최대진폭의 Fluctuation에 의한 유성음구간 Explicit 검출

○

배 명 진 — 천안 호서대학 전자공학과

안 수 길 — 서울대학교 전자공학과

An Explicit Voiced Speech Classification by using the Fluctuation of Maximum Magnitudes

Myungjin BAE — Hoseo College

Soongil ANN — Seoul National University

ABSTRACT

Accurate detection of the voiced segment in speech signals is important for robust pitch extraction. This paper describes an explicit detection algorithm for detecting the voiced segment in speech signals. This algorithm is based on the fluctuation properties of maximum magnitudes in each frame of speech signals. The performance of this detector is evaluated and compared to that obtained from manually classifying 150 recorded digit utterances.

I. 서 론

피치를 구할 때 필요한 아주 중요하고 어려운 문제중에 하나는 음성신호에서 유성음 구간을 정확히 결정하는 문제이다. 유성음 구간을 정확히 구하면 피치추출의 정확도에 직접적인 영향을 주기때문이다. 지금까지 유성음구간을 검출하기위해 사용하는 패러미터들로는 유성음에서 성대의 떨림에 의해 발생하는 거의 안정된 주기적 성질을 이용하거나, 성도에서 나타나는 공명현상을 이용하기 위해 에너지의 펄스를 사용하고 있다.

그렇지만 이러한 패러미터를 사용하여 유성음 구간을 분류하려고 하면 몇가지의 문제점이 일어난다. 안정된 주기를 측정하는 경우에 파열음이나 천이기간이 존재하는 음에서는 안정된 주기가 구해지지 않아 에러가 발생하게된다. 또한 에너지펄스를 사용하여 유성음 구간을 분류할 때는 에너지가 낮은 유성자음 구간에서는 배경잡음의 에너지와 구분이 어렵게 되고, 저주파수의 에너지와 고주파수의 에너지가 서로 구분이 되지않기 때문에 유성음 구간을 분류하기가 어려워진다.

따라서 유성음이 갖는 두가지의 성질을 함께 대별할 수 있는 새로운 패러미터의 창안이 필요하다. 이때문에 본논문에서는 샘플된 음성신호의 한 segment안에 포함된 주기성분과 에너지값을 동시에 대별해 주는 새로운 패러미터를 제안하고자 한다.

2. 유성음의 성질

음성신호는 발성모델에 따라 무성음, 유성음, 묵음(silence)으로 분류될 수 있다. 무성음은 불규칙한 잡음이 성대를 자극하는 입력으로 되어 성대를 통과하는 동안 성대의 협착점에서 공명이 발생한다. 따라서 무성음의 스펙트럼에서도 2500Hz 근처에서 주된 공명봉우리가 존재하게 된다.

유성음은 준주기적인 glottal 펄스가 성도를

통해감으로서 발생되기 때문에 유성음 각 음소 마다 성대에서 고유한 공명이 일어난다. 따라서 유성음의 스펙트럼은 음소 마다 고유한 공명봉우리를 갖게된다. 이러한 공명봉우리를 포먼트(formant)라하고 낮은쪽 주파수에서 부터 두드러진 포먼트들을 차례로 제일, 제이, 제삼포먼트 등으로 순번을 붙인다. 유성음의 스펙트럼에서는 보통 제일포먼트의 주파수가 250-750Hz에 존재한다. 또한 유성음은 공명현상 때문에 무성음에 비해 에너지가 크고, 성대의 진동에 의해 성도의 여기가 되기 때문에 준주기성을 띠게된다. 성대의 진동주기를 피치라고 하며 남녀노소 및 주변조건에 따라 다르지만 2.5 - 25 msec 정도가 된다.

유성음의 진폭 스펙트럼을 보면 제일포먼트의 성분이 다른 포먼트의 것보다 에너지가 10-db 이상 크다. 이러한 관계를 시간영역에서 살펴보면 유성음의 파형이 대략 제일포먼트의 주파수 2배 정도로 DC-점을 교차하게된다[2]. 유성음과 마찬가지로 무성음에서도 주된 공명봉우리가 존재하기 때문에 그 파형은 주된 공명봉우리 주파수의 약 2배 정도로 유성음에서 보다는 많이 DC-점을 교차하게 된다. 이때 음성파형 s(n)에서 N-개의 샘플 구간 마다 대표적인 파형값 r(n)을 추출해 낸다면,

$$r(n)=typ|s(n),s(n-1),\ldots,s(n-N-1)| \quad ---(1)$$

여기서 typ|.|는 괄호 안의 N개 샘플값 중에서 대표값을 나타낸다. aliasing효과를 무시할때 대표값 r(n)은 N-구간 동안에 averaging되어 N-구간 사이에 존재하는 주기적인 성분은 없어지게 된다. 음성신호의 샘플링주파수를 8KHz로 할때 N=8 (=1msec)로 하여 대표값 r(n)을 구하면 1KHz 이상 성분은 이 대표값들의 파형에 나타나지 않으므로 유성음의 주기성분과 에너지만 남게된다.

실제로 aliasing현상이 나타나지 않게 N-개 샘플구간 동안에 대표값 하나를 유지하려면 음성신호를 저역 통과 여파기에 통과시킨 후에 down sampling을 수행해야만 한다. 이러한 개념은 무성음의 주된 공명현상을 제거시켜 버리는 것이므로 대표값이 유성음의 성질만을 대표하는 것이 아니라 DC-offset이나 60-Hz 혹도 포함될 수 있다.

음성의 스펙트럼은 주파수가 증가함에 따라 급한 기울기로 감소되고 있으며 공명현상에 의한 주된 공명봉우리가 존재한다. 음성을 저역통과 여파기에 통과시키지 않고 down sampling을 수행하면 높은쪽 주파수의 성분이 저주파 영역에 영향을 미치게된다. 이때 스펙트럼의 최대 봉우리에는 에너지 비율로 볼때 그 영향이 상대적으로 작게 미치기 때문에 주된 봉우리의 위치(주파수)에는 영향이 작게된다. 스펙트럼 상에서 최대봉우리는 시간영역에서 신호파형의 주된 형태를 결정짓기 때문에 음성파형의 N-구간 동안에 최대 진폭값을 대표값으로 사용하면 aliasing의 영향을 최소화 시킬 수 있게된다. 즉, N-구간에서 대표값은

$$r(n)=max|s(n),s(n+1),\ldots,s(n+N-1)| \quad ---(2)$$

여기서 max|.|은 N-개 샘플값에서 최대값을 나타낸다.

으로 구해질 수 있다. 숫자음 /칠/에 대해 N=8을 식(2)에 적용하여 +진폭과 -진폭에 대해 각각 구한 대표값들의 파형을 그림1에 나타내었다.

3. 대표값의 fluctuation성질

식(2)를 통해서 계산된 N-구간의 대표값(최대값)은 N-주기를 초과하는 신호의 성분을 그최대값 레벨의 DC-값으로 만들어버린다. 또한 각 최대값들의 파형은 유성음의 제일포먼트의 주기가 1/N로 줄어든 형태로 되고 파형의 진폭은 최대진폭값을 유지하게 된다. 유성음의 이러한 성질을 추출하려면 각 대표값 사이의 Fluctuation을 조사하면 된다. fluctuation을 계산하는 간단한 방법은 한프레임(=20*N 혹은 20msec) 동안에 대표값들 사이의 미분값을 계산하면 된다. 즉, 한프레임 안에서 fluctuation값 fl(.)은,

$$fl(n)=\sum_{k=n*20}^{n*20+N-1} |r(k)-r(k-1)| \quad ---(3)$$

이된다. 이렇게 계산된 fluctuation값은 음성신호의 진폭에 따라 값이 서로 다르다.

식(3)을 통해 구한 fluctuation값을 유성음 구간을 분류하는데 사용하려면 유성음 구간에 대한 문턱값이

구해져야만 한다. 유성음의 제일포먼트의 주파수 중에서 가장 낮은주파수는 250Hz 정도이므로 1-msec 안에서는 진폭의 기울기가 최소한 8-번 바뀌게된다. 따라서 1-프레임 안에서는 20*N 혹은 160번 진폭의 변화가 일어날 수 있다. 그 프레임 내에서 최대 진폭값이 MX라고 할때, 유성음 구간이 되려면 fluctuation의 크기는,

$$Thr = MX * 160 * 80\% \quad ---(4)$$

는 넘어야 한다. 여기서 80%를 곱한것은 경험적으로 구해진것이며 fluctuation의 변동을 허용하는 범위이다.

시뮬레이션에 사용한 데이타는 남자 3-명과 여자 2-명이 각각 3-번씩 발음한 숫자음 150-개의 고립단어이다. 파형을 눈으로 보면서 조사한 유성음 구간을 기준으로 했을 때 식(1)-(4)의 과정을 통해 분류된 유성음 구간이 1-프레임 편차 이내로 분류한 율은 149/150(=99.3%)였으며, Gaussian잡음을 6-dB 섞었을 때 6' 분류율은 142/150(=94.6%)가 얻어졌다.

4. 결 론

음성신호에서 유성음 구간을 분류하는 문제는 피치를 구할때 정확도를 좌우하는 중요한 문제로 음성신호처리 분야에서는 아주 필요하고 어려운 일이다. 유성음분류에 대한 대다수의 논문들은 피치를 구하는 과정속에 보통 포함시키고 있으며 사용하는 패러미터도 유성음이 갖는 주기성이나 혹은 에너지의 크기에 제한시키고 있다. 그럼지만 본 논문에서는 음성신호에서 유성음 구간을 explicit하게 검출하기 위해 음성파형의 일정 구간 안에서 최대값을 대표값으로 사용하여 그들간의 fluctuation을 패러미터로 사용하였다. 이 패러미터는 음성신호의 발성모델을 근거로하여 제안되었고 무성음이나 DC-offset에서는 fluctuation성질이 나타나지 않기 때문에 처리과정이 간단하면서도 유성음의 성질을 잘 대변해주게 된다. 그러나 이 패러미터는 최대진폭값에 근거를 두고있기 때문에 임펄스성 잡음의 영향에 아주 민감하므로 이러한 문제를 해결할 수 있는 개선안이 필요해진다.

[REFERENCES]

[1] L.R. Rabiner and Schefer, Digital Processing of Speech Signals, Prentice-Hall inc., Englewood Cliffs, N.J., 1978.

[2] Myungjin BAE and Souguil ANN, "The High Speed Pitch Extraction of Speech Signals using the Area Comparison Method", KIEE, vol.22, no.2, pp101-105, Feb. 1985.

[3] B.Gold and L.R. Rabiner, "Parallel Processing Technique for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Amer., vol.46, pp.442-448, 1969.

[4] J.D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. Audio Electroacoust., vo- l. AU-20, pp.367-377, Dec. 1972.

[5] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. Mc Gonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-24, pp.399-418, Oct. 1976.

[6] M. Lahat, R.J. Niederjohn, and D.A. Krubsack, "A Spectral Autocorrelation Method for Measurement of the Fundamental Frequency of Noise-Corrupted Speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-35, no.6, june 1987.

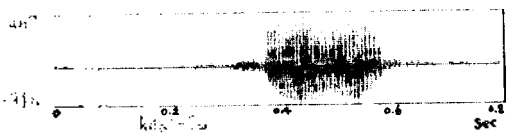

그림1 Waveform of the typical value per N=1 msec for speech /chil/.