Level Crossing과 DPCM을 사용한 유성음/무성음/묵음의 분류

°김 진영, 성 굉모
서울대학교 공과대학 전자공학과

Voiced/Unvoiced/Silence Classification of Speech Signal by Level Crossing and DPCM

Jin Young Kim, Koeng Mo Sung
Dept. of Electronics Eng., Seoul Nat'l Univ.

요 약

시간 영역에서 만들어진 음성신호의 파라미터을 이용하여 주어진 음성신호의 구간이 유성음, 무성음, 혹은 묵음인지를 분류하는 새로운 알고리듬을 제시하였다. 이에 사용한 파라미터은 구간내에서 샘플링된 값의 절대치 합과 일정한 level 이상의 peak 의 합(T-peak), T-peak와 절대치 합의 비 그리고, DPCM 의 절대치 합들이다. 이들 파라미터를 이용하여 간단히 유성음/무성음/묵음 구간을 분류 할 였다.

Abstract

This paper proposes new algorithm for classifying speech signal frame into voiced, unvoiced, silence frame, using the parameters extracted from time domain behavior of speech signal. The prameters used in this paper are absolute magnitude, the sum of peaks lager than reference level (T-peak), the ratio of T-peak to absolute magnitude and the magnitude of signal outputs of DPCM. Using this parameters, speech signal is more easily classified into voiced/ unvoiced/silence frame.

I. 서론

음성처리신호 시스템은 그 계산 시간과 처리 과정을 간단히 하기위해 전처리 과정으로서 음성신호를 유성음/무성음/묵음 구간으로 분류하는 기법을 사용한다.

지금까지 음성구간을 유성음/무성음/묵음으로 분류하기 위하여 여러가지의 방법이 제시되어 왔는데, 그 중 대부분은 pitch 분석을 수행하여 유성음/무성음을 분류하였으나 그 방법이 복잡하고 많은 시간이 소요되는 단점이 있다. 한편, 통계적 방법으로 영교차율, log energy, autocorrelation, LPC 계수, LPC error등을 이용한 알고리듬등이 시행되어 왔으나, 이 또한 LPC나 autocorrelation등 복잡한 계산이 필요하고 표준 패턴을 구하기위해 많은 data양과 노력이 요구되는 단점이 있다[1]. 따라서 유성음/무성음/묵음을 특징짓는 간단한 파라미터들을 사용하여 분류하는 알고리듬이 필요하겠다.

유성음은 고주파 영역보다는 저주파 영역에서 에너지가 큰 것이 특징이다. 따라서 유성음은 시간 영역에서 저주파 성분(1'st formant)의 영향이 많이 나타나 영교차율이 적은 편이고 에너지가 큰 특징을 갖고 있다. 한편, 무성음은 영교차율이 묵음과 유성음 구간보다 많으며,그 통계적 특성이 colored noise의 성질을 가지고 있다. 묵음은 white

gaussian noise로서 근사시킬 수 있으며 따라서 영교차율이 크다할지라도 정해진 level이상의 peak 수가 무성음에 비하여 적다.

그러므로, 묵음구간에서 결정된 level을 기준으로한 교차 즉, level crossing을 사용하면 묵음, 무성음, 유성음의 분류가 용이하여진다. 그런데, 무성음과 유성음의 분류에 있어서는 영교차율이 비슷한 경우가 있다. 이것을 극복하기위해 유성음이 무성음보다 correlation이 큰 성질을 이용한다. 즉, 선형 예측을 사용하여 잔류 에너지를 고려하면, 무성음의 스펙트럼이 강조되어 유성음과의 분류가 가능해진다. 이를 위하여 가장 간단한 선형 예측 즉 DPCM을 사용하면 충분하다.

본 논문에서는 위와같이 level crossing와 DPCM을 이용하여 유성음/무성음/묵음 분류를 하고자 한다.

II. 유성음/무성음/묵음의 분류

유성음,무성음,묵음의 대표적인 파형을 보면 그림1과 같다. 그림에서 보는 바와 같이 무성음은 에너지가 작고 유성음은 크며, 묵음은 에너지 level이 무성음에 비하여 작거나 같다.

그런데, level crossing의 해(solution)에 의하면 level crossing density 는

$$\lambda_a = f_x(a)E[|\dot{X}|] = \lambda_0 \exp[-a^2/R(0)] \quad --(1)$$

단, a:level

R(k):autocorrelation

이다.

여기서 λ_0는 zero crossing density로써

$$\lambda_0 = \frac{1}{\pi}\sqrt{[-\ddot{R}(0)/R(0)]} \quad --(2)$$

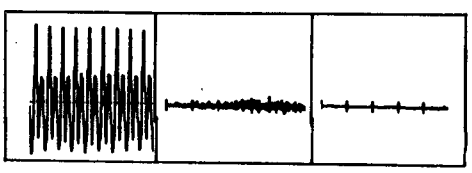


그림1. (a)유성음 (b)무성음 (c)묵음

$$\lambda_0^2 = -\ddot{R}(0)/R(0) = \frac{\int_{-\infty}^{\infty}\omega^2 S(\omega)d\omega}{\pi^2\int_{-\infty}^{\infty}S(\omega)d\omega} \quad --(3)$$

단, S(ω):입력신호의 spectral density

이다.

식 (1),(2),(3) 을 사용하면

$$\lambda_a^2 = \frac{\int_{-\infty}^{\infty}\omega^2 S(\omega)d\omega}{\pi^2\int_{-\infty}^{\infty}S(\omega)d\omega}\,EXP(-a^2/R(0)) \quad --(4)$$

이다.

주파수의 대역폭이 제한되어 있다면 (bandwidth=B)

$$\lambda_a^2 = \frac{\int_{-B}^{B}\omega^2 S(\omega)d\omega}{\pi^2\int_{-B}^{B}S(\omega)d\omega}\,EXP(-a^2/R(0)) \quad --(5)$$

이다.

식 (5)에서 분자가 ω^2으로 가중됨으로 고주파쪽이 강조되어진다. 그런데, 묵음은 white gaussian noise로 무성음은 colored noise로 성질이 표현됨으로 에너지가 같은 경우라면 고주파가 많은 무성음의 level crossing density가 크다. 예를 들어 묵음과 유성음의 spectral density가 그림 2와 같다면 다음과 같다.

$$\lambda_a^2(\text{묵음}) \propto 1/3 \times B^2$$

$$\lambda_a^2(\text{무성음}) \propto 1/3 \times (4(1-1/4^3)B^2$$

$$\lambda_a(\text{무성음})/\lambda_a(\text{묵음}) = 2$$

가 된다.

즉 무성음과 묵음의 에너지가 같더라도 특정 level 이상의 peak 수(또는 peak들의 합)가 무성음쪽이 더 크므로 무성음과 묵음을 분류할 수 있다. 위와 같은 성질은 무성음과 유성음의 분류에도 사용될 수 있으나, 유성음의 1st' formant

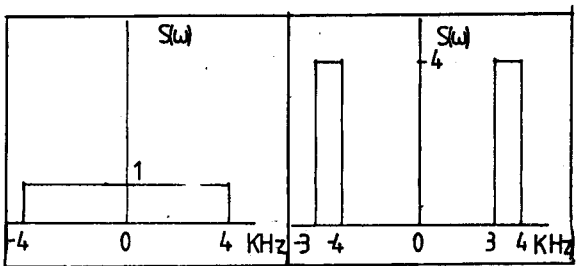


그림 2. (a) 묵음 (b) 무성음

주파수가 높고 큰 경우와 무성음의 기본음(ㅋ,ㄷ,ㅂ)과 같이 영고차율이 작은 경우 에는 분류가 어렵다. 따라서, 에너지의 도입이 불가피하다. 에너지 level을 강조하기 위해 1차의 선형예측을 사용하여 잔류 에너지를 고려하면 상관성이 작은 무성음이 강조된다. 여기서는 간단히 DPCM을 사용하면 충분하다. DPCM의 전달함수를 보면 z-영역에서 ($1-z^{-1}$)이 되어 고역 통과 필터인 것을 알 수 있다.

무성음은 대체로 2800Hz 부근에서 큰 에너지를 가지고 있다. 따라서, 심한 경우 8KHz 샘플링의 경우에 4KHz의 주파수를 갖고 있다면 DPCM의 후의 에너지는그림 3와 같이 2배 가까이 커진다.

그러므로, 무성음의 DPCM 후의 에너지는 원래 신호의 에너지보다 크거나, 작더라도 두 에너지의 차가 작다. 그러나 유성음의 경우는 DPCM의 에너지는 원 신호의 에너지보다 언제나 작다. 따라서 이 논문에서 사용하는 파라미터는 다음과 같다.

$$ASE=\sum_{i=1}^{n}|X(i)| \qquad ----(6)$$

단,n: frame의 길이

주어진 level이상의 peak의 합을 T 라하면

$$T\text{-}peak=\sum_{j=1}^{k}|X(j)| \qquad ----(7)$$

단,k: frame 안에서 level을 넘는 최대 peak의 합(그림 5 참조)

$$RATIO=ASE/T\text{-}peak \qquad ----(8)$$

$$DASE=\sum_{i=1}^{n}|X(i)-X(i-1)| \qquad ----(9)$$

$$DM=ASE-DASE \qquad -----(10)$$

이다.

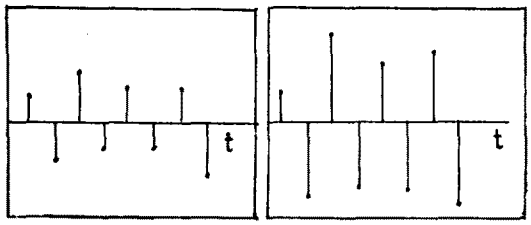


그림 3. (a) DPCM 전 (b) DPCM 후

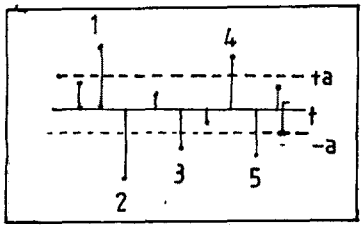


그림4. T-peak level =a

표1 유성음/무성음/묵음 분류

	ASE	T-peak	R	DASE	DM
묵 음	작다	매우 작다	작다	작다	
무성음	작다	작 다	크다	작다	>,<0
유성음	크다	매우 크다	작다	크다	>0

위의 파라미터를 사용 하면 유성음/무성음/묵음 분류를 할 수 있고 표1과 같다.

한가지 기억할 것은 DPCM은 white gaussian noise인 묵음을 무성음보다 더 억제하므로 묵음 검출에 병행시켜 사용할 수 있다.

III.알고리듬및 실험결과

위에서 제시한 유성음/무성음/묵음의 알고리듬은 다음과 같다.

(1) 샘플링 한 data를 frame 단위로 입음.

(2) ASE,DASE,T를 계산.

(3) DASE<thdase 이면 묵음으로 판정. (9)로 가라.

(4) T<tht 이면 묵음으로 판정. (9)로 가라.

(5) DM,R을 계산.

(6) DM<0 이면 무성음으로 판정. (9)로 가라.

(7) DM<thdm,R>thr 이면 무성음으로 판정. (9)로 가라.

(8) 유성음으로 판정.

(9) (1)로 가라.

위의 (1)-(9)와 같은 과정으로 유성음/ 무성음/ 묵음 분류를 한다. 위 알고리듬에서 tht, thdase는 처음 묵음 구간에서 정해진다. 따라서

묵음의 성질에 관계없이 분류가 정확해질수 있다.

본 실험에서는 50 개의 숫자음을 사용하였다. Data의 수집은 4KHz LPF를 통과시킨 후 8KHz로 샘플링하여 12bits D/A 변환기를 사용하였다. 한 구간은 128 샘플(16 msec)로 하였다.

실험 결과로서 그림 5은 'ㅅ' 음에 대한 DPCM전과 후의 파형을 나타내어 준다. 그림에서 보듯이 파형의 크기가 커진 것을 알 수 있다. ('ㅅ'의 zero crossing rate= 44-90/frame)

그림 6은 최종적인 결과로서 '삼' 음에 대한 무성음/유성음/묵음 분류를 보여주고 있다. 그림에서 보듯이 정확한 판정을 함을 알 수 있다.

50개의 숫자음에 대하여 실험한 결과 오분류는 대부분 종성의 비음에서 발생하였다. 그 이유는 음성의 끝 부분은 에너지가 작으며 비음의 경우는 autocorrelation이 커 DPCM 후의 에너지가 작기 때문이다.

IV. 결론

본 논문은 한국어 음성인식 시스템을 위하여 음성 구간을 유성음, 무성음, 묵음으로 분류하는 간단한 알고리듬을 제시하였다.

사용 파라미터는 일정 level이상의 peak, peak들의 합, DPCM의 잔류 에너지들이다. 이들 파라미터은 시간 영역에서 간단한 계산으로 구해짐으로 다른 기법에 비하여 빠른 속도로 분류할 수 있다. 또한 level crossing의 level과 분류의 threshold (tht, thdase)이 처음 입력되는 3개의 묵음 구간에서 정하여 짐으로 묵음구간의 통계적 성질이나 크기에 관계 없이 유성음, 무성음, 묵음 분류를 정확히 할 수 있다.

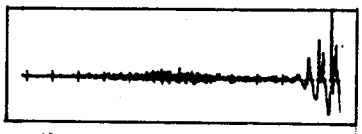
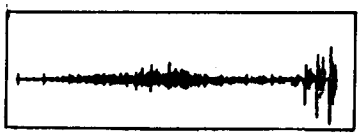


그림 5. 'ㅅ' (a) DPCM 전



(b) DPCM 후

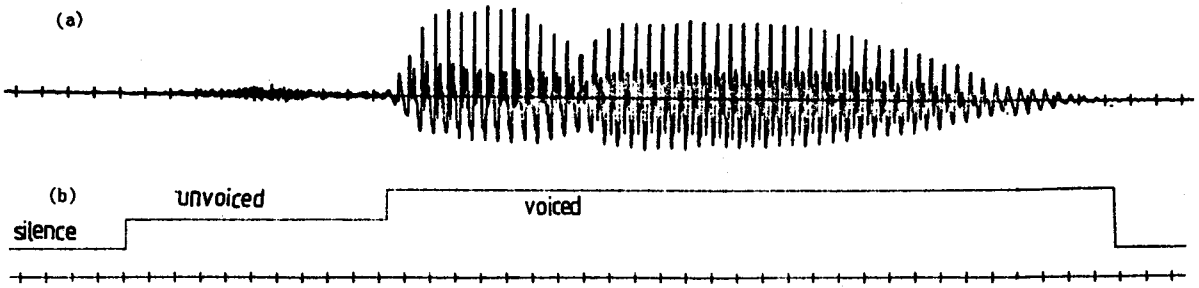


그림 6. (a) '삼' 음 (b) 무성음/유성음/묵음 분류 결과

V. 참고문헌

[1] Atal, B. S. and L. R. Rabiner, " A Pattern Reconition Approach to Voiced-Unvoiced-Silence Classification with Application to Speech Recognition," IEEE ASSP-24, pp. 201-212, 1976

[2] 은 종관, 김 현수," 잡음이 섞인 음성에서의 음성/무언의 구별," 한국음향학회지 제3권 제1호, pp. 35-42, 1984

[3] 배 명진, 안 수길," Spectrum 강조특성을 이용한 음성신호에서 Voiced- Unvoiced-Silence의 분류," 한국음향학회지 제4권 제1호, pp. 9-15, 1985