MÉTODO DE K-MÉDIAS PONDERADO APLICADO À ESTIMAÇÃO DA FREQUÊNCIA FUNDAMENTAL EM SINAIS DE FALA

MARCELO DE OLIVEIRA ROSA[1]

1. *LabSom, Depto. 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

Abstract— Here it is proposed an iterative method based on the K-means algorithm to estimate the fundamental frequency of the most predominant signal in short-time duration segments obtained from speech signals. All the used potential F_0s in such a non-supervised classification are calculated from an approximated greater common divisor algorithm applied over all combinations of frequencies of the local maxima of the power spectrum density of a speech segment located in a large spectral band. Considering the harmonic structure redundancy in the segment power spectrum, this strategy produced good results even when the signal-to-noise ratio was -5dB, or when the supraglottal tract amplified low-frequency harmonics of the human voice.

Keywords— Speech processing, pitch estimation, K-means clustering, GCD algorithm, non-supervised classification

Resumo— Propõe-se aqui um método iterativo baseado no algoritmo de K-médias para estimar a frequência fundamental do sinal mais proeminente em segmentos de curta duração extraídos de sinais de fala. Todos os potenciais F_0s usados nessa classificação sem supervisão são calculados através de um algoritmo de máximo divisor comum aproximado aplicado a todas as combinações das frequências dos máximos locais da curva densidade de potência espectral desse segmento em uma banda espectral larga. Considerando a redundância da estrutura harmônica no espectro do segmento, a estratégia produziu bons resultados mesmo quando a relação sinal-ruído foi de -5dB, ou quando o trato supraglotal amplifica harmônicos de baixa frequência da voz humana.

Palavras-chave— Processamento de fala, estimação de pitch, agrupamento por K-médias, algoritmo MDC, classificação não supervisionada.

1 Introdução

A estimação do pitch ou frequência fundamental (F_0) de sinais de fala é uma tarefa complexa devido a dois fatores preponderantes: os tratos glotal (laringe) e supraglotal (boca, lábios, língua, etc.) podem assumir diferentes configurações geométricas que modulam o sinal conhecido por sinal glotal, e a presença de outros sinais que são distintos do sinal de interesse (incluindo ruídos de ambiente) (Klapuri, 2003).

No geral, as análises consistem em dividir o sinal analisado em segmentos de curta duração (com ou sem sobreposição entre tais segmentos). Em seguida a F_0 é calculada para cada segmento para, posteriormente, determinar a curva completa de pitch do sinal. Assim, a precisão da estimação da F_0 inter-segmentos afeta diretamente tal determinação. A curva de pitch pode ser aplicada em análise de prosódia, avaliação de doenças da laringe, alteração de pitch em sinais de fala (*auto-tuning*), entre outras aplicações.

Os métodos de estimação do pitch em um segmento (ou janela) de sinal dividem-se em métodos temporais, espectrais ou híbridos. Um exemplo de método temporal consiste na busca do instante de máxima amplitude da função de autocorrelação do i-ésimo segmento ($r_i(\tau)$), que é o período fundamental do segmento ($T_0 = 1/F_0$). Os efeitos de modulação das estruturas supraglotais podem levar a estimações errôneas como calcular um período fundamental que é a metade do valor correto. Variações desse método (Cheveigne e Kawahara, 2002) lidam com tais problemas.

Os métodos espectrais baseiam-se na análise da curva densidade de potência (PS) dos segmentos da fala. Aqui as frequências harmônicas são identificadas e usadas para definir a F_0 dos segmentos (Sreenivas e Rao, 1979; Mitre *et al*, 2006) de modo análogo à aplicação um filtro do tipo pente (*comb-filtering*) adaptativo. Outra técnica espectral, baseada na transformada cepstral (Noll, 1966), permite isolar a excitação periódica do trato glotal (que é o gerador da periodicidade fundamental que se busca) da influência moduladora do trato supraglotal e facilitar a identificação da F_0.

A partir da estimação de coeficientes de um modelo representativo do trato supraglotal, pode-se eliminar a influência dessa estrutura pela filtragem inversa do sinal que gerou tal modelo. O sinal resultante (dito resíduo) contém informação relevante sobre a periodicidade do sinal, que permite a estimação da frequência fundamental de um segmento (Deller *et al*, 1993).

Analisando tais métodos determinísticos (que empregam algum tipo de *threshold* no processo de detecção do pitch, ou seja *hard decision*) constata-se que os mesmos exigiram a definição de regras específicas (baseadas em experiência) para definir a F_0 de um segmento. Alguns métodos (Hu e Wang, 2010; Chu e Alwan, 2012) contornam a geração dessas regras específicas via técnicas classificatórias (*soft decision*) aplicadas a potenciais candidatos a F_0 do segmento obtidos por técnicas de estimação simples. Com base em conhecimento prévio da F0 para alguns segmentos com diferentes envoltórias formantes (basicamente, diferentes fonemas) e diferentes relações sinal-ruído (fase de treinamento), tais métodos

encontram o valor correto da F_0 do segmento analisado.

Neste trabalho, considera-se que a estrutura harmônica da fala, de modo redundante, espalha-se por uma grande largura de banda na curva de densidade de potência espectral (PS), mesmo em alta frequência, de acordo com a teoria de percepção de pitch de Licklider (Chu e Alwan, 2012), que propôs que o cérebro analisa grupos sequenciais de harmônicos (frequência e amplitude) para estimar o pitch do sinal audível.

A estimação das F_0s candidatas foi realizada por um algoritmo aproximado de máximo divisor comum (MDC) aplicado a todos os pares de frequências dos harmônicos da PS. Tais frequências candidatas são agrupadas via um algoritmo de K-médias ponderado para obter o pitch correto do segmento.

O objetivo do uso combinado de MDC e K-médias é reduzir a complexidade da estimação do pitch em relação às técnicas mais recentes, usar informações contidas apenas no segmento analisado e reduzir o tamanho do segmento, que eleva a resolução temporal da curva de pitch do sinal.

As definições matemáticas desses algoritmos são apresentadas nas seções 2 e 3, respectivamente, com resultados e discussões apresentados na seção 4.

2 Potenciais F_0 via MDC Aproximado

O sinal analisado foi dividido em segmentos de igual tamanho, sendo modulados através da janela de Hamming. A curva de densidade de potência espectral (PS) consistiu na magnitude quadrada da Transformada Discreta de Fourier (DFT) do segmento modulado considerado ($|S[f]|^2$). Para aumentar a resolução da DFT, incluíram-se zeros ao final do segmento em quantidade igual a quatro (4) vezes o tamanho desses segmentos.

Assumindo que as frequências f_p associadas aos máximos locais $|S[f]|^2$ pertencem a componentes senoidais presentes no segmento, tais frequências foram obtidas no intervalo de 40 a 1000Hz a partir das seguintes inequações:

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

na qual Δ_f é a resolução temporal da PS.

Para refinar a estimativa das frequências f_p, aplicou-se uma interpolação parabólica na vizinhança centrada em tais frequências, obtendo-se as frequências $\widetilde{f_p}$. Isso é possível porque a interpolação parabólica se aproxima de uma interpolação espectral devido a inclusão de zeros no cálculo da DFT.

A partir de tais frequências ($\widetilde{f_p}$) determinaram-se as potenciais F_0s para posterior classificação usando um algoritmo de MDC para números reais (análogo ao procedimento de Sreenivas e Rao, 1979): assumindo duas frequências (f_a e f_b), o seu MDC aproximado é obtido pela seguinte recursão:

$$\begin{aligned} amdc(f_a, f_b, err) = f_a, f_b \leq err \\ amdc(f_a, f_b, err) \\ = amdc(f_b, f_a \% f_b, err), c.c. \end{aligned} \quad (2)$$

na qual *err* é o erro de aproximação e o símbolo % refere-se ao resto da divisão inteira. O erro de aproximação corresponde a um critério de parada da recursão ou um *threshold* para este estimador. Definiu-se *err* = 40Hz com base no intervalo de busca usada na Equação 1.

As potenciais F_0s são obtidos a partir da aplicação do algoritmo descrito pela Equação 2 sobre todos os pares de frequências $\widetilde{f_p}$, permutadas 2 a 2, gerando um vetor *s* de frequências a serem classificadas via K-médias.

3 Agrupamento por K-médias

O emprego do algoritmo de agrupamento sem supervisão K-médias (MacQueen, 1967) é ponto central da estimação apresentada aqui. Com baixa complexidade computacional em relação a outros métodos que empregam estratégias classificatórias (Klapuri, 2003; Chu e Alwan, 2012), consiste em determinar a F_0 do sinal mais proeminente no segmento ao invés de se determinar múltiplos F_0s que possam estar presentes no segmento analisado. Isso é feito devido à exigência de conhecimento prévio do número de subconjuntos contidos no conjunto de potenciais F_0s (pertencentes um conjunto chamado *s*) a ser classificado via K-médias.

O pitch do segmento é estimado assumindo que há apenas dois subconjuntos em *s*. Aplica-se então K-médias (com K=2) e o subconjunto de menor tamanho (número de elementos) é descartado. Sobre o subconjunto remanescente, aplica-se K-médias (com K=2) novamente. Tal iteração é repetida até que apenas um único subconjunto seja indicado pelo algoritmo. Assumiu-se assim que o sinal mais significativo do segmento é aquele com maior concentração de potenciais F_0s na vizinhança do pitch correto.

Visualmente a PS de qualquer segmento de fala indica que as magnitudes dos sinais ali presentes influenciam a identificação das F_0s destes sinais. Por exemplo, um sinal composto por 5 harmônicos de uma frequências F_0 e ruído colorido tal que a SNR seja 10 dB produziria um grande número de $\widetilde{f_p}$ devido a presença do ruído. Entretanto, pela proeminência dos harmônicos em relação ao conteúdo não-harmônico do ruído (ou do efeito numérico na DFT pelo próprio enjanelamento), seria evidente que apenas alguns dos potenciais F_0s (obtidos pelo método MDC apresentado) seriam significativos.

Para reforçar tal evidência, definiu-se primeiramente um peso *A[p]* para cada frequência $\widetilde{f_p}$ obtida. Considere P como sendo o número dessas frequências ($1 \leq p \leq P$) e *M[p]* como sendo a magnitude

em dB, na PS, para a frequência $\widetilde{f_p}$. Tal peso é definido por:

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

na qual $\min(M)$ é a menor magnitude (em dB) associada às P frequências $\widetilde{f_p}$.

No cálculo de um potencial F0 exigiram-se duas frequências. Se $f_a, f_b \in \{\widetilde{f_p}\}$, com pesos *A[a]* e *A[b]* obtidos da Equação 3, então o peso associado a tal frequência fundamental é definido pelo produto:

$$W_{p=\{a,b\}} = A[a]A[b] \quad (4)$$

Assim, para cada $\widetilde{f_p}$ tem-se um peso W_p que foi incorporado na definição do centroide dos K grupos do algoritmo K-médias (ao invés de uma média simples, usou-se uma média ponderada para definir esses grupos).

4 Resultados e Discussões

Para avaliar a desempenho da técnica, fixou-se o tamanho dos segmentos em três (3) vezes o período fundamental o sinal analisado. Tal período era conhecido antecipadamente e foi obtido a partir de avaliação manual dos sinais apresentados aqui, que foram cuidadosamente gravados e discretizados. O uso de segmentos com tal largura foi avaliar o comportamento da técnica em condição de elevada resolução temporal das estimativas (a DFT exige a definição de qual é o compromisso entre resolução temporal e espectral).

Os fonemas analisados foram /a/, /e/, /i/, /o/ e /u/. Particularmente o fonema /i/ apresenta-se como um desafio para estimação de frequência fundamental em segmentos de curta duração visto que a primeira frequência formante afeta a banda espectral entre 200 e 400 Hz, ou seja, próximo a potenciais F_0s de sinais de fala. Sua segunda frequência formante afeta a banda espectral entre 2 e 3 kHz, região onde se pode observar harmônicos sendo amplificados.

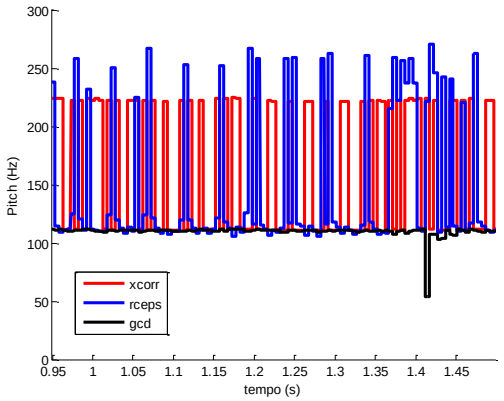

Figura 4. Exemplo de estimação de pitch usando diferentes técnicas para o fonema /i/.

Para exemplificar, a Figura 1 mostra como tal fonema prejudicou estimativas de F_0 usando alguns métodos (em várias situações, o método de autocorrelação indicou incorretamente que a frequência fundamental era o dobro do valor correto).

Cabe ressaltar que o método apresentado não requer qualquer tipo de filtragem. Nos resultados obtidos aqui, apenas para os métodos de autocorrelação (xcorr), por análise cepstrum (rceps) e por resíduos de filtragem LPC (resíduos) pré-filtrou-se os sinais de voz usando um filtro digital passa-banda do tipo Butterworth na banda de 40 a 500 Hz.

Para verificar a robustez do procedimento (MDC+K-médias) adicionaram-se ruídos gaussianos com diferentes SNRs aos sinais de voz apresentados já que suas F_0s eram conhecidas previamente.

Tabela 1. Estimação da frequência fundamental em condições de ruído. kmédias, xcorr, cepstrum e resíduos correspondem aos métodos MDC+K-médias, de autocorrelação, pelo método cepstrum e pelo méodo dos resíduos da filtragem LPC, respectivamente.

Método	Fonema	Percentual de erro (%) para diferentes SNRs			
		20dB	10dB	0dB	-5dB
kmédias	/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
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

O percentual de erro apresentado na Tabela 1 corresponde ao percentual de diferenças acima de 5% entre as F_0s estimadas e a correta, por segmento e ao longo do sinal de voz. Reforça-se que a frequência fundamental correta foi obtida manualmente.

Os resultados reafirmam que o fonema /i/ dificulta a estimação da F0 em todos os métodos. Mais ainda para SNRs elevados, nos quais parte da sua envoltória formante é destruída pela presença do ruído.

O método MDC+K-médias apresentou bons resultados comparando-o em relação a todos os fonemas e todas as SNRs. Os métodos em sua maioria estimavam frequências fundamentais iguais ao dobro

do valor correto (*double pitch*) e em poucos casos iguais à metade (*half pitch*).

Os problemas de estimação do método MDC+K-médias relacionaram-se com o fato de que a estratégia iterativa de classificação dos potenciais F_0s presentes no segmento opera exaustivamente sobre conjuntos com K=2, até que um dos subconjuntos obtidos seja vazio. O outro subconjunto geralmente possuía um único elemento – a F_0 atribuída pelo método para o segmento – ao final da iteração. Se K fosse conhecido, a priori, F_0 seria o centroide de um dos subconjuntos agrupados, e esse valor contabilizaria diversos potenciais vizinhos na estimativa (como uma interpolação).

5 Conclusão

O uso do algoritmo de K-médias para classificar sem supervisão os potenciais F_0s presentes no segmento analisado (que foram obtidas através do cálculo do MDC aproximado entre as frequências dos máximos locais da PS do segmento) é capaz de determinar a F_0 mesmo em sons vocálicos com baixo SNR. Inclusive é capaz de lidar com situações nas quais parte da PS é fortemente influenciada pela modulação provocada pela cavidade supraglotal (caso do fonema /i/). Isso é possível, pois a estratégica classificatória avalia as relações entre frequências harmônicas e não-harmônicas em uma grande largura de banda do segmento analisado.

O desempenho pode ser melhorado se forem eliminadas da análise conjunta MDC+K-means aqueles máximos locais relacionados a características numéricas da DFT relacionadas com a alta resolução temporal usada (pequeno comprimento dos segmentos). A adoção de método para estimar *a priori* o número de subconjuntos existentes no conjunto inicial de potenciais F_0s pode melhorar a resolução da estimativa de F_0 por segmento. Ambas as abordagens serão analisadas em trabalhos futuros.

Agradecimentos

O autor agradece ao CNPq e Fundação Araucária pelo financiamento provido para o desenvolvimento desta pesquisa.

Referências Bibliográficas

Cheveigne, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, Vol. 111, pp. 1917-1930.

Chu, W. and Alwan, A. (2012). SAFE: A statistical approach to F0 estimation under clean and noisy conditions. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, pp. 933-944.

Deller, J. R.; Proakis, J. G. and Hansen, J. H. (1993). Discrete-time processing of speech signals. MacMillan Co. Nova York, EUA.

Hu, G. and Wang, D. (2010). A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, pp. 2067-2079.

Klapuri, A. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. IEEE Transactions on Speech Audio Processing, Vol. 11, pp. 804-816.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, EUA, Vol. 1, pp. 281-297.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 34, pp. 744-754.

Mitre, A.; Queiroz, M. and Faria, R. (2006). Accurate and efficient fundamental frequency determination from precise partial estimates. Anais do 4º Congresso de Engenharia de Audio/10ª Convenção Nacional da AES/Brasil, São Paulo, Brasil.

Noll, A. (1966). Cepstrum pitch determination. The Journal of the Acoustical Society of America, Vol. 41, pp. 293-309.

Sreenivas, T. and Rao, P. (1979). Pitch extraction from corrupted harmonics of the power spectrum. The Journal of the Acoustical Society of America, Vol. 65, pp. 223-228.