論文 87-24-3-3

음성발생 모델로 부터의 G-peak를 이용한 음성에너지 추출에 관한 연구

(A Study on the Energy Extraction Using G-peak from the Speech Production Model)

裵 明 振*, 林 材 烈**, 安 秀 桔*

(Myungjin Bae, Jaeyeol Rheem and Souguil ANN)

要 約

음성 신호의 생성 모델에 의하면 유성음의 기본주파수 한 구간에서 처음의 positive peak가 에너지원을 대변하는 glottis와 첫째 포오만트의 영향을 주로 받게된다. 이러한 성질을 이용하여 한 기본주파수 구간에서 처음의 positive peak의 면적을 에너지 파라미터로 사용하는 방법을 연구하였다.

이 방법을 사용하면 에너지 파라미터가 window 크기에 무관해질 뿐만 아니라 기본주파수의 추출도 동시에 할 수 있어, 기본주파수 단위로의 에너지를 구할 수 있다.

Abstract

By the speech production model, the first positive peak in a pitch interval of the voiced speech is mainly affected by the glottis and the first formant component, known as a typical energy source of the voiced speech. From these characteristics, the energy parameter can be replaced by the area of the positive peak in a pitch interval, which parameter is generally used for classification of speech signals. In this method, the changed energy parameter is independent of window length applied for analysis, and the pitch can be extracted simultaneously. Furthermore, the energy can be extracted in the pitch period unit.

I. 序 論

連續音 認識過程에서 본격적인 認識過程을 簡單하게 處理하기 위하여 音聲信號를 事前에 分類하는 分類認識 過程은 전체 인식률을 좌우할 뿐만 아니라 단어 受容을 쉽게 해준다. 分類認識에서는 분류인식률이 높은 것보다는 音聲의 基本 性質을 잘 나타내는 파라미터를 利用하여 認識過程에 앞서 簡單하고도 빠른 處理가 要求된다.

時間領域에서 音聲信號를 有聲音과 無聲音으로 區分(V-UV decision)하는 分類過程은 에너지와 ZCR (zero crossing rate), 자기상관함수 등을 利用하여 入力信號를 音聲信號와 無音으로 區分하며, 音聲信號에 대해서는 다시 有聲音과 無聲音으로 分類하는 過程으로 이루어진다. 따라서 音聲信號를 分類해 내는데는 에너지가 중요한 파라미터로 作用하며, 音聲信號에서 에너지 파라미터를 抽出하기 위해서는 短時間 에너지(short-time energy)나 平均振幅(average magnitude)과 같은 方法이 有用하게 使用되어 왔다.[1] 그런데 이들 方法은 音聲信號를 自乘하거나 絶對値를 取하여 平均한 값들로써 音聲信號의 平均的 에너지를 잘나타내나 window 크기에 따라 敏感하게 변화한다는 사실이 잘 알려져 있다.[1]

한편 音聲信號를 음성신호의 生成모델[1]측면에서 考

**準會員, *正會員, 서울大學校 電子工學科
(Dept. of Elec. Eng., Seoul Nat'l Univ.)

接受日字 : 1986年 8月 23日

慮하면 그림 1에서 처럼 無聲音의 경우에는 random noise generator가 그 生成源이므로 주기성은 나타나지 않지만, 주로 3KHz 근방에서 共振 봉우리를 갖기 때문에 有聲音(voiced speech)에 비해서 平均 zero crossing rate가 크다.[2] 有聲音은 glottal pulse가 그 生成源이며 성대(vocal cord)의 振動에 따른 聲道(vocal tract)의 영향이 強調되어 나타나 일반적으로 振幅이 크고 準周期的인 性質(pitch)을 갖는다. 따라서 有聲音의 에너지원은 glottal 成分이라고 볼 수 있다.

本 論文에서는 有聲音의 生成모델에 根據하여 有聲音의 한 pitch 區間內에서 처음의 peak가 glottal 特徵을 代表하는 새로운 에너지 파라미터로 適用될 수 있음을 보이고, window 크기에 無關한 에너지 파라미터를 求하였다.

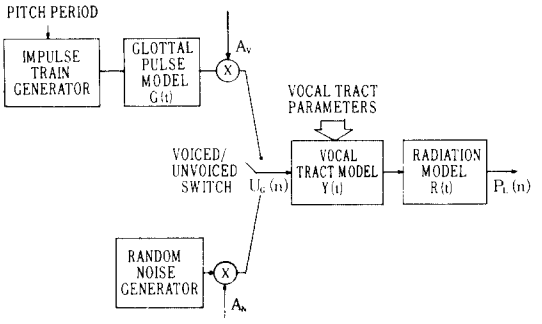


그림 1. 음성신호의 생성모델[1]
Fig. 1. Speech Production Model.[1]

Ⅱ. 에너지 파라미터

既存의 에너지 파라미터 抽出 方法인 短時間 에너지나 平均 振幅은 音聲信號의 特徵이 時間에 대하여 比較的 천천히 변화한다는 假定下에 無聲音의 진폭이 有聲音의 진폭보다 훨씬 작다는 性質을 利用한 것이다.[1]

短時間 에너지는 式(1)과 같이 定義되며, 유성음과 무성음의 진폭 변화를 잘 나타내 준다.

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \qquad (1)$$

여기서 W(n)은 window 함수를 意味한다.

短時間 平均 振幅은 식(2)로 定義되며, 이것은 短時間 에너지가 振幅이 아주 큰 신호에 敏感하다는 사실을 考慮한 것이다.[1]

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|W(n-m) \qquad (2)$$

위의 두 方法의 문제점은 短時間 window 함수의 길이가 音聲信號의 급격한 振幅의 變化를 수용할 수 있도록 충분히 작으면서, 원만한 에너지 곡선을 얻기 위해서 충분히 커야 한다는 것이다. 그림2에 rectangular window를 使用하여 window 크기에 따른 平均 振幅함수의 변화를 보였다.

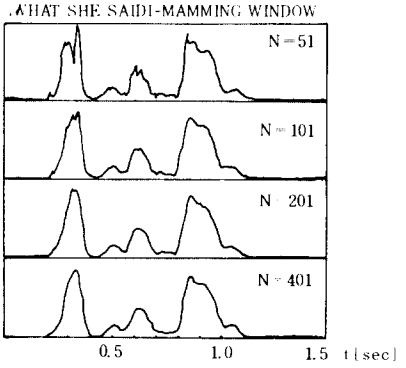


그림 2. Rectangular window length에 따른 平均振幅함수[1]
Fig. 2. Magnitude Function for Rectangular Window with Various Length[1].

그림 2에서 보면 window의 크기 N이 충분히 작으면 音聲信號의 振幅의 변화에 敏感하여 날카로운 平均振幅함수를 얻으며, N이 충분히 크면 평균하는(averaging) 효과가 나타나 완만한 振幅함수가 됨을 보여 준다.

따라서 短時間 에너지나 平均振幅함수가 分類認識의 에너지를 잘 代表한다고 하나 window 크기 設定의 問題를 가지고 있다. 특히 window 크기 N이 1개의 기본주파수 구간에 해당될 때 가장 바람직하다[1]는 것이 알려져 있으므로 바람직한 에너지 파라미터를 구하기 위해서는 pitch (기본주파수)에 따라 적응하는 window 크기를 갖는 window 함수가 必要하게 된다.

Ⅲ. G-peak의 定義 및 考察

일반적으로 有聲音은 성대의 진동과 성도의 共鳴 現象 때문에 時間領域(time-domain)에서 振幅이 크고 準周期的 性質을 갖는다. 이것을 주파수영역(frequency-domain)에서 살펴보면 그림 3에서 처럼 성도의 共振周波數(formant frequency)의 포락선(envelope) 위에 有聲音의 基本周波數가 겹쳐서 나타난다. 그리고 제1 포오만트 周波數F_1의 이득이 다른 포오만트보다 약 10dB이상 높으므로 F_1만의 포락선을 가지고 성대를 近似할 수 있다.[2,5]

그림 4 에서와 같이 F1의 포락선이 대역폭(band width) 내에서 cosine의 포락선을 갖는다고 하면 이것

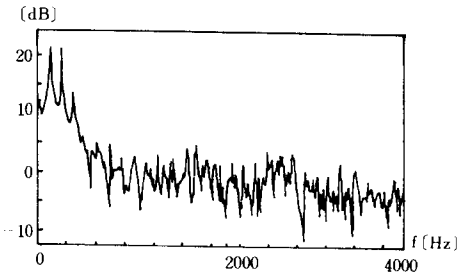


그림 3. "애" 모음의 스펙트럼
Fig. 3. Spectrum of Vowel /ε/.

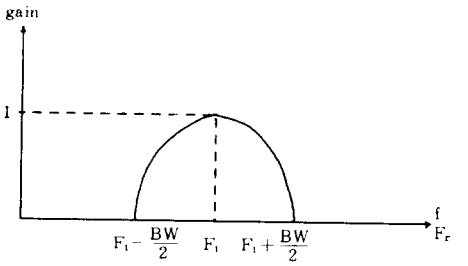


그림 4. 주파수영역에서 제 1 포오만트의 근사분석[5]
Fig. 4. First Formant Approximation in Frequency Domain.[5]

에 대한 時間領域에서의 波形은 이것을 inverse fourier transform하면 얻을 수 있다. (여기서 위상특성은 zero라고 假定한다.)[5]

$$h(t) = \int_{-\infty}^{\infty} F(f) e^{j2\pi ft} df$$

$$= \int_{-\frac{B_W}{2}}^{\frac{B_W}{2}} \cos\left(\frac{2\pi f}{2B_W}\right) e^{j2\pi ft} df \cdot 2\cos\{(2\pi F_1 t) - \frac{\pi}{2}\}$$

$$= \frac{4B_W}{\pi - 4\pi B_W^2 t^2} \cos(\pi B_W t) \cdot \cos(2\pi F_1 t - \frac{\pi}{2}) \quad (3)$$

여기에 glottal pulse 모양은 Rosenberg에 의하여 合成된 波形을 適用할 수 있다.[1,4]

$$g(n) = \frac{1}{2}\{1 - \cos(\pi n/N_1)\}, \quad 0 \le n \le N_1 \quad (4)$$

$$= \cos\{\pi(n - N_1)/2N_2\}, \quad N_1 \le n \le N_1 + N_2$$

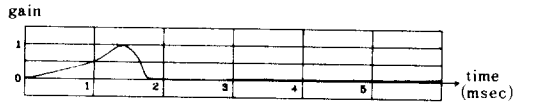
$$= 0 \quad , \text{otherwise}$$

따라서 音聲信號 S(n)은 (3)式과 (4)式의 時間領域에서의 convolution으로 近似될 수 있다.[1,2]

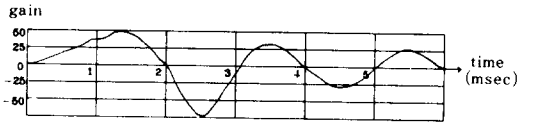
$$S(n) \fallingdotseq h(n) * g(n) \quad (5)$$

그림 5에 式(3), (4), (5)의 파형을 보였으며 특히 그림 5의 (c)에서 보면 有聲音의 한 기본주파수 區間에서 처음의 positive peak가 強調됨을 알 수 있다.[5] 이것은 제 1 포오만트 F_1이 대역폭을 가지고 있어서 減

(a) h(n) ; Approximated vocal tract waveform



(b) g(n) ; Glottal wave shape



(c) $S_V(n)$; Convolved voiced speech

그림 5. 유성음의 근사분석[5]
Fig. 5. Approximation Analysis for Voiced Speech.[5]

衰振動을 하고 glottal pulse가 그림 5의 (b)처럼 positive 쪽으로 치우쳐 있기 때문이다.[2] 따라서 그림 5의 (c)에서 처음의 peak가 glottal 成分과 F_1의 영향을 지배적으로 받는다 할 수 있다. 여기서 처음의 peak를 G-peak라고 定義한다. 이것은 glottal 成分이 지배적인 peak라는 意味이다.

한편 連續音聲信號에 대한 長時間(예를들면 1분이상) 전력밀도스펙트럼(power density spectrum)을 살펴보면 그림 6과 같이 500Hz 근처에서 peak가 나타난다.[1,3] 즉 音聲信號의 에너지는 주로 700Hz 이하에 몰려 있음을 알 수 있다. 이것은 성도의 제 1 포오만트 주파수 범위와 비슷하므로 F_1만의 포락선에 의한 성도의 近似가 妥當함을 보여준다. 그리고 일반적으로 기본주파수 F_0가 50~500Hz 사이에 나타나므로 이것을 包含하는 G-peak의 特性이 有聲音의 에너지 源이됨을 알 수 있다. 따라서 G-peak는 한 pitch區間에서 音聲

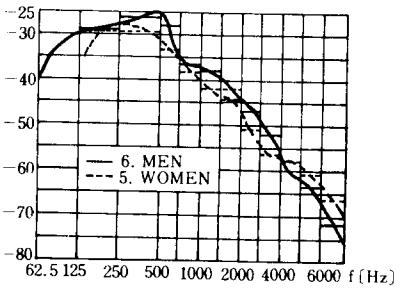


그림 6. 長時間 전력밀도스펙트럼[1]
Fig. 6. Long Time Power Density Spectrum[1]

信號의 glottal 成分과 F_1의 영향이 지배적으로 나타나는 peak임을 알 수 있어, G-peak의 변화는 음성신호에서의 glottal 成分의 변화와 F_1의 변화를 동시에 나타내게 된다.

또한 G-peak가 式(5)와 같이 glottal pulse의 波形과 F_1의 포락선이 서로 convolution된 信號의 첫 peak이기 때문에 G-peak의 밑변(interval)은 glottal pulse의 밑변보다는 길게되고 F_1의 포락선이 대역폭을 가지고 있어서 감쇄진동을 하기 때문에 첫 peak가 다른 peak들에 비해서 상대적으로 振幅이 크게 된다. 결국 G-peak의 면적은 한 pitch區間 안에서 平均振幅을 代表할 수 있게 된다. 따라서 分類認識에 있어서 G-peak의 면적을 考慮함으로써 分類認識에 使用된 平均振幅 함수를 대신할 수 있게 된다. 특히 G-peak의 면적은 한 기본주파수(pitch) 區間에서 定義되기 때문에, G-peak의 면적을 求하는 것은 결국 pitch를 구하는 것과 같다.

그림 7 에 있는 유성음 '오'에 대한 positive 면적 값들을 살펴보면 G-peak에 해당하는 면적이 한 pitch 구간에서 다른 positive side peak 면적보다 크다는 것을 알 수 있다. 따라서 주변의 면적 값들에 peaking 알고리즘을 적용하여 G-peak의 면적을 구할 수 있으며, 이렇게 얻은 G-peak 사이의 간격이 곧 기본주파수(pitch)의 간격(interval)이 된다.[5] 결국 G-peak 의 면적을 이용하여 平均振幅함수를 구하면서 pitch도 구할 수 있게 된다.

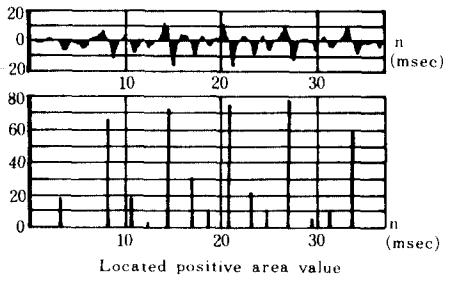


그림 7. 유성음'오'에 대한 positive 면적값 배치[5]
Fig. 7. Located Positive area for Speech /o/.[5]

式(2)에 의하여 平均振幅을 求할 때 信號 X(n)이 電壓 V_d만큼 DC-offset이 걸려 있다면, 平均振幅은 DC-bias 되어 나타나 이를 補正해 주어야 할 필요가 있다. 그러나 G-peak는 波形의 positive peak에 해당되므로 G-peak를 求할 때 예상되는 offset 전압이상에 대해서만 계산하면 이 問題를 解決할 수 있다. 즉 G-peak의 면적 $A(n_2)$는,

$$A(n_2)=\sum_{i=n1}^{n2}\{X(i)-V_d\} \qquad (6)$$

이 된다. 여기서 음성 표본치 X(i)는 $n_1\le i\le n_2$ 區間에서 offset 전압 V_d이상의 positive peak가 나타나는 경우이다(즉 $\{X(i)-V_d\}\ge 0$, $n_1\le i\le n_2$).

Ⅳ. 實驗 및 分析

G-peak의 變化를 관찰하기 위하여 12bit A/D converter를 利用하여 8KHz 표본화(sampling)를 했으며, 모든 데이타의 處理는 16-bit personal computer인 IBM-PC/XT를 使用하였다.

實驗의 目的은 短時間 平均振幅과 G-peak의 面積을 比較하여 G-peak의 面積이 에너지 파라미터로서 有效하게 使用될 수 있는지를 고찰하는데 있으며 부수적으로는 G-peak의 면적의 변화로 부터 音聲信號에서의 특별한 變化를 檢出할 수 있는가에 있다.

사용된 音聲信號 데이타는 3초간의 연속음과 1초간의 숫자음들이며, peaking 알고리즘은 ACM(area-comparison method)[5]을 수정하여 적용했다. 그 결과로써 그림 8 에 "서울대 전자공학과 음성신호 처리연구팀이다."라는 음성신호 데이타의 結果를 보였다. 그림 8의 (a), (b), (c)는 rectangular window 크기 N에 따른 平均振幅의 변화를 나타낸 것이며, (d)는 G-peak의 面積을 利用하고 pitch 간격에 맞추어 구한 것이다. 실험결과 (a)와 (d)는 거의 一致함을 알 수 있으며, 이것은 G-peak 面積의 간격으로서 구한 평균 pitch가 N=48이라는 사실을 고려할 때 G-peak의 面積을 利用하여 window 크기에 無關한 에너지 파라미터를 구할 수 있음을 意味한다. 특히 (a)와 (d)를比較하면, 곡선이 완만한 부근에서는 (d)가 더욱 완만하며 급격한 변화가 있는 곳에서도 역시 (d)가 더 날카로움을 관찰할 수 있다. 그림 9 에는 "예수님께서 천지 창조의 교훈을 말씀하셨다."라는 음성 신호에 대한 결과이다. 그림 9 의 화자는 평균 pitch가 N=47(47.3)이며 그림 8 의 화자와는 다른 사람이다. 그림10, 그림11에는 각각 숫자음 '3', '7'에 대한 실험결과이다.

위의 실험 데이타를 살펴보면, 무성파열음 /ㅌ/과무성파찰음 /ㅊ/과 같이 격하게 발음하는 곳에서 G-peak의 변화가 민감하고 독특함을 알 수 있다. 이것은 /ㅌ/, /ㅊ/이 단독으로는 발음되지 않고 /트/,/츠/와 같이 순간적으로 발음되어져 각 단어의 초기에 상대적으로 많은 에너지가 실리게 되기 때문이다. 그런데 window함수를 이런 무성음에서 유성음으로의 전이구간에 적용할 경우 averaging 효과에 의하여 두드러지지 않으나, G-peak를 이용하면 pitch의 전이에 따른 효과가 나타나기 때문이다. 또 그림 9 의 '께서'의 /ㅅ/

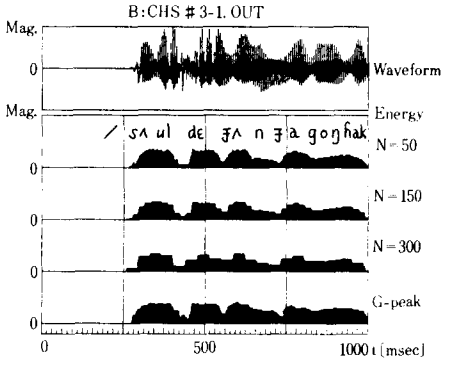


그림8(a). 실험data 1 "서울대 전자공학"
Fig. 8 (a). Experiment data 1 /sʌuldɛ ʤʌn ʤa goŋhak/.

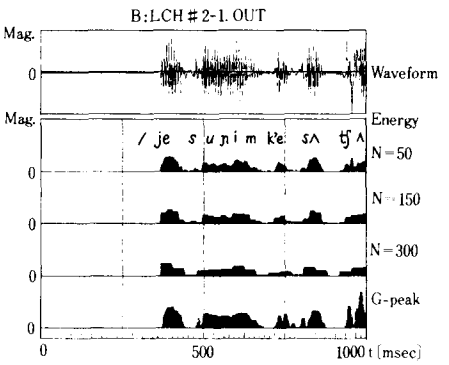


그림9(a). 실험data 2 "예수님께서처-"
Fig. 9 (a). Experiment data 2 /je su nim kesʌtʃʌ/.

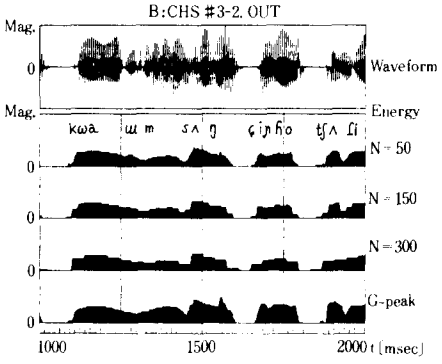


그림8(b). "과 음성 신호처리"
Fig. 8 (b). /kwa ɯm sʌŋ ɛiŋhо tʃʌ ʃi/.

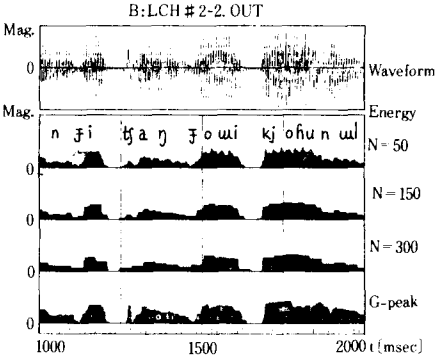


그림9(b). "-ㄴ지 창조의 교훈을"
Fig. 9 (b). /-n ʤi tʃaŋ ʤowi kjo hun wl/.

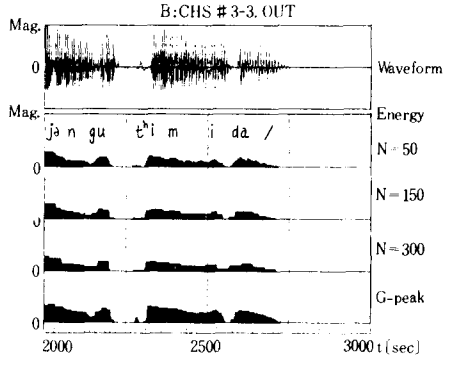


그림8(c). "연구 팀이다."
Fig. 8 (b). /jʌŋgu tʰim ida/.

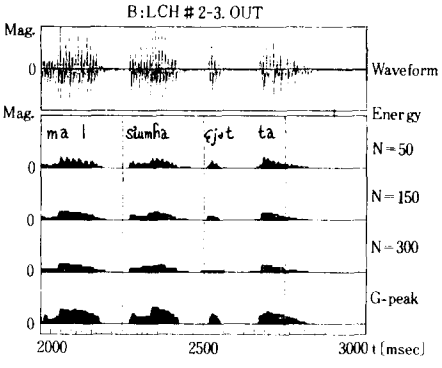


그림9(c). "말씀하셨다."
Fig. 9 (c). /mal s'ɯmha ɛjɔtta/.

과 그림10의 '삼'의 /ㅅ/발음의 독특함은 /ㅅ/ 단독으로 발음되지 않고 순간적으로 /스/와 같이 발음 되어져 /으/ 모음과의 조음현상에 의한 것이라고 해석할 수 있다. 실험 data에서 발음표기는 IPA(international phonetic association) 방식을 따랐음을 밝혀둔다.[12]

V. 結 論

音聲信號의 生成 모델에 根據하여 有聲音을 時間領域에서 살펴보면 한 pitch 區間中 처음의 positive peak가 glottal 性分과 성도(vocal tract)의 제1 포오만트

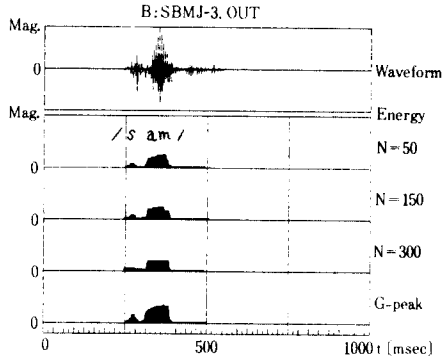


그림10. 실험 data3 숫자음"3"
Fig. 10. Experiment data3 digit/sam/.

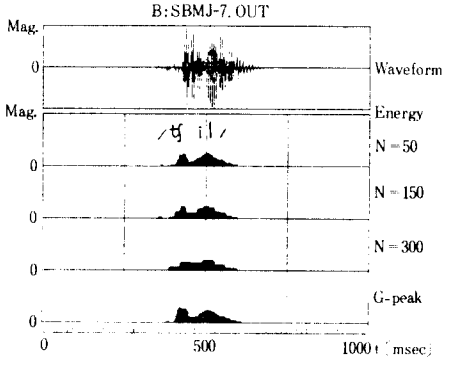


그림11. 실험data4 숫자음 "7"
Fig. 11. Experiment data4 digit/ʧil/.

(F_1)의 영향을 지배적으로 받는다. 이것은 한 pitch 區間에서 zero-crossing 간격은 거의 성도의 main resonance에 관계되고, 감쇠율은 F_1의 대역폭에 의한다[2,5]는 사실을 이용한 것이다. 그리고 여기에 glottal 波形이 convolution 되어 처음의 positive peak 部分이 強調되기 때문이다.[4] 이 처음의 peak를 G-peak라고 定義했다.

本 論文에서는 G-peak의 面積을 고려하여 분류인식의 에너지 파라미터를 고찰했으며, 그 결과 다음과 같은 結論을 얻었다.

첫째, G-peak는 有聲音의 한 pitch 區間에서 定義되어 window 크기에 무관한 에너지 파라미터를 구할 수 있다.

둘째, G-peak의 面積을 찾는 과정이 곧 pitch 추출 과정과 같으므로 G-peak를 이용하여 에너지 파라미터를 구하면서 pitch도 동시에 구하여진다.

세째, G-peak 계산이 덧셈과 선택 로직에 의해서 구하여지기 때문에 범용 μ-computer에서도 실시간 처리가 가능해진다.

參 考 文 獻

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

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

[3] H.K. Dunn and S.D. White, "Statistical measurements on conversation speech," *J. Acoust. Soc. Am.*, vol. 11, pp. 278-288, January 1940.

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

[5] Myungjin BAE and Souguil ANN, "The high speed pitch extraction of speech sinals using the area comparison method", KIEE, vol. 22, no. 2, pp. 101-105, Feb. 1985.

[6] Myungjin BAE and Souguil ANN, "A study on the fundamental frequency extraction of speech signals using second order rundown method", Seoul National university, *MA Paper*, Jan. 1983.

[7] Myungjin BAE and Souguil ANN, "The voiced-unvoiced-silence classification by emphasized spectrum of speech signals", *JASK*, vol. 4, no. 1, pp. 9-15, June 1985.

[8] Myungjin BAE and souguil ANN, "Low Pass filtering on the high speed pitch extraction", *JASK*, to be published, 1987.

[9] Myungjin BAE and Souguil ANN, "Inverse rate type filtering for the pitch extraction", *JASK*, vol. 5, no. 3, Sept. 1986.

[10] Myungjin BAE and Souguil ANN, "Data compression by elimination of redundancy in human speech signals, "Seoul national university engineering report, vol. 17, no. 1, pp. 129-133, April 1985.

[11] Myungjin BAE, Ikjoo CHUNG, and Souguil ANN, "The extraction of nasal sound by using G-peak in continued speech," KIEE, vol. 24, no. 2, pp. 92-97, March 1987.

[12] 이현복, 국제음성문자와 한글음성문자, 과학사, 1981. *