실시간 음성분석도구의 MatLab 구현*

박일서(창원대), 김대현(창원대), 조철우(창원대)

<차 례>

<Abstract>

Matlab Implementation of Real-time Speech Analysis Tool

Il-suh Bak, Dae-hyun Kim, Cheol-woo Jo

There are many speech analysis tools available. Among them real-time analysis tool is very useful for interactive experiments. A real-time speech analysis tool was implemented using Matlab. Matlab is a very widely used general purpose signal processing tool. In general, its computational speed is relatively lower than that of the codes from conventional programming languages. Especially, real-time analysis including input of signal and output of the result was not possible in the past. However, due to the improvement of computing power of PCs and inclusion of real-time I/O toolboxes in Matlab, real-time analysis is now possible in some extent by Matlab only. In this experiment, we tried to implement a real-time speech analysis tool using Matlab. Pitch and spectral information is computed in real-time. From the result it is shown that such real-time applications can be implemented easily using Matlab.

* 주제어: Matlab, 음성분석도구(speech analysis tool)

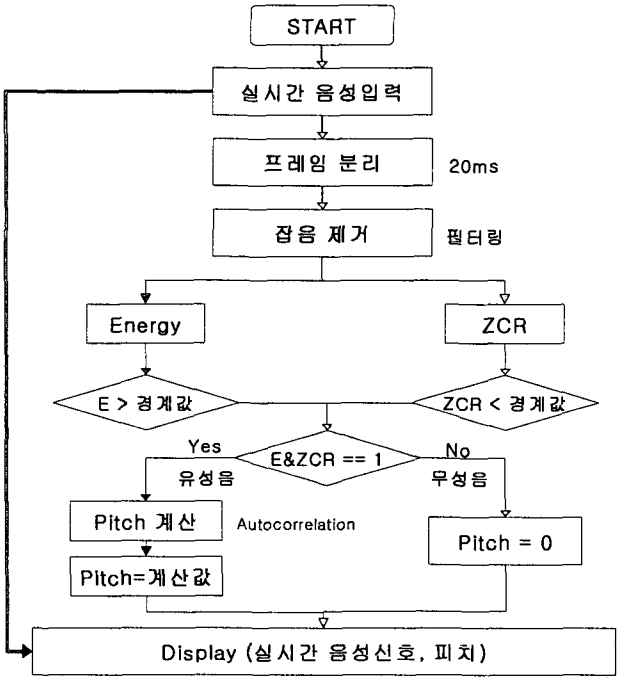
* 이 논문은 2001년도 창원대학교 교내연구비에 의하여 연구되었습니다. 지원에 감사드립니다.

1. 서 론

피치는 음성의 주기적 특성을 나타내는 특성으로 음성으로부터 여러 가지 의미를 추출해 내는데 중요한 요소로 작용한다. 이는 말하는 사람의 의도나 정서적 특징 심지어는 성대의 이상 유무 등 많은 의미를 지닌다. 그래서 피치는 공학적으로 음성의 인식·코딩·합성에, 음성학적으로 운율 정보를 나타내고 심리학적으로 정서 상태의 판정에 쓰인다. 또 음성 의학 분야에서 음성 장애 상태를 판단하는 것뿐만 아니라 성악에서 가창력과 성량 측정 및 개선 등에 이용된다. 그러므로 피치 검출기는 여러 가지 음성 처리 시스템에 필수적인 요소이다. 피치 검출기의 성능은 음성 합성 시에 여기원의 특성을 나타내어 음절의 자연성을 좌우하고, 검출된 피치 변화도는 화자 인식용 및 발음 장애자를 위한 보조 시스템용 파라미터로 널리 적용된다. 이처럼 피치 검출기는 거의 모든 음성 분석-합성(vocoder) 시스템에 널리 쓰이고 있다. 따라서 이러한 피치 검출에 대한 다양한 알고리즘들이 제안되어 왔는데 그 처리 영역에 따라 시간 영역, 주파수 영역, 시간-주파수 혼성 영역으로 분리해서 다루어지고 있다.[1][2][3][4][5][6] 이중에서도 시간 영역 피치 검출법은 시간 영역에서 직접 처리하기 때문에 다른 영역으로의 변환 과정이 불필요하며 합, 차, 비교 논리 등에 의해서만 처리가 가능하므로 쉽다. 또 피치의 범위가 보통 2.5ms에서 25ms로 알려져 있고 음성을 8㎑로 표본화하여도 그 범위가 20-200 표본 사이에 나타나기 때문에 시간 영역의 검출은 분해능이 높은 특징이 있다. 시간 영역 피치 검출법에는 병렬 처리법, 면적 비교법, 자기상관계수 이용법, ADMF 이용법 등이 있는데[2][5] 여기서는 가장 널리 사용하고 있는 자기상관함수에 의한 피치 검출법을 사용하여 실시간 피치 분석기를 구현해 보았다.

2. 피치추출방법

피치는 유성음에서 관찰되는 주기 성분으로 피치 주파수는 그 주기의 역수로 음성의 기본주파수가 된다. 인간이 말을 할 때 성대가 진동하여 음성이 발생하는데 그 주기성은 완전하지는 않으나(quasi-periodic) 보통 파형에서 눈으로 알아볼 수 있을 정도로 규칙적이다. 사람에 따라 성도 길이가 다른데, 어린이나 여자가 그 길이가 짧아서 주파수가 높듯이 모든 사람에게는 후두 구조에 의해 제약되는 피치 범위가 있다. 남자는 보통 50-250Hz, 여자는 120-500Hz 인데 이는 최대 범위로, 보통은 자연스럽게 말할 때 평균적으로 사용하는 습관적인 피치 레벨을 갖고 있다. 피치는 강세, 억양, 감정 등의 요인에 따라 변한다. 지금까지 알려진 바에 따르면 피치를 조절하는 가장 큰 요소는 성대 근육의 장력이다.



<그림 1> 실시간 피치분석기 구현의 순서도

2.1. 자기상관

음성에서 주기 신호의 주기를 구하는 방법은 자기상관(Autocorrelation), Cepstrum, FFT 등 여러 가지가 있으나 자기상관법이 일반적으로 이용되고 있다.

자기상관함수는 피치 검출, 유/무성음 결정 그리고 선형 예측 등에 사용되는데 신호의 고조파와 포만트 진폭, 주기 정보를 유지하고 위상 정보는 무시한다. 자기상관함수는 두 신호의 시간 지연으로 유사성을 측정하는 상호상관함수의 특수한 경우로 다음과 같은 식으로 정의된다.

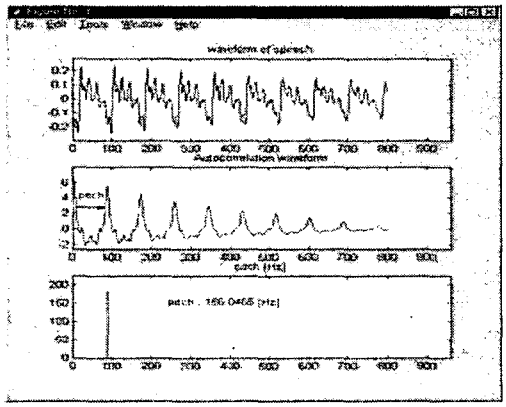
(1) $$\phi(k) = \sum_{t=-\infty}^{\infty} x(t)x(t+k)$$

자기상관함수는 우함수($\Psi(k)=\Psi(-k)$)로 k=0 일 때 최대값을 가지며 $\Psi(0)$는 이 신호의 에너지이다. 어떤 신호가 주기가 P인 주기신호라면 $\Psi(k)$ 역시 주기를 갖고 $\Psi(k)$의 최대값은 k=0, ±P, ±2P의 피치 주기에서 발생한다.

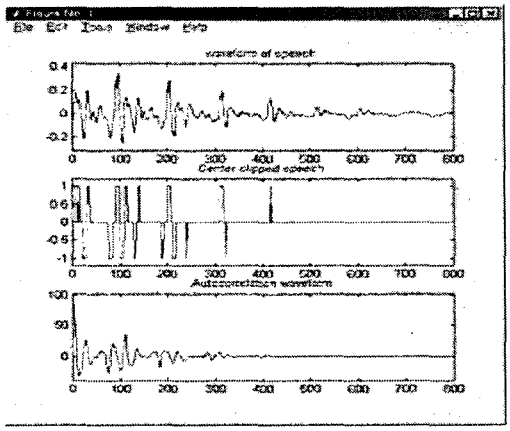
자기상관을 하기 전에 중앙 클리핑 과정을 거치는데 이를 수행하면 음성신호의 봉오리만 남게 되어 쉽게 피치를 검출할 수 있을 뿐만 아니라 성도 정보를 제거하여 더 정확하게 피치를 검출할 수 있다. 클리핑 레벨은 일반적으로 프레임 최

대값의 1/3 정도로 정한다.

여기서는 일정 구간에서 평균적인 피치를 구했는데 음성의 한 프레임(20ms)마다 자기상관을 수행하여 주기를 구하고 피치 주파수를 구했다. 평균피치는 구하기가 비교적 쉽고 음성의 강세 변화 등 전체적인 변화 양상을 관찰하는데 유용하다.



<그림 2> 자기상관함수를 사용한 주기검출



<그림 3> 중앙 클리핑된 음성신호와 자기상관
중앙클리핑 값 (최고값의 1/3) = 0.1066
이 구간의 피치값 = 110
피치주파수(기본주파수) = 145.4545

2.2. 유성음의 판단

음성은 크게 유성음과 무성음으로 나눌 수 있는데 유성음 부분을 살펴보면 준주기적인 파형이 반복됨을 알 수 있다. 이러한 파형의 주기가 피치인데 음성신호에서는 20ms의 한 프레임 내에 일반적으로 2~3개의 주기가 나타나며 무성음은 그 크기가 상대적으로 작고 비주기적이다. 따라서 분석 구간 내에서도 그 프레임이 유성음 구간인지 무성음 구간인지 파라미터에 의해 구별할 수 있어야 한다. 일반적으로 피치 검출의 목적은 피치 주기에 연관된 2가지의 모델계수를 얻는데 있다. 이것은 여기모드(excitation mode) 즉 유/무성음의 여부에 대한 값과 피치 주기의 값이다. 유성음에서만 준주기적인 성질이 나타나므로 우선 피치를 구하려면 음성을 무성음과 유성음 부분으로 나누어야 한다.

여기서는 유/무성음을 판단하기 위한 파라미터로 영교차율과 에너지를 사용하였다.

2.3. 영교차율

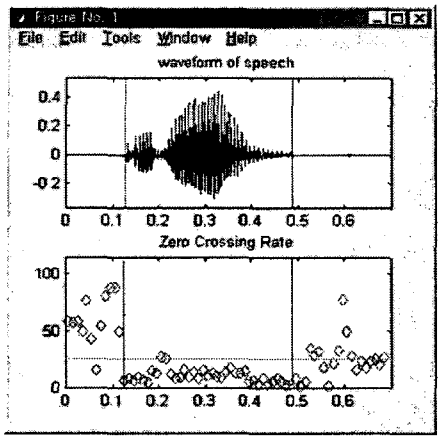
영교차율이란 말 그대로 음성 신호 파형의 위상이 중심축을 통과하는 회수를 말하는데 이는 적은 계산 양으로 음성이 유성음인지 무성음인지 판단할 수 있는 중요한 정보를 얻을 수 있다.

$$Z_n = \sum_{m=-\infty}^{\infty} |sign[s(m)] - sign[s(m-1)]| w(n-m) \tag{2}$$

대부분의 음성 에너지가 3㎑ 이하에 분포하는 반면 무성음의 경우에는 높은 주파수에 분포하므로 영교차율이 커진다. 파열음 같은 무성음의 경우 진폭이 크지 않고 불규칙적인 진동이 계속 있으므로, 일정한 만큼만 중심축을 통과하는 유성음보다 영교차율이 클 수밖에 없다.

프레임별 영교차율을 보면 유성음 구간에서 대체로 무성음 구간보다 영교차율이 적음을 볼 수 있다. 그러나 실제로 잡음이 심한 곳에서 실시간으로 피치를 검출할 경우 유성음 구간의 영교차율과 무성음 구간의 영교차율이 잘 구분되지 않음을 볼 수 있다.

여기서 알 수 있듯 영교차율은 잡음에 매우 민감하기 때문에 환경에 따라 영교차율 만으로 모든 유/무성음을 결정하는데는 한계가 있다. 그래서 영교차율과 함께 유/무성음을 결정하는 파라미터로 같이 사용한 것이 바로 에너지이다.



<그림 4> 신호프레임별 영교차율

2.4. 에너지

다음과 같이 표현되는 신호의 이산시간에너지는 신호의 전구간에 걸친 값의 제곱의 합이다.

(3) $$E = \sum_{m=-\infty}^{\infty} x^2(m)$$

반면에 아래와 같은 신호의 단시간 에너지는 표본에서 n까지 N 개의 표본의 제곱의 합으로 정의된다.

(4) $$E_n = \sum_{m=n-N+1}^{n} x^2(m)$$

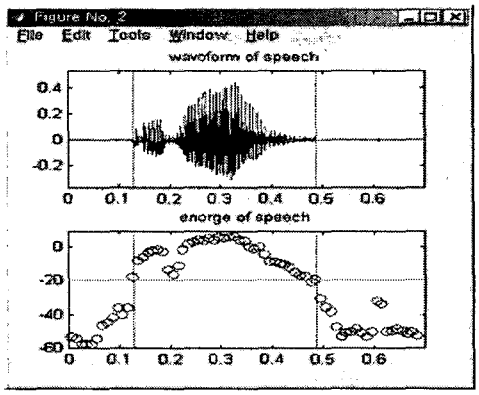
여기서는 더 간단한 방법으로 프레임의 에너지를 구했다.

(5) $$\phi(k) = \sum_{t=-\infty}^{\infty} x(t)x(t+k)$$

앞에서 자기상관을 수행하는 과정에서 k=0일 때의 값이 바로 에너지가 되므로 자기상관을 수행한 후에 시간축의 영점에서 그 크기 Ψ(0)를 사용하였다.

실제 음성 구간에서 프레임별로 에너지를 구하여 보면 잡음이 많이 포함되어

유/무성음이 잘 구별되지 않는 구간에서도 에너지를 통해 확연히 드러난다. 이를 보면 유성음 구간이 무성음보다 에너지가 훨씬 크다. 에너지를 dB 단위로 환산하여 그 크기가 작은 구간에서의 구별을 쉽게 했다.

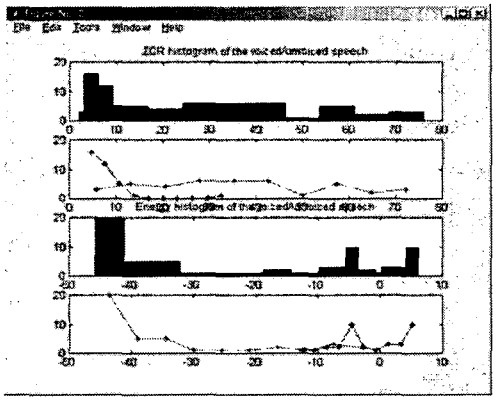


<그림 5> 신호 프레임별 에너지

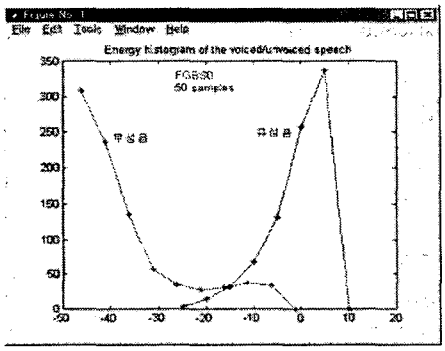
2.5. Threshold

앞에서 영교차율과 에너지를 구했으므로 이제 이것을 이용해서 유/무성음을 구분해야 한다. 우리가 파라미터로 정한 에너지와 영교차율에서 그 판단을 하기 위해 미리 경계값을 찾아야 한다.

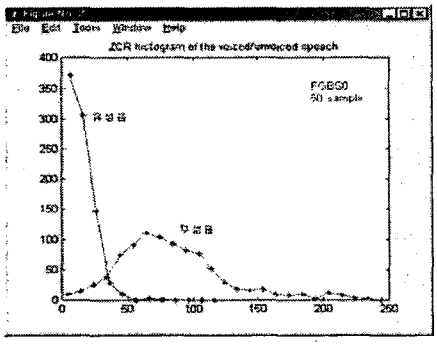
미리 남녀 음성 샘플 약 500여 개를 가지고 유성음 구간을 미리 입력하고 유성음에서의 에너지와 영교차율을 구해 보았다. 그리고 유성음 구간에서 히스토그램으로 구해진 에너지와 영교차율을 누적시켜서 그 경계치를 찾는 과정을 거쳤다. 하나의 음성 샘플에서 경계치를 찾지 않고 성별을 가리지 않는 여러 개의 음성 샘플에서 그 평균값을 찾도록 했다.



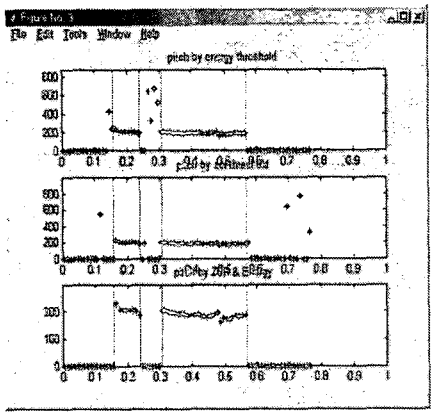
<그림 6> 유성음 구간의 영교차율과 에너지 히스토그램



<그림 7> 에너지 누적 히스토그램



<그림 8> 영교차율 누적 히스토그램
(유성음과 무성음의 교차점을 경계값으로 결정)



<그림 9> 에너지와 영교차율의 경계값으로 찾은 최종 유성음 구간과 피치값

위와 같이 미리 유성음이라 정해 준 구간에서 에너지와 영교차율을 구해 히스토그램을 그리고, 그 경계값을 찾아낸다.

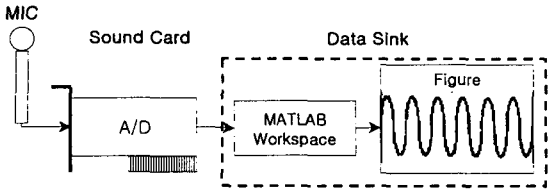
ZCR과 Energy에 의한 히스토그램에서 찾은 경계값을 기준으로 영교차율이 기준값보다 작을 경우 또 에너지가 기준값보다 클 경우 유성음으로 판단한다. 이 두 파라미터에 의해 유성음으로 판정된 음성 구간을 AND를 취해서 최종 유성음의 구간을 찾고 그 피치를 구했다. 이 구간으로 판별된 유성음 구간은 초기에 설정해 준 유성음 구간과 거의 일치함을 볼 수 있다.

그러나 이 경우 영교차율과 에너지가 구해지는 프레임에 의해 유/무성음 판단의 정확성이 좌우된다. 음성이 시작되는 구간을 정확히 잡아서 시작하지 않으므로 유성음이라 할지라도 그 주기가 충분히 드러나지 않게 프레임이 구성될 경우 판단은 틀릴 가능성이 많다. 또 유/무성음이 같이 섞여 있거나 유성음이라도 잡음이 많이 포함된 경우 단순히 히스토그램으로 얻어진 경계값 만으로 유/무성음을 확정짓는 데에는 문제가 있다.

3. 실시간 구현을 위한 방법

실시간으로 음성신호를 PC를 통해 받아들이기 위해서는 일반적인 사운드 카드를 인식할 수 있어야 한다. MATLAB에서 지원하는 Data Acquisition Toolbox를 이용하여 사운드 카드를 인식할 수 있으며, 또한 실시간으로 원하는 샘플링 주파수와 구간(duration)선택도 가능하다.

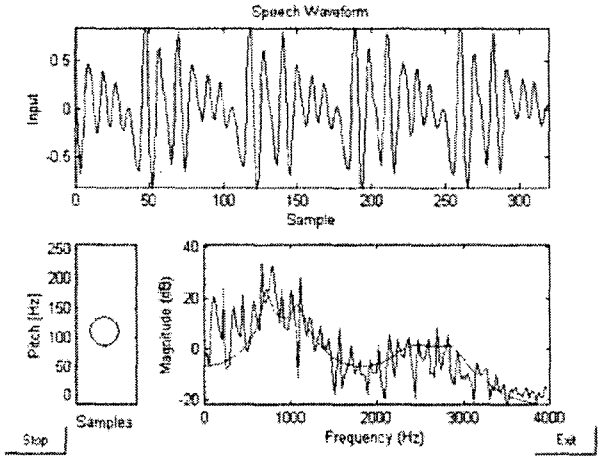
이러한 목적을 위해 설정된 하드웨어와의 연결은 그림 10과 같다.



<그림 10> Matlab의 실시간 음성획득을 위한 연결

4. 실시간 피치 분석기

실제로 마이크를 통해 입력받은 음성이 목소리의 높낮이와 유. 무성음의 여부에 따라서 피치가 구해짐을 볼 수 있다. 동시에 퓨리에 변환의 결과와 함께 LPC 스펙트럼을 표현함으로써 피치는 물론, 유성음일 경우 포만트의 분석도 실시간으로 가능하다.



<그림 11> 실시간 피치분석기를 실행한 /아/음성- realtime_pitch2.m

5. 결 론

지금까지 MatLab을 이용한 실시간 피치 분석기 구현에 대해 살펴보았다. 여기서는 시간 영역 피치검출방법인 자기상관 기법을 이용하여 실시간으로 입력되는 음성 프레임의 평균피치를 측정하고 에너지와 영교차율을 계산하여 이들을 미리 많은 샘플들로부터 구해진 경계치와 비교하여 유성음을 판단했다. 구현된 피치 분

석기는 강약, 고저에 의해 입력 음성의 유성음 피치를 어느 정도 정확히 나타내 준다.

그렇지만 대부분의 피치검출기는 잡음이 있는 환경, 대역폭이 제한된 경우, 주파수의 위상 특성이 변질되는 경우와 같은 복잡한 환경에서는 신뢰할 만한 결과를 얻지 못한다. 여기서 사용한 시간 영역 피치 검출법 역시 구현하기 쉽고 분해능이 높은 장점이 있는 반면에 특성전송 채널에 통과된 경우나 배경 잡음이 부가된 경우에는 이의 영향이 커져서 피치 검출 오차가 높아지는 단점도 있다.

음성신호의 피치주기를 정확하게 측정하는 것이 어려운 것은 성문의 여기 파형이 완전히 주기적인 파형이 아닐뿐더러 성도의 성문간의 상호작용 때문이다. 또 유성음 구간 동안 피치주기의 정확한 시작과 끝을 정의하기가 어렵고 무엇보다 여기서도 나타났듯이 무성음과 낮은 레벨의 유성음을 정확히 구별하기가 쉽지 않다는 것이다.

또 음성 입력 부분에 있어서 입력 신호는 음성의 크기, 주변 환경, 마이크의 성능 심지어는 사운드 카드에 의해서도 영향을 받는다. 그래서 잡음이 심하면 영교차율이 커져서 유성음도 무성음으로 판단되기도 하고 음성의 크기가 너무 작은 경우 에너지 레벨에 의해 잡음으로 판단되기도 한다. 그러므로 우선 전원 잡음을 제거하는 등 H/W적으로 잡음의 영향이 적도록 하고 더불어 잡음을 완전히 제거할 수 있는 알고리즘을 S/W적으로 보완해 주어야 할 것이다. 그리고 유/무성음의 판단에 있어서도 경계치 부근에서의 판단에 좀더 융통성을 줄 수 있도록 퍼지(Fuzzy) 알고리즘을 사용해 보는 것도 좋을 것이라 생각된다.

그리고 입력 신호는 실시간으로 들어오는데 반해 피치를 계산하기 위한 프로그램의 실행으로 시간 지연이 생기는데 이로 인해 완전한 실시간 구현이 되지 않았다. 이는 계산 과정의 최적화를 통해 보완해 줄 수 있을 것이다.

참 고 문 헌

[1] Sohubha Kadambe & G. Faye Boudreau (1992), Applications of Wavelet Transform for Pitch Detection of Speech Signals, *IEEE Trans on Information Theory 38.*

[2] W. Hess (1983), *Pitch Determination of Speech Signal*, Spriger-Verlag.

[3] Wong et al. (1979), Least squares glottal inverse filtering from the acoustic speech waveform, *IEEE. Trans. ASSP 27*, pp.350~355.

[4] Wilpon, J. G., L. R. Rabiner and T. B. Martin (1984), An improved word-detection algorithm for telephone-quality speech incorporating both synthetic and semantic constraints, AT&T Tech, pp.479~497.

[5] Rabiner et al. (1976), A comparative study of several pitch detection algorithm, *IEEE Trans. ASSP 24*, pp.399~413.

[6] Jo Cheol-woo et al. (1996), Improve glottal closure instant detector based on linear prediction and standard pitch concept, *Proceeding of ICSLP'98*, pp.1217~1220.

접수일자: 2002년 11월 15일
게재결정: 2002년 12월 12일

▶ 박일서 (Il Suh Bak)
주소: 641-773 경상남도 창원시 사림동 9번지 창원대학교 공과대학 제어계측공학과
소속: 창원대학교 제어계측공학과 음성 및 음향 신호처리 실험실
전화: 055) 279-7550
Fax: 052) 262-5064
E-mail: heapung@sarim.changwon.ac.kr

▶ 김대현 (Dae Hyun Kim)
주소: 641-773 경상남도 창원시 사림동 9번지 창원대학교 공과대학 제어계측공학과
소속: 창원대학교 제어계측공학과 음성 및 음향 신호처리 실험실
전화: 055) 279-7550
Fax: 052) 262-5064
E-mail: midas03@taegu.net

▶ 조철우 (Cheol-Woo Jo)
주소: 641-773 경상남도 창원시 사림동 9번지 창원대학교 공과대학 제어계측공학과
소속: 창원대학교 제어계측공학과 음성 및 음향 신호처리 실험실
전화: 055) 279-7550
Fax: 052) 262-5064
E-mail: cwjo@sarim.changwon.ac.kr