Estimação de F_0 via MDC aproximado e agrupamento por K-médias ponderado

Marcelo de Oliveira Rosa

***Resumo*— Dado um conjunto de frequências de sinais parciais presentes em segmentos de sinais enjanelados de curta duração, propõem-se um estimador de frequência fundamental calculado a partir do máximo divisor comum (MDC) aproximado de tais frequências parciais e da aplicação do método K-médias ponderado para sua classificação. Tal algoritmo permite uso de segmentos de curta duração, resultando em alta resolução temporal da estimação da curva da F_0. A estratégia produziu bons resultados quando a relação sinal-ruído é muito baixa (-5dB) ou quando as frequências formantes da voz localizam-se próximas à F_0 ou a seus primeiros harmônicos.**

***Palavras-Chave*— Estimação de pitch, MDC, agrupamento por K-médias**

***Abstract*— Given a set of spectrum partials obtained from shorter windowed signal frames, it is proposed a fundamental frequency estimation by calculating the approximate GCD of such partials and applying a weighted K-means cluster method to classify them. Such an algorithm allows the use of shorter frame sizes, resulting in high temporal resolution for the estimation of F_0 curve. The strategy produced good results when the SNR was very low (-5dB) or when the formant centers are closer to the f_0 or to its first harmonics.**

***Keywords*— Pitch estimation, GCD, K-means clustering**

I. INTRODUÇÃO

Uma das características importantes dos sinais de fala é sua frequência fundamental (ou *pitch*). Tal componente da voz diz respeito a vibração das cordas vocais do falante. Cabe ressaltar que algumas vezes encontrar-se-á definições musicais ou psicoacústicas para *pitch*. No presente texto, *pitch* e frequência fundamental são sinônimos para sinais de fala.

A presença ou não de estrutura harmônica (pela vibração ou não das cordas vocais) define os fonemas sonoros dos não sonoros. Como característica ligada a vibração das cordas vocais, pode ser útil na identificação do falante ou em codificação de fala. Aplicações musicais baseiam-se no *pitch* instantâneo para alterá-lo de forma a corrigir imperfeições do cantor/falante.

Determinar a frequência fundamental de sinais de fala é reconhecidamente um procedimento algorítmico difícil quando se considera a variabilidade das formas que os tratos glotal e supraglotal podem assumir. A presença de ruídos de ambiente ou sinais musicais no sinal analisado devem ser considerados pela sua relevância prática quando se analisa sinais de fala [1].

Métodos baseados no domínio do tempo têm sido desenvolvidos para identificar o período fundamental de sinais de fala: por exemplo, a busca do instante de máxima amplitude da função de autocorrelação de curta duração ($r_t(\tau)$) define o período fundamental ($T_0 = 1/F_0$) do sinal no instante t do sinal analisado. Algumas variações desse método têm lidado com seus problemas inerentes como *pitch* duplo, por exemplo [2].

Através da análise da densidade de potência espectral (PS) de curta duração, os componentes harmônicos podem ser identificados e usados para derivar a F_0 [3], [4]. Este tipo de análise, que é usado em procedimentos de análise por síntese [5] e que envolve implicitamente uma de filtragem do tipo pente (*comb-filtering*), isola as frequências harmônicas presentes no sinal analisado usadas para derivar a F_0 de um segmento de fala. De modo similar, através da separação não-linear entre a envoltória formante da fala e a excitação periódica do trato glotal, a análise cepstral pode ser considerada um tipo de análise espectral [6].

Neste breve resumo de métodos determinísticos de estimação da F_0, algumas dificuldades podem ser listadas: a influência da amplificação provocada pelas frequências formantes em harmônicos de alta frequência (como é o caso do fonema /i/), a redução ou eliminação da amplitude do sinal parcial relativo à própria F_0 devido a algum fenômeno espectral destrutivo, a presença de falsos F_0 com características espectro-temporais equivalentes ao F_0 correto, como *pitch* duplo (em que a F_0 estimada é o dobro da correta), e a presença de ruído colorido (que prejudica métodos de estimação de *pitch* baseados em filtragem inversa, por exemplo).

Alguns métodos estatísticos [7], [8] tem sido propostos como alternativa para superar as dificuldade de métodos baseados em *threshold* (chamados de métodos *hard decision*). Nesses há um procedimento de estimação dos potenciais candidatos a F_0 para um dado segmento e a aplicação de procedimentos classificatórios (*soft decision*) que analisa a característica espectral do segmento analisado (incluindo a relação sinal-ruído local) em associação com uma base de casos previamente levantadas (por treinamento) para definir o valor correto da F_0 do segmento.

O presente trabalho baseia-se na premissa de que a estrutura harmônica da fala tem grande largura de banda, espalhando-se pela curva de densidade de potência espectral (PS), mesmo em altas frequências, conforme teoria de Licklider sobre a percepção de *pitch* [8] que diz que o cérebro processa grupos de harmônicos sequenciais para estimação do *pitch* do sinal audível. Assim, o conjunto das frequências relativas aos máximos locais da PS contém informação redundante sobre a F_0 do segmento de sinal processado.

Para agrupar essas frequências (que contém harmônicas da F_0 desejada) empregou-se uma aproximação do algoritmo de máximo divisor comum (MDC) que determina um conjunto potencial de F_0s, e um algoritmo de K-médias ponderadas para

Marcelo de Oliveira Rosa, Departamento Acadêmico de Eletrotécnica, Universidade Tecnológica Federal do Paraná, Av. Sete de Setembro, 3165, Rebouças, 80230-901, Curitiba, Paraná, Brasil, E-mail: mrosa@utfpr.edu.br.
Este trabalho foi financiado pelo CNPq e Fundação Araucária

agrupar tal conjunto em termos de sua relação de amplitude e calcular a correta F_0 do segmento.

O desenvolvimento desta abordagem foi orientado pela: redução da complexidade da estimativa de *pitch* em relação às técnicas mais recentes, eliminação de *thresholds* nos algoritmos envolvidos (algo comum em técnicas de estimação de F_0) e redução do tamanho dos segmentos para aumentar a resolução temporal da curva de F_0 do sinal.

Nas seções II e III são apresentadas as definições matemáticas dos algoritmos fundamentais, enquanto na seção IV alguns resultados são discutidos.

II. MDC APROXIMADO

Na primeira etapa da estimação da F_0 extraiu-se os sinais parciais que compõem o segmento analisado. Cada segmento do sinal foi modulado por uma janela de Hamming e sua curva de densidade de potência espectral (PS) foi calculada via magnitude quadrada de sua transformada discreta de Fourier (DFT), ou seja $|S[f]|^2$. No cálculo de tal DFT foram acrescentados zeros ao final do segmento (em quantidade igual a 4 vezes o tamanho original do segmento).

Na sequência, todas as frequências (f_p, totalizando P frequências) dos sinais parciais contidas no segmento foram obtidas a partir dos máximos locais de $|S[f]|^2$ no intervalo de 40 a 1000Hz (para aproveitar a redundância da informação harmônica conforme [8]), ou seja:

$$\begin{aligned} &|S[f_p - \Delta_f]|^2 < |S[f_p]|^2 \geq |S[f_p + \Delta_f]|^2 \text{ ou} \\ &|S[f_p - \Delta_f]|^2 \leq |S[f_p]|^2 > |S[f_p + \Delta_f]|^2 \end{aligned} \quad (1)$$

onde Δ_f é a resolução espectral de acordo com o tamanho da DFT (5 vezes o tamanho do segmento).

Com o acréscimo de zeros, a PS do segmento gera frequências f_p muito próximas do máximo efetivo devido a discretização envolvida no cálculo da DFT. Assim, uma interpolação parabólica foi aplicada para determinar precisamente as frequências dos sinais parciais ($\widetilde{f}_p$), pois a adição dos zeros garante que a tal interpolação seja equivalente à interpolação espectral. Naturalmente as frequências envolvidas nas inequações 1 são muito próximas entre si. Neste trabalho não se considerou o caso em que $|S[f_p - \Delta_f]|^2 = |S[f_p]|^2 = |S[f_p + \Delta_f]|^2$.

No passo seguinte calculou-se o MDC de todas as combinações (2 a 2) das frequências $\widetilde{f}_p$ para estimar todos os potenciais F_0 para o segmento. O algoritmo para cálculo do MDC entre números reais é baseado no algoritmo de Euclides e em [3]. Dado duas frequências f_a e f_b (no caso, interpolações das frequências dos sinais parciais), o seu MDC aproximado é obtido pela recursão:

$$\begin{aligned} &\text{app\_gcd}(f_a, f_b, err) = f_a \qquad \text{se } f_b \leq err \\ &\text{app\_gcd}(f_a, f_b, err) = \text{app\_gcd}(f_b, f_a \% f_b, err) \end{aligned} \quad (2)$$

na qual err é o erro de aproximação e o símbolo $\%$ corresponde ao resto da divisão inteira.

A definição de err, que atua como critério de parada para a recursão (e também um *threshold* para este estimador de F_0) demanda um compromisso similar ao seu equivalente em [3]: pequenos valores de err produzem pequenos valores de MDC dependendo das características aritméticas de f_a e f_b. Aqui definiu-se $err = 40$Hz baseado no intervalo de busca da F_0 (que foi de 40 a 1000Hz). Assim, um MDC pequeno na recursão indica que duas frequências não são harmônicas entre si.

A. *Identificando a* F_0

Agora o problema consiste em definir as F_0s efetivas dos sinais presentes no segmento. Isso porque o MDC aproximado poderia ser usado para estimar múltiplas frequências fundamentais (por exemplo, das vozes de múltiplos falantes). A Figura 1 mostra um conjunto de amplitudes e frequências de sinais parciais obtidos a partir de dois sinais de fala hipotéticos (cujas F_0s são 100 e 120Hz respectivamente) presentes no segmento analisado.

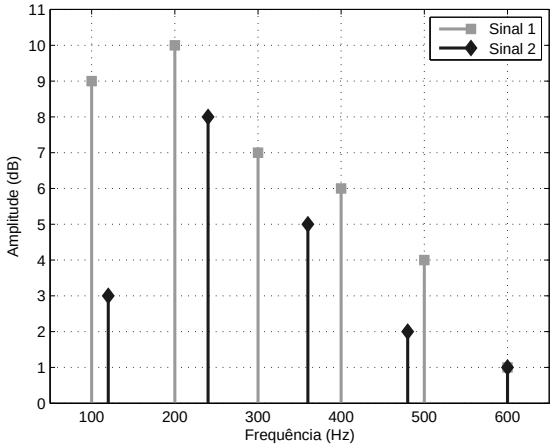

Fig. 1. Conjunto de frequências e amplitudes de sinais parciais de um segmento de fala hipotético (que é a soma de dois sinais - Sinal 1 e Sinal 2).

Após o cálculo do MDC aproximado de todas as 45 combinações das frequências desses sinais parciais (combinação das 10 frequências da Figura 1, tomadas 2 a 2, sem repetição), observou-se que o número de ocorrências dos valores de MDC para as frequências 100 e 120Hz foi significativo (Figura 2). Tais frequências corresponderam às F_0s dos sinais presentes no segmento analisado. Poder-se-ia escolher 100Hz como frequência fundamental do sinal mais significativo no segmento. Outra alternativa seria adotar o procedimento iterativo de [1], que consiste em escolher a F_0 do sinal mais significativo do segmento e remover tal sinal do segmento, progressivamente até que não haja mais estrutura harmônica no segmento.

Apesar dos resultados aparentemente promissores, perturbações numéricas na estimação das frequências dos sinais parciais tornam a distribuição dos potenciais F_0 mais uniforme, conforme mostra a Figura 3. Na sua construção, acrescentou-se ruído com distribuição normal N($\mu = 0$, $\sigma^2 = 25$) às frequências dos sinais parciais mostrados na Figura 1. Percebe-se que apesar da redução do número de ocorrências (reduzido em relação à Figura 2) as potenciais F_0s dos sinais parciais foram 100 e 120Hz, mesmos valores obtidos anteriormente. Nota-se uma grande quantidade potenciais (e falsos) F_0s

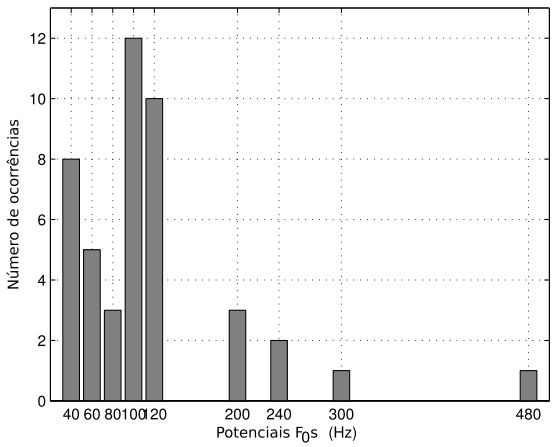

Fig. 2. Histograma dos MDCs aproximados dos pares de frequências de sinais parciais.

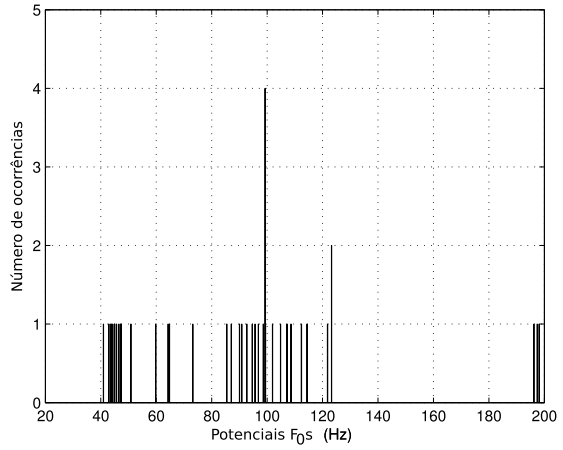

Fig. 3. Histograma dos MDCs aproximados dos pares de frequências de sinais parciais com ruído aditivo.

próximo a 40Hz, valor relacionado com a condição de parada (err) do algoritmo utilizado para cálculo aproximado do MDC entre frequências.

III. Agrupamento por K-médias

A estratégia de agrupamento baseada no método de K-médias [9] foi estruturada para determinar qual a F_0 predominante no segmento com baixa complexidade computacional (quando comparada a [1], [8], por exemplo), ao invés de se obter todas as F_0s presentes no segmento - estimação de múltiplas frequências fundamentais - por causa de uma característica importante do método de agrupamento: ele requer, previamente, a definição de quantos (K) subconjuntos existem em um conjunto de potenciais F_0s (que foram obtidos a partir do cálculo exaustivo de MDC entre as frequências interpoladas - $\widetilde{f}_p$ dos sinais parciais).

Assim a estimação da F_0 consistiu em aplicar o método de K-médias assumindo que existiam apenas 2 subconjuntos no conjunto s de F_0s potenciais analisados. Na sequência, o subconjunto de menor tamanho foi descartado e o método foi reaplicado. Tal procedimento iterativo continuou até que existisse um único subconjunto no conjunto de F_0s analisado. Tal procedimento definiu a frequência fundamental do sinal mais significativo no segmento de acordo com a concentração (número de ocorrências) de vários potenciais F_0s na vizinhança do valor correto.

Analisando a PS dos segmentos, percebe-se que a magnitude dos sinais parciais influencia a identificação das frequências fundamentais. O processo de agrupamento apresentado até aqui considerou que todos os potenciais F_0s foram obtidos a partir de sinais parciais cujas amplitudes são iguais. Assim, componentes não-harmônicas das F_0s presentes no segmento, muitas vezes de pequena magnitude (devido a efeitos numéricos nos cálculos da DFT ou do enjanelamento, presença de *jitter* e o compromisso entre as resoluções temporal e espetral, por exemplo), potencializam erros na estimação de F_0 via agrupamento por K-médias, pois tal algoritmo pode decidir por F_0s próximas a err.

Para incorporar a magnitude da componente espectral associada ($M[p]$) à frequência ($\widetilde{f}_p$) do sinal parcial como um peso que privilegie os mais proeminentes sinais parciais no processo de estimação de F_0, definiu-se o seguinte algoritmo: primeiramente atribuiu-se um peso para todas das frequências dos P sinais parciais presentes no segmento a partir da magnitude $M[p]$ (em dB, de $1 \leq p \leq P$), que é calculado por:

$$A[p] = \lceil M[p] - \min(M) + 1 \rceil \tag{3}$$

na qual $\min(M)$ é a menor magnitude (em dB) dos P sinais parciais identificados no segmento.

Assim, para cada combinação de duas frequências de sinais parciais ($f[a]$ e $f[b]$, com $1 \leq a \leq P$ e $1 \leq b \leq P$) presentes no segmento analisado, calculou-se o MDC aproximado para produzir o potencial F_0 a ser considerado pelo algoritmo de agrupamento. Cada F_0 potencial recebeu um peso representando quão forte é tal F_0 em comparação com as demais. Esse peso foi definido pelo produto:

$$W_{a,b} = A[a]\ A[b] \tag{4}$$

Com isso, as relações entre frequências parciais com maior magnitude espectral foram reforçadas.

Finalmente, o conjunto de frequências usadas no método de K-médias foi modificado para incorporar tais pesos: o MDC aproximado de cada par de frequências (f_a e f_b), que é o potencial F_0 do destas, foi replicado $W_{a,b}$ vezes no conjunto de dados analisado pelo método. Por exemplo, se $W_{a,b} = 4$ e $F_{0_{a,b}} = 100$Hz, sua contribuição no conjunto de dados seria $s = \{\cdots, 100, 100, 100, 100, \cdots\}$.

IV. Resultados e Discussões

O método apresentado não define se o segmento é vocálico, não-vocálico ou silêncio. Assim, os resultados foram orientados a análise de vogais sustentadas.

O tamanho do segmento foi fixado em aproximadamente 3 vezes o período fundamental do sinal analisado, para garantir que houvesse um vale entre sucessivos picos harmônicos na

PS (o intervalo entre sucessivos segmentos foi fixado em um período fundamental desse sinal). Tal período foi obtido a partir de observações do sinal analisado, que foi cuidadosamente gravado e discretizado. Essa escolha foi feita para que se tenha elevada resolução temporal na estimação da curva de F_0. Considerando a taxa de amostragem usada ($44,1$kHz) e que o corpus de fonemas usados na análise variou em torno de 80Hz (voz masculina) e 210Hz (voz feminina), o tamanho das janelas variou em torno de 3×550 e 3×210 amostras, respectivamente.

Ruído gaussiano foi adicionado aos sinais estudados para avaliar a performance do método para diferentes SNRs (já que se conhecia previamente a F_0 real dos sinais).

Os fonemas analisados foram /a/, /e/, /i/, /o/ e /u/ (cada fonema foi pronunciado uma vez por dois falantes em sustentação por 4-5 segundos). Entre eles, o fonema /i/ é um desafio para estimação de F_0 quando segmentos de curta duração são usados, pois seu primeiro formante localiza-se na faixa de 200 a 400Hz (apesar de ser um formante com menos potência que o segundo formante, localizado na faixa de 2 a 3kHz). Na Figura 4, nota-se que isso prejudicou métodos baseados na função de autocorrelação na descoberta do valor correto de F_0 (alguma vezes produziu $2F_0$ como frequência fundamental). [2] discutiu como contornar tal dificuldade, atenuando a magnitude de harmônicos de alta frequência para evitar que tais frequências sejam escolhidas como F_0.

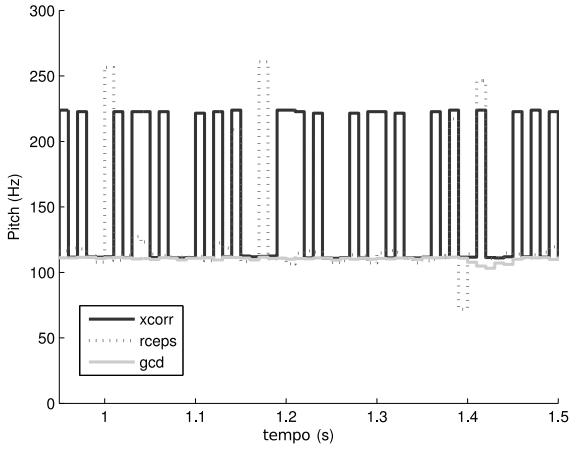

Fig. 4. Exemplo de estimação de *pitch* usando diferentes técnicas para o fonema /i/.

Em comparação com outros métodos (baseados em autocorrelação, em cepstrum, e em resíduos de filtragem LPC), o método aqui apresentado não requer qualquer pré-filtragem (apesar do uso da janela de Hamming, ela não é um requisito, pois o método é baseado nos máximos locais da PS): para os demais métodos usados para comparação neste estudo, os sinais foram pré-filtrados com um filtro digital passa-banda do tipo Butterworth na faixa de 40 a 500Hz.

A Tabela I mostra os resultados de 4 métodos de estimação de F_0. Em todos os métodos, implementou-se a avaliação de *pitch* em um segmento sem considerar o resultado obtido para os segmentos vizinhos (sem *tracking* de *pitch*). O percentual de erro mostrado nessa tabela corresponde ao percentual de

TABELA I

ESTIMAÇÃO DE F_0 EM CONDIÇÕES DE RUÍDO. MDC, XCORR, CEPSTRUM, E RESÍDUOS CORRESPONDEM AOS MÉTODOS MDC + K-MÉDIAS, DE AUTOCORRELAÇÃO, CEPSTRUM, E DOS RESÍDUOS, RESPECTIVAMENTE.

Método	Fonema	Percentual de erro (%) para diferentes SNRs			
		20dB	10dB	0dB	-5dB
xcorr	/a/	0.00	0.00	3.06	12.99
	/e/	0.00	0.00	0.00	1.21
	/i/	42.08	44.87	46.98	50.48
	/o/	0.00	0.00	1.03	4.12
	/u/	13.49	15.87	15.07	18.25
cepstrum	/a/	0.76	0.76	0.00	6.87
	/e/	0.00	0.60	4.82	7.23
	/i/	23.57	41.37	62.84	72.92
	/o/	1.03	0.00	5.15	23.72
	/u/	0.00	0.00	19.83	34.11
resíduos	/a/	0.00	0.00	6.88	0.00
	/e/	0.00	0.00	0.00	1.21
	/i/	0.00	0.00	30.86	68.01
	/o/	0.00	0.00	0.00	0.00
	/u/	0.00	0.00	0.00	8.72
MDC	/a/	0.00	0.00	6.88	14.52
	/e/	0.00	0.00	1.81	6.01
	/i/	1.40	0.70	2.10	29.45
	/o/	0.00	0.00	1.03	8.25
	/u/	0.00	0.00	0.00	7.19

diferenças acima de 5% entre as F_0s estimada e correta (conhecida previamente), ao longo do sinal de voz.

O fonema /i/, como esperado, causou dificuldades para os estimadores de *pitch*, especialmente nas condições de ruído (SNRs iguais a 0 e -5dB). O método MDC apresentou bons resultados quando comparado aos demais. Os erros dos demais métodos geralmente originaram-se do pequeno tamanho da janela e produziram F_0s iguais ao dobro do valor correto para alguns segmentos ao longo do sinal analisado.

Nota-se que a performance do método MDC não foi a melhor em todos os casos considerados, mas foi a mais regular. Isto ocorre porque o agrupamento ponderado por K-médias das potenciais F_0s oriundas da aplicação do algoritmo MDC aproximado não calcula a média das F_0s do maior conjunto: aqui o processo K-médias é executado iterativamente até que não seja mais possível separar o conjunto de dados (F_0s potenciais) em dois subconjuntos. Isso ocorre pelo fato de que não é possível determinar a priori o número de subconjuntos presentes no conjunto de dados [9].

V. CONCLUSÃO

O método apresentado (agrupamento via K-médias de todos os potenciais F_0 obtidos através do cálculo do MDC das frequências obtidas a partir dos máximos locais da curva de densidade de potência espectral) é capaz de determinar F_0 para sons vocálicos com baixo SNR. Pode inclusive ser empregado quando parte da curva PS foi destruída ou está sob influência da modulação formante, pois é baseado em relações aritméticas entre as frequências harmônicas, que estão espalhadas pela curva PS.

A performance pode ser melhorada eliminando máximos locais que se relacionem com características numéricas da DFT (pois se não forem múltiplos da F_0 real, afetam a estimação) e definindo um método para estimar a priori o

número de subconjuntos existentes no conjunto de F_0s a ser agrupado pelo método K-médias.

REFERÊNCIAS

[1] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 804–816, 2003.

[2] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917–1930, 2002.

[3] T. Sreenivas and P. Rao, "Pitch extraction from corrupted harmonics of the power spectrum," *J. Acoust. Soc. Am.*, vol. 65, pp. 223–228, 1979.

[4] A. Mitre, M. Queiroz, and R. Faria, "Accurate and efficient fundamental frequency determination from precise partial estimates," in 4^o *Congresso de Engenharia de Audio e* 10^a *Convenção Nacional da AES/Brasil*, São Paulo, Brazil, 2006.

[5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 744–754, 1986.

[6] A. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309, 1966.

[7] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, pp. 2067–2079, 2010.

[8] W. Chu and A. Alwan, "Safe: A statistical approach to f0 estimation under clean and noisy conditions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, pp. 933–944, 2012.

[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, USA, 1967, vol. 1, pp. 281–297.