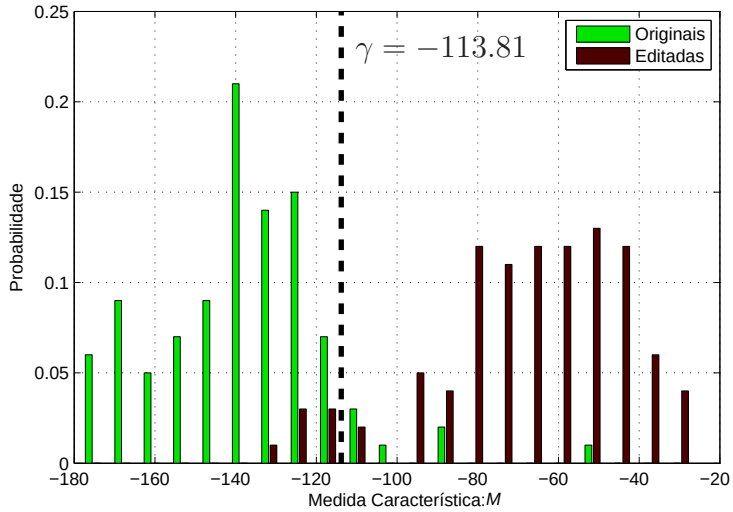

Fig. 3. Histograma normalizado da distribuição da medida característica M, para a banco de dados proveniente da Espanha.

é muito pequena ou nula. Em gravações realizadas na prática, num ambiente não-controlado, o sinal apresenta degradações causadas, por exemplo, pela inserção de ruído ou pela saturação do sinal (não-linearidade que também afeta a ENF). Além disso, uma gravação feita numa localidade onde a ENF varia mais rapidamente em torno do seu valor nominal pode prejudicar o desempenho da técnica proposta.

Com o objetivo de avaliar o mecanismo de discriminação automática em situações menos favoráveis, foi criado um segundo banco de dados com sinais digitalmente editados. Os sinais deste banco foram provenientes de ligações feitas no Rio de Janeiro, RJ. A ENF nominal é de 60 Hz. Os locutores convidados foram 50 homens e 50 mulheres, cada um gravando um sinal de áudio, num total de 100 sinais. Nessa base, há sinais que apresentam saturações, variações significativas da ENF e/ou ruído de fundo elevado. Esse banco, portanto, descreve situações reais, mais comuns nas gravações encontradas em aplicações forenses.

As correspondentes versões editadas dos sinais foram obtidas da mesma maneira do que no banco de dados proveniente da Espanha. O histograma da distribuição das mudanças de fase dos sinais editados é apresentado na Fig. 4, e indica distribuição uniforme entre $-180°$ e $+180°$.

Apresentamos na Fig. 5 os histogramas normalizados das distribuições da medida característica (Eq. 2) M para os sinais originais e editados deste banco de dados. O valor do limiar, γ, necessário para a decisão final foi de $-94,06$ e o erro do sistema de autenticação foi de 17%.

São mostrados a seguir dois exemplos de sinais deste banco de ligações feitas no Rio de Janeiro onde se constatam os problemas mencionados. A Fig. 6 apresenta um caso de edição de áudio onde eliminou-se um fragmento do sinal original. O sinal analisado está visivelmente saturado. Na curva de estimação de

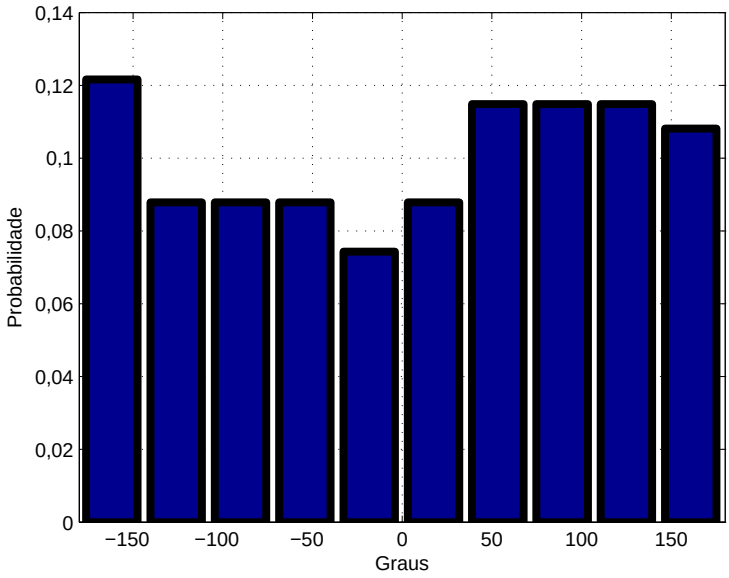

Fig. 4. Histograma normalizado da distribuição das mudanças de fase nos sinais editados para o banco de dados proveniente de ligações telefónicas feitas em Rio de Janeiro, RJ.

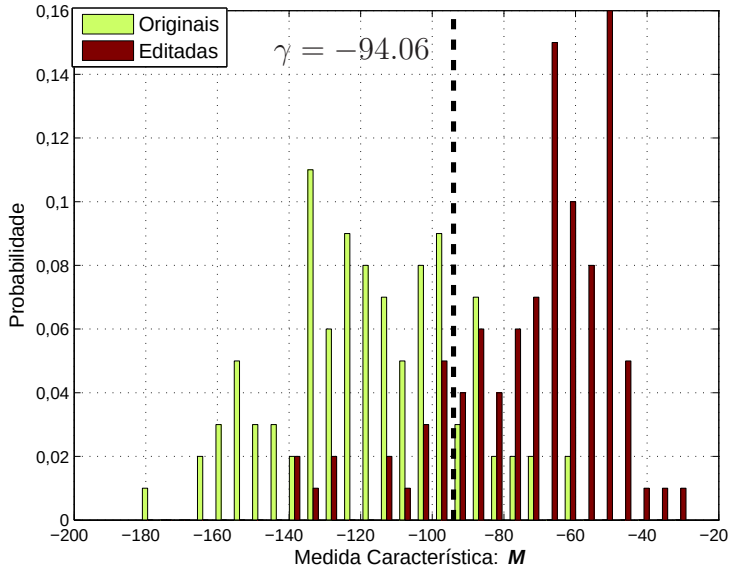

Fig. 5. Histograma normalizado da distribuição da medida característica M, para a banco de dados proveniente de ligações telefónicas feitas em Rio de Janeiro, RJ.

fase $\phi(n)$, além da mudança de fase provocada pela edição feita, existem outras variações que poderiam ser confundidas com possíveis edições. A curva $\phi'(n)$ apresenta um pico que sem corresponder a um ponto de edição, poderia ser tomado como tal. Como dado adicional, menciona-se que 21 dos 100 sinais da base apresentavam saturações significativas. A Fig. 7 apresenta outro caso de difícil análise de edição de

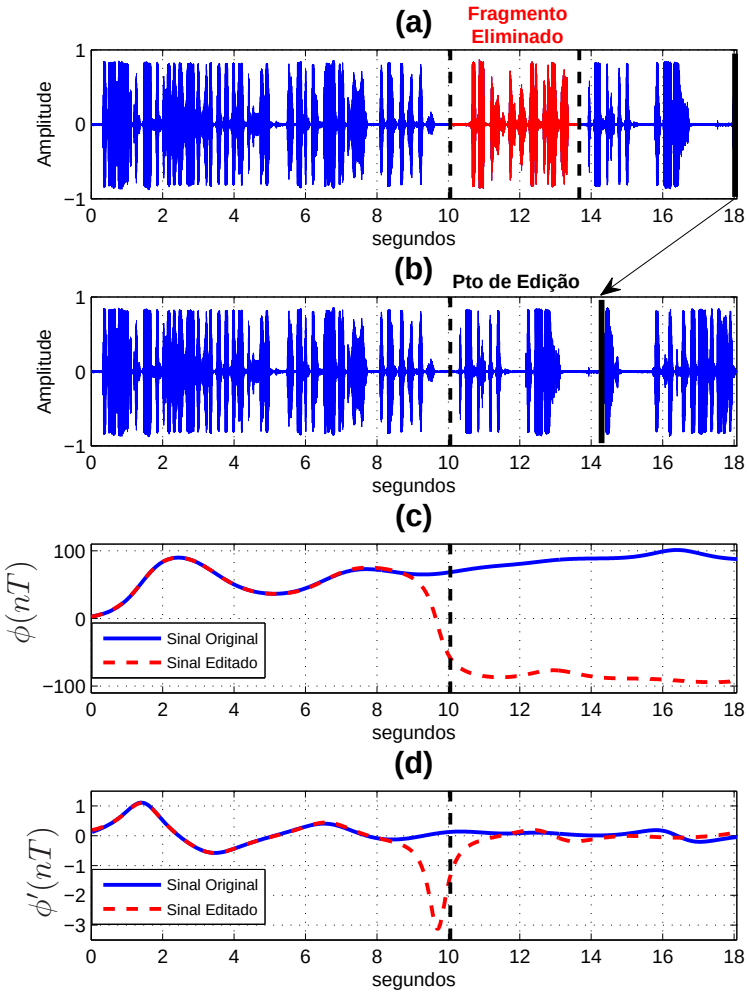

Fig. 6. Edição de áudio (Sinal saturado): eliminação de um fragmento do sinal. O valor nominal da ENF é de 60 Hz. $T = \frac{1}{60}$ segundo (1 ciclo da ENF nominal). (a) Sinal original; (b) Sinal editado; (c) Curva de estimação de fase na ENF, $\phi(n)$; (d) Curva da variação de fase entre duas amostras consecutivas, $\phi'(n)$.

áudio, onde as curvas de estimação de fase $\phi(n)$ tanto do sinal original quanto do sinal editado apresentam variações devidas às flutuações temporais da ENF. Neste caso, também é complicado determinar se o sinal sofreu ou não edição.

Para fazer uma análise mais sistemática do efeito da presença de saturações nos sinais na detecção, recorreu-se ao banco de dados proveniente da Espanha (com sinais livres de saturação) como ponto de partida. Primeiramente, aplicou-se um algoritmo de separação de regiões de voz ativa e de ruído de fundo (VAD-*Voice Active Detection*) a todos os sinais da base, para posteriormente provocar saturações somente nas regiões de voz ativa. Utilizou-se um limiar fixo para limitar a amplitude das amostras do sinal, cujo

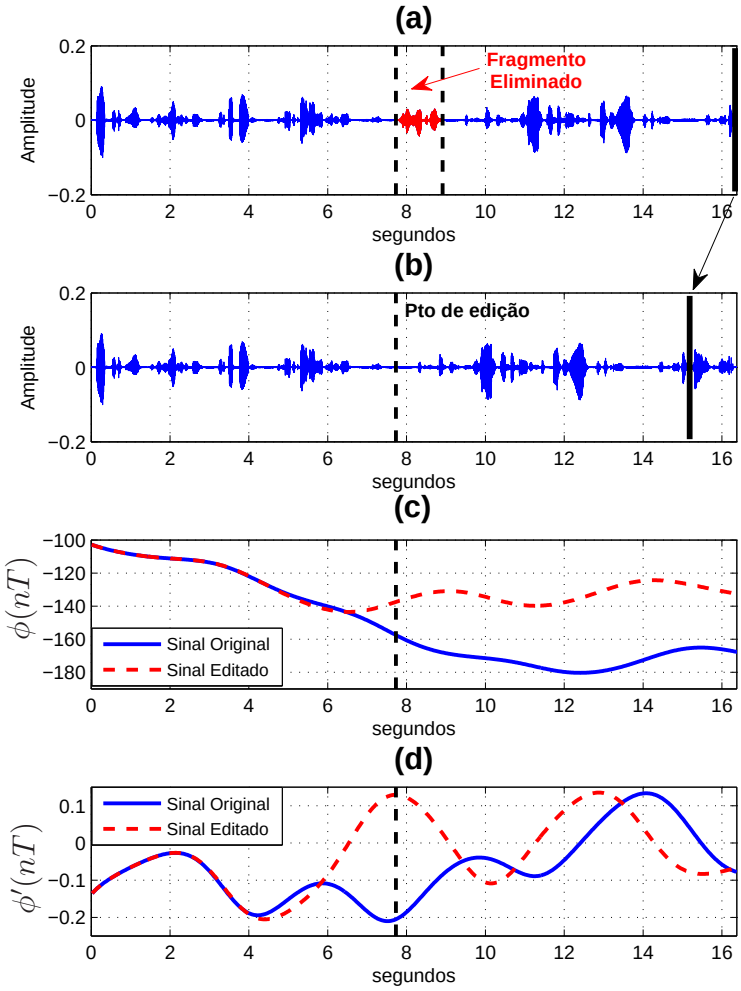

Fig. 7. Edição de áudio (com ENF variando no tempo): eliminação de um fragmento do sinal. O valor nominal da ENF é de 60 Hz. $T = \frac{1}{60}$ segundo (1 ciclo da ENF nominal). (a) Sinal original; (b) Sinal editado; (c) Curva de estimação de fase na ENF, $\phi(n)$; (d) Curva da variação de fase entre duas amostras consecutivas, $\phi'(n)$.

valor define implicitamente um determinado percentual de amostras saturadas nas regiões de voz ativa. Modificando a percentagem de saturação e aplicando a técnica de autenticação de áudio proposta para todos os sinais deste banco de dados (proveniente da Espanha), obteve-se a curva característica da taxa de erro de detecção em função da percentagem de saturação, mostrada na Fig. 8. Como se pode observar, níveis de saturação superiores a 0,5 % já afetam consideravelmente o método de autenticação proposto.

Por outro lado, analisando-se o comportamento da ENF nos sinais, foi verificado (mediante estimativas de frequência) que os sinais do banco de dados proveniente do Rio de Janeiro apresentavam uma variação bem maior da frequência em torno de seu valor nominal ao longo do sinal gravado do que no banco de

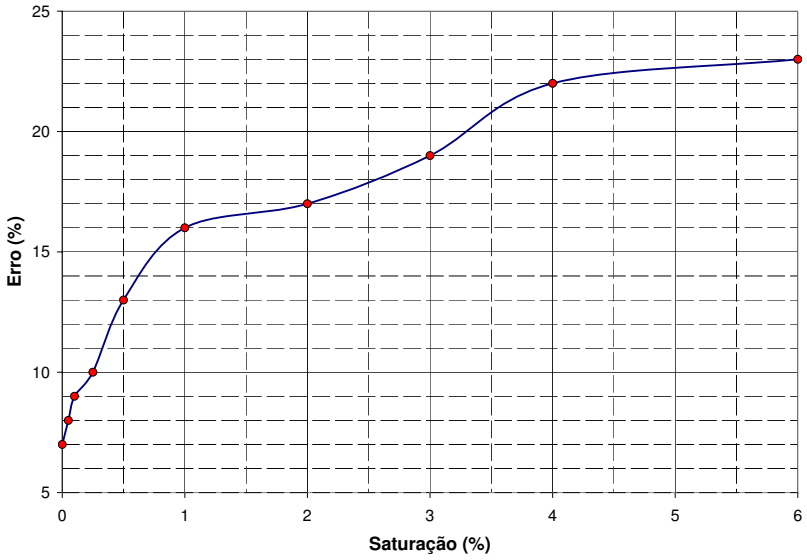

Fig. 8. Efeito de saturação: Erro de detecção na autenticação de áudio em função de percentagem de saturação. O erro do sistema sobre os sinais sem saturação é de 7 %.

dados proveniente da Espanha, que apresentava muito pouca oscilação da ENF.

Por fim, analisou-se a relação entre a SNR (razão sinal-ruído) e o mecanismo de autenticação proposto. Inicialmente, verificou-se por estimativas de potência de ruído que a SNR média para o banco de dados proveniente da Espanha é de 35 dB e para o banco de dados proveniente de Rio de Janeiro é de $29,77$ dB. Como se pode perceber, o banco de dados que possui maior SNR (menos ruidoso) resulta em menor erro na detecção. Para ter uma idéia quantitativa do efeito do ruído na detecção, adicionou-se ruído branco gaussiano a todos os sinais do banco de dados proveniente da Espanha, variando a SNR, para posteriormente ser aplicado o mecanismo de autenticação nos sinais e verificar o comportamento do erro de detecção. O erro inicial era de 7% a 35 dB. A Fig. 9 apresenta a curva do erro na detecção de autenticação de áudio em função da SNR. Como se pode observar, quanto menor for a SNR, maior será o erro de detecção, seguindo uma relação quase linear.

IV. Conclusões

A ferramenta de autenticação de áudio aqui apresentada mostrou-se bastante eficiente na detecção de edições em sinais gravados em condições favoráveis (pouca variabilidade da ENF, alta SNR e ausência de saturações). Esta técnica fornece informação visual que permite a localização de pontos onde o sinal foi possivelmente editado e a inferência do tipo de edição (se recorte ou inserção de áudio). Seguindo outra filosofia, o cômputo de uma *feature* permite que se discrimine de forma automática entre sinais editados e originais.

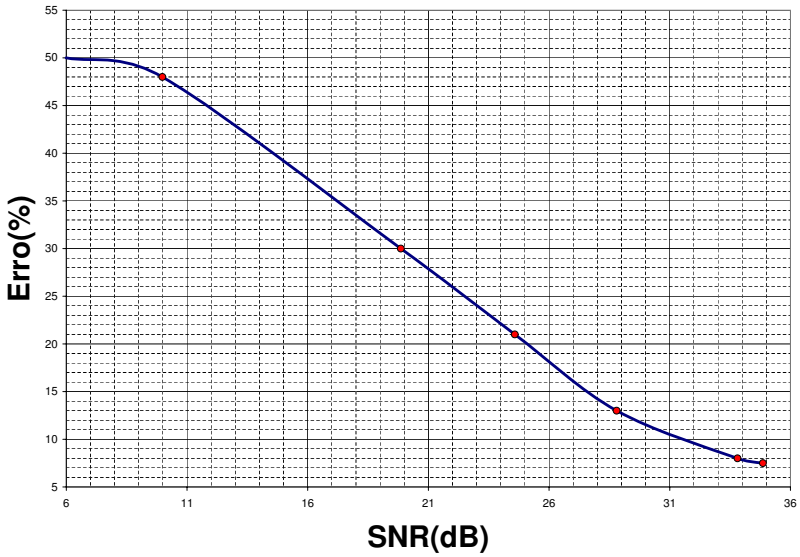

Fig. 9. Efeito de ruído: Erro de detecção na autenticação de áudio em função da SNR. O erro do sistema sem inserir ruído é de 7 %. A razão sinal-ruido inicial é 35 dB.

A taxa de acerto de 93% obtida na avaliação do método para o banco de dados gravado sob condições controladas é uma boa confirmação da eficácia do sistema. Além disso, considerando a presença da ENF na maioria das gravações, a ferramenta proposta pode ser de grande importância para aplicação em fonética forense. A importância deste estudo é reforçada pelo fato de diferentes localidades no Brasil possuirem variação temporal da ENF distinta da de outros países e, portanto, necessitarmos de uma ferramenta adaptada às condições nacionais. No Laboratório de Processamento de Sinais de Voz do IME, investiga-se atualmente como aumentar a robustez do método a variações da ENF e a ruído. Um outro tópico de pequisa seria avaliar o método em localidades no Brasil onde a ENF não possui um controle muito sofisticado (por exemplo, em regiões alimentadas por usinas geradora de menor porte).

AGRADECIMENTOS

Os autores agradecem ao CNPq, à FAPERJ, e à CAPES pelo apoio financeiro aos projetos de pesquisa.

REFERÊNCIAS

[1] B. E. KOENING, "Authentication of forensic audio recordings," *Journal of the Audio Engineering Society*, vol. 38, pp. 3–33, January/February 1990.

[2] E. B. BRIXEN, "ENF: quantification of the magnetic field," *AES* 33^{rd} *International Conference: Audio Forensic, Theory and Practice*, Denver, CO, USA, June 2008.

[3] R. W. SANDERS, "Digital audio authenticity using the electric network frequency," *AES* 33^{rd} *International Conference: Audio Forensic, Theory and Practice*, Denver, CO, USA, June 2008.

[4] A. J. COOPER, "The Electric Network Frequency (ENF) as an aid to authenticating forensic digital audio recordings – an automated approach," *AES* 33^{rd} *International Conference: Audio Forensic, Theory and Practice*, Denver, CO, USA, June 2008.

[5] *Edit Track, User Manual*. Speech Technology Center, St. Petersburg, Russia, 2005.

[6] D. NICOLALDE and J. A. APOLINÁRIO JR., "Evaluating digital audio authenticity with spectral distances and ENF phase change," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, April 2009.

[7] D. NICOLALDE and J. A. APOLINÁRIO JR. and L. W. PEREIRA BISCAINHO, "Autenticação de áudio digital com base na mudança de fase da frequência da rede elétrica," *XXVII Simpósio Brasileiro de Telecomunicações (SBrT'*09*)*, Blumenau, Brasil, Setembro 2009.

[8] J. ORTEGA-GARCIA, J. GONZALEZ-RODRIGUEZ, and V. MARRERO-AGUIAR, "AHUMADA, A large speech corpus in spanish for speaker characterization and identification," *Elsevier Speech Communication*, vol. 31, pp. 255–264, June 2000.

BIOGRAFIA

Daniel Patricio Nicolalde

Nasceu em Quito, Equador, em 1982. Possui o título de Engenheiro Eletrônico e Telecomunicações, outorgado pela Escuela Politécnica del Ejército (ESPE), Quito, Equador, 2007. Atualmente é aluno do Programa de Pós-graduação em Engenharia Elétrica do Instituto Militar de Engenharia (IME), Rio de Janeiro, RJ, Brasil. Suas áreas de interesse profissional incluem Processamento Digital de Sinais, Processamento de Áudio$\&$Voz e Programação Dinâmica.

José Antonio Apolinário Junior

Nasceu em Taubaté, SP, em 1960. Graduou-se na Academia Militar das Agulhas Negras (AMAN), Resende, RJ, em 1981 e Engenheiro Eletrônico pelo Instituto Militar de Engenharia (IME), Rio de Janeiro, em 1988. Recebeu os títulos de Mestre en Ciências pela Universidade de Brasília (UnB), Brasília, em 1993 e Doutor em Ciências pela Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, em 1998, ambos em engenharia elétrica. Coronel da Reserva do Quadro de Engenheiros Militares do Exército Brasileiro, ele é atualmente Professor Adjunto do Departamento de Engenharia Elétrica do IME onde já serviu como Chefe de Departamento, Pró-Reitor de Ensino e Pesquisa e Subcomandante. O Dr. Apolinário foi Professor Visitante da Escuela Politécnica del Ejército (ESPE), Quito, Equador, de 1999 a 2000, Pesquisador Visitante duas vezes e Professor Visitante na Helsinki University of Technology (HUT), Finlândia, em 1997, 1999 e 2004, onde também realizou um estágio pós-doutoral em 2006. Seu interesse em pesquisa compreendem vários aspectos de processamento digital de sinais, incluindo filtragem adaptativa, voz e processamento em arranjo de sensores. Ele organizou e foi o primeiro Presidente do Capítulo Rio de Janeiro da Sociedade de Comunicações do IEEE. Recentemente editou o livro "QRD-RLS Adaptive Filtering"(Springer, 2009). É membro da SBrT (Sociedade Brasileira de Telecomunicações) e senior member do IEEE (Institute of Electrical and Electronics engineers).

Luiz Wagner Pereira Biscainho

Nasceu no Rio de Janeiro em 1962. Graduou-se em Engenharia Eletrônica (magna cum laude) pela Escola de Engenharia (EE) hoje Escola Politécnica (Poli) da Universidade Federal do Rio de Janeiro (UFRJ) em 1985, e obteve seus títulos de M.Sc. e D.Sc. em Engenharia Elétrica pelo Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE) da UFRJ em 1990 e 2000, respectivamente. Dr. Biscainho é Professor Adjunto do Departamento de Engenharia Eletrônica e de Computação (DEL) da Poli e do Programa de Engenharia Elétrica (PEE) do COPPE, na UFRJ. Sua área de pesquisa é processamento digital de sinais, em particular processamento de áudio e sistemas adaptativos. É membro ativo do IEEE (Institute of Electrical and Electronics Engineers), da AES (Audio Engineering Society), e da SBrT (Sociedade Brasileira de Telecomunicações).