Uso de Técnicas Acústicas para Verificação de Locutor em Simulação Experimental

Aline Machado [1], *Plínio A. Barbosa*[2]

[1] Universidade Estadual de Campinas, Brasil
[2] Grupo de Estudos de Prosódia da Fala, Dep. De Linguística, Universidade Estadual de Campinas, Brasil

machadop.aline@gmail.com, pabarbosa.unicampbr@gmail.com

Resumo

Este artigo tem como objetivo investigar a eficácia de um conjunto de medidas acústicas no que concerne ao reconhecimento da fala de um indivíduo, denominado "criminoso", em um grupo de dez falantes da variação do português brasileiro (PB). Os parâmetros escolhidos, entre eles estão as frequências dos dois primeiros formantes, a frequência fundamental média, duração de unidades do tamanho da sílaba e da vogal, a taxa de dinamicidade de frequência do segundo o desvio padrão de durações de intervalos consonânticos (ΔC), das vogais do PB. Comparamos as amostras de fala extraídas de entrevistas realizadas em ambiente acústico de baixa qualidade (ao ar livre e ligações de telefone celular) com a gravação do criminoso (em um ambiente sem ruído). Através de técnicas de análise estatística – ANOVA de um fator e Teste de Duncan – concluímos que parâmetros rítmicos e temporais, tais como duração média das vogais, taxa de elocução, ΔC e taxa de movimento do segundo formante, mantiveram-se robustas com a troca de transmissão de gravação (direta e telefônica). Os resultados mostram também que os mesmos parâmetros foram os mais discriminantes para o reconhecimento do "criminoso".

Palavras-chave: fonética forense, verificação de locutor, transmissão telefônica

1. Introdução

O Reconhecimento de Locutor (RL) é caracterizado como "qualquer atividade pela qual uma amostra de fala é atribuída a uma pessoa com base em suas propriedades fonético-acústicas ou perceptuais" [1]. Esta tarefa é feita tanto pelo cidadão comum (Reconhecimento Leigo de Locutor), por exemplo, quando atendemos um telefone e reconhecemos quem está do outro lado da linha sem que haja uma identificação prévia; quanto para uma análise especializada por peritos treinados e/ou máquinas em ambientes acadêmicos ou também para investigações policiais.

Esta pesquisa tem como base o método acústico semiautomático de análise, isto é, através da extração automática de um conjunto de parâmetros acústicos das amostras de fala segmentadas e etiquetadas (manualmente) do *software* PRAAT [2], comparamos estatisticamente os dados obtidos para chegarmos no objetivo da pesquisa. Ressalta-se que o método auditivo, embora também usado para a tarefa de RL, não é aplicado neste estudo, pois (1) os sujeitos, excetuando-se o "criminoso", não são desconhecidos dos pesquisadores e (2) não apresentam grandes diferenças de sotaque e/ou outras características importantes para a discriminação nesta análise (i.e. patologia na fala, idioleto etc). O termo "verificação de locutor" foi escolhido como tipo de tarefa de reconhecimento de falantes partir da definição de Hollien [3]. Segundo o autor, nesta técnica de análise, todo o processo da pesquisa é controlado pelo pesquisador, a coleta e os participantes das gravações. Os sujeitos analisados são cooperativos, eles produzem várias amostras de sua fala para a comparação de voz sem adotar algum tipo de disfarce ou variações em sua fala.

2. A pesquisa

Usamos como protocolo para este trabalho o roteiro de um cenário típico encontrado na Fonética Forense: uma gravação de fonte desconhecida (ruidosa, podendo ser de ligação de celular), chamada de gravação questionada; é comparada com a gravação de um suspeito, ou seja, gravação de referência, gravada preferencialmente em cabines com isolamento acústico. Este foi o protocolo seguido para o trabalho efetuado.

2.1. O efeito do celular

Como dito anteriormente, em muitas situações forenses, cientistas tem em mãos como material de avaliação, isto é, como gravação questionada, escutas telefônicas que são, em sua grande maioria, de péssima qualidade das quais devem apresentar algum resultado substancial para o júri. Por isso, as gravações por telefone celular foram escolhidas para a pesquisa. No Brasil, por exemplo, há mais de 271 milhões de linhas de telefone celular de acordo com o último censo da ANATEL [4]. Foi evidenciado também que a gravação por telefone celular apresenta resultados menos robustos em comparação com as por telefone fixo para alguns parâmetros acústicos, como por exemplo, para as frequências do primeiro e segundo formantes [5, 6]. Alguns efeitos causados pelo telefone celular foram apresentados por Byrne e Foulkes [6] em seu artigo, como:

- Efeitos do ambiente: ligações de celulares podem ser realizadas em ambiente com alto nível de ruído de fundo (e.g. trânsito).
- Efeito dos falantes: o registro telefônico da voz muda consciente ou subconscientemente influenciando na taxa de locução, qualidade de voz e pronúncia. Um dos efeitos mais comuns é o aumento do volume da voz de um indivíduo enquanto fala ao telefone, afetando diretamente sua de taxa da frequência fundamental (F0).

2.1.1. Efeitos técnicos

Dois fenômenos muito importantes ocorrem através deste efeito causado pelo telefone celular, a distorção espectral e o deslocamento das taxas de frequências formânticas. O primeiro consiste no "apagamento" das frequências que se encontram no filtro de passa-baixa (abaixo de 300 Hz) e no filtro passa-alta (acima de 3.500 Hz). O que pode influenciar na perda de frequências de formantes em vogais altas, por exemplo. O "deslocamento de frequências formânticas" por outro lado, consiste no deslocamento inversamente proporcional à frequência do formante. Isto é, ao sofrer o efeito do filtro do canal telefônico, as taxas de frequências dos primeiros dois formantes tendem a aumentar, já as frequências mais altas, como a do quarto formante (F4) diminuem de valor. Isso pode fazer com que a qualidade de uma vogal mude, entre outras consequências.

Nessa pesquisa conseguimos constatar este último fenômeno técnico de celular. A frequência do segundo formante teve um aumento de 98 Hz em comparação com a gravação direta. Outros parâmetros acústicos como frequência fundamental e frequência *baseline* também tiveram um aumento de 4 Hz em seu valor, o que por sua vez é estatisticamente insignificante.

2.2. Parâmetros acústicos

Os parâmetros acústicos foram escolhidos através de uma extensa análise bibliográfica e a partir de uma série de um conjunto critérios desenvolvidos por Francis Nolan [7].

- Alta variabilidade interfalante: escolher um parâmetro acústico que exponha um grau de variação de um falante para o outro.
- Baixa variação intrafalante: o parâmetro acústico tem que apresentar um grau de variação intrafalante menor que a variação interfalante.
- Resistência à tentativa de disfarce ou imitação: o parâmetro tem que se manter robusto na mudança e variação da voz de um falante que tendenciosamente disfarça sua voz.
- Eficácia: O parâmetro acústico escolhido deve ocorrer de forma corriqueira na fala de um locutor. Um parâmetro que ocorre raramente na fala necessita de uma grande quantidade de dados em ambos *corpora* teste e de referência para análise.
- Robustez na transmissão: ele não pode ter sua informação perdida ou reduzida em diferentes formas de transmissão de gravação, i.e. gravação direta ou telefônica.
- Mensurabilidade: a sua extração não deve ser excessivamente difícil.

2.3. Metodologia

Este artigo é resultado de uma pesquisa de Mestrado [8] que teve como objetivo reconhecer um indivíduo pela voz em um grupo de dez falantes do PB divididos em quatro estados do Brasil, Bahia, Rio Grande do Sul, São Paulo e Pará. Coletamos gravações de seis participantes do estado de São Paulo (três da capital, um de Jundiaí, um de Campinas e um de Cordeirópolis); dois sujeitos da Bahia, ambos de Salvador; um sujeito de Santarém no estado do Pará; e, por fim, um de Pelotas no Rio Grande do Sul. Os sujeitos tinham faixa etária de 18 a 28 anos, com nível de educação mínimo de ensino superior incompleto e moraram a maior parte de suas vidas em suas respectivas cidades natais. As amostras de fala de todos os indivíduos foram gravadas em dois canais de gravação, sendo gravação direta e por telefone celular. Além disso, um indivíduo do grupo era aleatoriamente escolhido, sendo denominado como "criminoso" e sua fala fora gravada em um ambiente acusticamente tratado. O objetivo central da pesquisa era o de reconhecimento do "criminoso", por isso escolhemos realizar a comparação de vozes entre as gravações diretas com a gravação de estúdio, pois uma amostra de fala telefônica, como observado no projeto atual, apresentava ruídos e sofria de efeitos como citados na seção 2.1. Além disso, também analisamos quais dos parâmetros acústicos escolhidos eram estatisticamente invariáveis, ou seja, mantinham-se robustos ao aos efeitos do filtro do canal telefônico.

Foram realizadas vinte e uma gravações, dez delas usando um Mini Gravador Coby Cx-r190 ao ar livre (gravações diretas), dez por telefone celular e uma gravação direta em ambiente acusticamente tratado. Nas gravações telefônicas, utilizou-se um celular *Samsung Galaxy Young* com rede 3G da operadora TIM. Seguimos o seguinte esquema para as gravações de celular:

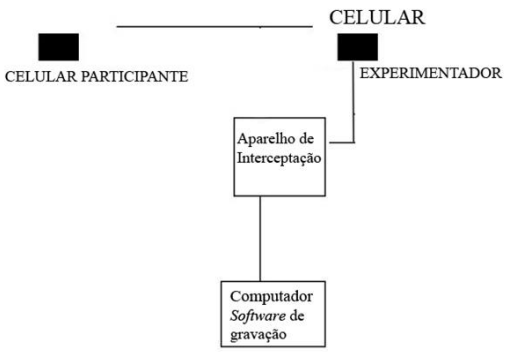

Figura 1: *Diagrama esquemático da gravação telefônica.*

O experimentador, permanecendo em um ambiente com nível mínimo de ruído de fundo, fazia a ligação para o participante que se encontrava em sua respectiva cidade natal. O aparelho de interceptação foi uma placa de áudio, U-Control UCA222, conectado ao telefone celular que, por sua vez, também estava ligado com o *desktop*. Cada conversa foi gravada pelo *software* Audacity. Os arquivos de áudio coletados foram do formato *.wav* e a frequência de amostragem de 8.000 Hz.

Todas as gravações foram segmentadas manualmente via *software* PRAAT e as medidas extraídas automaticamente pelo *script* ForensicDataTracking [9]. Este *script* extraiu os valores para frequência do segundo formante das vogais, taxa de movimento de formante para F2 de todas as vogais, frequência *baseline*, média de F0, duração média das vogais, média de unidades VV (média da duração de unidade do tamanho da sílaba), ênfase espectral e ΔC.

3. Análise de dados

Os parâmetros acústicos escolhidos passaram por séries de testes estatísticos, sendo eles, teste de ANOVA e um teste *post-hoc* de Duncan. Todos os testes estatísticos utilizados nesta pesquisa foram feitos a partir do *software* R [10].

3.1. ANOVA

Realizamos o teste estatístico de ANOVA um fator com os seguintes objetivos: (i) determinar se os parâmetros acústicos analisados permaneciam robustos com a mudança de canal de transmissão, ou seja, se havia variação de parâmetros da gravação direta para uma gravação de telefone celular, e (ii) se algum desses parâmetros acústicos conseguiriam determinar qual dos sujeitos analisados é o "criminoso".

3.1.1. Robustez de parâmetros acústicos

Os parâmetros acústicos que mais se mostraram robustos em relação à mudança de canal de transmissão foram a média da duração das vogais, taxa de elocução, ΔC e taxa do movimento do segundo formantes.

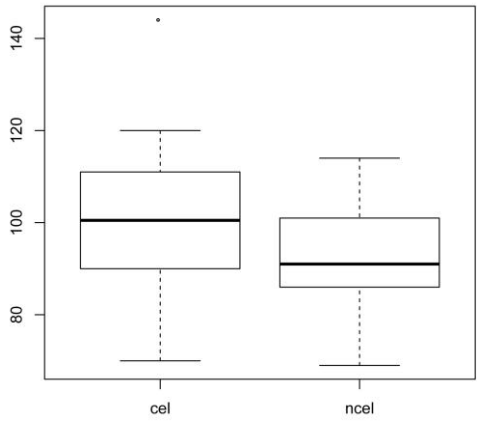

Figura 2: *Boxplot gerado a partir do teste de ANOVA da média de duração das vogais em unidades VV. Determina a mediana da duração deste parâmetro para gravações de celular (cel) e diretas (ncel).*

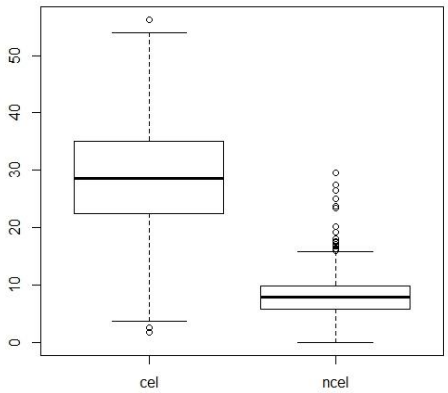

Figura 3: *Boxplot gerado a partir do teste de ANOVA para ênfase espectral das vogais do PB. Determina a mediana da duração deste parâmetro para gravações de celular (cel) e diretas (ncel).*

Segundo a Figura 2, podemos observar uma variação estatisticamente mínima entre os valores das medianas (os traçados horizontais em negrito) para a média da duração das vogais em unidades VV. Esta figura representa o p valor gerado pelo teste de ANOVA. Se o parâmetro acústico apresentou um valor de p 0,05 significa que ele não sofreu variação de canal de transmissão, portanto, um parâmetro robusto.

Tabela 1. *Valor de p para testes de condições de uso de ANOVA para média de duração das vogais, taxa de elocução e ΔC.*

Cel- Direta	Média de duração das vogais	Taxa de elocução	ΔC
Shapiro-Wilk	p-value = 0.9108	p-value = 0.9515	p-value = 0.822
Fligner-Killeen	p-value = 0.4227	p-value = 0.5611	p-value = 0.2825
ANOVA	p-value = 0.245	p-value = 0.36	p-value = 0.05265

Tabela 2. *Valor de p para testes de condições de uso de ANOVA para frequência de F2, taxa de F2, transição de F2, F0, frequência baseline e ênfase espectral.*

Cel – Direta	F2	Taxa de F2	Transição F2	F0	Baseline	Ênfase Espectral
Fligner-Killeen	p-value = 0.05298	p-value = 9.707e-05	p-value = 0.7776	p-value = 1.833e-13	4.435e-10	p-value < 2.2e-16
Kruskal-Wallis	p-value = 1.3e-09	p-value = 0.5911	p-value = 0.6792	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16

Como podemos observar, os parâmetros de média de duração das vogais, taxa de elocução, ΔC, e taxa de movimento do segundo formante mantiveram robustos na condição de mudança de canal de transmissão (Cel-Direta), apresentando o p valor acima de 5%. Utilizamos o valor de p para testes de condições de uso da ANOVA para $\alpha = 0,05$ e Kruskkal-Wallis para $\alpha = 0,05$ nos valores de F2, taxa de F2, F0, frequência *baseline* e ênfase espectral.

Assim como alguns dos parâmetros rítmicos escolhidos para o teste de acordo com a literatura, provou-se ser robusto com a troca do canal de transmissão.

A Figura 3 nos revela a o gráfico para a ênfase espectral. Este parâmetro devido ao ruído de transmissão e ao filtro revelou uma grande variação de transmissão. Em sua tese, Constantini [11], ao adicionar ruído em suas gravações e compará-las através deste parâmetro, obteve-se um aumento de 156% de gravações diretas para ruidosas.

3.1.2. Diferenças interfalantes

O próximo passo da pesquisa foi analisar quais dos parâmetros acústicos tiveram ou não variação em relação aos sujeitos. Ou seja, se um parâmetro acústico de um participante da pesquisa não apresentou variação com o "criminoso", poderemos dizer, a princípio, que são a mesma pessoa.

Os parâmetros acústicos que apresentaram menor variação interfalantes (entre os sujeitos e o criminoso) foram a taxa de movimento do segundo formante, ΔC, taxa de elocução e frequência *baseline*. Em seguida através da análise de *boxplots* que ilustravam os parâmetros de sujeitos que mais se

assemelhavam com o "criminoso", concluímos que o sujeito 4 foi o que mais apresentou semelhanças em seus parâmetros acústicos com o criminoso.

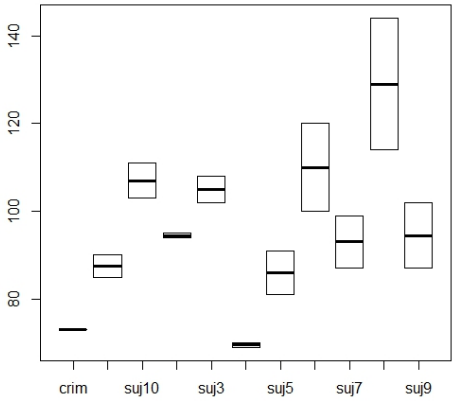

Figura 4: *Boxplot para média de duração das vogais para a variação interfalantes. A escala mostra na abscissa os sujeitos analisados: criminoso, sujeito 1, sujeito 10, sujeito 2, sujeito 3, sujeito 4, sujeito 5, sujeito 6, sujeito 7, sujeito 8 e sujeito 9.*

De acordo com o gráfico acima, podemos ver que o sujeito 4 é o sujeito que mais se assemelha ao "criminoso" a partir da comparação de suas medianas para a duração das vogais.
Podemos ver a seguir os valores de p para os parâmetros acústicos analisados. Utilizamos o teste Kruskal-Wallis para α = 0,05.

Tabela 3. *Resultados para testes de variação interfalantes para os parâmetros F2, taxa de F2, transição de F2, F0, frequência baseline e Ênfase espectral.*

Suj	F2	Taxa de F2	Transição de F2	F0	Baseline	Ênfase Espectral
Fligner-Killeen	p-value = 3.117e-15	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16
Kruskal-Wallis	p-value < 2.2e-16	p-value = 0.0002058	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16

Tabela 4. *Resultados para testes de variação interfalantes para os parâmetros média de duração das vogais, taxa de elocução e ΔC.*

Suj	Média de duração das vogais	Taxa de elocução	ΔC
Shapiro-Wilk	p-value = 1	p-value = 0.9744	p-value = 0.7885
Fligner-Killeen	p-value = 0.02925	p-value = 0.02925	p-value = 0.02925
Kruskal-Wallis	p-value = 0.06432	p-value = 0.1736	p-value = 0.5828

3.2. Teste de Duncan

Este teste estatístico *post-hoc* faz um agrupamento de valores semelhantes de todos os parâmetros acústicos analisados. Se duas médias não são estatisticamente diferentes, elas ficarão no mesmo grupo. Isto é, se um sujeito apresentar um maior número de médias paramétricas (em detrimento dos demais sujeitos) semelhantes com as do "criminoso", portanto seriam a mesma pessoa. Os sujeitos 5 e 7 apresentaram um número maior de médias semelhantes com as do "criminoso".

4. Conclusões

Os testes aplicados para a pesquisa mostraram resultados favoráveis aos objetivos propósitos. Segundo às análises estatísticas, os parâmetros que mais se mostraram robustos em relação à mudança de canal de transmissão foram a média da duração das vogais, taxa de elocução, ΔC e taxa de movimento de segundo formante.

Parâmetros acústicos como a frequência de formantes sofreram uma grande influência do canal de transmissão, tendo um aumento de 98 Hz da gravação direta à telefônica. Este fato é devido ao fenômeno de "deslocamento de frequências formânticas" [6], onde as taxas de baixa frequências sofrem um aumento ao passar pelo filtro do telefone celular. A frequência fundamental, assim como a frequência *baseline* apresentaram um aumento estatisticamente insignificante para a mudança de canal de transmissão, porém foram pouco determinantes para apresentarem informações idiossincráticas dos falantes.

Concluímos que os parâmetros acústicos de ritmo e tempo se mostraram promissores para futuras análises da Fonética Forense, com um bom desempenho discriminativo e robustez em canais de transmissão.

5. Agradecimentos

Agradecemos a CAPES por ter financiado esta pesquisa e à CNPq por financiar a continuação e aprimoramento da mesma.

6. Referências

[1] M. Jessen. "Forensic Linguistics". In: *Language and Liinguistics Compass*, n. 2, 2008, p. 671-711.
[2] P. Boersma, D. Weenink, Praat: doing phonetics by computer (Versão 5.2.25). [Computer program]. Disponível em: http://www.praat.org. Acesso em 2013.
[3] H. Hollien, *Forensic Voice Identification.* London: Academic Press, 2002.
[4] Agência Nacional de Telecomunicações. Disponível em: http://www.anatel.gov.br/institucional/. Acesso em 2012.
[5] H. J. Kunzel, "Beware of the telephone effect: The influence of telephone transmission on the measurement of formant frequencies". In: *Forensic Linguistics*, n 8, 2001, p. 80-99.
[6] C. Byrne, P. Foulkes. "The 'mobile phone effect' on vowel formants. In: *The International Journal of Speech, Language and the Law*, n 11, 2004, p. 83-102.
[7] F. Nolan, *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press, 2009.
[8] A. P. Machado, *Uso de técnicas acústicas para verificação de locutor em simulação experimental*. Dissertação de Mestrado (Linguística) - Universidade Estadual de Campinas, 2014.
[9] P. A. Barbosa. Forensic Data tracking. 2013.
[10] J. Chambers et al, R: The R Project for Statistical Computing (versão 3.2.3). [Computer program]. Disponível em: https://www.r-project.org/. Acesso em 2014.
[11] A. C. Constantini, *Individualização de características prosódicas em enunciados com baixa e alta relação sinal-ruído*. Tese de Doutorado (Linguística) – Universidade Estadual de Campinas, 2014.