음성 신호에서 유성음, 무성음, 묵음의 식별에 관한 연구

°김 학 윤, 황 영 수, 차 일 환
연세대학교 전자공학과

A Study on the Voiced, Unvoiced, and Silence Classification of Speech Signals

Kim, Hack-yoon, Whang, Young-soo, Cha, Il-whan
Dept. of Electronic Eng., Yonsei University

Abstract

In this paper, we describe a new algorithm for deciding whether a given segment of a siqnal is classified as voiced speech, unvoiced speech, or silence.

The excitation is represented by one of the two states: Voiced-more or less periodic, produced by vibration of the vocal cords, or unvoiced-noise-like, produced by forcing air past some constriction in the vocal tract.

The measured parameters for the voiced-unvoiced classification are the PARCOR(Partial Autocorrelation) and LPC(Linear Predictive Coding).

The employed parameters for the unvoiced-silence classification, also, are each of ZCR(Zero-Crossing Rate) and MAG(Magnitude) during four milisecond interval.

A classifier is obtained which, in limited tests, achieves 96.5 Percent classification accuracy on speaker dependent test, and 95 Percent accuracy on speaker independent tests.

Ⅰ. 서 론

음성 분석의 일반적 과정은 음성의 물리적 성질, 음향학적 Modeling, 언어학적 해석에 있다. 인간의 음성기관은 세가지로 구분할 수 있다. 1) 호흡기 : 호흡기는 언어음의 대부분을 생산하는데 필요한 기류를 공급한다.
2) 후두 : 음의 에너지를 생성한다.
3) 성문 : 공명체 구실을 한다.
이때 성문의 모양에 따라 유·무성음이 발성되며, 성대의 진동 속도에 따라 주파수가 달라진다. 우리나라 말은 우랄 알타이어족에 속하며 구조상으론 교착어이므로 모음조화현상이 뚜렷하다.

즉 자음 14자, 모음 10자로된 음소문자이다. 모음의 음색은 본질적으로 formant에 연관되어지며, 이 formant는 음성기관의 주된 두 공명감 인두와 입에 연관되어져 있다. 모음의 특징은 청각적으로 귀에 들리는 소음이 없다는 것과 조음의 견지에서 공기의 통로가 자유로우며, 자음은 소음이거나, 소음을 포함하고 있거나 하며, 공기의 흐름을 폐쇄하거나 좁히거나 해서 발음된다. 이로서 자음은 순간자음과 지속자음으로 구별된다.

일반적 음성분석 시스템에서는 시간축과 주파수축으로 분석을 행한다. 이때 계산 시간 및 처리 과정을 줄이기 위해 연속된 음성신호 X(n)를 Voiced, Unvoiced, Silence로 구분하는 것이 상당히 필요하다.
이렇게 구분하는 방법으로는 수msec동안 Speech Segment를 단위로 평균 ZCR, Average Energy, 인접 Samples들의 Correlation, LPC, PARCOR, Formant, 예측 에너지등을 사용할수 있다.
따라서, 본 연구에서는 음성신호가 갖는 특수한 성질을 이용하여 V/U/S를 결정짓는 새로운 Parameter를 제시하였다.

Ⅱ. B·P와 E·P결정 Algorithm

입력 음성신호 X(n)의 B·P(Beginning Point)와 E·P(Ening Point)의 검출은 불필요한 입력정보를 제거시키고 분석에 필요한 Signal만을 검출하여 처리과정을 줄이기 위해서는 중요한 과정이다.

1) Average Magnitude

음성신호 X(n)의 Magnitude function Mn은

$$Mn = \sum_{m=-\infty}^{\infty} |X(m)| W(n-m) \quad \cdots\cdots(1)$$

이고 Short-time Energy En은

$$En = \sum_{m=-\infty}^{\infty} [X(m)W(n-m)]^2 \quad \cdots\cdots(2)$$

이며 W(m)은 Window function이다. (1)식과 (2) 식을 비교하여 보면 Multiplication이 줄으므로 CPU계산 시간도 단축되며 입력 신호X(n)의 레벨이 클때 En은 자승항이라 Sensitive하므로 이런 단점이 없는 Magnitude function을 사용하였다.

2) ZCR (Zero-Crossing Rate)

입력 신호의 주파수가 f_0의 정현파일때 ZCR은 $2f_0$이다. 이때 Average ZCR Z는

$$Z = 2F_0/Fs \text{ [Crossings/Sample]} \quad \cdots\cdots(3)$$

어떤 임의의 점 n에서의 ZCR Zn은

$$Zn = \sum_{m=-\infty}^{\infty} |Sgn[X(m)]-Sgn[X(m-n)]| \quad \cdots\cdots(4)$$

이다.

일반적으로 성도의 특성에 따라 무성음의 주파수 대역은 3K정도로 유성음보다 상당히 높다. 그러므로 이와 같은 음향학적 특징에 따라 ZCR은 무성음이 유성음보다 높다.

본 연구에서는 수행시간이 빠른 이 ZCR을 이용하여 무성음과 유성음의 구별의 한 방법으로 택했다.

3) Algorithm

X(n)의 B·P와 E·P를 구하는 방법은 아래와 같다. 처음 음성 신호가 입력될때 어느 Sample까지는 Data가 입력되지 않고 그 Sample이상 부터 입력되었을 때 그 구간에서 ZCR과 MAG의 평균값 X_{ZCR}, X_{MAG}와 분산값 σ_{ZCR}, σ_{MAG}를 구한후 입력음성의 Sample값을 이동시키며 각 Sample당 ZCR′과 MAG′를 구한다.

이때

$$MAG' > X_{MAG} + K\sigma_{MAG} \quad (K : 상수) \quad \cdots\cdots(5)$$

일때의 memory번지를 검출한후 입력 Sample을 역으로 이동시키며

$$ZCR' < X_{ZCR} + K\sigma_{ZCR} \quad \cdots\cdots(6)$$

인 Momory번지를 검출하여 B·P로 잡았다.

E·P검출은 B·P를 구한 방법과 유사하게 행했다. 먼저 신호의 Peak치를 찾은후 그때의 X_{ZCR}, X_{MAG}, σ_{ZCR}, σ_{MAG}를 계산한 후

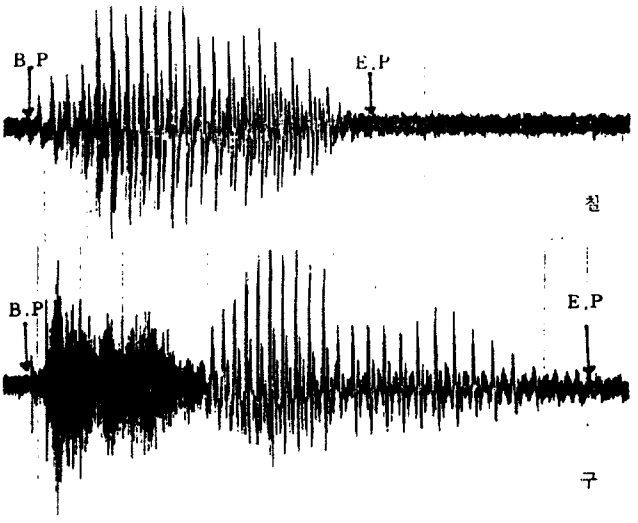
$$MAG' < X_{MAG} + \sigma_{MAG} \quad \cdots\cdots(7)$$

인 번지를 검출하여 그 번지부터 순방향으로 이동시키면서

$$ZCR' < X_{ZCR} + \sigma_{ZCR}$$

일때의 번지를 E·P로 잡았다.

<그림 1>에 숫자음 "칠, 구"의 B·P와 E·P 잡은 예를 나타내었다.



<그림 1> 숫자음 (칠, 구)의 B·P와 E·P

Ⅲ. 유성음, 무성음, 묵음 식별 방법

유성음을 주파수 영역에서 살펴보면 유성음은 quasi-periodic 한 신호이므로 Auto correlation을 취하면 주기가 나타나지만 무성음의 경우는 이산적이다.

이 예를 <그림 2>에 나타내었다.

1) Magnitude function

2) Normalized autocorrelation cofficient

$$\phi(n) = \sum_{i=1}^{N} S_i S_{i-1} / [\sum_{i=1}^{N} S_i^2 \sum_{i=0}^{N-1} S_i^2]^{1/2}$$

3) Linear Predictive coding normalized minimum error

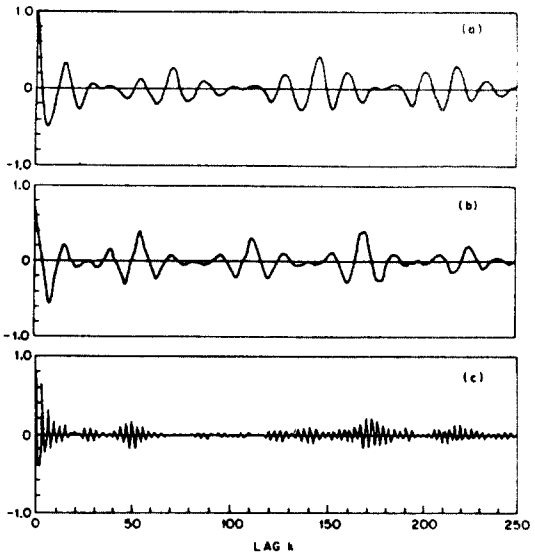
4) The first LPC predictor계수

5) ZCR

6) PARCOR

위에 서술한 6가지 방법에 의해 유성음, 무성음, 묵음을 구별하는 방법으로 삼았다. ZCR사이에 존재하는 area는 평균 MAG와 Zero Crossing interval에 비례하므로 결국 ZCR이 높고 Magnitude가 작은 무성음에서 구한 area값들은, 에너지가 크고 ZCR이 낮은 유성음에서 area에 비해 차이가 두드러지게 나타남을 알수 있으므로 유성음과 무성음 결정의 Threshold level를 잡기가 용이해진다. 그러므로 Back ground noise

로 부터의 영향을 최소화 할수 있다. 또한 무성음과 묵음의 식별은 무성음은 약 3KHz 부근에 에너지가 집중되어 있다. 이것이 noise와 다른점은 Colored noise 형태를 띤 점인데 이런 음의 구별은 고주파 대역을 강조시켜 분리하면 구별이 용이해진다.



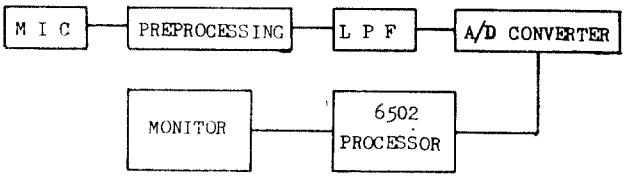
<그림 2> 유성음, 무성음의 Aucorrelation Function

Ⅳ. 실험 및 결과

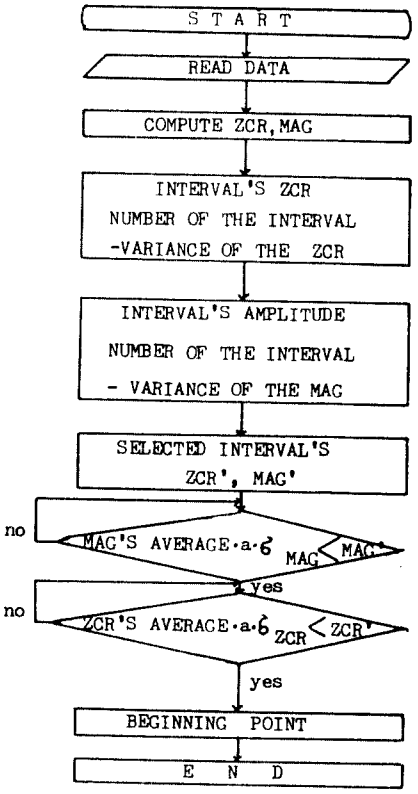
본 연구에서 사용한 전체 System 도는 <그림 3>에 표시하였다.

앞에서 제시한 방법으로 부터 MAG, ZCR을 비교하여 B·P와 E·P를 잡은 다음 MAG를 비교하여 산출된 threshold level점을 무성음과 유성음의 교차점으로 잡고 B·P와 이점사이를 무성음으로 구분 확인 하였다. 또한 무성음과 Silence구분은 Auto와 MAG를 고주파 강조를 사용 구분하였다.

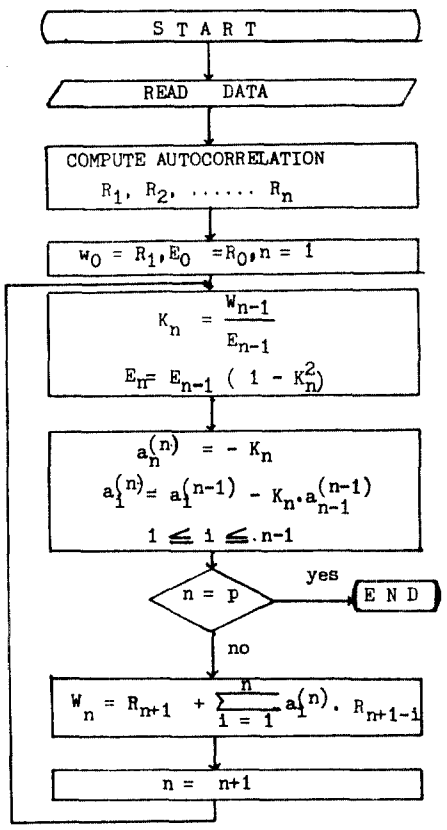
본 연구에서 사용된 ZCR, MAG의 flow-chart를 <그림 4>에 나타내었다. 실험대상은 학습받은 성인 2 명을 택해 Reference pattern을 잡은 후 학습받지 않은 성인 5명을 택하여 유성음, 무성음, 묵음을 식별하였다. 모음부 구별을 위해 Autocorrelation; LPC, PARCOR 방식의 flow-chart를 <그림 5>에 보여 준다.
또한 2장에서 설명된 B·P와 E·P의 예를 <그림 1>에 나타내었다.



<그림 3> 전체 System Block diagrum



<그림 4> B·P와 E·P검출 Flow-chart



<그림 5> PARCOR계수 산출 Flow-chart

V. 결 론

입력 음성신호X(n)의 B·P와 E·P추출 Algorithm을 토대로 시작점과 끝점 검출은 용이하게 이루어진다. 그렇지만 threshold의 차이로 인해 약간의 오차는 생겼으리라 추측된다. 이런 음성신호의 시작점과 끝점 검출은 음성인식 및 합성에 상당한 효과를 가지고 있다. 또한, ZCR, MAG, AUTO, LPC, PARCOR, F_0(First formant)를 사용한 Algorithm의 개발로 새로운 유성음, 무성음, 묵음의 식별을 시도하였다.

학습받은 화자인 경우 식별이 96.5% 식별이 이루어졌고 학습받지 않은 화자인 경우에는 95%의 식별이 이루어 졌다. 즉, ZCR으로 유성음과 무성음의 식별을 구분한 다음 여러가지 방법으로 행했으므로 좋은 식별율이 나왔다.

이와같은 방법을 사용하여 조음부 인식, 음성인식, 합성, 화자 식별이 좀더 용이해지리라 여겨진다.

본연구에서는 6502 CPU를 사용하여 식별을 피하였으므로 수행시간은 조금 오래 수행되었다.

참 고 문 헌

1. J.M.Baker, "A New Time-domain Analysis and other complex wave forms", Ph.D pillsburgh, 1975.
2. L.R.Rabinar and M.R.Sambur, "An Algorithm for Determining the Endpoints of Isolated Mtterances" Bell syst.Tech.J. Vol54 No.2 February,1975.
3. Nolan, "The phonetic bases of speaker recognition", Cambridge University Press, 1983.
4. Denes, P. & Mathews, M.V. "Spoken digit recognition using tins-freguency pattern Matching", JASA Vol 32, 1960.
5. Davis, K.H.Biddulph. "Automatic recogmition of spoken digit", JASA Vol 24, 1980.
6. LR.Rabiner/R.W Schater, "Digital processing of speech signals",prentice-Hall, 1978
7. B.S.Atal, "Effectiveness of linear prediction characteristies of speech wave for automatic speaker identification and veritication JASA, Vol55, 1974.
8. I. Kameny, "Automatic acoustic-phonetic analysis of vowels and sonorants", In proc IEEE ASSP, Apr. 1976.
9. 차일환, 음향 공학 개론, 한신문화사.
10. B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J.Acoust. Soc.Amer., Vol 55, 1974.
11. J.N.Maksym, "Real time pitch extraction by adative prediction of the speech waveform," IEEE, Audio Electroacoust., Vol 21, 1963