Técnica de Segmentação Multidimensional de Fala

Raissa Bezerra Rocha [†] [§], Wamberto José Lira de Queiroz [*§] e Marcelo Sampaio de Alencar[*§]
[*]Universidade Federal de Campina Grande – UFCG, Campina Grande, Brasil
[†]Universidade Federal de Sergipe – UFS, São Cristóvão, Brasil
[§]Instituto de Estudos Avançados em Comunicações – Iecom
E-mails: {raissa, wamberto, malencar}@iecom.org.br

***Resumo*— A segmentação de fala é uma etapa importante em várias aplicações que envolve o processamento do sinal de voz, como reconhecimento, síntese e codificação de fala, bem como utilizada como ferramenta para tratamentos fonoaudiológicos. Este artigo decreve um novo método de segmentação baseado na observação da energia do sinal da voz. Trata-se de um algoritmo dinâmico, que divide a locução em multi regiões e detecta os limiares fonéticos pela comparação da energia a cada curto segmento da fala com a energia média de cada região. Para otimizar o desempenho do segmentador, um sistema de refinamento usando o tamanho máximo de cada fonema é proposto. O desempenho do segmentador é aferido por testes objetivos, que indicam que a técnica proposta fornece resultados competitivos com os encontrados na literatura, apresentando uma taxa de 84,86% de segmentação.**

***Palavras-chave*— Segmentação de fala, energia, sinal de voz.**

I. Introdução

A fala é formada pela junção de pequenos sons denominados fones. Em diversas aplicações envolvendo processamento do sinal de voz, a segmentação é uma etapa fundamental no desenvolvimento do sistema de fala ou é utilizada para aumentar seu desempenho.

Um sistema de segmentação de voz tem o objetivo de determinar as fronteiras que separam os elementos essenciais da fala, como palavras, sílabas ou fonemas de uma determinada locução. Ele pode ser usado em algoritmos de codificação de voz, como é o caso dos codificadores fonéticos, assim como em sistemas de reconhecimento automático, síntese de fala e no auxílio para pacientes em tratamento fonoaudiológico.

De acordo com a literatura, os segmentadores de fala podem ser classificados de acordo com a presença ou ausência da categoria linguística e observações acústicas [1].

Categoria linguística é o conjunto de informações linguísticas, como a transcrição fonética da locução, que pode ou não ser apresentada como entrada para o sistema de segmentação. Por outro lado, as observações acústicas consistem em informações extraídas do sinal de fala, normalmente representadas por um vetor de parâmetros, com informações do sinal de fala, atribuídos a janelas de curto intervalo de tempo.

Os sistemas de segmentação de fala podem ser classificados como segmentação implícita ou segmentação explícita. A segmentação implícita acontece quando a categoria linguística não é considerada no processo de segmentação, sendo consideradas apenas observações acústicas para o sistema gerar as fronteiras de segmentação. A segmentação explícita utiliza a transcrição fonética (informações linguísticas) para gerar as marcas de segmentação. Dessa forma, nesse tipo de segmentação, as transcrições fonéticas da fala a serem segmentadas devem ser antecipadamente geradas e utilizadas como entrada para o sistema de segmentação [2].

Na literatura, a segmentação de fala é realizada, por exemplo, utilizando técnicas probabilísticas, como os Modelos de Markov Escondidos, além da sua combinação com técnicas como DTW (*Dynamic Time Warping*), SPM (*Score Preditive Model*), ou até mesmo usando informações visuais da fala, tais como o movimento dos lábios, língua e dentes [3], [14], [4], [5], [13], [12]. Outros trabalhos também propõem a segmentação por meio da observação do *pitch*, detecção de envoltória e estudo de regras fonéticas [2], [1], [6], [7], [8], [16], [17].

Este artigo apresenta um novo método de segmentação de fala, com ênfase na divisão fonética. Trata-se de uma técnica que secciona a locução em regiões e observa as variações de energia usando como referência a energia média de cada região. Para aprimorar o método, é implementado um sistema de refinamento, que elimina falsas demarcações e localiza fronteiras não detectadas anteriormente.

Além desta seção introdutória, este artigo está dividido em mais três seções. A Secão II descreve o método de segmentação fonética proposto, além do sistema de refinamento utilizado para o seu aprimoramento. A análise de desempenho da técnica de segmentação, bem como a comparação com outros segmentadores para o Português do Brasil está na Seção III. Por fim, a Seção IV apresenta as conclusões e trabalhos futuros.

II. Descrição da Técnica de Segmentação Multidimensional de Fala

A técnica de segmentação multidimensional de fala é caracterizada por permitir obter fronteiras em nível fonético por meio da divisão em múltiplas dimensões da locução a ser segmentada.

Diferentemente dos demais trabalhos encontrados na literatura sobre segmentação de fala, que utilizam métodos estatísticos, caracterizados pela alta complexidade de algoritmo e uma prévia etapa de treinamentos de modelos acústicos [20], [21], a técnica multidimensional apresenta um algoritmo de fácil implementação e não requer etapas de pré-processamento do sinal de voz para obtenção das fronteiras entre fonemas.

Recebido em março de 2017. Aceito em julho de 2017

A técnica multidimensional realiza inicialmente uma segmentação para a identificação de regiões audíveis e não audíveis por meio da energia de curta duração. Esse parâmetro é utilizado, pois apresenta valores significativamente maiores para regiões audíveis em uma locução, sendo possível distinguir a voz do silêncio.

De acordo com a literatura, a energia de curta duração é calculada por meio de janelas cuja duração deve variar entre 20, amostras para uma voz aguda, a 250 amostras, para voz grave. Na prática, para uma taxa de amostragem na ordem de 10 kamostras/s, deve-se utilizar uma janela entre 100 e 200 amostras $(10ms < t < 20ms)$.

A taxa de amostragem das locuções usadas no teste do sistema de segmentação é de 22050 amostras/s. O cálculo da energia é feito utilizando uma janela com 200 amostras, e um deslocamento de 20 amostras. Dessa forma, o algoritmo calcula o valor da energia em cada janela de duração pré-definida e determina o início de uma região audível se o valor da energia for maior que um limiar, ou seja, 0.002 V. A Figura 1 ilustra um exemplo de identificação de fronteiras que separam regiões audíveis e silêncio.

Fig. 1: Exemplo de obtenção de fronteiras entre silêncio e fala e vice-versa.

Em seguida, o método multidimensional identifica as marcas de transição entre fonemas separadamente para cada região audível, que pode representar uma palavra completa ou um fragmento dela.

Cada região audível detectada é selecionada pelo algoritmo para encontrar os limiares contidos na locução fraccionada. Nessa etapa, a segmentação multidimensional é caracterizada por ser implícita, pois o método não solicita antecipadamente a transcrição fonética para fornecer as demarcações entre fonemas.

As demarcações decorrentes da segmentação, que representam o ponto de início e fim de cada fonema, são dadas em número de amostras, sendo possível obter a duração de cada fonema pela divisão da quantidade de amostras encontradas pela taxa de amostragem.

As fronteiras entre fonemas ocorrem nos instantes em que características de um determinado fonema começam a desvanecer, à medida que atributos de outro fonema começam a ficar mais evidentes. Nos pontos de transição, o sinal de voz é caracterizado por exibir um certa diferença de energia. Dessa forma, o segmentador multidimensional faz o uso dessa característica para encontrar as marcas de segmentação.

A técnica multidimensional realiza a segmentação por meio de um único parâmetro prosódico do sinal de voz, a energia dada por

$$E = \frac{1}{N}\sum_{n=1}^{N} x^2(n), \qquad (1)$$

em que $x(n)$ representa amostras do sinal de voz e N a quantidade de amostras em cada janela.

Dessa forma, para cada região delimitada pelo silêncio, é calculada a energia utilizando um intervalo de duração pré-definido de 200 amostras. A Figura 2 ilustra o comportamento da energia (vermelho) em relação a uma locução (azul) para uma locução.

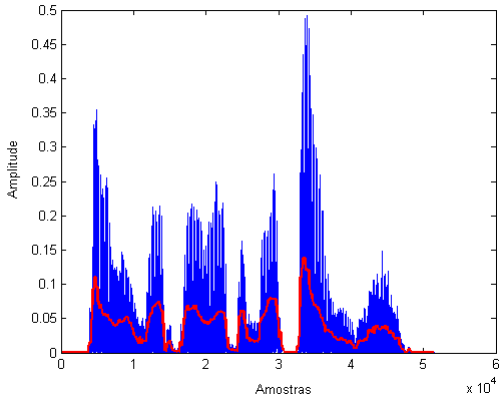

Fig. 2: Curva da energia em relação à locução.

De acordo com a Figura 2, é possível observar que a envoltória dos valores da energia se comporta de forma semelhante à curva do sinal de voz. Uma vez que há uma diferença de energia entre os fonemas, ou seja, em regiões nas quais a energia está crescendo ou decrescendo, os pontos de transição entre fonemas estão presentes no início e no fim dos vales presentes na curva de energia.

Para identificação dos instantes inicial e final de cada vale, é realizada uma codificação dos valores de energia. Inicialmente, o trecho a ser segmentado é dividido em duas regiões. Para cada região obtém-se a energia média. Esse parâmetro é utilizado como um limiar, sendo atribuído aos valores de energia acima dele, o código 1 e, para valores de energia abaixo do limiar, o código 0.

Como resultado desse procedimento, é obtido um vetor formado por regiões de zeros e uns. Para localização das fronteiras fonéticas é realizada uma busca da transição entre as regiões de uns e zeros, uma vez que ela representa o ponto em que há diferença de energia entre os fonemas, configurando em uma demarcação fonética.

Assim, ao encontrar a transição entre as regiões do vetor de energia codificada, o algoritmo tem a informação de quantos códigos 1 ou 0 estão à esquerda da transição e, ao multiplicar pelo valor pré-definido para a janela de cálculo da energia, informa, em número de amostras, o ponto de localização inicial ou final de um fonema.

O vetor de fronteiras resultantes é então armazenado. O algoritmo do sistema de segmentação é reiniciado, modificando o número de divisões em que o trecho da locução entre regiões de silêncio é dividida, resultando em um método de segmentação multidimensional, devido ao fato do algoritmo fraccionar o trecho a ser segmentado em várias dimensões.

A técnica subdivide o sinal em até quatorze vezes. Para cada vez que o método é reiniciado, um vetor de fronteiras é obtido e armazenado. A subdivisão em quartoze vezes é o meio de encontrar o maior número de fronteiras possíveis, justificada pelo fato de que há uma mudança do valor da energia média, sendo possível encontrar mais fronteiras.

Ao final, o sistema possui quatorze vetores de transição fonética. Esses vetores são reunidos em apenas uma e, em seguida, é realizado um procedimento de retirada de fronteiras

iguais ou que difiram por menos de 200 amostras, já que não há no português brasileiro um fonema cuja duração seja menor que tal tamanho.

A quantidade de vezes que o segmentador é reiniciado foi determinada de forma experimental. Dessa forma, observou-se que com mais do que quatorze interações não houve melhoria no resultado da segmentação.

A Figura 3 ilustra o fluxograma do algoritmo utilizado pelo método de segmentação multidimensional.

A. Técnica de Refinamento de Segmentação

A técnica multidimensional é capaz de identificar grande parte das localizações dos fonemas. Entretanto, devido à variação de energia que alguns fonemas têm na sua forma de onda, o algoritmo apresenta algumas falsas fronteiras, ou seja, marcas de segmentação que não representam instantes iniciais ou finais dos fonemas. Assim, é necessário um processo de refinamento com o objetivo de eliminar tais fronteiras, de forma que a técnica de segmentação multidimensional apresente em sua saída apenas as demarcações reais.

É verificado que vários dos falsos limiares fonéticos exibidos no vetor final de fronteiras são separados das delimitações reais por até mil amostras.

O refinamento é realizado inicialmente com a detecção de fronteiras adjacentes cuja distância não ultrapasse mil amostras. As fronteiras que estão dentro do limiar estabelecido são substituídas por uma nova marca de segmentação obtida a partir da média aritmética das fronteiras adjacentes. Esse procedimento resulta na eleminação das falsas demarcações, uma vez que fronteiras vizinhas são substituídas por um único limiar fonético.

O deslocamento das localizações verdadeiras dos fonemas, provocado pela alteração do seu valor inicial pela nova fronteira resultante da média de fronteiras vizinhas, não ocasiona erro na segmentação final, pois a fronteira obtida se mantém dentro da margem de erro aceitável na avaliação do segmentador fonético.

No entanto, o processo de refinamento elimina marcas de divisão de fonemas cuja duração é menor que o limiar designado para o refinamento. Para que o segmentador seja capaz de repor tais fronteiras, é necessário o uso da transcrição fonética, etapa que torna o divisor fonético proposto um sistema de segmentação explícita.

À vista disso, o segmentador requer o conhecimento prévio da transcrição fonética da locução a ser segmentada. Para restituir as fronteiras eliminadas, o sistema observa os fonemas localizados nas extremidades de cada região. Para evitar a presença de falsas fronteiras, não há refinamento na parte interior das locuções.

Com a utilização da transcrição fonética, o método de segmentação reconhece os fonemas contidos nas extremidades de cada trecho audível, cujo algoritmo deve encontrar sua localização. Assim, é verificado, para cada fonema, se a técnica multidimensional fornece uma demarcação cuja localização seja menor ou igual a uma duração máxima previamente estabelecida para os fonemas em questão. Caso haja um limiar de segmentação, o algoritmo passa a procurar a fronteira do fonema da próxima extremidade. Caso contrário, uma marca de divisão fonética é adicionada de acordo com um passo médio de duração do fonema em curso.

A Tabela I apresenta os fonemas utilizados na técnica de reposição de fronteiras, a quantidade máxima de amostras por fonema e passo médio designado para cada fonema.

III. Resultados

O desempenho de um sistema de segmentação de fala pode ser avaliado por meio de testes subjetivos e objetivos.

Em testes subjetivos, os segmentos de voz obtidos na simulação são escutados e julgados por um grupo de pessoas que atribuem notas de acordo com uma escala pré-definida, ou os identificam como algum fonema de uma lista pré-determinada.

Na avaliação objetiva, os limiares fonéticos encontrados no sistema de segmentação automático são comparados com as marcas de segmentação obtidas de forma manual. Neste caso, o erro entre tais fronteiras não deve ultrapassar 20 ms [1], [3], [2], [9], [11], [10], [18], [19].

A avaliação objetiva foi escolhida para aferir o desempenho do segmentador proposto neste artigo. Para isto, foi utilizado como referência às fronteiras ideais, um banco de dados de fala segmentado manualmente, apresentado em [1], [2]. Ele é composto por 200 frases, gravadas por um locutor paulista, do interior do Estado de São Paulo, não abordando, dessa forma, os regionalismos do País, assim como os diferentes tipos de pronúcia para algumas locuções. As sentenças foram gravadas a uma taxa de 22,05 kamostras/s e quantizadas com 16 *bits* por amostra. As locuções têm, em média, três segundos e foram gravadas com o mínimo de ruído possível.

Para realizar a avaliação objetiva, foram selecionadas aleatoriamente da base de dados de voz 50 locuções. Nenhum procedimento de pré-processamento foi realizado nas locuções.

O algoritmo de segmentação proposto foi capaz de localizar, com erro de até 20 ms, 1333 limiares de transição entre fonemas de um total de 1569 fronteiras, representando uma taxa de 84,86% de segmentação.

Além disso, em média, o método de segmentação multidimensional identifica uma delimitação fora da margem de erro a cada conjunto de duas locuções testadas. Por fim, o algoritmo não identifica, em média, 4,18 fronteiras e exibe uma média de 1,3 falsas delimitações fonéticas por locução.

O método de segmentação multidimensional foi avaliado para o idioma Português do Brasil, no entanto, seu algoritmo pode ser adaptável para outros idiomas. Na comparação com outros métodos de segmentação disponíveis na literatura, o método de segmentação multidimensional apresenta um algoritmo de baixa complexidade, em que os limiares fonéticos são determinados sem qualquer treinamento prévio, bem como pré-processamento inicial nas locuções a serem segmentadas, diferentemente dos segmentadores que utilizam modelos probabilísticos.

A técnica apresentada também requer uma menor dependência da transcrição fonética em relação a outros segmentadores encontrados na literatura, uma vez que necessita apenas do conhecimento dos fonemas localizados nas extremidades da locução. Por fim, o método multidimensional

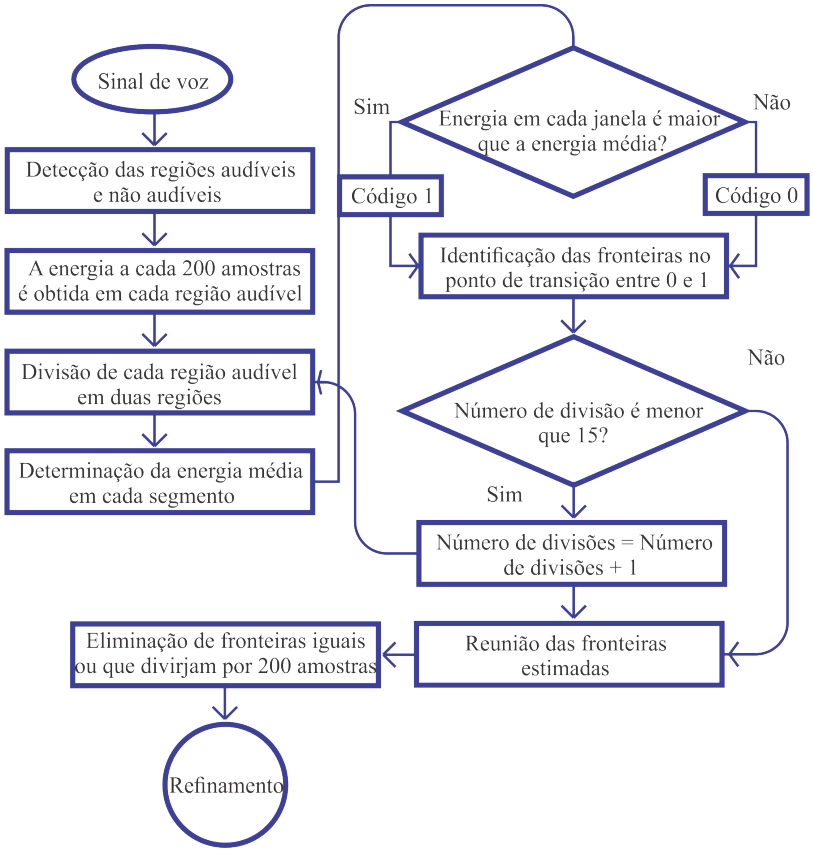

Fig. 3: Etapas iniciais do método de segmentação multidimensional de fala.

TABELA I: Fonemas Utilizados no Refinamento do Método Multidimensional.

Fonemas	Quantidade Máxima de Amostras	Passo Médio
p	1000	600
g	2000	1500
k, R, t, T, p	1000	600
D	2000	1500
Vogais (fim do trecho audível)	3000	2500
Vogais (início do trecho audível)	3500	3000

apresenta taxa de segmentação competitiva com outros segmentadores encontrados na literatura em nível fonético para português do Brasil, como os apresentados em [2], [1], [6].

IV. Conclusões

Este artigo aborda a segmentação de fala, com ênfase na obtenção dos limiares fonéticos pela observação da energia e pela transição fonética dos fonemas localizados nas extremidades das locuções.

O segmentador multidimensional apresenta baixa complexidade e pode ser utilizado no projeto de codificadores fonéticos, assim como para compor bancos de segmentos sonoros necessários para sintetizar a fala. Testes objetivos foram realizados para aferir o desempenho do segmentador e os resultados mostram que a técnica fornece taxa de 84,86% de segmentação, sendo competitiva com os demais segmentadores para o idioma, mesmo com uma menor dependência com a transcrição fonética.

Como trabalhos futuro, pretende-se realizar novos testes de desempenho, com locuções gravadas com diferentes taxas de amostras, bem como em outros idiomas.

Agradecimentos

Os autores agradecem o apoio da Universidade Federal de Campina Grande (UFCG), do Instituto de Estudos Avançados em Comunicações (Iecom) e da Universidade Federal de Sergipe (UFS).

Referências

[1] A. M. Selmini. Sistema Baseado em Regras para o Refinamento da Segmentação Automática de Fala. Tese de doutorado, Universidade Estadual de Campinas, Campinas, Brasil, Agosto de 2008.

[2] E. D. S. Paranaguá. Segmentação Automática do Sinal de Voz para Sistemas de Conversão Texto-Fala. Tese de doutorado, Universidade Federal do Rio de Janeiro, Março 2012.

[3] S. S. Park and N. S. Kim. On Using Multiple Models of Automatic Speech Segmatation. *IEEE Transaction on Audio, Speech, and Language Processing*, 15(8):2202-2212, November 2007.

[4] E. Akdemir and T. Çiloglu. Using Visual Information in Automatic Speech Segmentation. *Signal Processing, Communications and Applications Conference*, 2008.

[5] H. Talea and K. Yaghmaie. Automatic Visual Speech Segmentation. *Communitations Software and Networks* (ICCSN), 2011.

[6] R. B. Rocha, V. V. Freire, F. M. B. Junior e M. S. de Alencar. Sistema de Segmentação de Fala Baseado na Observação do *Pitch*. *Revista de Tecnologia da Informação e Comunicação*, 4(1):36-42, 2014.

[7] M. Acioli, N. Azevedo, R. Fonte, R. Caiado e W. Cavalcanti I. R Barros, K. H. Efken. *Ensino, texto e discurso.* Editora CRV, Curitiba, 2014.

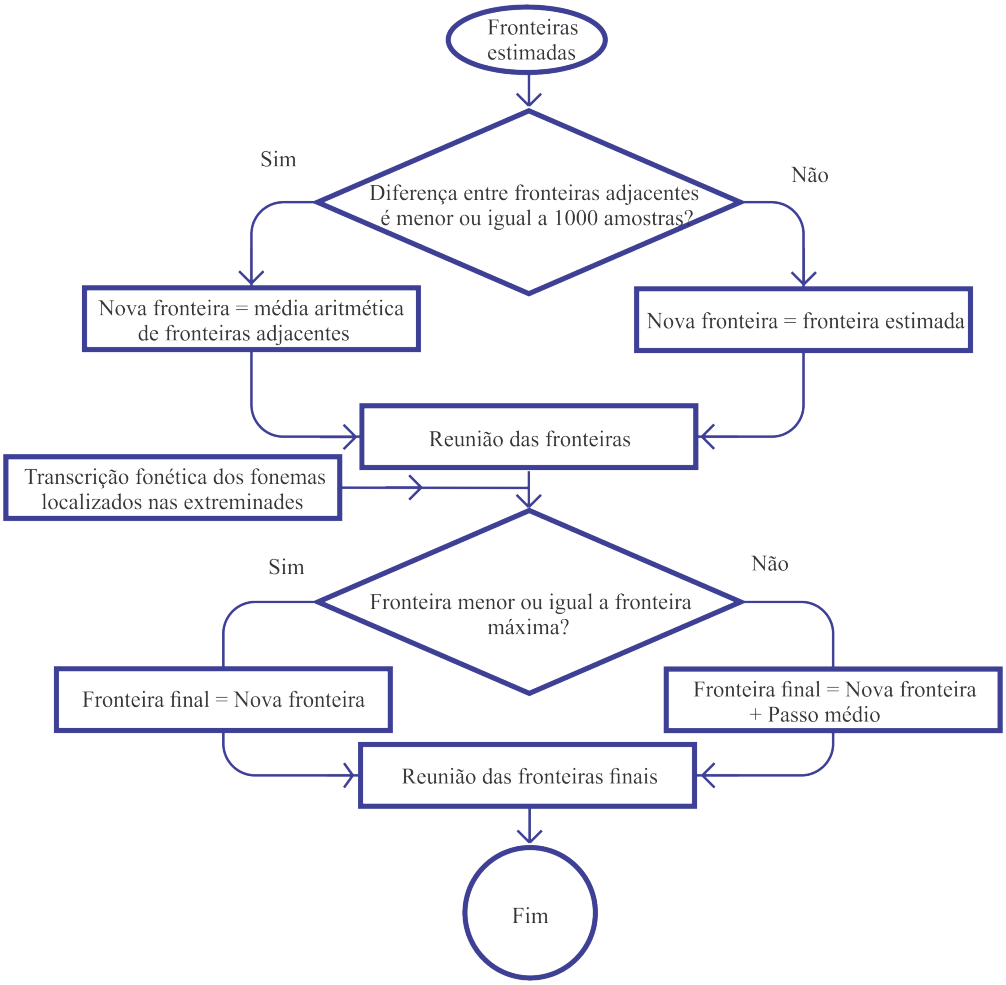

Fig. 4: Etapas para o refinamento do método de segmentação multidimensional de fala.

[8] E. L. F. da Silva. Estimativas de Comportamento Vocálicos de Locutores e um Novo Sistema de Separação Silábica. Dissertação de mestrado, Universidade Federal de Pernambuco, Recife, Brasil, 2012.

[9] S. Jarifi and D. Pastor and O. Rosec. A Fusion Approach for Automatic Speech Segmentation of Large Corpora with Application to Speech Synthesis. *Speech Communication*, (50):67-80, 2008.

[10] L. A. D. M. Figueira. Medidas de Confiançana Segmentação Automática de Fala. Dissertação de mestrado, Universidade Técnica de Lisboa, Novembro 2008.

[11] D. T. Toledano and L. A. H. Gomez and L. V. Grande. Automatic Phonetic Segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617-625, November 2003.

[12] K. Prahallad and A. W. Black. Segmentation of monologues in audio books for building synthetic voices. *IEEE Transaction on Audio, Speech, and Language Processing*, 19(5):1444- 1449, July 2011.

[13] S.Harish, P.Vijayalakshmi, and T. Nagarajan. Significance of Segmentation in Phoneme Based Tamil Speech Recognition System . 2011.

[14] C. Lin and J. R. Jang. Automatic Phonetic Segmentation by Score Predictive Model for the Corpora of Mandarin Singing Voices. *IEEE Transaction on Audio, Speech, and Language Processing*, 15(7):2151-2159, September 2007.

[15] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort. Automatic phone alignment: A comparison between speaker-independent models and models trained on the corpus to align. *Springer Berlin Heidelberg*, 7614:300-311, 2012.

[16] D. C. Costa and G. A. M. Lopes and C. A. B. Mello and H. O. Viana. Speech and Phonema Segmantation Under Noisy Environment Through Spctrogram Image Analysis. *IEEE International Conference on Systems, Man, and Cybernetics*, pages 1017-1022, October 2012.

[17] D. Pekar and S. Tsikhanenka. Speech Segmentation Algorithm Based on an Analysis of the Normalized Power Spectral Density. *Journal of Telecommunications and Information Technology*, pages 44-49, 2010.

[18] S. Paulo and L. C. Oliveira. Automatic Phonetic Alignment and Its Confidence Measures. *4th International Conference*, EsTAL 2004, 2004.

[19] D. T. Toledano and L. A. H. Gómez and L. V. Grande. Automatic Phonetic Segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6), 2003.

[20] H. Kawai and T. Toda. An Evaluation on Automatic Phone Segmentation for Concatenative Speech Synthesis. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2004.

[21] L. F. M. P. Coelho. Etiquetagem Automática de Sinais de Fala Segmentação e Classificação Fonética. Dissertação de mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, Fevereiro de 2005.