論文 85-22-2-3

面積 比較法에 의한 高速 PITCH 抽出

(The High Speed Pitch Extraction of Speech Signals Using the Area Comparison Method)

裵 明 振*, 安 秀 桔*
(Myungjin BAE and Souguil ANN)

要 約

음성신호의 기본주파수를 구하는 새로운 방법인 area comparison method를 제안 하였다. 음성신호는 생성모델에 의해 한 pitch 구간에서 첫 봉우리의 면적이 강조 된다. 이 논문은 이러한 성질을 이용하여 pitch period를 구하는 것으로 기존의 방법에 비해 많은 장점을 갖는다.

시간 영역에서 취급하고 알고리즘이 간단하기 때문에 고속이다. 또한 sample의 개개가 아닌 면적을 적용하므로 impulse성 잡음이나, 전처리 filtering이 필요없게 된다.

Abstract

In this paper, a new pitch extraction method, the area comparison method, is proposed. By the speech production model, the area of the first peak on a pitch interval of speech signals is emphasized.

By using the above characteristics, this method have more advantages than the others for pitch extraction.

The defective decision caused by an impulsive noise is minimized and the pre-filtering is not necessary for this method, because the integration of signals takes place in the process.

Ⅰ. 序 論

基本周波數 抽出에 關한 硏究가 오랫동안 進行되어 왔음에도 불구하고 이에 對한 完全한 方法은 아직 發見되지 않고 있다. 지금까지 硏究된 方法에는 크게 3가지로 나 눌 수 있는데, time domain pitch extraction type, autocorrelation type, spectral analysis가 있다.[1-2,4,10] 그러나 대부분의 基本周波數 抽出 알고리즘들은 많은 計算時間을 要한다.

本 論文에서는 special speed up hardware 없이, micro-computer에 의해 高速이며 正確히 pitch period를 抽出하는 새로운 area comparison method를 提示하였다.

音聲信號의 有聲音 區間에서는 signal이 基本周波數 外에도 많은 高周波를 包含하여 그림 1 처럼 true pitch의 peak에서 다음 Pitch의 Peak가 나타날 동안 peak의 振幅이 time domain에서 漸次的으로 減少되고 있다.[8] 即, zero level 以上의 ware form을 살펴보면 true pitch가 存在하는 area가 存在하지 않는 area보다 크다는 것을 알 수 있다. 따라서 이들 wave form에 對해 positive level 以上의 area들을 서로 比較함으로써 가장 correlation coefficient가 높은 區間을 求하여 音聲信號의 基本周波數를 求한다.

Ⅱ. Voiced Signal Analysis

音聲信號는 音聲源에 따라 voiced, unvoiced, plosive signals으로 區分할 수 있다. voiced signal은 肺에서 올라온 空氣를 glottis를 通하여 排出 시킴으로서 生成

*正會員, *正會員, 서울大學校 電子工學科
(Dept. of Electronics Eng., Seoul National Univ.)
接受日字 : 1984年 6月 1日

되므로 vocal cord 振動을 隨伴한다. 그리고, vocal-tract에서의 共鳴으로 인해 그림1처럼 energy가 크고 準週期的인 形態의 signal이 된다.

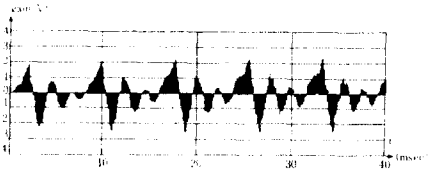

그림 1. 유성음 "아"의 파형
Fig. 1. Weveform of voiced speech "a"

이를 frequency domain에서 살펴보면 그림2와 같이 vocal tract resonance의 envelope에 音聲信號의 基本周波數 F_0가 세세하게 나타나고 있다.[8] vocal-tract resonance의 봉우리에 該當하는 frequency 들을 formants라하고 가장 낮은 주파수를 갖는 봉우리를 first formant, F_1라 한다.[4][8]

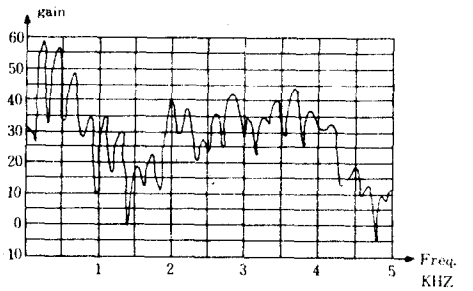

그림 2. 유성음 "이"에 대한 spectrum
Fig. 2. Spectrum for voiced speech "i".

Voiced speech signal에서는 vocal tract의 resonance인 F_1이 다른 formant들 보다 에너지가 약10dB 以上 높다. 때문에 이를 Time domain으로 表現하면 F_1의 影響이 主로 나타나며 한 pitch區間에서 zero crossing interval의 逆數는 F_1의 frequency와 거의 같게된다. 그리고 formant들은 band-width를 갖게 되므로 time domain의 한 pitch區間에서는 감쇠진동을 하게 된다.[8]

F_1이 주파수 영역에서 다른 formant들보다 훨씬 높은 에너지 봉우리를 갖기 때문에 F_1만을 고려하여 근사적인 방법으로 vocal tract를 분석할 수 있다. 그림2에서 처럼 F_1의 magnitude가 band-width 내에서 cosine봉우리를 갖는다고 하면 이에 의한 시간영역에서의 파형은 그림2를 Inverse Fourier Transform 하면 된다. (여기서 위상특성은 zero라 가정한다)

$$h(t)=\int_{-\infty}^{\infty}F(f)e^{j2\pi ft}\,df$$
$$=\int_{-\frac{BW}{2}}^{\frac{BW}{2}}\cos\left(\frac{2\pi f}{2B_w}\right)e^{j2\pi ft}df\cdot 2\cos\left((2\pi F_1 t)-\frac{\pi}{2}\right)$$
$$=\frac{4B_w}{\pi-4\pi B_w^2t^2}\cos(\pi B_w t)\cos\left((2\pi F_1 t)-\frac{\pi}{2}\right)\quad(1)$$

여기서 F_1=first formant frequency 및 $B_w=F_1$ 이 갖는 대역폭이다.

(1)식을 살펴보면 마지막 두 因子가 시간 영역에서 oscillation을 결정하는데, 여기서 $F_1\gg B_w$라면 oscillation은 F_1에만 의존하게 된다. 또한 (1)식의 첫 항은 감쇠인자로 작용하는데, slope는 B_w에 관계됨을 알 수 있다. (1)식에서, F_1=500Hz, B_w=50Hz, 및 표본주기를 0.1mSec의 간격으로 한pitch區間을 6m-sec 동안으로한 waveform을 그림4에 表示하였다.

한편 voiced signal의 glottal wave를 發生하는 慣例的인 方法을 그림4에 보였다.[4]

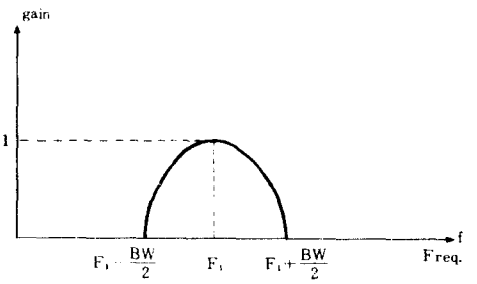

그림 3. 주파수 영역에서 first formant 근사분석
Fig. 3. First formant approximation in frequency domain.

Impulse train generator는 pitch period로 할당된 unit impulse의 sequence를 發生한다. 다음에 이 signal은 impulse response가 glottal wave shape인g(n)을 excite한다. g(n)의 形態는 단적으로 特徵지을 수 없지만, rosenberg에 의해 合成 pulse waveform形態로 提示되었다.[3]

$$g(n)=\frac{1}{2}\{1-\cos(\pi n/N_1)\},\quad 0\leqq n\leqq N_1$$
$$=\cos\{\pi(n-N_1)/2N_2\},\quad N_1\leqq n\leqq N_1+N_2$$
$$=0,\ \text{otherwise}\qquad(2)$$

(2)式에서 N_1=13, N_2=4 및 n=0.1m sec(표본주기)로 하였을때 waveform을 그림4에 表現 하였다.

g(n)이 finite length이므로 All-pole model이 바람직하게 되며, G(Z)=z[g(n)]에 대해 two-pole model로 普通 modeling하고 있다. 그리고 Lip radiation의 效果는 $R(z)=R_0(1-z^{-1})$로 나타낼 수 있으며, 이는

High pass filter로 動作하여 vocaltract의 Main resonance를 強調 시키고 있다.[4][8]

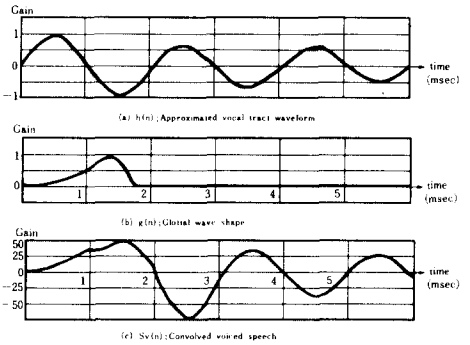

그림 4. 유성음의 근사분석
Fig. 4. Approximation analysis for voiced speech.

結局 Voiced speech signal $S_v(n)$은 (1)式과 (2)式이 Time domain에서 Convolution으로 나타난다.

$$S_v(n) \fallingdotseq h(n) * g(n) \qquad (3)$$

(3)式은 그림4(a)와 4(b)를 convolve한 것으로 그림4(c)로 나타난다. 이로서 한 pitch 區間에서 처음 positive peak가 다른 peak들 보다 強調 되고 있음을 알 수 있다.

Ⅲ. Area Comparison Method

(3)式과 같은 voiced speech signal, $S_v(n)$은 한 pitch 區間에서 감쇠진동을 하는 形態로 주어진다. 이러한 性質을 利用하여 pitch period를 extraction 하려는 硏究가 시도되었으며 그中에 하나가 "parallel processing method로 發表 되었다.[1] 그러나 強調된 peak와 다른 peak들의 差異가 顯著히 두드러지게 나타나지 않음으로서 true peak를 더 強調하기 위해 pitch period를 抽出하기 前에 signal을 자승하거나[5] 或은 vocaltract의 影響을 除去시키는 simple inverse filtering, tracking SIFT方法이 提示되었다.[2] 그렇지만 이렇게 함으로서 computation size가 增加되어, 이를 real time化 시키기에는 아직도 speed up hardware의 도움이 要求된다. 따라서 本 論文에서는 voiced speech signal이 한 dpitch period 區間에서 true peak가 強化됨은 勿論 이러한 peak의 zero crossing interval도 다른 peak들 보다 크다는 點을 着眼하였다. 이러한 두개의 變數를 하나의 variable로 代置를 시키면 이는 面積이 된다. sampled data에서 주어진 區間의 area는 sampled data들의 summation으로 表現되며, 이로서 speed도 增大된다.

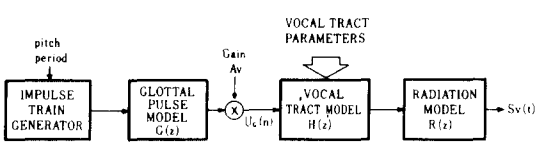

그림 5. 유성음의 생성모델
Fig. 5. Speech production model for voiced signals.

먼저 可能한 true pitch peak들에 對한 zero crossing interval사이의 area를 求하고 이들 area에 對해 peak picking algorithm을 適用 함으로써 pitch period를 求한다. 이러한 processing을 area comparison method라 이름 붙이겠다. 比較하고자 하는 봉우리들이 갖는 area 값이 位置하는 곳은 다음과 같이 表示된다.

$$A(n_2) = \sum_{i=n1}^{n2} S_v(i) \qquad (4)$$

[[여기서 sampled data의 한zero crossing interval $= (n_1, n_2)$]

이 方法은 voiced signal을 integration하는 效果를 얻을 수 있으므로 pitch period extraction 時에 impulse성 noise로 부터의 攪亂이 最少화 될 수 있다. 더욱이 이는 speech sample에 對해 low pass filter, LPF를 遂行하는 것과 같으므로서 pitch period를 extraction하기 위해 別途의 LPF와 smoothing을 遂行하므로서 나타나는 計算時間과 正確度를 save 하게 된다. 또한 zero crossing사이의 area는 zero crossing rate, ZCR에 反比例 하므로 (4)式에서 求한 area는 ZCR 情報까지 包含하게 된다. 따라서 speech signal의 全區間에 對해 이들 area를 求하여 配置하면, ZCR이 높고 energy가 작은 unvoiced signal에서 求한 area값들은 voiced signal에서 求한 area들 보다는 상당히 작게 된다. 即, voiced와 unvoiced signal의 ZCR의 比를 1 : 5, 및 peak의 比를 5 : 1이라

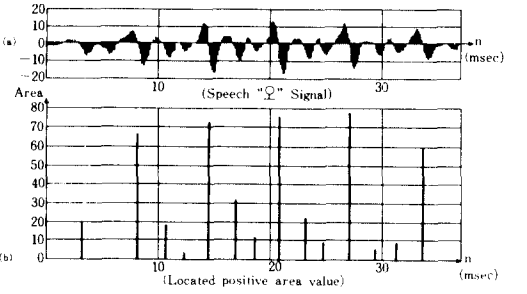

그림 6. 유성음 "오"에 대한 positive area값 배치
Fig. 6. Located positive area for speech "o".

하면 area의 比는 25 : 1 程度로 큰 差異가 나타난다. 從前의 方式은 ZCR과 magnitude들을 使用하여 voiced-unvoiced, V-UV decision을 遂行한 後에 voiced signal에 對해서만 pitch period extraction을 遂行하였다. 그러나 area들은 ZCR 의 逆數와 Magnitucle의 곱에 比例함으로서 voiced와 unvoiced 사이의 area 差異가 두드러지게 나타나기 때문에, V-UV decision 의 threshold level을 잡기가 容易해 진다. 따라서 別途의 V-UV decision을 事前에 遂行할 必要가 없게 된다.

Ⅳ. Decision Algorithm

Glottal wave shape에 의해 한 pitch 區間에서는 처음 peak가 zero crossing interval이 길어지고, magnitude가 強調된다.[3] 이러한 glottal wave는 (2)式과 같이 positive value만 갖게 됨으로서, 強調되는 true peak는 speech wave form의 positive level이 된다. 따라서 area comparison method는 positive arae들에 對해 適用해야 한다.

그러나 사람의 귀는 位相變化를 거의 분간 할 수 없게 되어있다. 따라서 一部 audio recorder는 recording에서 比해 play-back때의 位相이 180° 變化되어 販賣되고 있다. 따라서 recorder에 貯藏된 음성신호 데이타를 利用하여 area comparison method를 適用하기 前에, 우선 speech signal에서 positive level을 判定할 필요가 있다.

入力으로 들어온 speech signal에 對해 glottal wave 特性이 나타나는 positive level을 decision 하려면 (4)式에 의한 方法으로 各 area들을 求한다. 이들 area 中 가장 두드러지게 큰 값이 나타나면 이것의 zero crossing interval을 求한다. 다음 옆에 붙어있는 negative value들의 zero crossing interval을 각각 求한 다음 이들을 positive value와 比較한다. 이때 positive area에 對한 interval이 더 크면 現在의 positive value에 對해서 適用하고, 두 negative value 보다 interval이 작다면 glottal wave는 negative level쪽으로 나타난다. 이 경우는 negative area에 對해 Area comparison method를 適用해야 한다. Voiced signal "오"에 對해 (4)式을 適用하여 area를 配置한 것을 그림 5 에 表示했다.

다음에 이들 Area들에 對해 true pitch를 求해야 한다. true area가 隣近해 있는 area들(한 pitch interval에서)보다는 두드러지게 差異가 남으로서 true area(true pitch)를 決定하는 decision logic도 簡單하다.

Un-voiced signal의 影響을 除去하기 위해 바로 前에 찾아진 true area value값의 $\frac{1}{4}$以上이 되는 값들에 對해서만 true pitch의 有, 無를 檢討한다. true pitch의 유, 무 판정은 방금 전에 찾아진 true pitch area의 $\frac{1}{2}$이상인 값으로 정하였다.

연속음 "인수다"에 대해 area comparison method로 決定한 pitch period의 變化와 파형을 보면서 눈으로 찾아낸 結果를 그림 6 에 提示하였다.

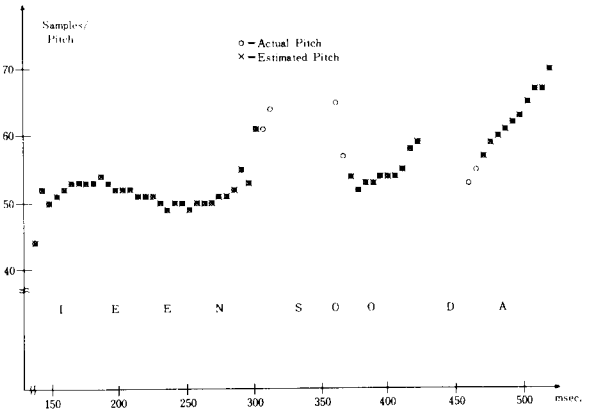

그림 7. 실제의 pitch와 검출된 pitch의 비교 (연속음 "인수다")

Fig. 7. Comparison between actual and estimated pitch period for speech "leen soo da".

Ⅴ. 結 論

Voiced speech signals을 time domain에서 살펴보면 한 pitch 區間에서 zero-crossing interval은 거의 vocal-tract의 main resonance에 關系되고, decay ratio는 band width에 의한다.[4] 그리고 glottal wave shape가 convolve되어 처음 positive peak部分이 強調 된다.[3]

本 論文에서는 이러한 性質을 利用하여 area comparison method를 提安하였다. 이렇게 하여 pitch period analysis를 遂行한 結果는 다음과 같다.

첫째, 이 方法은 area를 求하는데 addition만 遂行하므로서 計算 speed가 크다. 이로서 micro-processor에 의해서도 real time化 할 수 있다.

둘째, area ualue가 ZCR의 逆數와 magnitude 의 곱에 比例하므로서 unnvoiced signal과 voiced signal에서 두드러진 差異를 보이므로 pitch를 extraction 하기 前에 voiced-unvoiced, V-UV decision을 別途로 할 必要가 없다.

셋째, 이 方法은 適用하는 過程에서 Integration을 遂行하는 것이므로 impulse 성 noise로 부터의 攪亂을

補償할 수 있다. 또한 이것은 low pass filter LPF를 의미하므로서 pre-processing filtering을 별도로 할 必要가 없다.

넷째, true pitch area가 다른隣近 area 보다 상당한 差異를 나타냄으로써 desision logic도 簡單해 진다.

參 考 文 獻

[1] B. Gold and L.R. Rabiner, "Parallel processing technique for estimating pitch periods of speech in the time domain," *J. Acoustic SOC. Am.*, vol. 46, no. 2, Pt.2, pp. 442-448, August, 1969.

[2] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE trans, on audio and electroacostics,* vol. Au-20, no. 5, pp. 367-377, December 1972.

[3] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am,* vol. 49, pp. 583-590, 1971.

[4] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals.* Prentica-Hall, Inc. 1978.

[5] Myung-Jin BAE, *A Study on the Fundamental Frequency Extraction of Speech Signals Using Second Order Rundown Method.* Seoul National University, Graduate Thesis, Jan. 1983.

[6] Souguil ANN, "Linear prediction of speech," *Seoul National University Engineering Report* vol. 12, no. 2, Oct. 1980.

[7] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust. Speech and Signal Proc., vol.* ASSP-22, pp. 535-362, Oct. 1974.

[8] J.D. Markel and A.H. Gray, *Linear Prediction of Speech.* Springer-Verlag, Berlin Heidelberg, New York, 1980.

[9] L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech and Signal Proc.,* vol. ASSP-25, no. 1, pp. 24-33, Feb. 1977.

[10] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. Mc Gonegal, "A Comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, and Signal Proc.,* vol. ASSP-24, no. 5, pp. 399-418, Oct. 1976.

[11] C.K. Un and S.C. Yang, "A Pitch Extraction Algorithm Based on LPC Inverse Filtering and AMDF," *IEEE Trans. on ASSP,* vol. ASSP-25, no. 6, pp. 565-572, Nov. 1977.