Desviando da própria fala: implicações para a verificação de locutor por falantes e não-falantes de português brasileiro

Renata Regina Passetti, Plínio Barbosa
Grupo de Estudos de Prosódia da Fala
Universidade Estadual de Campinas
Campinas, Brasil
re.passetti@gmail.com; pabarbosa.unicampbr@gmail.com

Anders Eriksson
Departamento de Linguística
Universidade de Gotemburgo
Gotemburgo, Suécia
anders.eriksson.gu@gmail.com

***Resumo*— Este trabalho tem como objetivo avaliar a taxa de reconhecimento de locutor entre grupos de falantes e não-falantes do português brasileiro e investigar qual o tipo de informação (acústica e/ou lexical) empregada durante a tarefa de verificação de locutor, além de tecer considerações sobre possíveis pistas acústicas que estariam interferindo na decisão destes falantes. Os resultados obtidos pela análise comparativa da duração de unidades VV e do grupo acentual e, posteriormente, pela análise da mediana de *f0*, entre pares de amostras de fala (locutor-alvo e distrator falsamente escolhido), apresentam uma possível explicação para a escolha incorreta dos ouvintes.**

Palavras-chave: verificação de locutor; percepção da fala; disfarce vocal

***Abstract*— This research aims at analyzing some of the perceptual features used by speakers and non-speakers of Brazilian Portuguese for speaker verification in Brazilian Portuguese speech samples. The focus of the study was to evaluate the identification rate and to analyze which type of information (acoustic and/or lexical) was used by listeners during the speaker verification task. The results of a comparative analysis of duration-related acoustic salience and speech rate in VV units/s and, furthermore, the analysis of the speakers' *f0* median could be one explanation for listeners' wrong choices, since they could be relying their choices on these rhythmic cues.**

Keywords: speaker verification; speech perception; vocal disguise

I. INTRODUÇÃO

O desvio e manipulação da identidade fonético-fonológica do locutor para a realização de disfarces vocais provocam modificações prosódicas e acústicas nos padrões habituais da fala, com o intuito de ocultar a identidade do locutor em situações de interação com ouvintes leigos. O exame das estratégias perceptuais de ouvintes interessa aos estudos de verificação de locutor na medida em que contribui de maneira significativa para a compreensão dos processos de raciocínio em testemunhas auditivas.

As amostras de fala desempenham um papel fundamental no processo de verificação de locutor por testemunhas auditivas. Eriksson [5] afirma que os fatores que primeiramente influenciam a precisão na verificação de locutor são a duração da amostra de fala e a qualidade acústica da amostra. Em estudos iniciais de Pollack, Pickett & Sumby [7], analisou-se o efeito de diversas variáveis no desempenho da verificação de locutor. Os autores observaram que, até aproximadamente 1,2 segundo, a precisão da identificação de falantes aumentava conforme o tamanho da amostra, para palavras monossilábicas. Para as amostras longas, a variação fonética foi assumida como o fator mais importante para a identificação de um locutor em um conjunto fechado. Os autores concluíram que a duração apenas se mostra importante na medida em que admite uma amostragem estatística maior ou menor do repertório de fala do locutor.

Bricker e Pruzansky [3], ao analisarem os efeitos na verificação de locutor de estímulos com variações fonêmicas e duracionais, descobriram que a taxa de identificação somente aumentava com a duração se os estímulos mais longos também contivessem uma maior variação fonêmica. O aumento na precisão da identificação também estava diretamente relacionado ao número de fonemas na amostra de fala, mesmo quando a duração era controlada.

Ainda no campo de estudos sobre verificação de locutor, a influência de sotaques e línguas estrangeiras tem sido objeto de estudo de diversas pesquisas. Schiller e Köster [8], com o intuito de revisar determinados aspectos de uma investigação experimental sobre a influência do conhecimento da língua materna na prática de verificação de locutor, utilizaram gravações de seis falantes nativos de alemão para testar falantes nativos de inglês americano sem conhecimento de alemão, falantes de inglês americano com algum conhecimento de alemão e falantes nativos de alemão. Os resultados indicaram que a não familiaridade com a língua-alvo afeta a habilidade de reconhecimento de locutor. Sujeitos sem conhecimento de alemão obtiveram, significativamente, mais erros de identificação e sujeitos com algum conhecimento de alemão obtiveram resultados semelhantes aos dos falantes nativos de alemão. Os autores argumentam que a prática de reconhecimento de locutor parece não envolver apenas características puramente fonéticas, mas também a incorporação de informação linguística.

Köster e Schiller [6] replicaram este experimento [8] em ouvintes nativos de espanhol e chinês. Os resultados mostraram que ouvintes espanhóis e chineses com algum conhecimento de

Esta pesquisa é resultado do Estágio de Pesquisa no Exterior de Renata Regina Passetti, sob orientação dos Professores Doutores Plínio A. Barbosa e Anders Eriksson e realizada mediante bolsa concedida pela Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP, Processo 2012/10909-3, no período de 30/08/2012 a 29/09/2012.

alemão obtiveram melhores taxas de reconhecimento de locutor do que ouvintes espanhóis e chineses sem conhecimento da língua-alvo. Porém, quando comparados com falantes nativos de alemão e ouvintes nativos de inglês americano com algum conhecimento de alemão, os ouvintes espanhóis e chineses com algum conhecimento de alemão obtiveram resultados mensuravelmente piores. A partir desses estudos, concluiu-se que, em testes de reconhecimento de locutores, o conhecimento prévio da língua-alvo pelos sujeitos ouvintes aumenta a confiabilidade dos resultados de reconhecimentos.

O desempenho de reconhecimento de locutor em estudos envolvendo língua e sotaques estrangeiros na verificação de locutor parece estar mais propenso a ser afetado pela familiaridade do ouvinte com a língua. O desempenho também se mostra sensível à duração da amostra de fala, como atestado anteriormente por [7] e [3]. Com base nos estudos desenvolvidos acerca da verificação de locutor envolvendo amostras de falas de línguas e sotaques distintos dos ouvintes, o principal objetivo desta pesquisa foi avaliar a taxa de reconhecimento de locutor em amostras de fala do português brasileiro apresentadas a ouvintes nativos de sueco e a ouvintes nativos de português brasileiro que vivem na Suécia e analisar qual o tipo de informação (acústica e/ou lexical) empregada durante a tarefa de verificação, além de tecer considerações sobre as possíveis pistas acústicas que estariam interferindo na decisão de falantes e não-falantes do português brasileiro.

II. METODOLOGIA

A. *Corpus*

O *corpus* utilizado neste estudo foi composto por 7 gravações de locutores brasileiros, três homens e quatro mulheres, com idade média de 21 anos. As gravações consistiam em diferentes sessões de leitura de um discurso do apresentador de televisão Sílvio Santos, evocando dois tipos diferentes de elocução: primeiramente, com o estilo natural de fala dos sujeitos e depois utilizando o disfarce vocal "lápis na boca", pela utilização de um lápis posicionado firmemente entre os dentes frontais. As gravações foram utilizadas na elaboração de 6 filas de reconhecimento, três com amostras de fala masculinas e três com amostras de fala femininas. Cada fila de reconhecimento era composta por uma amostra de referência (voz disfarçada), com duração de 40 segundos, separada por um *beep* de outras 6 amostras de fala não-disfarçadas, cujas durações variavam entre 6,5 a 8 segundos que, por sua vez, estavam separadas entre si por um período de silêncio de 4 segundos. Das amostras de fala não-disfarçadas, uma consistia na voz do locutor-alvo, ou seja, aquela ouvida anteriormente com o disfarce vocal e as outras 5 consistiam em vozes distratoras.

B. *Ouvintes*

Os participantes do teste de percepção pertenciam a dois grupos: (1) ouvintes nativos de sueco, sem nenhum conhecimento em português brasileiro (10 homens e 10 mulheres) e (2) ouvintes nativos de português brasileiro, que moravam na Suécia há, pelo menos, 1 ano (7 homens e 3 mulheres).

C. *Procedimentos experimentais*

O teste de percepção foi apresentado por meio de uma apresentação de slides, elaborada nos moldes propostos por Broeders e Amelsvoort [4], iniciada pela simulação de um crime, no qual esses ouvintes haviam sido supostamente vítimas, seguida de instruções sobre a condução da tarefa de verificação de locutor. A tarefa dos ouvintes era, inicialmente, ouvir a amostra de fala disfarçada com a voz do locutor-alvo e, em seguida, tentar descobrir qual dentre as seis amostras de fala sem disfarce vocal pertencia ao locutor-alvo.

As gravações utilizadas nas amostras de fala das filas de reconhecimento consistiam em diferentes sessões do mesmo texto enunciado por diferentes locutores e diferentes amostras de fala de um mesmo locutor (denominadas "versões"). Dessa forma, um locutor que havia desempenhado o papel de "locutor-alvo" em uma fila de reconhecimento, poderia ser distrator em outra fila com amostras de fala de locutores do mesmo sexo.

O teste de percepção foi aplicado no estúdio de fonética do Departamento de Filosofia, Linguística e Teoria da Ciência da Universidade de Gotemburgo, Suécia. A duração total do teste era de, aproximadamente, 15 minutos.

III. RESULTADOS

A análise do desempenho dos ouvintes no teste de percepção considerou três etapas. Primeiramente, as respostas dos ouvintes foram avaliadas de acordo com 4 diferentes tipos de análises: (1) análise da taxa de reconhecimento do locutor-alvo pela porcentagem de identificações falsas e corretas em cada fila de reconhecimento; (2) análise das falsas identificações por distrator; (3) análise das falsas identificações pela posição da amostra de fala na fila de reconhecimento e (4) análise das falsas identificações pelas amostras de fala de um mesmo distrator (versão). A segunda etapa de análise consistiu em um estudo acerca da duração relativa do grupo acentual e da taxa de elocução em unidades Vogal-Vogal (unidades VV) por segundo entre as amostras de fala do locutor-alvo e do distrator erroneamente escolhido com maior frequência para cada uma das filas de reconhecimento. Esta análise tinha como objetivo investigar se estes parâmetros poderiam ser considerados como pistas acústicas, nas quais os ouvintes estariam apoiando suas falsas escolhas. Por fim, foi conduzida uma análise da mediana da frequência fundamental (doravante *f0*) entre os pares de locutores (alvo e erroneamente escolhido) para cada fila de reconhecimento, cujo objetivo era aprimorar os resultados estabelecidos para a duração das unidades VV e para a duração do grupo acentual e também auxiliar na compreensão das falsas identificações feitas por ambos os grupos de ouvintes.

A. *Primeira etapa de análises: desempenho de ambos os grupos de ouvintes pela análise de identificações corretas e falsas e porcentagem de falsas identificações por distrator, pela posição na fila de reconhecimento e pelas versões de um mesmo distrator.*

Quanto ao desempenho entre os dois grupos de ouvintes (suecos e brasileiros), os resultados mostraram que não houve diferenças significativas em relação à média percentual de identificações corretas (m_{suecos}= 36,6%, $m_{brasileiros}$= 45%). Apesar de os ouvintes brasileiros terem obtido um desempenho melhor, os resultados revelaram que o teste mostrou-se difícil para ambos os grupos de ouvintes, atestando nível baixo de concordância entre os ouvintes de um mesmo grupo ($kappa_{suecos}$= 0,12; $kappa_{brasileiros}$= 0,02). A fila de

reconhecimento identificada corretamente com maior frequência, para ambos os grupos, foi a número 2, com porcentagem de acertos de 70% para os ouvintes suecos e 80% para os ouvintes brasileiros. O melhor desempenho dos ouvintes para esta fila de reconhecimento pode ser explicado pela posição do locutor-alvo na fila de reconhecimento. Este se encontrava na primeira posição, logo após a amostra de referência (voz disfarçada), o que pode ter auxiliado na decisão dos ouvintes, devido a maior retenção de características perceptuais da voz do locutor-alvo na memória acústica desses ouvintes. O distrator que obteve o maior número de falsas identificações para os ouvintes suecos foi o locutor 1, tendo a versão 1.3 escolhida mais vezes dentre suas outras versões. Para os ouvintes brasileiros, o distrator falsamente escolhido mais vezes foi o número 4, em sua versão 4.3. A posição erroneamente escolhida na maioria das vezes foi a número 3 para ambos os grupos, representando para os ouvintes suecos, aproximadamente 26% sobre a quantidade total de posições erroneamente escolhidas e para os ouvintes brasileiros, aproximadamente 34%.

B. Análise da duração relativa do grupo acentual e da taxa de elocução em unidades VV/s entre as amostras de fala do locutor-alvo e dos distratores erroneamente escolhidos.

A análise da taxa de reconhecimento do locutor-alvo pela avaliação das falsas identificações do locutor-alvo pôde ser aprimorada por uma análise comparativa da duração relativa da saliência duracional e da taxa de elocução em unidades VV por segundo. Barbosa [1] descreve a unidade VV como um componente prosódico, cuja evolução duracional ao longo do enunciado está diretamente associada ao ritmo da fala, revelando sua estrutura rítmica. Os valores das unidades VV em uma passagem permitem avaliar tanto a taxa de elocução, bem como a taxa da duração normalizada das saliências duracionais nessa passagem. Estes são importantes parâmetros a ser considerados em análises de fala inter-locutores, pois permitem avaliar se a diferença dos valores médios entre os distratores escolhidos e o locutor-alvo são suficientes para justificar uma falsa escolha.

A análise comparativa da duração de unidades VV e da duração dos grupos acentuais, delimitados pelos picos locais de duração relativa, foi calculada pelo *script* SGdetector [1], a partir dos intervalos VV marcados em *textgrids* no programa PRAAT [2]. A distância entre dois picos locais de durações normalizadas de unidades VV define um grupo acentual, cuja duração é computada pelo programa.

A análise comparativa da duração de unidades VV e da duração relativa da saliência duracional foi calculada entre as amostras de fala do locutor-alvo e do distrator falsamente escolhido com maior frequência, em cada fila de reconhecimento, para ambos os grupos de ouvintes (suecos e brasileiros). Um teste T de variáveis independentes, com nível de significância de 0,05, foi conduzido com a finalidade de testar a hipótese nula de que as duas amostras comparadas possuíam a mesma média para a duração das unidades VV e para a duração do grupo acentual. A tabela 1 mostra os resultados do teste T para a análise comparativa entre os dois grupos de ouvintes.

TABELA 1: RESULTADOS DO TESTE T PARA A ANÁLISE COMPARATIVA DA DURAÇÃO DAS UNIDADES VV (VV) E DA DURAÇÃO DO GRUPO ACENTUAL (DUR_{GA}).

Resultado do teste T: probabilidade p				
Fila de reconhecimento	**Ouvintes suecos**		**Ouvintes brasileiros**	
	VV	***Dur*$_{GA}$**	***VV***	***Dur*$_{GA}$**
1	**1.10^{-3}**	0,27	**0,04**	0,43
2	0,98	0,26	**0,04**	0,52
3	0,83	0,74	0,83	0,74
4	0,17	0,11	0,53	0,69
5	**5.10^{-4}**	0,08	**1.10^{-4}**	0,12
6	0,55	0,11	0,94	0,08

O teste T não apresentou resultados que corroborassem para a aceitação da hipótese nula ($p > 0,05$) para as durações médias das unidades VV na fila de reconhecimento 1, para os ouvintes suecos ($p < 0,001$) e para os ouvintes brasileiros ($p < 0,04$), também na fila de reconhecimento 5, com um p-valor menor que 5.10^{-4} e 1.10^{-4} para ouvintes suecos e brasileiros, respectivamente. O teste T também não mostrou resultados significativos na fila de reconhecimento 2 para os ouvintes brasileiros ($p < 0,04$). Todos os resultados do teste T para a duração do grupo acentual não apresentaram diferenças significantes entre os valores estabelecidos para os locutores-alvo e os distratores ($p > 0,05$).

Os resultados mostraram que, na maioria das filas de reconhecimento (exceto nas filas de reconhecimento 1 e 5, para ambos os grupos de ouvintes, e 2, para os ouvintes brasileiros), o locutor-alvo e o distrator comumente escolhido (falsa identificação) produziram a mesma quantidade de unidades VV por segundo. Todos os locutores comparados (alvo e distratores comummente escolhidos) produziram as mesmas taxas de duração do grupo acentual. A conservação da hipótese nula, em ambos os parâmetros de análise, pode explicar a escolha incorreta dos locutores, já que eles podiam apoiar suas escolhas nestas pistas rítmicas.

A consideração de outro parâmetro – a mediana de *f0* (Hz) – para todas as amostras de fala selecionadas e sua comparação entre os dois locutores escolhidos (alvo e distrator) também auxiliou na compreensão das taxas de falsas identificações pelos ouvintes. Os resultados da mediana de *f0* serão apresentados na seção seguinte.

C. Análise da mediana de f0 (Hz) entre as amostras de fala selecionadas (locutor-alvo e distratores).

A mediana de *f0* (Hz) foi calculada para cada par de locutores (alvo e distrator erroneamente escolhido com maior frequência) em todas as filas de reconhecimento. A análise deste parâmetro pôde aprimorar os resultados estabelecidos para a duração das unidades VV e para a duração do grupo acentual, e também auxiliar na compreensão das falsas identificações de ambos os grupos de ouvintes. A tabela 2 apresenta a taxa da mediana de *f0* para todos os pares de locutores comparados para ambos os grupos de ouvintes, em todas as filas de reconhecimento.

TABELA 2: VALORES DA MEDIANA DE *F0*, EM HERTZ, PARA TODOS OS PARES DE LOCUTORES (ALVO E DISTRATORES).

Taxas da mediana de *f0* (Hz)				
Fila de reconhecimento	**Ouvintes suecos**		**Ouvintes brasileiros**	
	Locutor-alvo	***Distrator***	***Locutor-alvo***	***Distrator***
1	132	164	132	130
2	140	164	140	118
3	255	267	255	267
4	236	258	236	230
5	223	222	223	266
6	143	127	143	117

Os valores para a mediana de *f0* mostraram-se similares entre os pares de locutores nas filas de reconhecimento 3 e 5 para os ouvintes suecos. Um ponto interessante a ser notado, é que, pra esse grupo de falantes, o resultado do teste T para a duração das unidades VV por segundo (taxa de elocução) não foi significante para as amostras de fala da fila de reconhecimento 5, mas uma possível explicação para a escolha errônea desses locutores pode ser dada pela similaridade nos valores estabelecidos para a mediana de *f0* desses locutores, que é praticamente a mesma (223 Hz, para o locutor-alvo e 222 Hz para o distrator erroneamente escolhido). A mesma hipótese pode ser utilizada para algumas comparações entre as escolhas dos ouvintes brasileiros. Os valores da mediana de *f0* apresentaram resultados similares para as amostras de falas entre os locutores nas filas de reconhecimento 1, 3 e 4. Para este grupo de ouvintes, o teste T para análise da taxa de elocução das unidades VV/s não apresentou resultados significantes para as amostras de fala da fila de reconhecimento 1. Por outro lado, a similaridade entre os valores da mediana de *f0* para o locutor-alvo e o distrator comumente escolhido, na fila de reconhecimento 1, pode explicar a falsa identificação desses ouvintes.

IV. CONCLUSÃO

Apesar de os resultados não terem apresentado diferenças significativas na média percentual de escolhas corretas entre ambos os grupos de ouvintes, o fato dos ouvintes brasileiros apresentarem uma tendência em alcançar um melhor desempenho no teste de percepção, pode concordar com discussões anteriores de [6] que afirmam que a prática de reconhecimento de locutor parece não envolver apenas características puramente fonéticas, mas também a incorporação de informação linguística, desde que esse grupo de ouvintes possua um conhecimento prévio da língua do locutor-alvo.

A análise da taxa de reconhecimento de locutor pela avaliação das falsas identificações do locutor-alvo foi aprimorada por uma análise comparativa da duração do grupo acentual e da taxa de elocução em unidades VV por segundo. Os resultados de um teste T de variáveis independentes (α = 0,05) mostraram que a conservação da hipótese nula em ambos os parâmetros, na maioria das filas de reconhecimento comparadas, poderia explicar a escolha errônea dos ouvintes, visto que eles poderiam estar apoiando suas escolhas nestas pistas rítmicas. A análise da mediana de *f0* ajudou a compreender as taxas de falsas identificações e também a aprimorar os resultados das durações médias das unidades VV (inverso da taxa de elocução em VV/s) e das durações dos grupos acentuais. O resultado mais importante da análise da mediana de *f0* diz respeito aos valores entre os pares de locutores da fila de reconhecimento 5 (para os ouvintes suecos) e na fila de reconhecimento 1 (para os ouvintes brasileiros). Para essas filas de reconhecimento, os valores atestados para este parâmetro mostraram algumas similaridades entre os pares de locutores escolhidos, o que auxiliou na compreensão das falsas identificações dos ouvintes, visto que os resultados apresentados para o teste T para a taxa de elocução em unidades VV/s foram inconclusivos para essas filas de reconhecimento.

REFERÊNCIAS

[1] BARBOSA, P. A. *Incursões em torno do ritmo da fala*. Campinas, Brasil: Pontes/FAPESP, 2006, p. 170.

[2] BOERSMA, P. & WEENINK (2012). *Praat: doing phonetics by computer* (Versão 5.3.16) [Programa computacional]. Disponível em: <http://www.fon.hum.uva.nl/praat/>

[3] BRICKER, P.D.; PRUZANSKY, S. *Effects of Stimulus Content and Duration on Talker Identification*. In: The Journal of the Acoustical Society of America, v.40, p. 1441-1450, 1966.

[4] BROEDERS, A.P.A; VAN AMELSVOORT, A.G. *Lineup construction for forensic earwitness identification: a practical approach*. In the Proceedings of the 7th International Congress of Phonetic Sciences. San Francisco, p. 1373-1376, 1999.

[5] ERIKSSON, A. *Tutorial on Forensic Speech Science*. Paper presented at tutorial session on Forensic Speech Science, Interspeech (9th European Conference on Speech Communication and Technology), Lisbon, Portugal, 2005.

[6] KÖSTER, O.; SCHILLER, N.O. *Different influences of the native language of a listener on speaker recognition*. In: Forensic Linguistic, v.4, n.1, p.176-185, 1997.

[7] POLLACK, I.; PICKETT, J.M.; SUMBY, W.H. *On the identification of speakers by voice*. In the Journal of the Acoustical Society of America, v.26, n.3, p.403-412, 1954. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] SCHILLER, N.O.; KÖSTER, O. *Evaluation of a foreign speaker in forensic phonetics: a report*. In: Forensic Linguistics, v.3, n.1, p. 176-185, 1996.