TEMA Tend. Mat. Apl. Comput., **8**, No. 2 (2007), 191-200.

Um Novo Método Usando Autocorrelação para Extração da Freqüência Fundamental em Sinais de Voz

A.S. BRANDÃO[1], E. CATALDO[2], F.R. LETA[3], Universidade Federal Fluminense, Rua Passos da Pátria, 156, São Domingos, Niterói, RJ, Brasil.

Resumo Este artigo descreve o algoritmo de extração da freqüência fundamental do sinal de voz usado na implementação do programa P-NAV (Programa Neuro Analizador Vocal), por Brandão (2006). O método proposto toma como base o algoritmo descrito por Boersma (1993), que usa o método da autocorrelação, e desenvolve quatro algoritmos obtendo, com isso, um método mais robusto para marcar corretamente os períodos do sinal de voz, mesmo em trechos severamente perturbados e diplofônicos.

1. Introdução

A detecção da freqüência fundamental do sinal de voz (ou *pitch*) é fundamental no processamento de sinais de voz, pois indica diretamente a presença de vocalização, ou seja, que as cordas vocais estão vibrando. A maioria das medidas acústicas depende da detecção do *pitch* e a diferença entre os valores das medidas acústicas nos programas existentes é devida, em grande parte, à diferença entre os diversos algoritmos usados [10]. Um dos algoritmos bem conhecidos é o de Boersma [1], que usa o método da autocorrelação. Porém, esse algoritmo apresenta diversas falhas, que propomos corrigir com o método que será descrito no decorrer desse artigo.

2. Janelas e Detecção de Trechos Vocalizados

O primeiro ajuste feito, em relação ao algoritmo de Boersma [1], é relacionado à escolha das janelas que deveriam ser descartadas da análise do período do sinal de voz, baseada no critério de preenchimento da janela. Por exemplo, dada uma janela em uma posição do arquivo de som, ela será analisada pelo algoritmo apenas se estiver pelo menos 3/4 preenchida com trecho vocalizado. A Fig. 1 ilustra os casos.

O algoritmo proposto determina, em cada posição da janela, se há vocalização ou não, através da verificação da existência de pelo menos um valor de amplitude

[1]Programa de Pós-Graduação em Engenharia Mecânica; brandaoalexandre@ig.com.br
[2]Departamento de Matemática; Programa de Pós-Graduação em Engenharia de Telecomunicações, ecataldo@im.uff.br
[3]Departamento de Engenharia Mecânica; fabiana@vm.uff.br

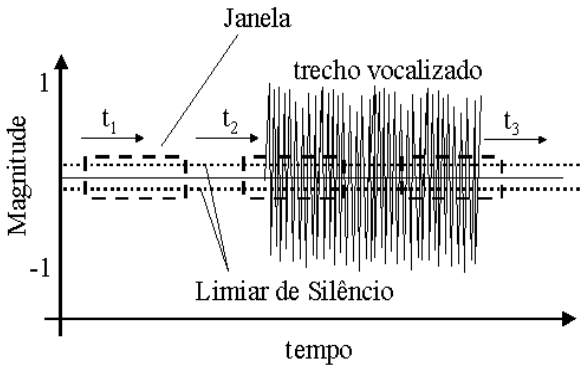


Figura 1: Determinação de trechos vocalizados. t_1 - não vocalizado, t_2 e t_3 - vocalizados

acima do limiar de silêncio (outro parâmetro do programa P-NAV) [3], para cada quarta parte da janela. Caso encontre vocalização em pelo menos três quartos da janela, a função de autocorrelação é calculada para determinação do período da freqüência fundamental. Caso contrário, avança para a próxima janela e reinicia o procedimento.

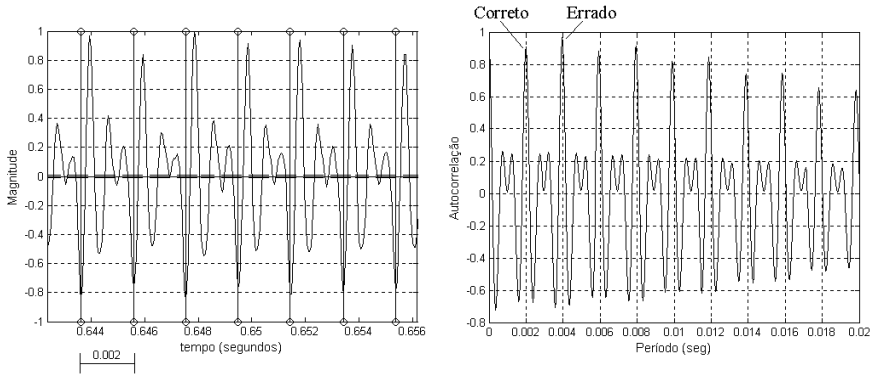
3. Algoritmo para Seleção do Pico da Função de Autocorrelação

Todo algoritmo que depende apenas do cálculo da função de autocorrelação para definir os períodos do sinal de voz está sujeito à escolha dos picos errados nesta função, podendo atribuir para a freqüência fundamental valores maiores ou menores que o valor correto.

Quando a janela atravessa trechos de grande variação na forma de onda do sinal (trechos silenciosos para trechos vocalizados ou variação brusca da freqüência fundamental entre dois trechos), a autocorrelação se reduz e seus picos tendem a se reduzirem, com exceção dos mais próximos da origem, fazendo com que o algoritmo escolha um valor de período menor do que deveria.

Para os casos em que o problema é a mudança da freqüência fundamental de forma brusca entre dois trechos do arquivo, o primeiro algoritmo aqui proposto define um critério de seleção do pico da função de autocorrelação. Quando a forma do sinal de voz tende a ser muito parecida com a de um sinal periódico, pode enganar a função de autocorrelação. Por exemplo, de dois em dois períodos o sinal também se repete, fazendo com que o algoritmo escolha um pico (o máximo) localizado a uma distância da origem que pode ser múltipla da distância de outro pico (de valor menor), que deveria ser o escolhido. A Fig. 2-(a) mostra o trecho do sinal marcado corretamente pelo algoritmo do P-NAV.

Repare que a diferença entre os instantes de tempo das linhas verticais corres-



(a) (b)

Figura 2: (a) Períodos corretos marcados. (b) Picos da Função de Autocorrelação.

ponde a 0.002 s, que é o período correspondente ao pico da função de autocorrelação mais próximo da origem, Fig. 2-(b).

Visando corrigir o problema descrito acima, dada a função de autocorrelação do trecho do sinal, foi implementado o seguinte critério de escolha:

1. Localize o primeiro máximo P, que estará a uma distância T da origem. T é o valor do primeiro candidato a período local.

2. Verifique se existe algum candidato a máximo mais provável a uma distância maior que T.

2.1 Localize o próximo máximo P_2 (que, se encontrado, estará a uma distância T_2 maior que T).

2.2 A altura deste máximo deve ser maior ou igual a 50 % (atualizado em relação à referência [3]) do primeiro máximo e T_2 deverá ser superior ao triplo de T, caso contrário rejeitar P_2. **Obs**: Isso evita a escolha de picos próximos demais da origem, causados por trechos de baixa autocorrelação do sinal.

2.3 Se P_2 foi rejeitado ou escolhido, o passo 3 a seguir vale para P ou P_2, respectivamente.

3. Verifique se existe algum candidato a máximo mais provável, a uma distância menor que T.

3.1 Localize o máximo anterior P_2 (que, se encontrado, estará a uma distância T_2 menor que T).

3.2 A altura deste máximo deve ser maior ou igual a 50% (atualizado em relação à referência [3]) do primeiro máximo e T_2 deverá ser aproximadamente a metade de T, caso contrário rejeitar P_2. **Obs**: Isso evita a escolha de picos distantes demais da origem causados por funções de autocorrelação geradas devido a trechos com formas muito próximas da periodicidade do sinal. Ver Fig. 2-(b).

As modificações acima são capazes de eliminar alguns erros, mas não todos.

Detectar o pico correto da função de autocorrelação influi de maneira decisiva na marcação dos períodos do sinal de voz. A magnitude desses picos decai com a distância da origem. Baseado nesse decaimento foi estipulado o critério de seleção do próximo candidato a período dos passos 2.2 e 3.2.

Outros artigos utilizam técnicas diferentes (cepstrum [5], correlação cruzada normalizada [11], etc.) para extração da freqüência fundamental. A maioria busca uma estimativa média desta freqüência em um trecho do sinal de voz, com a preocupação maior de descartar os trechos do sinal com detecção ruim, não sendo, portanto, passíveis de comparação com o algoritmo deste trabalho, cujo objetivo é definir e marcar a posição correta de cada pulso do sinal de voz, mesmo nos trechos difíceis ou impossíveis, pois, como será visto, nestes trechos foi usada interpolação linear para estimar os valores. Motliček [9] usa otimização em grafos, para melhorar métodos de detecção existentes. Na referência [7] é usado um método para estimar diferentes freqüências fundamentais em arquivos com sons de instrumentos musicais. A idéia básica é estimar a freqüência independentemente para diferentes faixas do espectro e depois ver qual tem a maior probabilidade de ser a fundamental. Possui boa taxa de acerto, mas não marca os pulsos graficamente e não usa interpolação para os trechos de difícil detecção. Além disso, os sons de instrumentos musicais têm periodicidade muito melhor definida do que a voz humana.

4. Algoritmo de Extração do Pitch

Este algoritmo é a base da análise do sinal de voz feita pelo programa P-NAV [3]. Ele varre o arquivo do início ao fim, obtendo os valores locais do período da freqüência fundamental, bem como os valores locais da medida HNR. Seus resultados são usados posteriormente pelo algoritmo que faz a marcação dos pulsos.

1. Definir o tamanho da janela (seis vezes o período da mínima freqüência detectável). Isso mantém o erro na ordem de 10^{-6} para sinais com essa freqüência conforme referência [1].
2. Definir a velocidade da janela, que é o passo de avanço da região de análise ao longo do arquivo de som.
3. Percorrer o arquivo de som do início ao fim, avançando trecho selecionado e para cada trecho:
   3.1 Calcular autocorrelação do sinal.
   3.2 Localizar o máximo na função de autocorrelação (algoritmo da subseção anterior).
   3.3 Pela posição desse máximo, definir o período local.
   3.4 Pela magnitude desse máximo, definir o valor para o HNR local.

O passo 3 gera um vetor com os diversos períodos locais e valores de HNR locais determinados para os pontos em que a janela se fixou, alguns certos, alguns errados, caso a forma de onda do sinal gere baixa autocorrelação em alguns pontos. Ver Figura 3-(a). Para sinais que, mesmo perturbados, possuam uma certa periodicidade, os valores de período corretos serão sempre em maior quantidade que os

errados, mas ainda assim existem erros.

5. Algoritmo de Filtragem de Períodos com Valor Errado

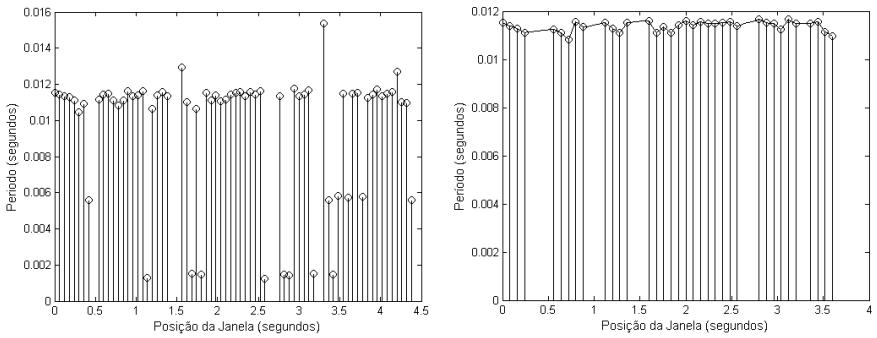
A solução foi proceder à eliminação dos valores errôneos. Cada valor de período em que ocorre erro é geralmente um múltiplo ou um divisor do valor correto. Então, entre dois períodos consecutivos há um salto. Baseado neste detalhe foi criado o seguinte algoritmo:

1. Percorrer o vetor de períodos.
2. Comparar o período atual e o seguinte.

2.1 Se o módulo da diferença entre os dois for superior a 27% (valor experimental).

2.1.1 Eliminar (atribuir valor zero) a todos os períodos que tenham valor 5% em torno daquele que tiver menor incidência no vetor.

Após o término da filtragem dos períodos errôneos, o vetor de períodos ficará com algumas componentes eliminadas (com valor zero). Para colocar o valor correto nessas componentes é feita a interpolação linear dos valores das componentes diferentes de zero. Ver Figura 3-(b).



(a) (b)

Figura 3: (a) Erros de Extração do Período pela Função de Autocorrelação. (b)Períodos Interpolados.

A escolha dos valores de 27% e 5% para os passos 2.1 e 2.1.1 partem de um princípio muito simples. Estamos supondo que, quando uma pessoa emite uma vogal sustentada, não pode haver variação superior a 27% entre dois períodos consecutivos sem que haja erro na detecção de um dos valores. Para não eliminarmos muitos períodos, caso a detecção de erro tenha sido equivocada, estabelecemos o valor de 5% para o passo 2.1.1. Isso torna o algoritmo mais poderoso, capaz de suportar variações bruscas de freqüência, Figuras 4-(a), (b) e (c).

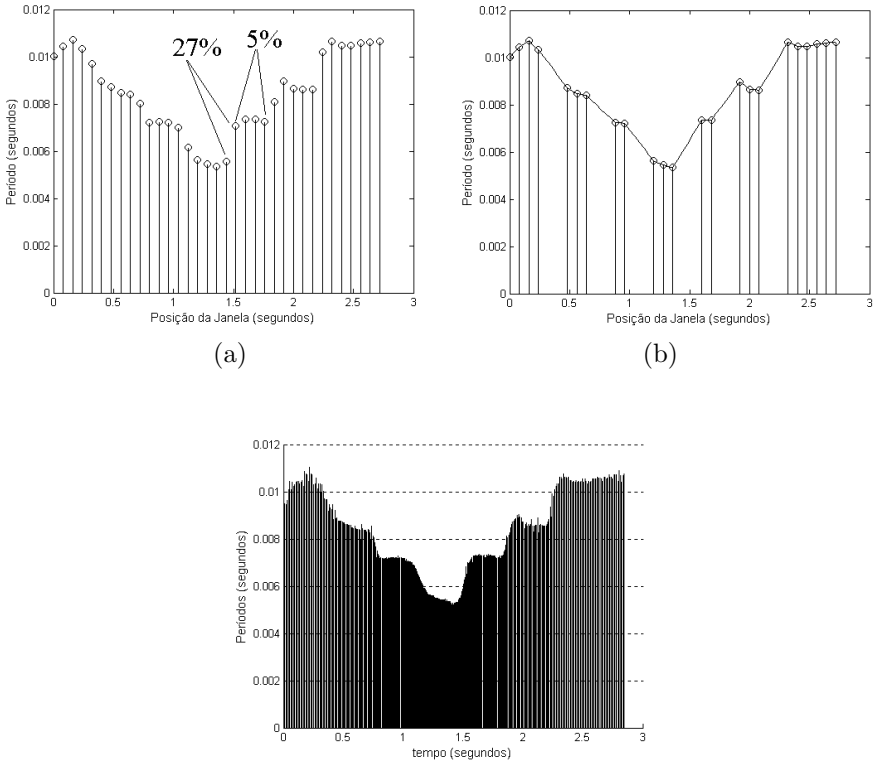


Figura 4: (a) Períodos extraídos. (b) Filtragem e interpolação. (c) Períodos corretos marcados pelo algoritmo da próxima seção.

A próxima etapa é marcar os períodos no gráfico do arquivo nas posições corretas. O algoritmo de marcação dos pulsos, descrito a seguir, utiliza o vetor de períodos calculados anteriormente. Sendo assim, a filtragem dos períodos errôneos, seguida da interpolação dos valores, traz grandes benefícios. Essa interpolação permite que o algoritmo de marcação dos pulsos funcione muito bem, pois sempre existirá uma referência correta para o valor do período local, permitindo acertar a marcação em sinais com certa periodicidade, porém com muitas ondulações e distorções na forma de onda, que geram valores de autocorrelação que, algumas vezes, não permitem a detecção do período local (Fig. 5).

6. Algoritmo de Marcação dos Pulsos

O algoritmo para marcar os períodos é descrito a seguir:

1. Percorrer o arquivo de som com velocidade de janela igual a 1 (passo de avanço igual ao tamanho da janela) até encontrar o primeiro valor P acima do parâmetro limiar de silêncio.

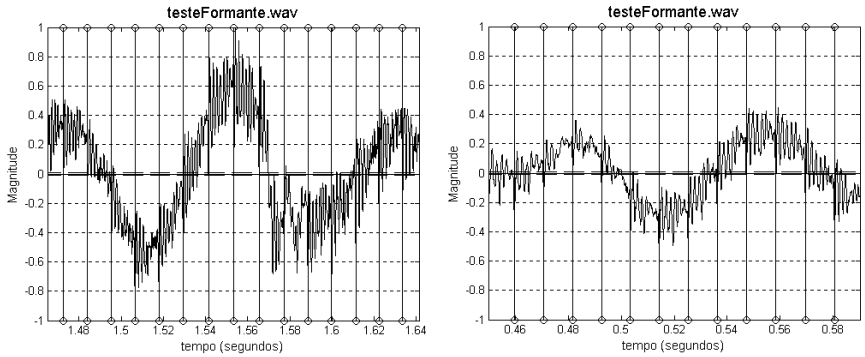


Figura 5: Marcação dos Pulsos (trechos diferentes do mesmo arquivo).

2. Encontrado esse valor P (e a sua posição no arquivo), encontre o valor de período determinado anteriormente para uma janela localizada numa posição que contenha a posição em que o valor P se encontra.

3. A partir da posição de P no arquivo, até a posição de P mais valor do período local, encontre um mínimo M_1 (e a sua posição no arquivo), encontre o valor de período determinado anteriormente para uma janela localizada numa posição que contenha a posição em que o valor M_1 se encontra.

4. A partir da posição do mínimo M_1, o próximo mínimo é encontrado numa faixa em torno de 10% do valor da posição de M_1 no arquivo, mais o período local. Esses 10% correspondem ao parâmetro fator de busca que também pode ser ajustado na interface do programa, sendo 10% um valor padrão, que foi obtido experimentalmente.

5. Os mínimos vão sendo encontrados e marcados seqüencialmente até que o arquivo ou o trecho vocalizado se acabem.

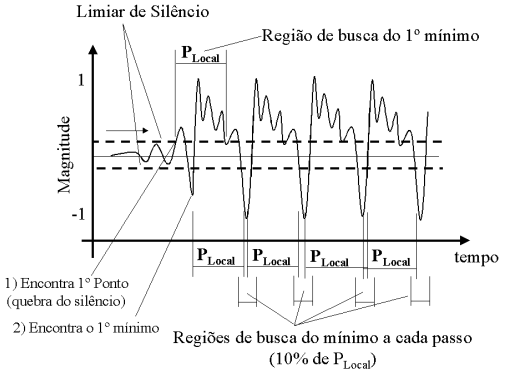


Figura 6: Algoritmo de Marcação dos Pulsos.

A Figura 6 ilustra os passos do algoritmo de marcação dos pulsos do sinal de

voz.

A partir dos valores dos pulsos marcados no algoritmo descrito anteriormente, são calculados os períodos reais do sinal de voz. Assim, é obtido um novo vetor de períodos, onde cada componente é formada pela diferença entre as posições no arquivo de cada dois pulsos consecutivos. Segundo o artigo de Rabiner et al [10] é extremamente difícil detectar a freqüência fundamental para pacientes diplofônicos, mesmo nas melhores condições. Nestes casos, a maioria dos algoritmos detectores de freqüência fundamental tendem a marcar o período como sendo a distância entre os períodos alternados e não entre os períodos adjacentes, fornecendo como resultado a metade do valor real da freqüência fundamental.

7. Robustez contra Diplofonia

Um sinal de voz possui características de diplofonia quando os pulsos alternados são iguais em período e amplitude, ao passo que os pulsos consecutivos são diferentes. Algoritmos que usam apenas autocorrelação tendem a marcar os períodos com o dobro do tamanho real, porque a autocorrelação de dois em dois períodos tende a ser maior que de um em um. Porém, com o critério de seleção, filtragem e marcação de pulsos implementado, foi possível marcar corretamente os períodos. Uma pergunta que surge é: como saber se a freqüência fundamental está sendo detectada com a metade do valor real, ou se o valor real é esse mesmo? (Pois o sinal também se repete de dois em dois períodos). O fato é que, neste caso, a voz era de um paciente do sexo feminino e a freqüência estava sendo detectada como 93 Hz no trecho diplofônico, valor característico de voz grave (masculina). A Figura 7 mostra o gráfico dos períodos detectados antes do ajuste nos parâmetros. Repare que o trecho inicial do arquivo não é diplofônico, correspondendo a um período médio de 0.0053 s (188.67 Hz), enquanto que de 0.5 s em diante o sinal passa a se tornar diplofônico, registrando um período médio correspondente a 0.0107 s (93 Hz).

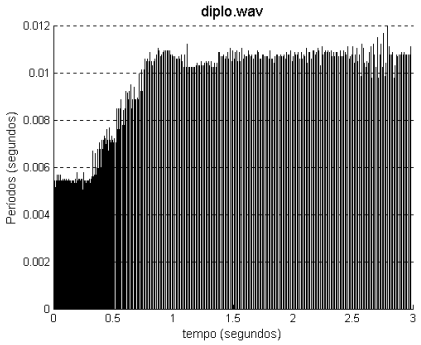
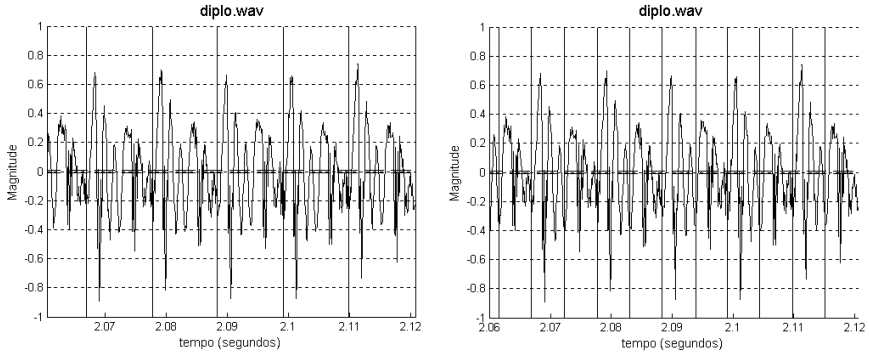


Figura 7: Períodos detectados (método convencional).

O gráfico dos períodos na Fig. 7 deveria ter sua média em torno de 0.0053 s, que é o valor correto do período médio para o paciente em questão. Este algoritmo

permite que isso aconteça. As Figuras 8-(a) e (b) mostram um mesmo trecho do arquivo, na região diplofônica, para que se possa visualizar melhor a diferença de marcação dos períodos.



(a) (b)

Figura 8: (a) Marcação dos pulsos no trecho diplofônico (antes). (b) Marcação dos pulsos no trecho diplofônico (depois).

8. Conclusões

Um novo método para extração da freqüência fundamental é proposto. O método foi testado e comparado com resultados obtidos por outros métodos, e os resultados apresentaram-se melhores, principalmente em relação à determinação de freqüência fundamental para vozes diplofônicas. Embora o método tome como base o cálculo da função de autocorrelação para a marcação dos períodos, a novidade está no fato de que não é essa a única ferramenta usada e, dessa forma, maior robustez é obtida.

Referências

[1] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *IFA Proceedings*, **17** (1993), 97-110.

[2] A. Brandão, F.R. Leta, Usando redes neurais para classificação de padrões de voz, em "XXVII CNMAC - Congresso Nacional de Matemática Aplicada e Computacional", SBMAC, 2005.

[3] A. Brandão, "Classificação de Vozes Naturais e de Vozes Sintetizadas através de Modelos Mecânicos de Laringe e de Trato Vocal usando Redes Neurais", Dissertação de Mestrado, Universidade Federal Fluminense, Niterói, RJ, 2006.

[4] A. Brandão, E. Cataldo, R. Sampaio, "Análise e Processamento de Sinais", Apostila, SBMAC, 2005.

[5] J. Cernocky, "Speech Processing Using Automatically Derived Segmental Units", PhD Thesis, ESIEE, France, 1998.

[6] M.P. Karnell, Laryngeal perturbation analysis: minimum length of analysis window, *Journal of Speech and Hearing Research*, **34** (1991), 544-548.

[7] A.P. Klapuri, Multiple fundamental frequency estimation based on harmonicity and spectral smoothness, *IEEE Transactions on Speech and Audio Processing*, **11**, No. 6 (2003).

[8] P. Lieberman, Perturbation in vocal pitch, *Journal of the Acoustical Society of America*, **33** (1961), 597-603.

[9] P. Motlíček, L. Burget, "Reliability Improvement of Speech Pitch Detetion Using Paths", Institute of Radio Electronics, Faculty of Electrical Engineering, TU Brno, 2000.

[10] L.R. Rabiner, et al., A comparative performance study of several pitch detection algorithms, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-24**, No. 5 (1976).

[11] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT). Speech Coding and Synthesis". New York, Elsevier, 1995.

[12] D. Wong, R. Lange, I. Titze, C.G. Guo, Mechanisms of Jitter-Induced Shimmer in a driven model of vocal fold vibration, in "NCVS Status and Progress Report", pp. 33-41, 1995.