DOI: 10.5769/C2015003 or http://dx.doi.org/10.5769/C2015003

Método para Análise Acústica e Reconhecimento de Vogais em Exames de Comparação de Locutores

Andréa Alves Guimarães Dresch, Hugo Vieira Neto, André Eugênio Lazzaretti e Rubens Alexandre de Faria

***Resumo*— Exames de Comparação Forense de Locutores apresentam características complexas, demandando análises demoradas quando realizadas manualmente. Propõe-se um método para reconhecimento automático de vogais com extração de características para análises acústicas, objetivando-se contribuir com uma ferramenta de apoio nesses exames. A proposta baseia-se na medição dos formantes através de LPC, seletivamente por detecção da frequência fundamental, taxa de passagem por zero, largura de banda e continuidade. Realiza-se o agrupamento das amostras através do método *k-means*, com centros iniciais determinados a partir dos histogramas dos primeiros formantes. Experimentos preliminares com uma base de dados pré-classificados forneceram resultados promissores, com localização de regiões correspondentes às vogais anteriores e posterior média-baixa, propiciando a visualização do comportamento do trato vocal de um falante.**

***Palavras-Chave*— Fonética Forense, Exame de Comparação de Locutores, Análise Acústica, Trapézio Fonético, software Praat.**

***Abstract*— Forensic Speaker Comparison exams have complex characteristics, demanding a long time for manual analysis. A method for automatic recognition of vowels, providing feature extraction for acoustics analysis is proposed, aiming to contribute as a support tool in these exams. The proposal is based in formant measurements by LPC, selectively by fundamental frequency detection, zero crossing rate, bandwidth and continuity. The k-means method is used for clustering, with initial centers determined from the first formants' histograms. Preliminary experiments, using a pre-classified database, have shown promising results, in which regions corresponding to front and lower-middle back vowels were successfully detected, providing visualization of a speaker's vocal tract behavior.**

Keywords*— *Forensic Phonetics, Forensic Speaker Comparison Exam, Acoustic analysis, Phonetic Trapezium, Praat software.

I. Introdução

A produção de provas por meio de registros de áudio, em especial após a promulgação da Lei 9296/96 que trata das interceptações telefônicas [1], tem crescido e consequentemente intensificado as demandas da área forense referente às perícias audiovisuais para atribuição de autorias.

O exame de Comparação de Locutores (CL) tem por finalidade verificar se dois registros de voz e fala foram produzidos por um mesmo indivíduo, consistindo na comparação entre um registro de áudio denominado questionado (sobre o qual pairam dúvidas quanto à autoria das falas) e um registro padrão (registros de fala de identidade conhecida). A importância desse exame reside na possibilidade de associar ou desvincular um indivíduo a um fato delituoso materializado através de um registro de áudio [2].

Relatórios de diagnóstico da Segurança Pública e da Perícia Criminal brasileira apontam a carência de peritos criminais [3] [4], que consequentemente culmina em passivo de laudos nos Institutos de Criminalística. O represamento de materiais a serem examinados prejudica a celeridade necessária para a produção de provas, o que, de acordo com Vargas e colaboradores [5], contribui para a morosidade de um processo penal.

Nesse contexto, agravado pela complexidade das análises envolvidas, uma vez que o exame de CL requer um tempo de execução muito superior à média dos demais exames periciais, a gestão de recursos humanos de Seções de Perícias Audiovisuais é dificultada, analogamente ao constatado por Vrubel e colaboradores em relação à Seção de Computação Forense [6].

Principalmente devido à interdisciplinaridade inerente a esse exame [7], e à construção de conhecimento que exige, o número reduzido de peritos criminais alocados para o mesmo é insuficiente. É desejável, portanto, que se busque o aperfeiçoamento das técnicas adotadas, para se otimizar a realização do exame - qualitativa e quantitativamente.

Em pesquisa realizada por Gold e French [8], foi efetuado um levantamento das técnicas utilizadas internacionalmente para esse exame em 13 países, sendo constatada a preponderância da utilização das análises classificadas como perceptivo-auditiva e acústico-instrumental. Os autores observaram que mesmo quando algum sistema de reconhecimento automático é utilizado, algum tipo de análise humana é feita, e que no Brasil são adotados os métodos perceptivo-auditivo e acústico-instrumental combinadamente.

A análise perceptivo-auditiva requer um profissional capacitado para identificar propriedades da qualidade da voz, traços linguísticos, padrões articulatórios, entre outros atributos. Por sua vez, a análise acústico-instrumental, ou simplesmente análise acústica, engloba medições de curto e de longo termo, nos domínios temporal e espectral.

Para realização dessa tarefa o software Praat é amplamente difundido [9], tanto no ambiente acadêmico como no forense. Contudo, algumas análises requerem extensiva segmentação de trechos com fonemas a serem submetidos à extração de parâmetros, o que, dependendo do volume do material, pode tornar o exame extremamente laborioso.

Sendo assim, a proposta desta pesquisa é o desenvolvimento de uma ferramenta para auxílio de análises acústicas que facilite a visualização de características úteis para o exame

Andréa Alves Guimarães Dresch, ICPR (Instituto de Criminalística do Paraná) e UTFPR (Universidade Federal Tecnológica do Paraná). Hugo Vieira Neto, UTFPR. André Eugênio Lazzaretti LACTEC (Instituto de Tecnologia para o Desenvolvimento). Rubens Alexandre de Faria, UTFPR. Curitiba-PR, Brasil. Emails: andrea.dresch@ic.pr.gov.br, hvieir@utfpr.edu.br, lazzaretti@lactec.org.br, rubens@utfpr.edu.br.

de CL (energia, frequência fundamental, frequência e banda de formantes, taxa de subida ou descida de formantes em um trecho).

Embora o foco seja forense, a ferramenta proposta também pode ser utilizada em outras áreas de linguística ou de fonoaudiologia. O intento é o reconhecimento de trechos vozeados de uma gravação, sem a obrigatoriedade de pré-segmentação manual, além da disponibilização de gráficos com possibilidade de seleção de áreas a serem reavaliadas com a visualização de oscilograma e espectrograma, com os trechos de interesse concatenados ou simplesmente etiquetados.

Tal funcionalidade seria útil, por exemplo, em análises do comportamento formântico a longo termo do trato vocal de um dado falante. Porém, nos casos em que tal hipótese não se confirme devido à interferência agressiva de ruído ou a particularidades da voz em questão, ou mesmo no caso de *outliers*, o analisador teria a possibilidade de confirmar perceptivamente o que ocorreu. Além disso, um padrão visual auxiliaria em análise intra e inter-sujeito, pois se espera em uma CL que sejam encontrados elementos estáveis o suficiente e que denotem similaridades em falas pertencentes a um falante, mas que não sejam comuns a outros indivíduos.

A medição dos formantes é feita pela técnica LPC (*Linear Predictive Coding*), conforme o método de Burg [10], com posterior ponderação de custos para determinação final dos valores de cada formante (com base na frequência e na banda). Serão descartados os pontos em que não houver detecção de F_0 (frequência fundamental), calculados através de autocorrelação nos *frames* (trechos em análise) com energia acima de limiar estabelecido e taxa de passagem por zero abaixo de limiares pré-estabelecidos.

Propõe-se ainda o reconhecimento de agrupamentos de pontos (ou *clusters*) referentes às regiões das vogais (anteriores/posteriores/centrais, altas/médias/baixas), idealmente encontrando a região de cada vogal (/a/,/e/, /ɛ/, /i/, /o/, /ɔ/ e /u/). Embora, como constatado por Escudero e colaboradores [11], no Português Brasileiro (PB) tal determinação possa ser feita por meio de várias combinações de parâmetros, a combinação dos formantes $F_1 \times F_2$ é a que melhor evidencia a distribuição das vogais. Para o reconhecimento serão realizados experimentos com os algoritmos *k-means* e distribuição Gaussiana.

O mecanismo desenvolvido deve permitir a análise das vogais, principalmente com base em seus valores de formantes, com medições realizadas sem necessidade de segmentação prévia. Neste trabalho são exploradas algumas estratégias para seleção dos instantes com valores válidos de formantes, tais como a detecção de frequência fundamental, determinação de limiares de taxa de passagem por zero e de energia de curto termo e continuidade de valores em amostras subsequentes, objetivando minimizar a interferência de fonemas consonantais.

A escolha do aproveitamento de interfaces do software Praat se deve pelo mesmo ser um software livre, e também pela familiaridade dos profissionais que trabalham com fonética. Pretende-se incorporar no futuro rotinas do software R [12], para a realização de cálculos estatísticos.

II. Fundamentação Teórica

A. Produção de Voz

A fala é um dos principais recursos de comunicação humana. Inicia-se por um processo interno do falante, que mentalmente formula a mensagem a ser transmitida, ocorrendo em seguida a ativação motora dos músculos e órgãos do aparelho fonador para a articulação da fala.

Após emissão da mensagem pelo falante e transmissão através do meio (o próprio ar ou um canal telefônico, por exemplo), terá vez o processo de percepção dos sons de fala pelo ouvinte. Tal processo é mais complexo do que a simples detecção de sinais acústicos (como tons puros ou ruído), pois é necessário identificar, categorizar e reconhecer esses sons em sua forma, para atribuir à fala seu significado (mensagem) [13] [14].

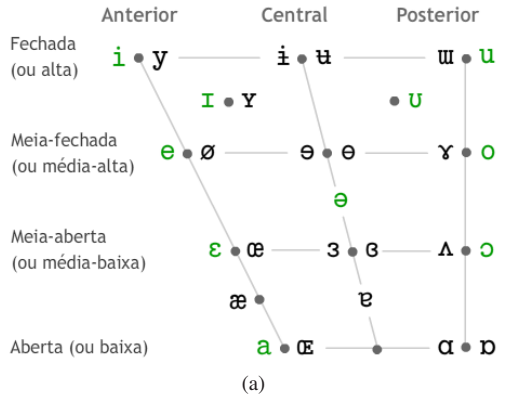
A voz é gerada pela conversão do fluxo contínuo de ar egresso dos pulmões em pulsos de ar (pulsos glóticos), quando ocorre a vibração das pregas vocais, responsável pela característica de vozeamento de vogais e de algumas consoantes. A frequência dessa vibração corresponderá à frequência fundamental (F_0), que possui como correlato acústico o *pitch* [15] [16].

As características anatômicas e fisiológicas do trato vocal provocam ressonâncias nos sons originados dos pulsos glóticos, conforme descrito no modelo fonte-filtro, que considera o sistema de geração do sinal de voz como uma composição de uma fonte de excitação (pulsos glóticos) acoplado a um filtro modelado pela anatomia do trato vocal. Durante a produção de fonemas vocálicos, as frequências amplificadas resultam nos formantes (F_1, F_2, F_3,...,F_n) [14]. Os primeiros formantes, F_1 e F_2, têm relação direta com a altura e o recuo da língua [13], sendo que a sua representação gráfica é normalmente realizada através do diagrama de Vogais Cardeais, também chamado de Trapézio Vocálico [17].

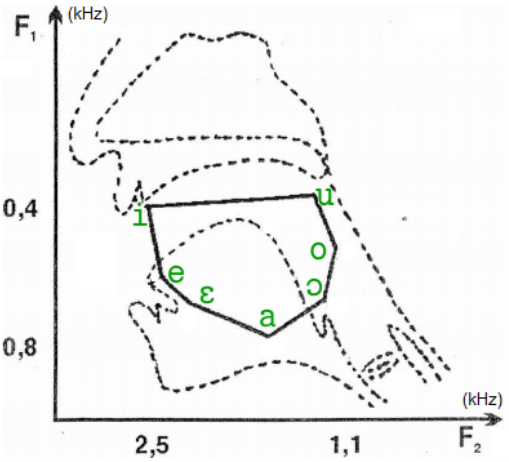
B. Fonemas do Português Brasileiro (PB)

As unidades linguísticas que organizam uma determinada língua são denominadas fonemas. No PB são subdivididos em vogais, semivogais ou *glides*, e consoantes.

- **Vogais:** representam o único tipo de segmento que pode atuar como núcleo silábico. São segmentos vozeados ou sonoros, devido à vibração das pregas vocais que sempre ocorre durante a sua articulação. Outro ponto importante para sua caracterização é que durante a sua produção o fluxo de ar não sofre obstruções no trato vocal, e como consequência os segmentos vocálicos geralmente apresentam maior energia que os segmentos consonantais [13] [14].
  Na Figura 1(a) é apresentado o trapézio fonético das vogais, em que as barras verticais e horizontais são alusivas à posição da língua nos respectivos eixos durante a produção de cada vogal. Dessa forma, cada vogal corresponde a uma configuração do trato vocal, interferindo diretamente nos valores dos formantes. O formante F_1 diz respeito à posição da língua no eixo vertical e F_2 à sua posição no eixo horizontal, conforme Figura 1(b), que ilustra as posições da língua durante a produção das



(a)



(b)

Fig. 1. Ilustração das posições das vogais no trapézio fonético e no gráfico $F_1 \times F_2$. (a) Trapézio fonético das vogais de acordo com o IPA (*International Phonetic Alfabet*), com destaque em verde nas vogais que ocorrem no PB – reproduzido de [18]; (b) Sobreposição do gráfico de $F_1 \times F_2$ (eixos com valores descrescentes para facilitar a análise) e de uma ilustração com a posição da língua durante a produção das vogais orais tônicas – adaptado de [13].

vogais orais tônicas **/a/** - "a", **/e/** - "ê", **/ɛ/** - "é", **/i/** - "i", **/o/** - "ô", **/ɔ/** - "ó" e **/u/** - "u" [13] [16].

- **Semivogais ou *glides*:** também são fonemas vozeados similares às vogais, porém com menor duração. No PB conectam-se a vogais para formar ditongos ou tritongos, podendo ser: /j/, como na palavra [paj] - "pai"; e /w/, como na palavra [maw] - "mau".
- **Consoantes:** ao contrário das vogais, durante a produção de uma consoante, o fluxo de ar egresso dos pulmões sofre obstrução total ou parcial. Assim, em decorrência do tipo de obstrução as consoantes podem ser: plosivas ou oclusivas (exemplos: /p/, /t/ e /g/ em [**p**ato] e [**g**ato]); fricativas (exemplos: /f/ em [**f**ɔka] - "foca", /s/ em [**s**apo], /ʃ/ em [**ʃ**ato] - "chato"); africadas (exemplo: /tʃ/ em [**tʃ**ia] - "tia"); nasais (exemplos: /m/ em [**m**ato] e /ɲ/ em [so**ɲ**o] -"sonho"); laterais (/l/ em [**l**ate]); tepes (exemplo: /ɾ/ em [ca**ɾ**o] - "caro"); e vibrantes (exemplo: /ř/ em [ca**ř**o] - "carro") [13] [16].
  Outra classificação diz respeito ao ponto da articulação: bilabial, labiodental, dental, alveolar, alveopalatal, palatal, velar ou glotal. As consoantes ainda podem ser vozeadas ou desvozeadas, sendo que na análise espectral de consoantes com mesmo ponto e modo de articulação (como por exemplo [f] e [v], de faca e vaca), a diferença pode ser observada através da barra de vozeamento (para o [v]).

C. Frequência fundamental

Estimadores de frequência fundamental procuram o componente frequencial que se sobressai em um trecho do sinal, valor que deverá ser equivalente ao período entre pulsos glóticos. Duas abordagens bastante utilizadas são a autocorrelação e a análise cepstral. Neste trabalho, optou-se pelo método de autocorrelação, por se mostrar mais robusto à presença de ruído [19].

O algoritmo nativo do software Praat calcula a autocorrelação de cada bloco de sinal submetido a uma janela de *Hanning* ou Gaussiana, sendo o resultado obtido pela divisão da função de autocorrelação do sinal pela autocorrelação da própria janela, como demonstrado na Equação (1), em que $r_x(\tau)$ corresponde à autocorrelação resultante, $r_{xw}(\tau)$ à autocorrelação do sinal janelado e $r_w(\tau)$ à autocorrelação da janela utilizada. Dessa forma, evita-se que harmônicos sejam confundidos com a frequência fundamental [20].

$$r_x(\tau) \approx r_{xw}(\tau)/r_w(\tau) \tag{1}$$

O algoritmo possui ainda refinamentos, com limiares de silêncio e de vozeamento e a atribuição de custos para transições de vozeamento/desvozeamento, valor de oitava e salto de oitava entre dois *frames* consecutivos. O tamanho da janela de análise também está atrelado ao limite inferior para busca de frequência (*pitch floor*) [9].

D. Formantes

Uma forma de reconhecer as regiões vocálicas de um sinal de voz é através da obtenção dos formantes, que pode ser feita pela aproximação do envelope espectral desse sinal através de uma análise de predição linear, ou LPC (*Linear Predictive Coding*), exemplificado na Figura 2. Tal técnica consiste em separar o sinal de excitação da resposta do trato vocal, extraindo justamente a informação de formantes que é de interesse para a análise [14].

A análise de predição linear parte do pressuposto de que cada amostra do sinal de fala é, aproximadamente, uma combinação linear das amostras anteriores. Normalmente é feita através de métodos de covariância ou de autocorrelação [21]. Uma representação deste modelo pode ser visualizada na Equação (2), em que $s[n]$ representa o sinal de saída, $x[n]$ o sinal de entrada e m o número de coeficientes que corresponderá a ordem do sistema. Uma vez que o sinal de entrada é desconhecido, o valor $\hat{s}[n]$ na Equação (3) seria uma estimativa do valor da amostra atual. O objetivo da análise preditiva é a determinação dos coeficientes $(a[i]|i = 1, \ldots, m)$ de forma que o erro de predição $e[n]$, constante na Equação (4), seja o menor possível.

$$s[n] = \sum_{i=1}^{m} a_i s[n-i] + x[n] \tag{2}$$

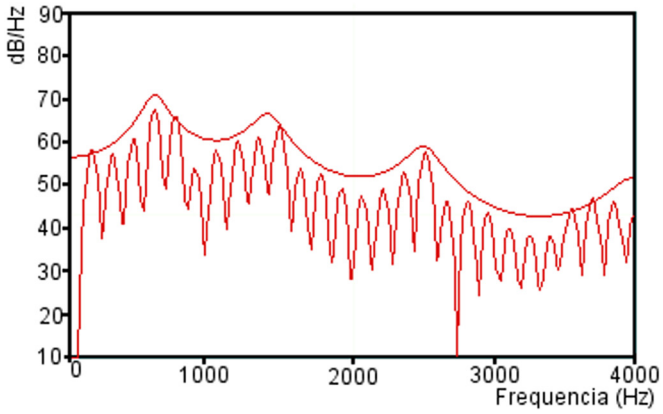


Fig. 2. Figura contemplando o espectro LPC (mais suavizado, em contraste com espectro de Fourier na parte inferior) de uma vogal /a/ produzida por um falante adulto do sexo masculino.

$$\hat{s}[n] = \sum_{i=1}^{m} a_i s[n-i] \tag{3}$$

$$e[n] = s[n] - \hat{s}[n] \tag{4}$$

Para este trabalho foi escolhido o algoritmo de Burg, por ser considerado um modelo estável e que apresenta bons resultados para gravações de curta duração [10]. O algoritmo de Burg considera, além da predição referente às amostras anteriores, y_n na Equação (5), também a referente às amostras posteriores, z_n na Equação (6). A escolha dos coeficientes é feita de forma a minimizar o erro de ambos os sentidos. A quantidade máxima de número de coeficientes, na prática, é determinada pelo valor da frequência de amostragem (em kHz) mais dois [21].

$$y_n = -\sum_{i=1}^{m} a_i x_{[n-i]} \tag{5}$$

$$z_n = -\sum_{i=1}^{m} a_i x_{[n+i]} \tag{6}$$

III. Materiais e Métodos

A. Amostras de dados

Para os experimentos preliminares, estão sendo utilizados arquivos de áudio produzidos em pesquisa realizada pelo Grupo de Estudos de Sons da Fala da UTFPR [22]. Trata-se de gravações de oito pesquisadoras, com a leitura de texto, apresentando aproximadamente 60 segundos de duração.

Os fonemas alvo daquele estudo eram interplosivos e presentes em sílabas tônicas, totalizando quatro repetições para cada uma das vogais orais tônicas do PB, as quais foram manualmente etiquetadas, servindo de referência.

Em uma próxima etapa do projeto serão utilizados os corpora *Spoltech* e "C-ORAL". O primeiro é um corpus compilado através do projeto "*CORPORA from CSLU: The Spoltech Brazilian Portuguese v1.0*" [23], que apresenta 8.080 trechos, com falas de 477 falantes, consistindo de leituras de sentenças foneticamente balanceadas e de respostas a perguntas. O segundo é um corpus compilado através do projeto "C-ORAL", desenvolvido pelo Núcleo de Estudos em Linguagem, Cognição e Cultura da Universidade Federal de Minas Gerais [24], que apresenta registros com fala espontânea, trazendo uma proximidade maior de situações reais.

B. Algoritmos utilizados

Para esta etapa está sendo utilizado, academicamente, o software de análise matemática *Matlab*, e sua *toolbox* de Processamento de Sinais. Após a seleção no Praat do arquivo ou do trecho a ser submetido à análise, inicia-se o processamento conforme ilustrado no fluxograma apresentado na Figura 3, cujos blocos principais estão enumerados e são descritos na sequência.

1) **Pré-processamento:** nesta etapa o sinal é reamostrado a uma taxa de 8kHz, e o nível DC removido através da subtração do nível médio do trecho.
2) **Rotina para Cálculo de ZCR:** o sinal é divido em frames (janelas) de 25ms de duração, sendo efetuado o cálculo do número de vezes que há alteração do sinal do valor da amostra (mudança de sinal de positivo para negativo e vice-versa). Após a finalização do processo acima, o resultado de todas as janelas são divididas pelo valor máximo para fins de normalização.
3) **Detecção de Frequência Fundamental:** no Praat é utilizada a opção "*To Pitch (ac)...*", por permitir a configuração dos parâmetros de inicialização, que incluem a definição das frequências mínima e máxima, além da escolha do tipo de janela (opção *Very accurate* para janela Gaussiana). O tamanho da janela não é definido, por ser uma função da frequência mínima. Neste primeiro momento mantiveram-se os valores de custo *default*. O objeto resultante é convertido para *PitchTier* e, em seguida para tabela, permitindo o armazenamento na forma de arquivo.
4) **Cálculo de Formantes:** no Praat é utilizada a opção "*To Formant (Burg)...*", que possibilita a escolha do número máximo de formantes a ser buscado, e o valor máximo da frequência. A largura da janela é configurada em 25ms, por ser um valor considerado (empiricamente) razoável para este tipo de análise. Mantem-se em 50 Hz o valor do filtro de pré-ênfase, que corresponde ao valor inicial em que o filtro atuará para corrigir a combinação da atenuação de altas frequências provocada pelo trato vocal e a amplificação associada à radiação (do som através da abertura dos lábios).
   Em seguida a matriz obtida é submetida à função "*Formant Track*", que considera os valores obtidos para cada *frame* como um candidato, ao qual é atribuído um custo referente ao valor da frequência, à banda, e à transição entre oitavas. O número máximo de formantes será menor, porém com maior exatidão dos valores obtidos. Após conversão para tabela, é realizada ainda uma limpeza de valores "*–undefined–*", para que o arquivo salvo possa ser corretamente carregado no Matlab.
5) **Seleção de amostras:** no Matlab as tabelas geradas

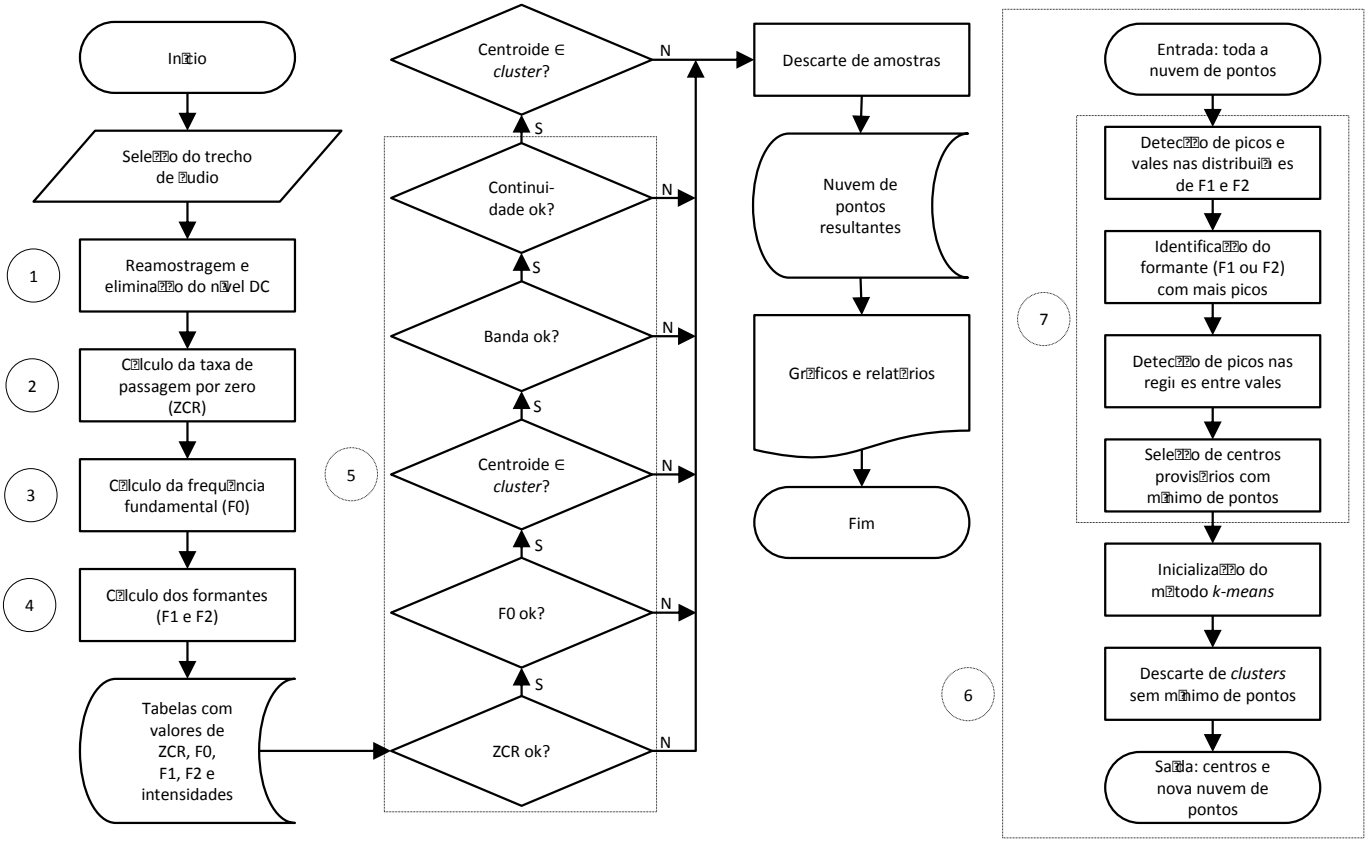


Fig. 3. Fluxograma dos algoritmos implementados. A numeração indicada ao lados dos blocos corresponde aos itens descritos na subseção III-B.

pelas rotinas do Praat são carregadas e salvas em matrizes. Realiza-se em seguida a etapa de seleção das amostras a terem os valores de formantes considerados. Inicialmente são eliminadas as que apresentem taxa de passagem por zero (ZCR) maior que 0,5 (50% do valor máximo), por apresentarem baixa probabilidade de serem voz.

Nos blocos não descartados anteriormente, mas nos quais não houve detecção de frequência fundamental, os valores dos formantes são desconsiderados. Em seguida, utiliza-se a função de busca por centroides (função explicada a seguir), em que só são mantidos os pontos que pertençam a um *cluster* (agrupamento) de tamanho mínimo . Dessa matriz são consideradas apenas as amostras que apresentem valores de banda de F_1 e F_2 menores que a média obtida para cada caso, e que com continuidade, ou seja, que o valor da amostra presente não tenha uma variação maior que 15% em relação aos blocos vizinhos. É feita uma nova busca por *clusters*, que resulta em uma matriz final e nos centroides identificados.

6) **Modelo de agrupamento:** a busca de centroides é realizado através do método *k-means*, o qual consiste no agrupamento sobre os padrões de entrada (formantes F_1 e F_2) em k grupos, sendo k um parâmetro definido *a priori*. O algoritmo é executado em duas etapas principais. Na primeira etapa, cada padrão de entrada é atribuído ao agrupamento mais próximo que esse se encontra, sendo a medida de proximidade representada pela distância euclidiana entre cada padrão e o centro do agrupamento. Na segunda etapa, é realizado o cálculo dos novos centros através da média aritmética entre a localização de todos os pontos associados a cada centro definido na primeira etapa. O processo se repete até que nenhuma nova alteração seja verificada nos agrupamentos, ou se um determinado número de iterações tenha ocorrido. No fim do algoritmo, cada padrão de entrada está associado a um dos agrupamentos definidos. Esse processo garante que a minimização da distância intra-grupos seja atingida no final das iterações, sendo essa a principal motivação da escolha deste método no contexto deste trabalho.

7) **Inicialização dos centroides:** nesta função chamada durante a seleção de amostras inicialmente são gerados histogramas suavizados (de forma a evidenciar máximos e mínimos das distribuições) para as matrizes de F_1 e F_2. É considerado aquele com maior número de máximos (o qual, intuitivamente mas não necessariamente, distinguiria melhor as regiões das diferentes vogais). A seguir o gráfico é subdividido em regiões (horizontais se $F1$ tem mais picos, ou verticais caso contrário), nas quais a geração de histograma suavizado é repetida. Com os valores desses máximos obtidos são determinados centroides temporários, para uma área delimitada pelos mínimos locais. Se essa área apresentar pelo menos 10% do número total de amostras, esse centro é considerado

válido. Caso o número de centros obtidos seja nulo, a função é repetida para o formante que inicialmente apresentou menor número de máximos.
Os centroides obtidos são utilizados para alimentar a função *k-means*, que na ausência de valores iniciais, a função estabeleceria os primeiros centros aleatoriamente, de forma que mesmo que houvesse um resultado convergente, este seria diferente a cada execução. Contudo, como há fornecimento dos valores iniciais no procedimento adotado, conforme descrição acima, a função torna-se determinística, fornecendo sempre os mesmos resultados sempre que executada.

IV. Resultados Preliminares

Nas análises do comportamento formântico foi observado que a simples remoção das amostras em que não houve detecção de frequência fundamental já resulta em um gráfico $F_1 \times F_2$ mais próximo do trapézio vocálico, conforme a Figura 1(a). Tal efeito pode ser visualizado na Figura 4(b), obtida a partir do processamento da nuvem de pontos da Figura 4(a), e na qual o contorno resultante se assemelha a um trapézio.

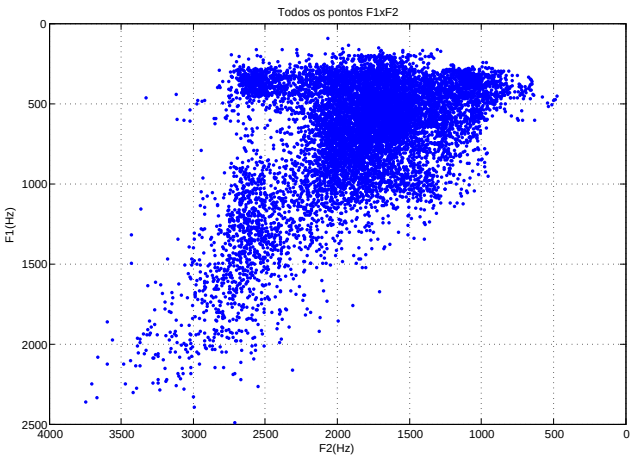
Outra forma de visualizar esse resultado é através da sobreposição da curva de formantes ao espectrograma. Conforme demonstrado na Figura 5, em que se observa em (a) a forma de onda de um trecho de áudio em análise com a sobreposição das funções do STE (energia) e de ZCR, em (b) um espectrograma de banda estreita com a sobreposição dos valores de frequência fundamental resultantes e em (c) o gráfico resultante para os valores dos formantes com a delimitação inicial dos trechos vozeados. É possível observar que os trechos considerados vozeados correspondem àqueles em que houve detecção frequência fundamental, apresentam uma energia relativa maior e baixa ZCR.

Conforme demonstrado na Figura 4(b), o gráfico resultante ainda apresenta pontos de frequências mais altas, possivelmente devido a efeitos de coarticulação, o que exigiu a aplicação dos demais algoritmos apresentados para que o conjunto resultante fosse mais consistente.

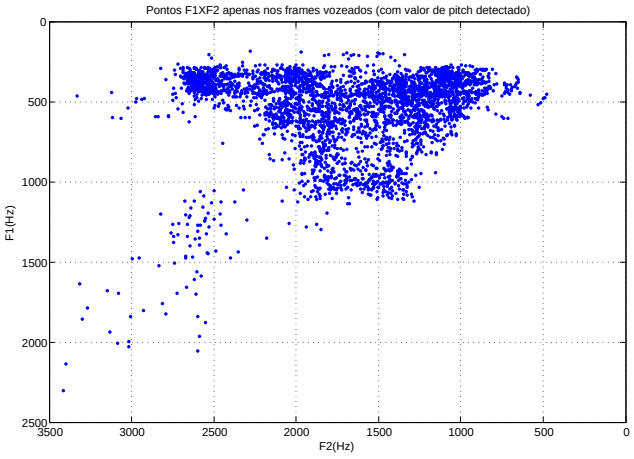
Após realização da etapa de busca de *clusters*, conforme fluxograma apresentado na Figura 3, obteve-se para as amostras da UTFPR um máximo de quatro centroides, com uma média de três.

Um exemplo de um dos gráficos resultantes é apresentado na Figura 6, no qual se observa a distribuição dos valores das amostras em um formato próximo a um trapézio. Os centroides obtidos durante a aplicação do método estão identificados pelos pontos em preto, enquanto que os valores de referência estão indicados pelos pontos vermelhos. É possível observar a proximidade dos centros com os valores de referência correspondentes, da esquerda para direita, às vogais **/e/**, **/ɛ/** e **/ɔ/**.

Na Tabela I são apresentados os valores de F_1 e F_2 obtidos para cada centro. Tais valores foram comparados com os valores de referência (Tabela II), referentes aos resultados da pesquisa realizada pelo Grupo de Estudos de Sons da Fala da UTFPR [22]. Para cada centro foi calculado, através de



(a)



(b)

Fig. 4. Exemplo de gráficos com resultados obtidos para uma das amostras de áudio do grupo de pesquisa da UTFPR. (a) Gráfico $F_1 \times F_2$ com todos os valores de formantes; (b) Gráfico $F_1 \times F_2$ com os valores de formantes nas janelas de análise com valor de frequência fundamental.

distância euclidiana, o valor de referência mais próximo, sendo determinada a vogal correspondente.

Calculou-se então, conforme apresentado na Tabela III, os percentuais das diferenças entre os formantes de cada centro e dos respectivos valores de referência, em relação aos próprios valores obtidos para os formantes. Tais percentuais variam entre 0,2% e 49,5% para F_1 e entre 0,1% e 16,1% para F_2. Entretanto, quando se leva em consideração, não apenas o ponto central, mas também a área da região correspondente através de seu desvio padrão das amostras referentes a cada cluster (Tabela IV), a distância entre o limite das regiões às referências é consideravelmente baixo, conforme pode-se observar nos valores da Tabela V.

Constata-se que em todos os casos houve centros identificados coincidindo com a referência para vogal posterior média-baixa (**/ɔ/**) e com a referência para vogal anterior alta ou média-alta (**/i/** ou **/e/**). Também se observa que centros coincidindo com a vogal anterior média-baixa (**/ɛ/**) ocorreram em nove amostras, coincidindo com vogal central baixa (**/a/**) foram menos frequentes, ocorrendo em apenas três casos (e

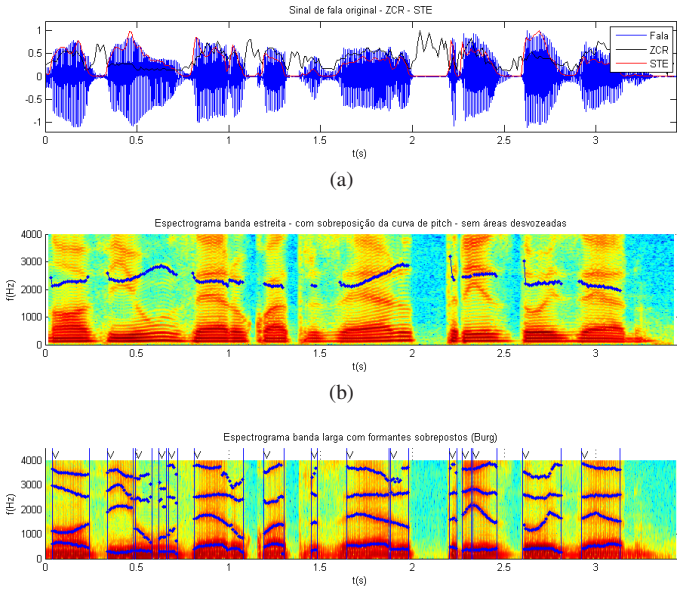


Fig. 5. Teste preliminar com um dos arquivos de "*Spoltech*", com a repetição: "nove, um, zero, quatro, zero, três, dois, zero". (a) Forma de onda com sobreposição dos gráficos de energia de curto prazo (STE) e taxa de passagem por zero (ZCR); (b) Espectrograma de banda estreita com sobreposição dos pontos de frequência fundamental; (c) Espectrograma de banda larga com sobreposição dos formantes das áreas vozeadas.

TABELA I

RESULTADOS DOS CENTROS ($F1, F2$) OBTIDOS PELO MÉTODO DE AGRUPAMENTO.

Amostra	Formante	Centro 1	Centro 2	Centro 3	Centro 4
GC1	F1 (Hz)	516,6	480,5		
	F2 (Hz)	1209,5	2286,5		
GC2	F1 (Hz)	455,2	400,1	848,5	
	F2 (Hz)	1190,4	2355,1	1663,1	
GC3	F1 (Hz)	527,7	697,8	612,1	
	F2 (Hz)	1254,7	2450,5	1812,8	
GC4	F1 (Hz)	611,4	705,2	442,0	
	F2 (Hz)	1117,5	1630,0	2183,7	
GC5	F1 (Hz)	608,6	523,1	595,2	
	F2 (Hz)	1297,7	2569,5	1793,7	
GC6	F1 (Hz)	548,3	476,5	595,4	608,9
	F2 (Hz)	1066,1	2401,4	1383,7	1878,1
GC7	F1 (Hz)	508,1	401,7	733,7	654,5
	F2 (Hz)	1044,9	2293,5	1415,1	1892,5
GC8	F1 (Hz)	501,1	691,2	454,8	
	F2 (Hz)	1239,3	1805,5	2446,6	
GC9	F1 (Hz)	533,2	517,8	697,4	
	F2 (Hz)	1090,8	2449,7	1729,6	
GC10	F1 (Hz)	596,0	493,8	386,0	967,5
	F2 (Hz)	1178,3	1991,6	2539,5	1938,8

naqueles em que houve quatro centros). Não houve nenhum centro coincidente com as vogais posteriores alta ou média-alta (**/u/** e **/o/**).

Tal comportamento sugere que a articulação das vogais anteriores média-alta ou alta e da posterior média-baixa, para os falantes em questão, seriam mais estáveis. Contudo, ressalta-se que faz-se necessária a investigação mais pormenorizada de cada região para verificar influência de coarticulação, de fonemas nasais, assim como de fonemas vocálicos átonos.

TABELA II

REFERÊNCIA - VALORES DE REFERÊNCIA OBTIDOS EM PESQUISA REALIZADA PELO GRUPO DE ESTUDOS DE SONS DA FALA DA UTFPR [22].

Amostra	Formante	/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
GC1	F1 (Hz)	354,5	397,5	628,0	996,3	623,0	407,0	381,8
	F2 (Hz)	2442,0	2413,3	1937,3	1373,5	1055,3	736,0	828,3
GC2	F1 (Hz)	327,5	434,0	703,0	1036,5	680,5	421,5	349,8
	F2 (Hz)	2286,5	1649,3	1755,5	1520,3	1041,3	882,5	941,5
GC3	F1 (Hz)	328,5	429,8	555,8	886,0	619,0	442,5	412,8
	F2 (Hz)	2314,5	2369,3	2055,0	1624,0	1187,3	912,5	1003,5
GC4	F1 (Hz)	362,3	544,3	709,0	906,8	653,8	459,0	420,0
	F2 (Hz)	2323,5	2118,0	1808,8	1386,0	1007,5	817,8	860,3
GC5	F1 (Hz)	321,3	433,8	609,0	750,5	658,3	430,5	335,0
	F2 (Hz)	2388,8	2175,8	1979,0	1533,3	1197,3	830,3	768,8
GC6	F1 (Hz)	389,3	435,5	631,3	760,3	579,3	437,5	388,5
	F2 (Hz)	2251,0	2326,5	2085,3	1440,8	1042,5	704,5	859,0
GC7	F1 (Hz)	330,8	472,0	655,5	877,5	663,8	431,5	378,3
	F2 (Hz)	2205,0	2098,3	1767,5	1455,0	1043,8	904,5	863,8
GC8	F1 (Hz)	370,0	474,8	673,5	1063,5	681,8	478,8	442,3
	F2 (Hz)	2453,8	2175,8	2095,5	1653,5	1136,3	867,0	986,5
GC9	F1 (Hz)	290,8	464,0	719,8	894,0	687,3	471,3	386,0
	F2 (Hz)	2305,5	2303,5	1594,3	1387,3	1050,3	895,3	840,8
GC10	F1 (Hz)	299,0	474,8	612,8	1021,3	614,3	403,8	385,5
	F2 (Hz)	2620,3	2303,0	2057,0	1768,3	1023,8	800,5	785,3

TABELA III

DISTÂNCIA ENTRE OS CENTROS OBTIDOS E O VALOR DE REFERÊNCIA MAIS PRÓXIMO.

Amostra	Formante	Centro 1	Centro 2	Centro 3	Centro 4
GC1	Referência	/ɔ/	/e/		
	$\Delta F1/F1$	20,6%	17,3%		
	$\Delta F2/F2$	12,8%	5,5%		
GC2	Referência	/ɔ/	/i/	/ɛ/	
	$\Delta F1/F1$	49,5%	18,1%	17,2%	
	$\Delta F2/F2$	12,5%	2,9%	5,6%	
GC3	Referência	/ɔ/	/e/	/ɛ/	
	$\Delta F1/F1$	17,3%	38,4%	9,2%	
	$\Delta F2/F2$	5,4%	3,3%	13,4%	
GC4	Referência	/ɔ/	/e/	/ɛ/	
	$\Delta F1/F1$	6,9%	23,1%	0,5%	
	$\Delta F2/F2$	9,8%	3,0%	11,0%	
GC5	Referência	/ɔ/	/i/	/ɛ/	
	$\Delta F1/F1$	8,2%	38,6%	2,3%	
	$\Delta F2/F2$	7,7%	7,0%	10,3%	
GC6	Referência	/ɔ/	/e/	/a/	/ɛ/
	$\Delta F1/F1$	5,7%	8,6%	27,7%	3,7%
	$\Delta F2/F2$	2,2%	3,1%	4,1%	11,0%
GC7	Referência	/ɔ/	/i/	/a/	/ɛ/
	$\Delta F1/F1$	30,6%	17,7%	19,6%	0,2%
	$\Delta F2/F2$	0,1%	3,9%	2,8%	6,6%
GC8	Referência	/ɔ/	/i/	/ɛ/	
	$\Delta F1/F1$	36,1%	18,6%	2,6%	
	$\Delta F2/F2$	8,3%	0,3%	16,1%	
GC9	Referência	/ɔ/	/e/	/ɛ/	
	$\Delta F1/F1$	28,9%	10,4%	3,2%	
	$\Delta F2/F2$	3,7%	6,0%	7,8%	
GC10	Referência	/ɔ/	/i/	/a/	/ɛ/
	$\Delta F1/F1$	3,1%	22,5%	5,6%	24,1%
	$\Delta F2/F2$	13,1%	3,2%	8,8%	3,3%

V. CONSIDERAÇÕES FINAIS

Os resultados preliminares obtidos foram promissores, com a identificação média de centros correspondentes a três vogais no espaço $F_1 \times F_2$, delineando um padrão próximo ao trapézio vocálico esperado.

Considerando as necessidades forenses, espera-se que o sistema proposto possa ser efetivamente utilizado como uma ferramenta de apoio em exames de registros de áudio, principalmente em exames de Comparação de Locutores. Apesar da aplicação estar restrita a amostras de áudio de apenas um falante, ou que contenham arquivos de delimitação entre os

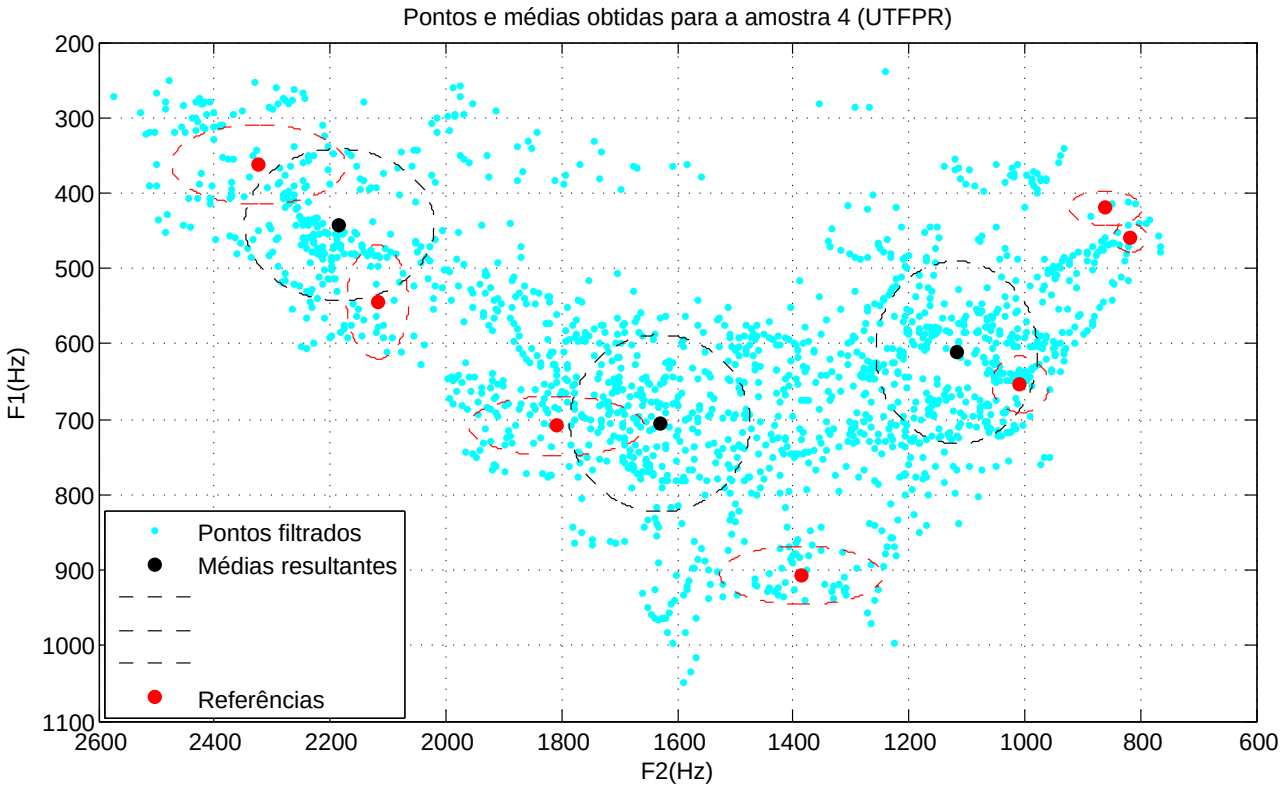


Fig. 6. Gráfico com pontos $F_1 \times F_2$ e centroides resultantes para a amostra 4 do grupo da UTFPR. Os valores de referência estão indicados pelos pontos vermelhos, enquanto que os centroides obtidos no trabalho estão indicados pelos pontos pretos. Os círculos maiores pontilhados representam a área das regiões obtidas.

TABELA IV

DESVIOS PADRÃO DOS VALORES PERTENCENTES À REGIÃO COM RELAÇÃO AO SEU CENTRO.

Amostra	Formante	Centro 1	Centro 2	Centro 3	Centro 4
GC1	F1 (Hz)	198,1	180,1		
	F2 (Hz)	221,0	301,7		
GC2	F1 (Hz)	172,5	129,2	140,9	
	F2 (Hz)	183,8	278,0	178,1	
GC3	F1 (Hz)	102,8	490,6	137,4	
	F2 (Hz)	137,9	156,8	167,9	
GC4	F1 (Hz)	121,0	117,2	100,4	
	F2 (Hz)	140,1	155,7	164,0	
GC5	F1 (Hz)	112,0	357,4	136,3	
	F2 (Hz)	153,7	202,8	148,0	
GC6	F1 (Hz)	64,4	160,7	114,1	131,5
	F2 (Hz)	107,8	118,2	104,8	134,3
GC7	F1 (Hz)	135,3	113,8	134,8	125,6
	F2 (Hz)	118,7	113,5	116,9	112,9
GC8	F1 (Hz)	97,9	224,3	152,7	
	F2 (Hz)	131,8	140,6	172,0	
GC9	F1 (Hz)	178,9	347,6	150,9	
	F2 (Hz)	159,1	207,4	226,3	
GC10	F1 (Hz)	183,0	130,9	82,2	184,9
	F2 (Hz)	205,2	166,3	139,4	235,1

TABELA V

DIFERENÇA PERCENTUAL DA DISTÂNCIA ENTRE A FRONTEIRA DE CADA REGIÃO E O VALOR DE REFERÊNCIA MAIS PRÓXIMO, QUANDO NÃO LOCALIZADO NO INTERIOR DA REGIÃO.

Amostra	Formante	Centro 1	Centro 2	Centro 3	Centro 4
GC1	F1	-	-		
	F2	-	-		
GC2	F1	-11,61%	-	-0,55%	
	F2	-	-	-	
GC3	F1	-	-	-	
	F2	-	-	-4,10%	
GC4	F1	-	-	-0,44%	
	F2	-	-1,41%	-	
GC5	F1	-	-	-	
	F2	-	-	-2,08%	
GC6	F1	-	-	-8,51%	-
	F2	-	-	-	-3,88%
GC7	F1	-4,00%	-	-1,23%	-
	F2	-	-	-	-0,64%
GC8	F1	-16,52%	-	-	
	F2	-	-8,27%	-	
GC9	F1	-	-	-	
	F2	-	-	-	
GC10	F1	-	-	-1,23%	-
	F2	-	-	-	-

turnos, acredita-se que terá utilidade em muitos casos.

Dadas as condições da maioria dos materiais encaminhados para esses exames, posteriormente será imprescindível a validação do mesmo também em condições degradadas, tais como: presença de ruído, compressões e limitações de frequência de canais telefônicos.

O presente trabalho ainda encontra-se em desenvolvimento, sendo que trabalhos futuros envolvem a integração com o software *R*, a possibilidade de interação com os gráficos para navegação dos trechos do áudio indicados através de pontos ou regiões selecionadas, a geração de relatórios e de registros de eventos (*logs*) para facilitar o elaboração de laudos e garantir a reprodutibilidade das etapas realizadas, assim como

o reconhecimento automático dos centroides equivalentes a cada vogal ou grupo de vogal (em função de suas classificações relativas à anterioridade e altura). Além dos valores de F_0, F_1 e F_2, o intuito é futuramente acrescentar outras dimensões para análise, tais como a banda de cada formante, e a variação de seus valores no decorrer da produção de um fonema.

É necessária ainda a realização de experimentos com diferentes durações de fala exclusiva, de um mesmo falante, para determinação da mínima duração a fim de se evidenciar graficamente o padrão formântico. Para tais ensaios serão utilizados os corpora *Spoltech* e "C-ORAL".

Também é importante permitir formas de validação dos resultados, como a possibilidade de treinamento ou de particionamento de uma amostra de áudio para verificar a coerência entre os resultados obtidos para cada segmento. O mecanismo desenvolvido deve propiciar análises de variações intra e inter-sujeito, importantíssimas de serem diferenciadas em exames de CL, permitindo que o examinador exclua elementos não-servíveis (isto é, que tenham grande variação intra-sujeito).

Posteriormente tais códigos serão migrados para rotinas do Praat ou outra linguagem que permita que todos os pacotes programados estejam em plataforma de software livre. A finalização de um ambiente de testes requer ainda que o examinador possa salvar um projeto com configurações realizadas, e a disponibilização de relatórios com as rotinas executadas e parâmetros utilizados.

Agradecimentos

Os autores agradecem a Denise de Oliveira Carneiro e Marilisa Exter Koslovski, peritas criminais do Instituto de Criminalística do Paraná, pelas importantes discussões acerca de ferramentas úteis para apoio ao exame de Comparação Forense de Locutores, a Eduardo Tondin Ferreira Dias e Philipe Ambrózio Dias, colegas do Laboratório de Processamento de Imagens e Sinais da UTFPR, pelas críticas e sugestões para melhoria do conteúdo do presente artigo, e ainda ao Grupo de Estudos dos Sons da Fala da UTFPR, liderado pela professora Maria Lúcia de Castro Gomes, pela cessão de amostras de áudio utilizadas neste trabalho.

Referências

[1] Brasil, "Lei nº 9.296 (lei das interceptações telefônicas), de 24 de julho de 1996." *Diário Oficial da República Federativa do Brasil*, 1996.

[2] A. C. M. Braid, *Fonética Forense*, 2nd ed., ser. Tratado de Perícias Criminalísticas. Campinas, SP: Editora Millenium, 2003.

[3] SENASP, "Diagnóstico da perícia criminal no brasil," Secretaria Nacional de Segurança Pública, Tech. Rep., 2012.

[4] ENASP, "Relatório nacional da execução da meta 2: um diagnóstico da investigação de homicídios no país," Conselho Nacional do Ministério Público, Tech. Rep., 2012.

[5] J. D. Vargas, I. Blavatsky, and L. M. L. Ribeiro, "Metodologia de tratamento do tempo e da morosidade processual na justiça criminal," Secretaria Nacional de Segurança Pública, Tech. Rep., 2006.

[6] A. Vrubel, A. Brondani, M. Silva, and L. Grochocki, "Modelo matemático para a gestão de recursos humanos baseados em controles estatísticos de demanda e produtividade," *Anais do VI Congresso CONSAD de Gestão Pública*, 2013.

[7] M. L. C. Gomes, L. Richert, and J. Malakoski, "Identificação de locutor na área forense: a importância da pesquisa interdisciplinar," in *Anais do X ENCONTRO DO CELSUL*, Cascavel, 2012.

[8] E. Gold and P. French, "International practices in forensic speaker comparison," *The International Journal of Speech, Language and the Law*, vol. 18, pp. 293–307, 2011.

[9] P. Boersma and D. Weenink, "Praat, doing phonetics by computer (version 5.4.08)," 2015. [Online]. Available: http://www.praat.org/

[10] C. Collomb, "Burg's method, algorithm and recursion," 2009. [Online]. Available: http://ccollomb.free.fr/

[11] P. Escudero, P. Boersma, A. S. Rauber, and R. A. H. Bion, "A cross-dialect acoustic description of vowels: Brazilian and european portuguese," *Journal of the Acoustical Society of America*, vol. 126, pp. 1379–1393, 2009.

[12] R. Core-Team, "R: A language and environment for statistical computing," 2013. [Online]. Available: http://www.R-project.org/

[13] I. Russo and M. Behlau, *Percepção da Fala: Análise Acústica do Português Brasileiro*. Editora Lovise, 1993.

[14] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Pearson, 2011.

[15] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 2000.

[16] A. P. P. F. Engelbert, *Fonética e Fonologia da Língua Portuguesa*. Curitiba: Ibpex, 2011.

[17] T. Cristófaro-Silva, *Dicionário de Fonética e Fonologia*. Editora Contexto, 2011.

[18] T. Cristófaro-Silva and H. C. Yehia, *Sonoridade em Artes, Saúde e Tecnologia*. Belo Horizonte: Faculdade de Letras, 2012. [Online]. Available: http://fonologia.org

[19] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 727–730, 2001.

[20] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *IFA Proceedings*, vol. 17, 1993.

[21] L. M. J. Barbosa, *Processamento de Sinais em Fonética Forense*, Departamento da Polícia Federal, 2012.

[22] M. L. C. Gomes, "An acoustic description of vowels brazilian portuguese in normal and disguised voice," in *IAFPA 2013 Annual Conference*, 2013.

[23] M. C. Schramm, L. F. R. Freitas, A. Zanuz, and D. Barone, "A brazilian portuguese language corpus development." in *International Conference on Spoken Language Processing 2000*. ISCA, 2000.

[24] T. Raso and H. Mello, *C-ORAL BRASIL I - Corpus de referência do português brasileiro falado informal*. Belo Horizonte, MG: Editora UFMG, 2012.