Análise de parâmetros prosódicos em diferentes relações sinal-ruído

Ana Carolina Constantini
Instituto de Estudos da Linguagem – Universidade Estadual de Campinas
e-mail: carolconstantini@gmail.com

Plinio Almeida Barbosa
Instituto de Estudos da Linguagem – Universidade Estadual de Campinas
e-mail: pabarbosa.unicamp@gmail.com

***Resumo*— Na fonética forense, a tarefa de identificar falantes por meio de amostras de fala de suspeitos de cometer um crime é muito comum. Uma das dificuldades que os profissionais atuantes na área podem se deparar é com a grande quantidade de ruído presente na amostra de fala disponível para a análise. É interessante saber quanto o ruído presente em amostras de fala pode perturbar os resultados obtidos pela análise dos parâmetros prosódicos, visto que estes são amplamente utilizados para auxiliar na identificação de falantes. Oito parâmetros prosódico-acústicos foram analisados em amostras de fala espontânea com diferentes relações sinal-ruído, para posterior comparação dos valores obtidos. A ênfase espectral e a mediana de frequência fundamental foram os parâmetros que mais apresentaram mudanças quando há inclusão de ruído aditivo na amostra de fala. A análise da estrutura rítmica das amostras de fala mostrou-se robusta e indicada para análise de parâmetros prosódicos em situações adversas, como por exemplo, a adição de ruído.**

Palavras chave-prosódia;relação sinal-ruído; fonética forense

***Abstract:* Speaker identification using speech sample is a common task in forensic phonetics. The presence of noise in the speech sample is a challenge for phoneticians because it can interfere or modify the parameters that are used to identificate the speaker. This study aims to analyze the impact of noise in spontaneous speech by measuring eight prosodic parameters in differents noise-to-ratio signals: high signal-to-noise ratio (recording made in a quiet room) and low signal-to-noise ratio (original recording with adition of gaussian noise). Spectral emphasis and median of fundamental frequency showed differences in its values when the signal-to-noise ratio had changed. The analysis of different parameters related to rhytmical structure of the speech samples were more resistant to noise presence and that parameters seem to be sufficiently robust to be used when the signal-to-noise ratio is low.**

Key-words: prosody, signal-to-noise ratio, forensic phonetics

I. INTRODUÇÃO

A tarefa de identificação de falantes, comum na Fonética Forense, ocorre quando há uma amostra de fala como prova de um crime, como por exemplo, uma gravação de um telefonema de um suspeito fazendo ameaças à outra pessoa. Durante o processo de identificação do falante, é preciso fazer a correlação entre a amostra de fala questionada (gravação do telefonema com a ameaça) e amostras de fala de referência (gravações com amostras de fala de suspeitos de terem feito ameaças), sendo que as últimas geralmente são feitas em situações totalmente diferentes das primeiras, podendo ser realizadas nas dependências de uma delegacia de polícia, por exemplo.

Os profissionais que realizam este tipo de atividade podem se deparar com uma série de dificuldades quando obtém a gravação da amostra de fala de referência para análise. As dificuldades encontradas podem estar, por exemplo, relacionadas à simulação da qualidade vocal ou disfarce vocal realizado pelo suspeito durante o telefonema, distorção do sinal em que a amostra de fala foi gravada, barulho de fundo como o som do vento, barulho do trânsito e interferências no sinal de fala devido ao uso de telefone celular [1,2].

Todas as situações descritas alteram a relação sinal--ruído da amostra de fala e podem comprometer a identificação correta de um indivíduo ao modificarem o sinal de fala e, consequentemente, as medidas acústicas que são amplamente utilizadas para auxiliar na identificação de falantes, no campo da fonética forense.

Dentre os parâmetros acústicos apontados como mais robustos na identificação de falantes estão a frequência fundamental [3] e variáveis presentes no estilo de fala, como taxa de elocução [4,5].

Por isso, julgamos importante conhecer mais sobre o comportamento de parâmetros prosódico-acústicos quando há ruído em uma amostra de fala. O objetivo deste trabalho é comparar determinados parâmetros prosódicos em duas situações: alta relação sinal-ruído (gravação realizada em cabine acústica) e baixa relação sinal-ruído (gravação com inserção de ruído tipo gaussiano).

II. METODOLOGIA

A. Sujeitos e Dados

Foram analisadas amostras de fala espontânea, do tipo entrevista livre, de 35 sujeitos, do sexo masculino, com idade média de 35.4 anos e grau de escolaridade variando entre ensino superior completo e pós-graduação completa. Os sujeitos eram de sete localidades brasileiras diferentes.

As gravações têm em média 5 minutos de duração e foram realizadas em cabine acústica, que caracteriza a condição alta relação sinal-ruído.

B. Tratamento dos dados e parâmetros analisados

As gravações de cada um dos sujeitos, com duração total de cinco minutos foram divididas em trechos menores, com aproximadamente 100 segundos cada trecho. Algumas gravações foram divididas em três trechos (com 100 segundos cada) e outras foram divididas em quatro trechos. A divisão em trechos menores foi feita para favorecer o uso de *scripts* que seriam utilizados no software Praat, nas etapas posteriores da pesquisa. O total de trechos de fala, levando em conta os 35 sujeitos analisados, foi 109.

Todos os 109 trechos de fala foram segmentados em unidades vogal-vogal (unidades VV) com o auxílio do *script* BeatExtractor [6]. A unidade VV é uma sílaba fonética, que compreende o segmento que vai do *onset* de uma vogal e se estende até o *onset* da vogal seguinte. Uma vasta literatura tem mostrado que os picos locais da duração normalizada dessas unidades são correlatos do acento frasal (ver [6] para revisão) e que são unidades mínimas de processamento do ritmo da fala [7,8].

Após segmentar as amostras de fala em unidades VV, o *script* ProsodicDescriptorExtractor [9] foi utilizado para extrair oito parâmetros prosódico-acústicos que contemplassem medidas de duração, frequência fundamental e intensidade, que são os parâmetros acústicos clássicos que compõem a prosódia. Os parâmetros analisados foram: taxa de elocução (unidades VV/s), média de z-score suavizado de duração de unidade VV, desvio-padrão de z-score suavizado de duração de unidade VV, assimetria de z-score suavizado de duração de unidade VV, taxa de saliência duracional (picos de z-score/s), mediana da frequência fundamental (Hz), ênfase espectral (dB) e taxa de unidades VV não proeminentes/s.

O cálculo de z-score assinala a duração normalizada das unidades VV ao longo do enunciado em termos de afastamento da média expresso em unidades de desvio-padrão. Assim, o z-score especifica o afastamento do valor medido (da duração de cada unidade VV) em relação a uma média de duração dos fones do português brasileiro (PB) que compõem a unidade VV. A opção por trabalhar com o z-score suavizado e seus três primeiros descritores estatísticos (média, desvio--padrão e assimetria) foi feita porque a suavização atenua os efeitos da realização local do acento lexical, assinalando apenas unidades salientes no enunciado. Os valores de média, desvio-padrão e assimetria do z-score suavizado revelam como se estruturam o grau de proeminência e a força das fronteiras no enunciado e são cruciais para fazer uma descrição completa do ritmo da fala de um indivíduo [10].

A taxa de unidades VV não proeminentes por segundo, é próxima à taxa de articulação, pois não utiliza em seu cálculo as pausas silenciosas. As unidades VV não proeminentes são aquelas que não possuem picos de duração normalizada para as respectivas unidades VV. A taxa de unidades VV não proeminentes é calculada utilizando o número total de unidades VV não proeminentes em um trecho de fala e dividindo-o pelo tempo total do trecho de fala, que é delimitado pelo primeiro *onset* vocálico do trecho até o último *onset* vocálico.

Em relação à frequência fundamental (F0), os valores de mediana foram utilizados porque são mais resistentes a erros de medição do que à média. A ênfase espectral é definida como a diferença de intensidade (em dB) do segmento a ser analisado entre toda a faixa espectral (0 a 11 kHz) e a região de frequências baixas, que é delimitada considerando a frequência fundamental multiplicada por 1.5. De acordo com [11], a razão por delimitar a região de frequências baixas com o cálculo de F0 x 1.5 é separar a frequência fundamental dos outros harmônicos e obter uma medida normalizada da energia nas altas frequências.

As medidas obtidas dos oito parâmetros prosódico-acústicos foram extraídas das amostras de fala das gravações originais e armazenadas em uma tabela. Após esta etapa, foi feita a inclusão de ruído aditivo do tipo gaussiano, por meio do software Praat, nas amostras de fala originais, com a fórmula descrita abaixo (1), em que a primeira posição entre parênteses é a média do ruído (nesse caso 0) e a segunda é o desvio-padrão em unidades arbitrárias de pressão sonora (nesse caso 0.01):

$$\text{Original} + \text{randomGauss}(0,0.01) \quad (1)$$

Duas magnitudes de ruído foram utilizadas: 0.01 (relação sinal-ruído baixa) e 0.02 (relação sinal-ruído muito baixa). O objetivo de aumentarmos ainda mais a magnitude do ruído foi testar como os parâmetros se comportariam quando diminuímos ainda mais a relação sinal-ruído.

Na etapa seguinte, rodamos o *script* ProsodicDescriptorExtractor novamente a fim de obtermos os resultados para os oito parâmetros prosódico-acústicos nas gravações com baixa relação sinal-ruído.

C. Análise estatística

Para a comparação dos resultados obtidos na análise dos parâmetros nas diferentes relações sinal-ruído estudadas, o teste escolhido foi o Teste T, com nível de significância de 0.016, que leva em conta as três situações que foram consideradas para análise (nível de significância 0.05/3, que garante que todas as comparações feitas tenham um máximo de 0.05 de significância quando tomadas conjuntamente). As três situações consideradas para análise são as seguintes:

1) Comparação dos resultados obtidos na análise da gravação original (alta relação sinal-ruído) com os resultados obtidos na análise da gravação com ruído aditivo de magnitude 0.01 (baixa relação sinal-ruído);
2) Comparação dos resultados obtidos na análise da gravação original (alta relação sinal-ruído) com os resultados obtidos na análise da gravação com ruído aditivo de magnitude 0.02 (relação sinal-ruído muito baixa);
3) Comparação dos resultados obtidos na análise da gravação com ruído aditivo de magnitude 0.01 (baixa relação sinal-ruído) com os resultados obtidos na análise da gravação

Pesquisa financiada pelo Fundo de Amparo à Pesquisa no Estado de São Paulo - FAPESP

com ruído aditivo de magnitude 0.02 (relação sinal-ruído muito baixa).

III. RESULTADOS

Alguns parâmetros analisados apresentaram mudanças em suas medidas após a adição de ruído na amostra de fala. São eles: ênfase espectral, mediana de frequência fundamental, taxa de saliência duracional, média, desvio-padrão e assimetria de z-score de duração de unidade VV. É importante ressaltar que, os quatro últimos parâmetros citados, apresentaram mudanças somente em seis trechos de fala (o total de trechos de fala analisados, considerando todos os sujeitos foi 109). A ênfase espectral e a mediana de frequência fundamental apresentaram mudanças em todos os trechos analisados, o que é esperado, visto que, a adição do ruído interfere na captação dessas duas medidas especificamente.

As discretas mudanças observadas nos parâmetros – taxa de saliência duracional e média, desvio-padrão e assimetria de z-score suavizado de duração de unidade VV ocorreram devido ao efeito natural que o ruído causa na extração de medidas através de fórmulas do Praat. O efeito do ruído dificulta a extração das medidas, influenciando principalmente a mediana de frequência fundamental e ênfase espectral. É preciso considerar que, para os outros parâmetros que apresentaram mudanças nos valores extraídos, a dificuldade apresentada pelo script é mínima, já que as mudanças ocorreram em poucos trechos e com mudanças pouco significativas (da ordem de uma casa decimal).

A tabela 1 apresenta os resultados obtidos na análise estatística ($p<0.016$).

	Ruído Gaussiano 0,01	Ruído Gaussiano 0,02
Mediana de F0		
Alta Relação Sinal-Ruído	p < 0,001 *	p < 0,001 *
Ruído Gaussiano 0,01	-	p < 0,001 *
Ênfase Espectral		
Alta Relação Sinal-Ruído	p < 0,001 *	p < 0,001 *
Ruído Gaussiano 0,01	-	p < 0,001 *

Tabela 1. Resultados do Teste T para as análises das diferentes relações sinal-ruído estudadas. *p<0.016

A tabela 1 mostra que, todas as situações possíveis de comparação entre os resultados da mediana de F0 e da ênfase espectral foram estatisticamente significativos. Porém, é interessante saber de que ordem foi a mudança nos valores dos parâmetros após a inclusão de ruído aditivo. Para isso, calculamos a média da ênfase espectral e de mediana de F0 nos três momentos analisados (tabela 2).

	Gravação original	Ruído 0,01	Magnitude de mudança	Ruído 0,02	Magnitude de mudança
$\mathbf{M}_{\mathbf{f0}}^{\mathbf{media\ na}}$	125 Hz	126 Hz	0,8%	128 Hz	2,4%
$\mathbf{M}_{\mathbf{espectral}}^{\mathbf{ênfase}}$	1.18 dB	1.8 dB	55%	3 dB	154%

Tabela 2. Médias de ênfase espectral e de mediana de frequência fundamental e magnitudes de mudança após adição de ruído.

IV. DISCUSSÃO

O aumento da intensidade do ruído nas amostras de fala resultou em mudanças significativas nos valores obtidos para dois parâmetros prosódico-acústicos: ênfase espectral e mediana de F0.

A tarefa de comparar duas amostras de fala, uma com baixa e outra com alta relação sinal-ruído é corriqueira na área da fonética forense e por isso, a análise realizada neste trabalho é de fundamental importância contribuir para a resolução deste tipo de problema.

A frequência fundamental é apontada, juntamente a frequência dos informantes e o estilo de fala, como uma das principais pistas acústicas que os ouvintes utilizam para identificar outros falantes, segundo estudos como [5] e [3]. Nossos resultados mostraram que, mesmo com a análise estatística apontando para significância das mudanças ocorridas nos valores encontrados para a mediana de frequência fundamental, os valores médios aumentaram no máximo 3 Hz (passando de 125 Hz na gravação original para 128 Hz na gravação com mais ruído), por isso, podemos considerar que para o tipo de ruído utilizado, mesmo com significância estatística, o aumento nos valores provavelmente não resultaria em um erro de análise, por exemplo, na identificação dos sujeitos.

A ênfase espectral vem sendo utilizada em pesquisas na área da análise prosódica por ser considerada um parâmetro mais confiável que a intensidade. Utilizar isoladamente o parâmetro intensidade pode diminuir a confiabilidade dos resultados visto que, a intensidade pode ser altamente suscetível a efeitos relacionados à distância do sujeito ao microfone, por exemplo [11]. Nossos resultados indicaram mudanças de até 154% de aumento dos valores de ênfase espectral quando há presença de ruído na amostra de fala e por isso, seu uso com a finalidade de identificar falantes em amostras de fala em que a presença de ruído deve ser feito com cuidado ou até mesmo evitado.

A análise destes parâmetros em diferentes tipos de ruído pode implicar na obtenção de diferentes resultados e de diferentes magnitudes de mudanças nos parâmetros. Por isso, é importante que novos estudos se proponham a continuar estudando os possíveis efeitos de diferentes tipos de ruído em parâmetros fonético-acústicos diversos.

A estrutura rítmica é uma escolha acertada para a análise de amostras de fala com ruído. De acordo com [10], a estrutura rítmica de um enunciado pode fazer com que os enunciados se pareçam semelhantes ou diferentes. O trabalho de [10] mostrou que os sujeitos levam em consideração, para

identificar corretamente diferenças na "forma de falar", pelo menos dois parâmetros relacionados à taxa de elocução em unidades VV/s e duração dos grupos acentuais. Como nossos resultados mostraram, houve mudança pontual em apenas alguns trechos analisados e mesmo assim, tais mudanças foram de pequena magnitude. Visto que os ouvintes parecem utilizar parâmetros rítmicos para decisões corretas acerca da forma de falar de outros indivíduos e dos resultados obtidos. Neste trabalho sugerimos que, a análise de parâmetros prosódicos associados ao ritmo, como os que foram utilizados neste trabalho, seja mais utilizada em tarefas de identificação de falantes no campo de pesquisa da fonética forense, principalmente quando há a necessidade de comparar amostras de fala com diferentes relações sinal-ruído.

V. CONCLUSÃO

A análise das amostras de fala com diferentes relações sinal-ruído mostrou que a ênfase espectral e a mediana de frequência fundamental foram os parâmetros que mais mudaram quando a relação sinal-ruído aumentou. No entanto, a ênfase espectral apresenta mudanças mais bruscas, por isso, seu uso com a finalidade de identificar falantes em amostras de fala com ruído não é recomendado.

A análise da estrutura rítmica do enunciado é a escolha mais eficaz quando é preciso comparar amostras de fala com diferentes relações sinal-ruído com a finalidade de identificar falantes.

REFERENCIAS

[1] Hollien, H. Forensic voice identification. San Diego, CA: Academic Press. 2002

[2] Eriksson, A. The disguised voice: imitating accents or speech styles and impersonating individuals. In C. Llamas & D. Watt (Eds.), Language and Identities. Edinburgh: Edinburgh University Press. 2010. p. 86–96

[3] Atal, B.S. Automatic speaker recognition based on pitch contours. Journal of the Acoustical Society of America. 1972. v. 52: 1687-1697

[4] Öhman, L. et. al. Mobile phone quality VS direct quality: how the presentation format affects earwitness identification accuracy. The European Journal of Psychology Applied to Legal Context. v. 2. 2010. p.161-182

[5] Kahn, J. Caractéristique propres au locuteur : Traitement automatique et distance perceptive. Université-Stendhal-Grenoble 3. Dissertação de Mestrado. 2008

[6] Barbosa, P.A. Incursões em torno de ritmo da fala. Campinas: Editora Pontes. 2006

[7] Pompino-Marschall, B. The syllable as a prosodic unit and the so-called P-centre effect. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, 29; 1991. p. 65-123.

[8] Dogil, G; Braun, G. The PIVOT model of speech parsing. Verlag der Österreichischen Akademie der Wissenschaften. Viena; 1988

[9] Barbosa, P.A. Script ProsodicDescriptorExtractor. Disponível com autor.

[10] Barbosa, P.A.;Silva, W. A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches. In: Computational Processing of the Portuguese Language 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012.

[11] Traunmüller, H. and Eriksson, A., "Acoustic effects of variation in vocal effort by men, women, and children". Journal of the Acoustical Society of America, Vol. 107(6), 2000, p. 3438-3451.