음성신호의 상위 포만트에 대한 ZCR-파라미터 검출에 관한 연구

유 건 수(*), °김 건 명(*), 이 성 우(*), 배 명 진(**)
(*) 호 서 대 학 교 (**) 숭 실 대 학 교

(On a Detection of the ZCR-Parameter for Higher Formants of Speech Signals)

Geonsoo Ryoo(*), °Geonmyung Kim(*), Sungwoo Lee(*), Myungjin Bae (**)
(*) Hoseo University (**) Soongsil University

Abstract

In many applications such as speech analysis, speech coding, speech recognition, etc., the voiced-unvoiced decision should be performed correctly for efficient processing. One of the parameters which are used for voiced-unvoiced decision is zero-crossing. But the information of higher formants have not represented properly with general zero-crossing. In this paper, so we proposed a higher formants ZCR-parameter which is represented as the zero-crossing rate for higher formants of speech signals.

I. 서론

음성분석, 음성코딩, 음성인식 등에서 음성정보를 효과적으로 처리하기 위해서는 음성구간을 정확하게 검출할 수 있어야 하고 유-무성음 구간을 정확하게 분류해야 한다. 현재 제안되어져 있는 기법으로는 에너지에 기본된 것, 패턴인식을 이용하는 것, 혹은 순서적 배열을 이용하는 것(rank-order statistics) 등이 있다.

지난 10년 동안 끝점검출에 가장 폭넓게 적용된 방법은 L.R.Rabiner와 M.R.Sambur에 의해 제안된 에너지-기본된 독립(explict)검출법이다. 이 검출법은 한 프레임마다 단시간 평균 에너지에 의해 유성음 구간을 찾고 이 유성음 구간으로부터 영교차율(zero-crossing rate)을 이용하여 무성음 구간을 검출하는 방법이다.

에너지를 이용하여 음성구간을 검출하거나 유-무성음을 판단할 경우 진폭의 빠른 변화에 대하여 민감한 단구간 에너지 함수가 필요하다. 따라서 창의 길이가 너무 짧으면(한 피치주기보다 짧으면) 에너지에는 세세한 국부봉우리가 많이 나타나게 된다. 또한 창의 길이가 너무 길면(몇 개의 피치주기를 포함하면) 에너지의 변화가 스므딩되어 음성신호의 변화특성을 충분히 나타내지 못한다. 영교차율을 이용하는 경우에는 기울 등에 의한 여기(excitation)나 dc 바이어스가 존재하면 ZCR 측정시에 이들의 영향을 많이 받게 된다. 따라서 음성신호의 주된 포만트(제 1 포만트)는 근사적으로 구할 수 있지만 유성음의 경우 제 1 포만트가 지배적이므로 상위 포만트 영역의 정보를 얻기에는 다소 어려움이 따르게 된다.

따라서 본 논문에서는 음성신호에서 상위 포만트 신호의 영교차율을 검출하여 음성신호의 유-무성음을 검출하거나, 끝점을 검출하는 전처리용 상위 포만트 ZCR-파라미터를 새로이 제안하고자 한다. 2절에서 4절까지는 단구간 평균 영교차율과 음성신호의 스펙트럼 및 음성신호의 자기 상관관계를 살펴보고 5절에서 본 논문에서 제안하는 상위 포만트에 대한 ZCR-파라미터에 대해 알아보기로 한다. 그리고는 제안한 ZCR-파라미터를 실제의 음성신호에 적용하여 산출된 처리결과를 검토하고 결론짓는다.

II. 단구간 평균 영교차율

이산신호에서 영교차는 연속적인 샘플의 부호가 다를 경우에 발생한다. 신호의 주파수는 간단히 영교차가 발생하는 비율로써 측정할 수 있다. 예를 들어 F_0 주파수의 정현신호를 F_s의 비율로 샘플링한다면 한 사이클당 F_s/F_0 샘플이 존재하게 된다. 각 사이클에서 2번의 영교차가 발생하기 때문에 평균 영교차율은 두배의 F_0/F_s가 된다.

단구간 평균 영교차율을 이용하여 스펙트럼 특성에 대한 대략적인 예측을 할 수 있다. 근사적인 정의식은 아래와 같다:

$$Z_n = \sum_{m=n+N-1}^{n} | sgn[x(m)] - sgn[x(m-1)] | w(n-m) \qquad (2\text{-}1)$$

여기에서

$$sgn[x(n)] = 1, \quad x(n) \geq 0$$
$$= -1, \quad x(n) < 0 \qquad (2\text{-}2)$$

이고

$$w(n) = \frac{1}{2N}, \quad 0 \leq n \leq N\text{-}1$$
$$= 0, \quad otherwise \qquad (2\text{-}3)$$

이다.

유성음과 무성음의 음성생성에 대한 모델을 살펴보면 성문파형에 의한 스펙트럼의 감소 때문에 유성음의 에너지는 약 1KHz 이하에 집중되어 있다. 반면 무성음의 경우에는 대부분의 에너지가 2kHz 이상의 주파수에 집중되어 있다. 주파수가 높다는 것은 영교차율이 높다는 것을 의미하고 주파수가 낮다는 것은 영교차율이 낮다는 것을 나타낸다. 따라서 영교차율이 높으면 음성신호는 무성음이고, 반대로 영교차율이 낮으면 음성신호는 유성음이 된다고 하는 것이 일반화되어 있다. 그러나 영교차율은 A/D 변환기의 dc 옵셋값, 신호에 있는 60Hz 험(hum), 기음에 의한 음성신호의 여기, 잡음 등에 의해 영향을 받는다.

III. 음성신호의 스펙트럼

유성음은 성도가 진동해서 발생하는 것이기 때문에 성도의 공명 특성에 따른 영향을 받아 공명 주파수에 에너지 봉우리가 나타난다. 대략 4kHz 이하에서는 2-4개가 관찰되는데 낮은 주파수로부터 제 1 포만트(F1), 제 2 포만트(F2) 등으로 부른다. 제 1 포만트(F1)는 다른 포만트들에 비해 약 10dB 정도 높게 나타나며 0 주파수에서 제 1 포만트(F1)까지 에너지가 배가 되는 봉우리 형태를 띠게 된다. 그림 1에 유성음의 스펙트럼을 나타내었다.

IV. 음성신호의 자기 상관관계

자기 상관관계는 현재표본과 과거표본이 얼마나 유사한 관계에 있는가를 나타낸다. 단구간에 대한 자기 상관관계함수는 다음과 같다:

$$R(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m+k)w(n-k-m) \qquad (4\text{-}1)$$

여기에서 W은 창함수, N은 창함수의 길이, k는 시간지연 차수를 나타낸다.

식 (4-1)에 의해 얻어지는 자기 상관관계값은 창의 길이, 신호의 레벨에 따라 크기가 다르게 된다. 따라서 실제의 경우에는 다음과 같이 규준화하여 사용한다:

$$R_n(k) = \frac{\sum_{n=1}^{N} S(n)S(n-k)}{(\sum_{n=1}^{N} S^2(n))(\sum_{n=1}^{N} S^2(n-k))} \qquad (4\text{-}2)$$

여기에서 규준화된 자기 상관관계계수 $R_n(k)$의 범위는 -1과 1사이이다. $R_n(k) = 1$은 두 신호가 동일하다는 것을 나타내며, $R_n(k) = 0$은 불규칙한 백색잡음만이 존재하기 때문에 자기 상관관계가 존재하지 않는다는 것을 나타낸다. 그러므로 이 계수값은 음성신호의 유-무성음 구간 검출용 파라미터로 적용된다.

또한 자기 상관관계를 이용하여 미지의 신호를 모델링하거나 미지의 신호를 예측하는데 적용할 수 있다. 신호 예측법은 시간지연된 과거신호의 선형조합에 의해 현재신호를 예측한다. 선형 예측법 중에서는 시간지연된 과거신호 $S(n-1)$에 의해 현재신호 $S(n)$을 예측하는 다음과 같은 전극형(all-pole type) 시스템 모델링이 주로 사용된다:

$$S(n) = \sum_{n=1}^{M} a(k)S(n-k) \qquad (4\text{-}3)$$

여기서 M은 모델링 차수이고 $a(k)$는 k차 지연된 상관관계계수이다.

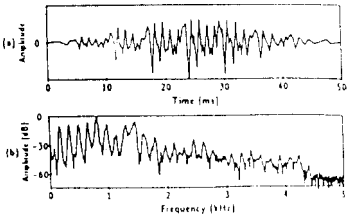

그림 1. 유성음의 파형 및 스펙트럼
FIG. 1. Waveform and Spectrum of voiced speech

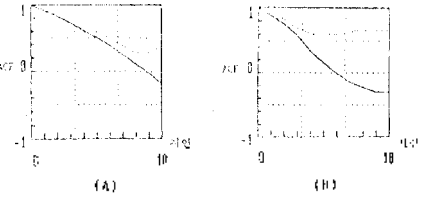

그림 2. 음성신호의 자기상관함수
FIG. 2. Autocorrelation function of speech waveform

(A) 여성화자
(B) 남성화자

현재신호를 정확하게 예측하려면 시스템의 모델링 차수를 증가시켜야 하는데 시간지연에 따라 계수값이 지수함수적으로 감소하게 된다. 또한 현재신호의 정확한 예측을 위해 예측차수를 크게 하면 시스템이 응답에 수렴하는 시간이 길어진다.

음성발성 모델링에서 음성신호의 생성은 인간두뇌의 음소조합에 의해 물리적 조음기관을 통해 이루어진다. 이전의 조음상태에서 현재 조음상태로의 변화를 나타내는 상관관계는 시간의 지연에 따라 다르게 나타난다. 그림 2는 8KHz로 양자화된 음성신호의 자기 상관관계를 나타낸다. 실선은 시간지연에 따른 과거신호들의 상관관계를 나타내고, 점선은 시간지연에 따른 과거신호와 미래신호의 평균에 대한 현재신호와의 상관관계를 나타낸다. 그림에서 알 수 있듯이 인접한 신호간의 상관관계가 다른 신호에 비해 아주 높고, 시간지연이 길어질수록 신호간의 상관관계는 아주 빠르게 감소된다. 따라서 현재신호으로부터 시간지연된 몇 개의 신호보다는 바로 인근한 두 신호를 이용하여 현재신호를 예측함으써 모델링 차수의 문제점을 해결할 수 있다. 따라서 본 논문에서는 신호간의 상관관계가 가장 높은 인근한 두 신호를 이용하여 저주파 신호를 예측하였다.

인근한 두 신호의 평균(합)은 저 주파수를 강조한 것이므로 예측은 저주파영역 성분 중 두드러진 포만트를 예측하게 된다.

V. 상위 포만트에 대한 ZCR-파라미터

음성신호에서 고주파 신호만을 분리해 내기 위해서는 저주파 신호를 예측하여야 한다. 따라서 현재신호에 대해 자기상관관계가 가장 높은 한 샘플 지연된 과거신호 $S(n-1)$과 한 샘플 후의 미래신호 $S(n+1)$와의 평균을 취하여 식 (5-1)과 같이 저주파 신호를 예측한다:

$$S'(n) = \frac{S(n-1) + S(n+1)}{2} \qquad (5\text{-}1)$$

식 (5-1)에 의해 예측된 신호와 현재신호의 차로 잉여에러 $E(n)$을 식 (5-2)와 같이 구할 수 있다:

$$E(n) = S(n) - \frac{S(n-1) + S(n+1)}{2}, \quad for \;\; 1 \le n \le N \qquad (5\text{-}2)$$

따라서 식 (5-2)에서 구한 잉여에러 $E(n)$은 음성신호의 고주파 신호로 나타낼 수 있다. 그림 3에는 남성화자가 발성한 숫자음 /삼/에 대한 음성신호 및 잉여에러의 스펙트럼을 나타내었다. 그림 3-(D)에 잉여에러에 대한 음성신호의 스펙트럼을 나타내었다. 그림 3-(D)의 스펙트럼을 보면 제 2 포만트와 제 3 포만트가 그림 3-(C)의 원래음성 스펙트럼보다 상대적으로 강조되어 있으며 낮은쪽 주파수가 감쇄되어 나타난다. 그러므로 잉여에러 $E(n)$은 음성신호에서 저주파 신호를 분리해 낸 고주파 신호임을 알 수 있다.

음성신호에 식 (5-2)를 적용하여 구한 잉여에러 $E(n)$의 영교차율을 구하면 음성신호의 제 1 포만트 주파수 성분은 감소되고 제 2 - 제 3 포만트의 성분이 갖고 있는 주파수에 비례된 값이 얻어진다. 무성음인 경우 불규칙한 잡음신호와 유사한 모양을 가지고 있으므로 영교차율이 높고, 유성음인 경우 음성피형이 준주기적인 성질을 가지고 있으므로 무성음보다는 영교차율이 낮게 나타난다.

VI. 실험 및 결과

이상의 과정을 컴퓨터 시뮬레이션하기 위해 IBM PC/386에 마이크로폰 입력이 가능하도록 12-비트 아날로그-디지탈 변환기를 인터페이싱하였다. 발성음은 남성화자와 여성화자를 통해 다음 문장을 발성하게 하면서 8KHz의 표본화율로 저장하였다.

발성 1) 28세 남성화자 : "숫자음 일 - 구"
발성 2) 23세 여성화자 : "숫자음 일 - 구"

본 실험에서는 한 프레임의 길이를 256으로 하였다. 먼저

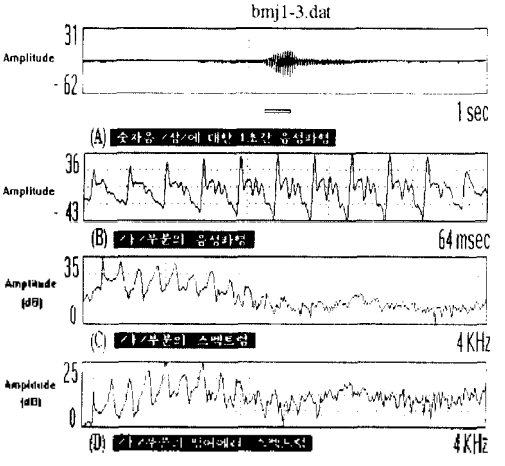

그림 3. 상위 포만트 ZCR-파라미터
FIG. 3. Higher formants ZCR-paramater

(A) 1초간 파형
(B) /ㅏ/부분의 512 샘플 음성파형
(C) /ㅏ/부분의 스펙트럼
(D) /ㅏ/부분의 잉여에러 스펙트럼

음성신호에서 고주파 성분을 대변하는 잉여에러 $E(n)$을 구하기 위해 현재신호 $S(n)$의 한 샘플 과거신호 $S(n-1)$과 한 샘플 미래신호 $S(n+1)$를 평균하여 저주파 신호를 예측한 후, 현재신호 $S(n)$과의 차로 잉여에러를 구하였다. 잉여에러의 영교차를 수행하여 음성구간 검출과 유-무성음을 분류하였다.

그림 4은 여성화자가 발성한 고립단어 숫자음 /삼/에 대한 처리결과이다. 그림 4-(A)에 고립단어 숫자음 /삼/의 1초간 파형을 나타내었다. 그림 4-(B)와 그림 4-(C)는 각각 숫자음 /삼/에서 /ㅅ/부분의 512 샘플 신호 및 스펙트럼이다. 그림 4-(D)는 숫자음 /삼/에 대한 /ㅅ/부분의 잉여에러 $E(n)$의 스펙트럼을 나타낸다.

그림 5는 여성화자가 발성한 그림 4와 같은 숫자음 /삼/에 대해서 에러신호의 규준화된 에너지비, 상위 포만트 ZCR-파라미터, 영교차율, 에너지 변화도를 처리한 결과이다. 그림 5-(D)에 있는 종래의 영교차율은 음성이 아닌 구간에서 무성음으로 판단할 수 있는 영교차율이 발생하여 무성음 구간과 묵음에 대한 구별이 쉽지 않으나, 본 논문에서 제안한 상위 포만트 ZCR-파라미터(그림 5-(C))인 경우 /ㅅ/부분에서 영교차율이 약 140개 이상 발생하여 무성음 구간임을 알 수 있으며, 유성음 구간과 무성음 구간사이의 영교차율이 70개 이상 발생하여 유성음과 무성음을 구별할 수 있었다.

VII. 결론

음성분석, 음성코딩, 음성인식 등에서 음성정보를 효과적으로 처리하기 위해서는 음성구간을 정확하게 검출할 수 있어야 하고 유-무성음 구간을 정확하게 분류하여야만 한다.

에너지를 이용하여 유-무성음 검출을 할 경우 창의 길이에 따라 영향을 받아 음성신호의 변화특성을 충분하게 나타내지 못한다. 시간영역의 음성신호를 주파수영역으로 변환시키면 고주파와 저주파의 경계영역은 창의 영향을 지배적으로 받게 된다. 또한 음성신호에 기움으로 인한 여기나 dc 바이어스가 존재하는 경우 종래의 영교차는 고주파 신호를 거의 검출할 수 없기 때문에 상위 포만트 성분을 제대로 나타낼 수 없다.

본 논문에서는 음성신호에서 고주파 신호의 영교차율을 검출하는 전처리용 상위 포만트 ZCR-파라미터를 제안하였다. 종래의 영교차는 제 1 포만트를 대변하므로 기음이나 dc 바이어스에 대한 영향을 받아 음성구간 검출이나 유-무성음을 잘못 판정하기 쉽지만, 상위 포만트에 대한 ZCR-파라미터는 음성신호의 상위 포만트 성분을 대변하는 고주파 신호에 대한 영교차율을 나타내므로 기음으로 인한 여기나 dc 바이어스의 영향을 받지 않는 장점을 가지고 있다.

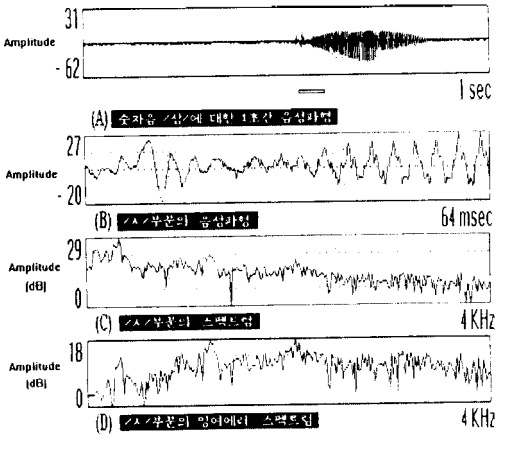

그림 4. 숫자음 /삼/에 대한 처리결과 (A)
FIG. 4. Result of utterance /3/ (A)

(A) 1초간 파형
(B) /ㅅ/부분의 512 샘플 음성파형
(C) /ㅅ/부분의 스펙트럼
(D) /ㅅ/부분의 잉여에러 스펙트럼

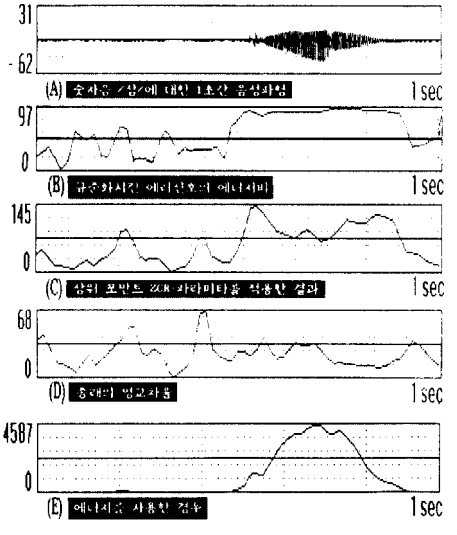

그림 5. 숫자음 /삼/에 대한 처리결과 (B)
Fig. 5. Result of utterance /3/ (B)

(A) 1초간 파형
(B) 규준화시킨 에러신호의 에너지비
(C) 상위 포만트 ZCR-파라미터를 적용한 결과
(D) 종래의 영교차율
(E) 에너지를 사용한 경우

REFFERENCE

[1] L.R.Rabiner & M.R.Sambur, "An Algorithm for Determining the Endpoint of isolated Utterearances", Bell Syste. Tech.j., Vol 54, pp.247-315, Feb 1975.

[2] L.R Rabiner & R.W.Schafer, *Digtal processing of speech Signals*, Prentice-Hall, Inc., Englewood Cliffis, New Jersey, 1978.

[3] B.V.Cox & L.K.Timothy, "Nonparametric Rank-Order Statistic Applied to Robust Voiced-Unvoiced Slience Classification", IEEE Trans, Acoust., Speech Signal Processing, Vol.ASSP-28, NO.5, October 1980.

[4] J.A Jankowski,Jr., "A New Digtal Voiced-Actived Switch, "Comsat Tech.Rev., Vol.43, no4, spring, 1976.

[5] L.F Lamel, L.R.Rabniner, A.E Rosenberg, and J.G. Wilpon, "An Imporved Endpoint Detection for Isolated Word Recognition", IEEE Trans, Acoust., Speech Signal Processing, Vol.ASSP-29, NO.4, august 1981.

[6] 정 영창, 배 명진, "배경잡음이 있는 음성신호에서의 Explicit 끝점검출", 호서대학교 논문집 제 6권, 1987.

[7] 이 미숙, 배 명진, 안 수길, "쌍 1차 차분을 통한 파형 코딩의 예측기에 관한 연구", 한국음향학회 11권 1호, 1991.

[8] 김 재우, 은 종관, "Forward/Backward 적응필터를 이용한 음질향상에 관한 연구", 한국음향학회 5권 1호, 1986.

[9] 유 건수, 김 건명, 배 명진, "쌍 자기 상관관계에 의한 음성신호의 끝점검출", 제 9회 음성신호 및 신호처리 워크샵 논문집(제 SCAS-9권 1호), 1992.