Uso de técnicas acústicas para verificação de locutor em simulação experimental

Aline de Paula Machado & Plínio Almeida Barbosa

Universidade Estadual de Campinas (UNICAMP)

Abstract. *The aim of this research was to investigate the efficiency of a set of acoustic measures to identify one randomly pre-selected individual, named "the criminal", out of a group of ten speakers of Brazilian Portuguese. Among the measures used were the first two formants of vowels, fundamental frequency, the duration of syllables and vowels, formant movement rate, intensity and the standard deviation of consonantal interval durations (ΔC). We analyzed all the Brazilian Portuguese vowels. All the subjects were speakers of Brazilian Portuguese from the states of São Paulo, Rio Grande do Sul, Pará and Bahia. All the samples were extracted from interviews recorded in low quality acoustic environments. The samples were divided into two groups i) interviews and ii) recorded telephone conversations (mobile phone to mobile phone). The parameters that were most robust were rhythm and timing, such as duration, speech rate, ΔC and formant movement rate for the second formant. Because of their interspeaker variability, timing measures proved to be highly discriminative. The statistical tests showed that the speech of three of the subjects contained similarities to that of "the criminal".*

Keywords: *Forensic phonetics, speaker verification, acoustic phonetics.*

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 em um grupo de dez falantes do português brasileiro. Um sujeito desse grupo foi sorteado e nomeado o "criminoso". Entre as medidas usadas na pesquisa estão as frequências dos dois primeiros formantes, a frequência fundamental média, a duração de unidades do tamanho da sílaba e da vogal, a dinamicidade dos formantes e o desvio padrão de durações de intervalos consonânticos (ΔC). Analisamos todas as vogais do português brasileiro. Todos os entrevistados eram falantes do português brasileiro de regiões dos estados de São Paulo, Rio Grande do Sul, Pará e Bahia. Todas as amostras foram extraídas de entrevistas gravadas em ambientes acústicos de baixa qualidade. Os trechos escolhidos para essa análise foram divididos em dois grupos, (i) entrevistas e (ii) gravações telefônicas (de celular para celular). Os parâmetros mais robustos foram ritmo e tempo, tais como duração, taxa de elocução, ΔC, e taxa de movimento do segundo formante. As medidas*

temporais da pesquisa, por serem as mais variáveis intersujeitos, tiveram grande poder discriminador. Os testes estatísticos apontaram que três dos indivíduos estudados, apresentavam semelhanças com o "criminoso".

Palavras-chave: *Fonética forense, verificação de locutor, fonética acústica.*

Introdução

O reconhecimento de locutor se caracteriza por "qualquer atividade pela qual uma amostra de fala é atribuída a uma pessoa com base em suas propriedades fonético-acústicas ou perceptuais" (Jessen, 2008: 671). Entra em jogo a fonética forense: a aplicação de técnicas de análise fonética em contextos policiais jurídicos. Essa é uma área que vem crescendo desde a década de 1960 no Reino Unido e que tem tido sua importância disseminada para todo o globo desde então (French, 1994).

No Brasil, essa subárea da fonética não é demasiadamente promovida nas universidades, e suas técnicas de análise pela polícia não são, de modo geral, semelhantes às usadas em outros países cujo sistema judicial demanda esse tipo de análise. No exterior, normalmente o especialista que faz as análises das amostras de fala trazidas pela polícia é um foneticista ou um profissional com extenso *background* fonético-linguístico. Nota-se, então, a relação estreita que existe entre departamento policial e a universidade, o que facilita essa troca de serviços. No Brasil, a análise das gravações colhidas é feita por um perito (não necessariamente um foneticista) utilizando métodos auditivos e de análise acústica. O profissional também tem como opção o uso de método automático de análise utilizando um *software* de computador, por exemplo, o Batvox (Gold e Fench, 2011). Nos países da Europa, como Inglaterra, Suécia e Alemanha, entre outros, o uso de sistemas automáticos é acompanhado por *insights* de um profissional com conhecimentos em fonética ou em linguística, tal como ocorre na Universidade de Gotemburgo, onde o *software* utilizado é o ALIZE SpkDet e os resultados obtidos pelo programa são combinados com análise acústico-auditiva tradicional (Eriksson, 2012).

Diante da necessidade de trabalhos de pesquisa no Brasil, este trabalho teve como objetivo reconhecer um indivíduo através de sua fala dentro de um grupo de dez falantes do português brasileiro, assinalando quais parâmetros acústicos são relevantes para a análise desse reconhecimento. Com relação ao método de análise, usou-se o método acústico semiautomático. Ressalta-se que o método auditivo 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).

Verificação de locutor e método de análise

Escolhemos utilizar em nossa pesquisa o termo "verificação", em vez de "identificação de locutor". Segundo Hollien (2002), na verificação de locutor é a identidade da pessoa que está em questão, ou seja, a voz é utilizada para acessar uma conta de banco por telefone ou alguma informação privilegiada. Essa análise é controlada por analistas e feita por computadores que comparam a voz questionada com uma voz já armazenada, o que permite que a verossimilhança seja verificada. O falante a ser avaliado, portanto, é cooperativo: ele produz várias amostras de sua fala para a comparação de voz sem, provavelmente, adotar algum tipo de disfarce ou variações em sua voz. Embora a fonética forense seja associada à tarefa de identificação de locutor, ou seja, à identificação

de uma única pessoa (desconhecida) em uma população — reconhecimento indireto de um sujeito —, na prática, a identificação de locutor acaba sendo verificação, já que o trabalho forense, na maioria das vezes, toma um número finito de suspeitos para o reconhecimento de um criminoso a partir da comparação entre gravações questionada e de referência.

As bases para a pesquisa

A Fonética Forense é constituída tanto pela aplicação de conhecimentos, teorias e métodos da fonética geral em tarefas práticas em contexto de trabalho policial ou de apresentação de evidência em tribunal, quanto pelo constante desenvolvimento de novos métodos, teorias e conhecimentos (Jessen, 2008).

Em uma situação forense há geralmente o seguinte cenário: uma gravação que pode vincular ou desvincular um indivíduo a uma atividade criminosa a partir da comparação com uma gravação de referência. A primeira, ou gravação questionada, costuma ser feita por interceptação telefônica, situação em que o indivíduo tende a falar espontaneamente, não sabendo que está sendo gravado. A segunda gravação geralmente é realizada em ambiente acusticamente tratado e os peritos pedem ao suspeito que leia um texto de forma clara para um microfone posicionado a sua frente. Porém, essa situação é apenas ideal em papel, não acontecendo, em grande parte, como metodologia adotada no Brasil. O sujeito pode, por exemplo, apresentar um nível de alfabetização muito baixo, não tornando viável a tarefa de leitura de texto. Esse tipo de técnica de análise acaba se tornando um ponto que dificultará o trabalho de perícia, pois são gravações em contextos diferentes, (1) uma situação de fala espontânea, com o discurso fluente e (2) em laboratório, com material lido. Com isso, palavras que foram encontradas na primeira gravação podem não estar presentes na segunda. O nível de estresse e a naturalidade da fala também afetam a produção de palavras, o que prejudica a precisão necessária para realizar a comparação das análises de cada indivíduo (Byrne e Foulkes, 2004). Outro motivo que pode dificultar a análise são os efeitos que o telefone celular pode causar na gravação.

O efeito do celular

Em muitas situações forenses, cientistas têm em mãos como material de avaliação escutas telefônicas que são, em sua grande maioria, de péssima qualidade, (embora geralmente sejam essas as únicas fontes sonoras para a extração de parâmetros acústicos) por meio das quais eles devem apresentar algum resultado substancial para o júri. Por isso, trazemos tal situação para a pesquisa, simulando casos de interceptação telefônica.

Primeiramente, escolhemos o celular, em lugar do telefone fixo, pois aquele aparelho é de grande uso pelos criminosos; além disso, sabe-se que no Brasil, há mais de 271[1] milhões de linhas de telefone celular. Foi evidenciado também que a gravação por telefone fixo, em comparação ao telefone celular, apresenta resultados mais robustos, principalmente para o primeiro formante (Kunzel, 2001; Byrne e Foulkes, 2004). Kunzel (2001) considera os efeitos do telefone fixo no sinal de fala para calcular quais consequências a diferença de canal de transmissão (no caso, telefone celular) pode causar nas frequências dos formantes nas gravações. Um dos fatores que influenciam a análise de dados a partir de gravação telefônica é a questão da perda do sinal, além de ruído ambiental — no caso do celular, a distorção do próprio aparelho (pela compressão provocada pelo *vocoder*) é o elemento mais crítico para análise fonética.

Alguns efeitos causados pelo telefone celular foram evidenciados por Byrne e Foulkes (2004). Esperava-se encontrar durante a pesquisa uma degradação do sinal da fala das gravações coletadas advinda da combinação desses efeitos.

i. Efeitos do ambiente: Um dos efeitos mais comuns que telefones podem causar no sinal de fala está ligado ao ambiente físico, isto é, por vezes ligações telefônicas podem acontecer em ambiente de alto nível de ruído de fundo, como no trânsito. Assim, esse tipo de efeito deve ser levado em consideração em análises forenses, pois os ruídos podem afetar informações cruciais no sinal da fala.
ii. Efeito dos falantes: Os próprios falantes influenciam nos dados da conversação telefônica, dado que eles tendem a modificar seu comportamento ao falar por telefone e a tornarem-se, por exemplo, mais formais (no caso do inglês britânico, segundo os autores). Já em interceptações telefônicas de crimes brasileiros, pode-se ocorrer o oposto, i. g. conversas entre integrantes de quadrilhas, apresentando tons informais, gírias etc. O registro telefônico da voz muda consciente ou subconscientemente influenciando na taxa de elocução, na qualidade da voz e, como dito anteriormente, na pronúncia. Um dos efeitos mais comuns que pode ser mencionado é o aumento do volume da voz do indivíduo enquanto esse fala ao telefone, o que afeta diretamente a frequência fundamental do falante (F0).
iii. Efeitos técnicos: Ocorre o que é chamado de "distorção espectral", isto é, o aumento das frequências que se encontram acima do filtro passa-baixa (300 Hz) e a diminuição das frequências que se encontram ligeiramente abaixo do filtro passa-alta (3500 Hz). Em outras palavras, as frequências que estão abaixo de 300 Hz e acima de 3.500 Hz são "apagadas" pelo filtro do telefone celular. Um outro exemplo de efeito técnico é o fenômeno conhecido como "deslocamento de frequências": quanto menor a frequência (por exemplo, o primeiro formante), mais atenuada ela fica pelo canal telefônico em comparação a uma gravação direta (Kunzel, 2001; Byrne e Foulkes, 2004). O contrário também acontece, causando a perda dos componentes de alta frequência, algo que é destrutivo para a identificação forense de falante, já que um grande número de informações (qualidade de voz, por exemplo) é codificado em faixa de frequências mais altas das vogais.

Um dos motivos que nos leva a acreditar que ocorra expressiva diferença entre telefone fixo e celular é que os telefones celulares estão sujeitos a um maior alcance de influências ambientais que os telefones fixos. Pelo fato de os primeiros poderem ser usados em qualquer lugar, muitos tipos diferentes de ruído de fundo serão encontrados nas gravações. Além disso, a telefonia celular utiliza taxas de transmissão inferiores, com compressão e codificação maiores do que a de telefonia fixa.

Parâmetros acústicos estudados para a pesquisa

Frequência fundamental e frequência *baseline*

A frequência fundamental é o correlato acústico da frequência de vibração das pregas vocais na produção de voz (Jessen, 2008). Ela é um parâmetro útil para a comparação interfalantes no ambiente forense. Suas medidas de distribuição de longo-termo, como sua média, são sempre sugeridas por pesquisadores da área (Eriksson, 2012; Rose, 2002). Segundo Eriksson (2012), o seu cálculo depende diretamente da duração da amostra de fala, ou seja, é necessário um tempo mínimo de trecho de fala para a extração de seu valor. Enquanto alguns autores sugerem durações de 14 segundos (Horii, 1975 *apud* Eriksson,

2012), outros sugerem de 60 segundos (Nolan, 1997) ou até de 2 minutos (Baldwin e French, 1990). Nesta pesquisa extraímos a frequência fundamental global do trecho com gravações que tiveram duração mínima de 50 segundos.

Algumas causas podem influenciar na variação da frequência fundamental, como fatores fisiológicos e emocionais do falante (idade, tabagismo, doença, intoxicação, estresse etc.), além de elementos externos, como ruído na amostra de gravação (Braun, 1995 *apud* Eriksson, 2012). Um outro fator que pode influenciar na variação desse parâmetro é o disfarce, pois indivíduos tendem a aumentar ou diminuir sua frequência fundamental para disfarçar a voz em algumas situações de crime (Kunzel, 2001).

Em meio a essa variação, que pode causar uma distorção na medida de F0, Lindh e Eriksson (2007) desenvolveram uma forma de representação para a frequência fundamental chamada de *baseline*. A frequência *baseline* se fundamenta na proposta de um nível de frequência fundamental neutro. Esse nível é um ponto estável estimado como 1,43 desvios-padrão de F0 abaixo da média de F0. Ela foi testada em diferentes materiais de fala que variavam quanto ao estilo de fala, esforço vocal e qualidade de gravação. Esta última condição consistia em gravações usando diferentes canais de transmissão, gravador digital e também telefone celular. Os resultados foram robustos para todos os contextos de gravação.

Frequência de formantes

Formantes são frequências de ressonância no trato vocal. Eles são constituídos por formas e volumes de diferentes cavidades do trato vocal (Fant, 1960).

É possível observar uma nova situação no que concerne o uso de canais de transmissão e sua relação com formantes: sabe-se que, hoje em dia, a maioria das chamadas telefônicas que têm conexão com crimes são feitas usando telefones celulares. Da mesma maneira, investigadores indicam que um número substancial de casos envolvendo fala gravada em celular está crescendo vertiginosamente (Öhman *et al.*, 2010). Assim, Byrne e Foulkes (2004) mostram como a transmissão por celular tem um efeito significativo nos formantes e, de maneira similar, Kunzel (2001) mostrou grandes efeitos causados pelo telefone fixo nos primeiros formantes.

Kunzel (2001) realizou um experimento no qual os participantes — 10 homens e 10 mulheres com idade de 20 a 59 anos – faziam uma leitura do texto "The North Wind and the Sun" em alemão, com taxa de elocução e altura de fala normais. As leituras duraram entre 35 a 40 segundos. O sinal de fala foi gravado simultaneamente em gravador e telefone. Foram analisados cerca de 25 contextos fonológicos de 13 vogais. O autor revelou que encontrou problemas com a própria metodologia do seu experimento, uma vez que o algoritmo usado ocasionava erros e, por exemplo, um formante mais alto do que deveria acabava sendo escolhido, situação que ocorreu principalmente nos dados telefônicos. Os resultados do experimento mostraram que todos os sujeitos apresentaram diferenças significativas para o primeiro formante em gravação telefônica, embora não houvesse diferenças significativas para o segundo formante. Outro dado expressivo foi que o valor da frequência do primeiro formante de cada vogal foi maior na transmissão telefônica do que por gravação direta. A diferença é maior para vogais fechadas como [i] e [u], média para vogais como [e] e [o] e menor ou zero para vogais abertas como [ɔ, a]. Com sua pesquisa, Künzel pôde concluir que os valores das frequências dos formantes

baixos das vogais de falantes masculinos e femininos são deslocados para cima (*formant shifted upwards*), causando erros de medidas.

Alguns anos depois, Byrne e Foulkes decidiram testar o efeito do telefone celular no sinal de fala como resposta ao experimento de Künzel, comparando os achados deste com a realidade da nova transmissão usada (via celular). O experimento consistia em 12 voluntários falantes do inglês, seis homens e seis mulheres, entre 20 e 39 anos. Enquanto esses sujeitos liam o texto "The story of Arthur the rat", duas gravações ocorriam simultaneamente. As gravações diretas foram realizadas por um microfone posicionado diretamente na frente do locutor, conectado em um gravador. Um segundo gravador foi conectado com o propósito de interceptar a chamada recebida na sala do experimentador. Os dados foram armazenados em um computador para análise acústica. Os resultados obtidos por Byrne e Foulkes indicam que: devido ao efeito de filtro da transmissão telefônica as frequências de F1 para a maioria das vogais foram maiores que seus homólogos nas gravações diretas; já as frequências do primeiro formante foram maiores do que as por telefone fixo apresentadas por Kunzel (2001); e as frequências do segundo formante não foram afetadas significativamente pelo canal telefônico. A frequência fundamental também foi comparada entre os dois contextos e obteve-se um aumento de 217Hz para a transmissão por telefone celular em relação ao fixo.

Dinamicidade de formantes de parâmetros do domínio de tempo

Um outro exemplo de estudo de formantes, só que relacionado a sua dinamicidade, foi proposto por Nolan *et al.* (2006). Os autores sugerem que as diferenças individuais em movimentos articulatórios podem ser usadas para a comparação de locutor. Seu experimento mostrou que esse parâmetro acústico apresenta informações idiossincráticas dos locutores, sendo calculado entre a diferença da frequência no contorno do formante e da sua área de transição até o centro do formante. Em seu experimento, valores ligados ao movimento das frequências do segundo formante apresentaram resultados determinantes para a discriminação de locutores. A medida foi feita da seguinte maneira: a partir do segmento de uma vogal, por exemplo, /u:/, foram feitas medidas do ponto médio dos contornos das frequências do primeiro e segundo formantes de cada segmento de /u:/ a partir do "formant tracker" do PRAAT. Um script foi usado para calcular a duração de cada segmento, dividindo-o em dez intervalos iguais. Um outro script mediu o centro das frequências dos formantes a cada passo, normalizando cada contorno formântico.

Outra medida de duração que também foi estudada com o objetivo de comparação de locutor é o ΔC, ou seja, o desvio padrão da duração de intervalos consonânticos. Dellwo e Koreman (2008), em estudo que consistia na gravação de dez falantes do alemão, avaliaram dados de diferentes taxas de elocução ao gravar seus sujeitos variando tais taxas de normal até rápida. O teste mostrou que parâmetros de tempo como o ΔC conseguiam capturar informações idiossincráticas dos sujeitos, mantendo-se robusto em diferentes condições de fala.

Ênfase espectral

Traunmüller e Eriksson (2000) tratam a ênfase espectral como a diferença entre a intensidade acústica do sinal integral e a intensidade do sinal submetido a um filtro passa-baixa com um limite de banda superior definido pela expressão 1, 5 F0, em que F0 é a média da frequência fundamental na vogal sendo analisada. Esperamos desse parâmetro uma

grande variação para o canal telefônico devido ao ruído e ao filtro. Segundo Constantini (2014), a ênfase espectral, em seu experimento, apresentou um aumento de 156% em gravações com ruído, inserido artificialmente pelo PRAAT, em relação às gravações originais.

Taxa de elocução

A taxa de elocução (*speech rate*), é o número de unidades da fala produzidas por minuto ou por segundo. As notações mais comuns são palavras por minuto e sílabas por segundo (Eriksson, 2012). Neste trabalho, ela é medida a partir da média da duração das unidades V-V, unidade do onset de uma vogal até o *onset* da vogal imediatamente seguinte. Pode ser medida automaticamente — no caso de boa qualidade na amostra de fala estudada — ou manualmente, quando há baixa qualidade na gravação. Em outras palavras, a ideia deste parâmetro é contar quantas unidades existem em um determinado trecho, medir a duração deste mesmo trecho e dividir o primeiro número pelo segundo. Esse cálculo resulta em uma taxa, um número x de unidades de fala (sílabas, V-V etc.) por unidade de tempo (em geral, segundos).

Segundo Eriksson (2012), a taxa de elocução apresenta um baixo poder de discriminação interfalantes, apresentando uma variação intrafalante alta. Testaremos nesta pesquisa como ela é afetada pelo canal telefônico, uma vez que a detecção do início da vogal pode ser prejudicada pelo canal. Neste trabalho, levando em consideração que a média da duração de unidade do tamanho da sílaba é o inverso da taxa de elocução e que, portanto, diferenças entre essas médias assinalam diferenças nas taxas, tomaremos a duração média da unidade V como medida de taxa de elocução.

O procedimento e os resultados da pesquisa

Este artigo é resultado de uma pesquisa de mestrado que teve como objetivo identificar um indivíduo pela voz em um grupo de dez falantes do português brasileiro divididos em quatro estados, São Paulo, Rio Grande do Sul, Bahia 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 uma faixa etária de 18 a 28 anos, com nível de educação mínimo de ensino superior incompleto (completando a Graduação) e moraram a maior parte da vida (mais do que a metade) em suas respectivas cidades natais.

Para realizar essa tarefa, analisamos os seguintes parâmetros acústicos, de todas as vogais do português brasileiro, de cada falante: frequência dos dois primeiros formantes, frequência fundamental média, taxa de elocução, frequência *baseline*, ênfase espectral, a dinamicidade dos formantes e o desvio padrão de durações de intervalos consonânticos (ΔC).

As amostras de todos os indivíduos foram gravadas em dois canais de gravação, gravação direta e gravação por telefone celular; essa última simula a dificuldade encontrada pelos peritos ao analisar gravações de baixa qualidade, tal como pode ser observado, por exemplo, em uma interceptação telefônica, cujo áudio apresenta ruído e deterioração. Além disso, o indivíduo escolhido, ao qual nos referimos como "criminoso", teve também sua fala gravada em ambiente acusticamente tratado para uma análise comparativa mais robusta. Simulamos um caso forense habitual, de crime, tendo como objetivo prin-

cipal o reconhecimento do "criminoso" dentro do grupo de falantes, além de mostrar qual método de análise estatística e parâmetros acústicos são mais eficazes para essa tarefa.

Nesta pesquisa, a gravação em estúdio da qual o indivíduo participa não foi feita através da leitura de um texto, mas em formato de entrevista conduzida pelos pesquisadores. Isso se deu com o objetivo de inserir os mesmos assuntos discutidos na primeira gravação e de deixar o entrevistado o mais à vontade possível para que sua fala fosse fluente e espontânea. Tentando atingir um grau mais próximo de fala espontânea, foi feita uma gravação de cada locutor simulando uma conversa corriqueira na qual eram abordados assuntos do cotidiano, como trabalho, plano para as férias etc.

Foram feitas vinte e uma gravações, dez usando o Mini Gravador Coby Cx-r190 ao ar livre, dez por telefone celular e uma gravação direta em ambiente acusticamente tratado. Nas gravações telefônicas, utilizou-se um celular *Samsung Galaxy Young* pela rede da operadora TIM. 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 se conectava com o *desktop*, e 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 de interesse extraídas automaticamente pelo *script* ForensicDataTracking, desenvolvido e disponibilizado por Barbosa (2013).

Apresentamos na Tabela 1, os dados das gravações dos sujeitos participantes da pesquisa, assim como a origem de cada um.

O *script* automaticamente extraiu medidas para frequência do segundo formante (F2) das vogais, taxa de movimento de formante para o segundo formante, frequência *baseline*, média da frequência fundamental, duração das vogais, inverso da taxa de elocução (média da duração de unidade do tamanho da sílaba), ênfase espectral e ΔC.

Métodos de análise estatística e resultados

Para este experimento decidimos utilizar os testes estatísticos ANOVA e teste de Duncan. A seguir, explicaremos os resultados obtidos nas gravações através deles.

ANOVA

Todos os testes estatísticos utilizados nesta pesquisa foram feitos a partir do *software* R^2. O teste estatístico de ANOVA, ou análise de variância, é a técnica estatística que permite avaliar afirmações sobre as médias de populações. Ele verifica se existe uma diferença significativa entre as médias e se os fatores exercem influência em alguma variável dependente (Dowdy *et al.*, 2004).

Para a pesquisa, estudamos a ANOVA com o seguinte intuito: (i) determinar se os parâmetros acústicos analisados permaneciam robustos com a mudança de canal de transmissão, ou seja, de uma gravação direta por gravador digital para uma gravação por telefone celular, e (ii) se algum desses parâmetros acústicos conseguiriam determinar qual dos sujeitos analisados é o "criminoso". Para a realização desse teste é preciso seguir algumas condições, tal como verificar se os resíduos compõem uma distribuição normal, o que pode ser identificado pelo teste estatístico Shapiro-Wilk, assim como verificar a

Sujeito	Naturalidade	Canal de comunicação	Duração (min)	Número de segmentos (vogais)
1	Bahia	Ar livre	2:15	229
		Celular	3:26	461
2	São Paulo	Ar livre	3:40	515
		Celular	2:21	279
3	São Paulo	Ar livre	1:50	152
		Celular	1:50	193
4	São Paulo	Ar livre	1:05	102
		Celular	2:07	185
5	São Paulo	Ar livre	01:40	180
		Celular	00:56	50
6	São Paulo	Ar livre	03:10	405
		Celular	01:36	245
7	São Paulo	Ar livre	01:27	148
		Celular	02:38	207
8	São Paulo	Ar livre	02:53	297
		Celular	02:57	296
9	Pará	Ar livre	01:55	217
		Celular	01:40	174
10	Rio Grande do Sul	Ar livre	02:10	250
		Celular	02:24	245
Criminoso	Desconhecida	Estúdio	9:40	2181

Tabela 1. Lista com os sujeitos participantes da pesquisa, contexto de gravação (celular ou não) cidade natal, duração de cada gravação e número de vogais estudadas de cada um.

homogeneidade das variâncias dos grupos através do teste Fligner-Killeen. Em seguida, é feita a análise de Kruska-Wallis, que é o correspondente não-paramétrico da ANOVA.

As tabelas 2 e 3 mostram os parâmetros acústicos estudados na pesquisa para o contexto de gravação telefônica e de gravação direta. Neste caso, se o parâmetro acústico apresentou um valor de $p > 5\%$ significa que ele não sofreu variação de canal de transmissão, portanto se mostrando um parâmetro robusto para a pesquisa. Em outras palavras, este é um bom parâmetro acústico para a comparação de trechos por diferentes canais. Podemos perceber através dos testes que os seguintes parâmetros acústicos aceitaram a hipótese nula, apresentando-se robustos para a transmissão telefônica: duração das vogais, taxa de elocução, ΔC e taxa de movimento do segundo formante (F2).

Celular x Gravação direta	MeanV	MeanVV	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 da ANOVA (normalidade e homogeneidade de variâncias) e do teste ANOVA, para $\alpha = 0{,}05$, para a condição de gravações por celular e direta. Resultados para a média da duração das vogais (MeanV), taxa de elocução (MeanVV) e ΔC.

Celulax x Gravação direta	F2	Taxa de F2	Taxa de transição de 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

Tabela 3. Valor de p para testes de condições de uso da ANOVA (normalidade e homogeneidade de variâncias) e do teste ANOVA, para $\alpha = 0{,}05$, para a condição de gravações por celular e direta para transição de F2; e Kruskal-Wallis, para $\alpha = 0{,}05$, para os valores de segundo formante (F2), taxa de F2, frequência fundamental (F0), frequência *baseline* e ênfase espectral.

Em seguida, analisamos quais dos parâmetros acústicos tiveram ou não variação em relação aos sujeitos. Ou seja, se um parâmetro acústico de um sujeito não apresentou variação com o "criminoso", poderemos dizer, a princípio, que são a mesma pessoa. Como podemos ver nas tabelas 4 e 5, os seguintes parâmetros acústicos apresentaram baixa variação em entre os sujeitos e o criminoso: taxa de movimento do segundo formante, ΔC, taxa de elocução e frequência *baseline*.

Sujeitos x Criminoso	MeanV	MeanVV	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

Tabela 4. Kruskal-Wallis, para $\alpha = 0{,}05$, para a variação interfalante. Resultados para a média da duração das vogais (MeanV), taxa de elocução (MeanVV) e ΔC.

Sujeitos x Criminoso	F2	Taxa de F2	Taxa de 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 5. Kruskal-Wallis, para $\alpha = 0{,}05$, para a variação interfalante. Resultado para os valores de segundo formante (F2), taxa de F2, transição de F2, frequência fundamental (F0), frequência *baseline* e ênfase espectral.

Os parâmetros acústicos que apresentaram um valor de p>0,05 foram duração média das vogais, taxa de elocução e ΔC. Através da análise por boxplots, o sujeito 4 foi o que mais apresentou semelhanças, em relação aos demais indivíduos, em seus parâmetros acústicos — os parâmetros de taxa de movimento do segundo formante, ΔC, taxa de elocução e frequência baseline — com o criminoso.

Teste de Duncan

Este teste faz um agrupamento de valores semelhantes baseado nas médias de cada parâmetro analisado. Se duas médias não são estatisticamente diferentes, elas ficarão no mesmo grupo.

De acordo com essa análise estatística, os sujeitos 5 e 7 apresentaram um número maior de médias semelhantes com as do "criminoso". O primeiro para os parâmetros de frequência fundamental, taxa de movimento do segundo formante, taxa de transição do segundo formante, frequência *baseline*, média da duração das vogais, taxa de elocução e ΔC; já o sujeito 7 apresentou semelhança com o "criminoso" nos seguintes parâmetros, frequência fundamental, frequência do segundo formante, taxa de movimento do segundo formante, frequência *baseline*, média de duração das vogais, taxa de elocução e ΔC. Em seguida, os sujeitos 1, 2, 3 e 10 apresentaram seis parâmetros acústicos com média semelhante ao do "criminoso"; logo após, os sujeitos 4 e 9, com cinco semelhantes; e, por fim, os sujeitos 6 e 8, com apenas 3 combinações, foram os que menos se assemelharam com o "criminoso".

Discussão e conclusão

Segundo os resultados analisados, os parâmetros acústicos que mais se mostraram robustos em relação à mudança de canal de transmissão foram: média da duração das vogais, taxa de elocução, ΔC e taxa de movimento de segundo formante. Com base na literatura da área, parâmetros de tempo conseguem capturar informações idiossincráticas do falante (Dellwo e Koreman, 2008) e foi isso que os resultados confirmaram.

A taxa de elocução, por sua vez, embora tenha se apresentado como um parâmetro acústico sem variação entre os sujeitos — reiterando Eriksson (2012) de que essa apresentaria um baixo poder de discriminação interfalantes —, foi um dos parâmetros que se manteve robusto na mudança de canal de transmissão, não tendo variação para o canal telefônico em relação à gravação direta.

A frequência fundamental também obteve um resultado esperado ao ser afetada pelo telefone celular. Esse parâmetro teve um aumento de seu valor de 4% em relação à gravação direta, valor estatisticamente pequeno para a variação. Conforme apontam Byrne e Foulkes (2004), através da transmissão GSM (2G) há um aumento de até 217 Hz em relação a gravação direta em seu experimento.

Outros parâmetros acústicos, como a frequência do segundo formante, também sofreram influência da mudança de canal de transmissão. Considera-se que as frequências formânticas são parâmetros que devem ser evitados ao realizar uma tarefa de comparação de voz (Kunzel, 2001; Byrne e Foulkes, 2004) por serem suscetíveis à variação. Nos resultados deste trabalho, a frequência do segundo formante teve uma diminuição de 7% em seu valor. Este é um efeito curioso para a transmissão telefônica, pois, segundo alguns autores (Kunzel, 2001; Byrne e Foulkes, 2004), formantes mais baixos, como os três primeiros, tendem a sofrer um fenômeno de "deslocamento para cima", ou seja, ao passarem pelo canal telefônico, os valores de suas frequências tendem a aumentar.

A frequência *baseline*, segundo Lindh e Eriksson (2007) se manteria robusta em diferentes tipos de canais de transmissão, incluindo canal telefônico. Porém, de acordo com nossos resultados essa frequência sofreu o impacto do efeito do celular, tendo uma diminuição de 4% em seu valor, mesma porcentagem que a frequência fundamental.

De acordo com o teste estatístico ANOVA, através de uma comparação para determinarmos quais parâmetros não apresentam variação interfalantes, é possível dizer que aqueles que se caracterizaram como menos variáveis entre os sujeitos foram média de duração das vogais, taxa de elocução e ΔC.

Já os demais parâmetros apresentaram uma variância entre os sujeitos. O sujeito 4, por exemplo, através da taxa de movimento do segundo formante, ΔC, taxa de elocução e frequência *baseline*, mostrou-se o mais semelhante com o "criminoso". Acreditamos, apoiados na literatura (Eriksson, 2012), que parâmetros de tempo como o ΔC, e um parâmetro que analisa a dinamicidade formântica, como a taxa de movimento para o segundo formante, são parâmetros que conseguem capturar informações idiossincráticas dos falantes. Com isso, o resultado da análise por meio dos *boxplots* apontaria o sujeito 4 como um possível candidato ao "criminoso", seguido pelos sujeitos 5, 7, 1, 2, 3, 10 e 9.

O "criminoso" deste experimento foi escolhido pelo orientador da pesquisa, sendo revelado somente após a análise de resultados. Soube-se, então, que o sujeito 4 era o "criminoso". No teste de Duncan, o sujeito 4 teve médias semelhantes às do "criminoso" para cinco parâmetros acústicos, a saber, a frequência fundamental, a taxa de movimento do segundo formante, a média de duração das vogais, a taxa de elocução e o ΔC. Isso nos mostra que os mesmos parâmetros que capturam informações idiossincráticas de falantes, também apontaram o sujeito 4 como sendo o "criminoso".

Os sujeitos 5 e 7, de acordo com o mesmo teste estatístico, apresentaram um total de sete médias de parâmetros acústicos similares ao "criminoso".

O que podemos concluir da pesquisa é que nenhum dos parâmetros acústicos foi definidor para a identificação precisa do "criminoso", objetivo principal do experimento. Porém, conseguimos demonstrar que os parâmetros acústicos que mais se caracterizam como robustos pela literatura internacional para a identificação interfalante, também apresentaram valor significativo para o trabalho, já que ΔC e a dinamicidade dos formantes foram essenciais para mostrar traços idiossincráticos dos indivíduos.

Também verificamos a robustez dos nove parâmetros acústicos analisados na mudança de canal de transmissão da fala. Obtendo resultados sólidos através do teste ANOVA, pode-se dizer que a média da duração das vogais, a taxa de elocução e a taxa de movimento do segundo formante foram os que não apresentaram variação do canal de gravação direta para o telefone celular.

A taxa de movimento do segundo formante foi o parâmetro acústico que apresentou melhores resultados na pesquisa. Sendo assim, sugerimos a utilização do mesmo para as pesquisas em fonética forense que caminham com metodologia análoga a nossa. É um parâmetro que será usado e melhor explorado em futuras pesquisas.

Assim como para Kunzel (2001), os nossos resultados para as demais frequências de formantes, incluindo a frequência fundamental, apresentaram grande variação para o canal de telefone celular. Assim como o autor, sugere-se evitar o uso das frequências dos formantes como formantes discriminadores para a comparação interfalante no contexto telefônico.

Agradecimentos

Gostaria de agradecer ao apoio do meu orientador, Plínio Almeida Barbosa, pela parceria e ensinamentos.

Notas

[1]http://www.anatel.gov.br/

[2]http://www.r-project.org/

Referências

Baldwin, J. e French, P. (1990). *Forensic Phonetics*. London: Pinter.

Barbosa, P. A. (2013). Forensic data tracking. programa de computador.

Braun, A. (1995). Fundamental frequency - how speaker specific i it? In A. Braun e J.-P. Koster, Orgs., *Studies in forensic phonetics*, 9–23. Trier: WVT Wissenschaftlicher.

Byrne, C. e Foulkes, P. (2004). The "mobile phone effect" on vowel formants. *The International Journal of Speech, Language and the Law*, 11(1), 83–102.

Constantini, A. C. (2014). *Caracterização prosódica de sujeitos de diferentes variedades de fala do português brasileiro em diferentes relações sinal-ruído*. Tese de doutorado em linguística, Unicamp, Campinas, SP.

Dellwo, V. e Koreman, J. (2008). How speaker idiosyncratic is measurable speech rhythm? In *Anais de IAFPA 2008*, Lausanne: Swiss Federal Institute of Technology Lausanne Disponível em http://www.hf.ntnu.no/isk/koreman/Publications/2008/IAFPA2008abstract_DellwoKoreman.pdf, último acesso em 30/09/2014.

Dowdy, S., Wearden, S. e Chilko, D. (2004). *Statistics for Research*. New York: John Wiley & Sons.

Eriksson, A. (2012). Aural/acoustic vs. automatic methods in forensic phonetic case work. In A. Neustein e H. Patil, Orgs., *A Forensic Speaker Recognition: Law Enforcement and Counter-terrorism*, 41–69. New York: Springer-Verlag.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Haia: Mouton.

French, P. (1994). An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics*, 1(1), 169–181.

Gold, E. e Fench, P. (2011). International practices in forensic speaker comparison. *The International Journal of Speech, Language and the Law*, 18(2), 293–307.

Hollien, H. (2002). *Forensic Voice Identification*. London: Academic Press.

Horii, Y. (1975). Some statistical characteristics of voice fundamental frequency. *Journal of speech and hearing research*, 18(1), 192–201.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2(4), 671–711.

Kunzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, 8(1), 80–99.

Lindh, J. e Eriksson, A. (2007). Robustness of long time measures of fundamental frequency. In *Proceedings of INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association*, 2025–2028, Antwerp, Belgium.

Nolan, F. (1997). Speaker recognition and forensic phonetics. In W. Hardcastle e J. Laver, Orgs., *A Handbook of Phonetic Science*. Oxford: Blackwell.

Nolan, F., McDougall, K., de Jong, G. e Hudson, T. (2006). A forensic phonetic study of 'dynamic' sources of variability in speech: The DyViS project. In P. Warren e C. Watson, Orgs., *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, 13–18, Auckland: Australasian Speech Science and Technology Association.

Öhman, L., Eriksson, A. e Granhag, P. A. (2010). Overhearing the planning of a crime: do adults outperform children as earwitnesses? *Journal of Police and Criminal Psychology*, 26(2), 118–127.

Rose, P. (2002). *Forensic-Phonetic parameters*. New York: Taylor and Francis.
Traunmüller, H. e Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107(6), 3438–3451.