Journal of Korea Multimedia Society Vol. 8, No. 1, January 2005(pp. 56-61)

연속음성인식을 위한 음성구간과 피치검출에 관한 연구

김태석†, 장종칠††

요 약

본 논문은 연속음성인식을 위한 음성구간과 피치를 검출하는 알고리즘을 제안한다. 이것은 연속음성을 입력받아 프레임단위로 자/모음을 구분하며, 구분된 유성음에서 피치를 검출하는 방법이다. 실제 잡음 환경에서 음성을 입력받아 적당한 문턱치 에너지를 사용함으로써 잡음환경에서 강인한 음성구간 추출이 가능하였고 추출한 음성구간에서 프레임단위로 영교차율과 단구간에너지를 이용한 알고리즘으로 유성음의 피치를 검출함과 동시에 자/모음을 구분하는 개선된 방식이다.

A Study on Speech Period and Pitch Detection for Continuous Speech Recognition

Tai-Suk Kim†, Chang jong chil††

ABSTRACT

In this thesis, propose speech period and pitch detection for continuous speech recognition. This mathod is distinguishes between vowel and consonant to frame unit in continuous speech, for distinguishable voice. Powerful extraction of speech period could threshold energy make use of input signal to real noise environment. Also algorithm of this method distinguish between vowel and consonant at the same time in voice make use of zero crossing rate and short time energy to extractible speech period.

Key words: Detecting Speech Period and Pitch(음성구간과 피치검출), Continues Signal(연속음성)

1. 서 론

음성을 통한 인간과 기계사이의 통신은 다른 정보통신 수단에 비해서 친숙하여 사용하기에 편리하나, 아직까지 만족할만한 결과를 얻지 못하고 있으며, 연속음성 인식의 경우에는 주위의 잡음환경에 의해 음성구간의 검출, 유성음과 무성음의 판별에도 상당한 어려움이 있다. 또한 대단위 단어의 연속음성인식에 있어서 단어 단위로 인식을 하면, 표준패턴 설정을 위한 메모리 사용과 계산시간이 상당히 요구될 것이다. 그러나 음소나 음절 단위로 인식을 한다면 아주 우수한 성능의 음성인식 시스템이 구현될 것이다. 그러기 위해서는 연속음성의 시작점과 끝점을 정확히 추출해야 하며, 유성음과 무성음의 구분도 명확하게 이루어져야 할 것이다.

음성은 성대의 진동의 유무에 따라서 유성음과 무성음으로 구분되며, 무성음은 공진이 발생하지 않을 정도로 빠른 속도로 공기를 압축하고 성도의 입구를 좁히면서 또는 한번 압축해서 난기류를 만들어 내는 소리로 성대의 떨림이 없으며 유성음에 비해서 많은 공기가 입으로 나온다. 유성음은 허파에서 압축된 공기가 성대(vocal folds)를 통과하면서 공진이 발생하

※ 교신저자(Corresponding Author) : 김태석, 주소 : 부산시 부산진구 엄광로(가야 산24번지)(614-714), 전화 : 051)890-1707, FAX : 051)890-1724, E-mail : tskim@deu.ac.kr
접수일 : 2004년 5월 4일, 완료일 : 2004년 7월 1일
† 종신회원, 동의대학교 공과대학 소프트웨어공학과 교수
†† 동의대학교 소프트웨어공학과 겸임교수

고 성대의 주기적인 떨림을 일으키면서 발생되는 소리이다.[1-4]

이러한 성대의 떨림을 피치(Pitch)라하며 음성신호 중에서 가장 기본이 되는 주파수로 시간축에서 보면 커다랗게 나타나는 피크들의 주파수를 의미한다. 피치는 인간의 청각에 매우 민감하게 반응하는 파라미터로서 음성신호의 화자를 구분하는데 사용한다. 따라서 정확한 피치의 검출은 복원 음질에 결정적인 역할을 하며 유성음과 무성음을 판별하는 파라미터로도 사용한다.[5]

본 논문에서는 단구간 에너지(STE:short time energy)를 이용하여 연속음성인식에서 기본적으로 요구되는 음성의 시작점과 끝점을 추출하였고, 단구간 에너지와 영교차율(ZCR:zero crossing rate)을 이용해서 유성음과 무성음을 구분한다.[6,7] 기존의 특정단어 인식이나 음절인식에서 벗어나 대화체 문장을 인식하는 연속음성인식분야에 이러한 알고리즘을 사용한다면 적은 메모리 사용과 처리속도가 빨라 실시간에 가까운 음성신호처리에 유용할 것이다.

2. 주기검출

2.1 실시간에 가까운 음성구간 검출

연속음성 인식에서는 우선적으로 입력된 음성에서 음성구간과 비음성 구간을 구분해야 하는데, 연속음성의 특성상 잡음환경에서도 정확한 음성구간의 검출이 가능한 강인한 알고리즘이 요구된다.

본 논문에서는 마이크로폰을 통해서 컴퓨터로 입력된 음성을 프레임 단위로 입력받아, 실험적인 단구간 에너지 문턱값(임계치, Threshold)를 적용하여 입력 신호가 먼저 음성인지를 구분하였고, 음성이면 단구간 에너지와 영교차율을 이용하여 유성음/무성음을 구분하였다.

음성신호의 일반적인 해석은 음성신호의 처리에 이용되는 정보가 저주파수 영역에 있다는 것을 착안하여 단구간 해석법이 사용된다. 음성신호의 주파수가 성대, 혀, 입술 등의 물리적 요인에 의해서 변화하므로 단구간 해석법을 이용한다. 이러한 단구간(10~30[ms])을 프레임이라 하며 단구간 내의 음성은 정상적으로 볼 수 있다.

음성과 비음성의 구별에 사용된 단구간 에너지의 문턱값은 잡음 환경에서도 검출이 가능하도록 입력되는 신호의 첫 3프레임의 합을 사용하였고 첫 3 프레임을 사용한 이유는 실험치로써 잡음환경에서 잡음신호만이 문턱값 계산에 포함되는 것을 막기 위한 것이다.

$$E_T = \sum_{i=1}^{3} S(i)^2 \tag{1}$$

또한 유성음/무성음 구분을 위한 단구간 에너지(STE)와 영교차율(ZCR)의 문턱값 역시 반복적인 실험을 통해서 가장 좋은 결과가 나온 신호의 첫 3프레임을 사용하였고 식 (2)와 (3)에 나타내었다.

$$STE_T = 3\sum_{i=1}^{3} S(i)^2 \tag{2}$$

$$ZCR_T = 2\sum_{i=1}^{3} N(i) \tag{3}$$

이러한 실험적인 문턱값의 적용방법은 실제 입력신호를 이용하였기 때문에 신호에 아주 큰 잡음이 없는 경우 대부분의 신호에서 실시간에 가까운 음성구간검출과 유성음/무성음의 구분이 가능하다. 그림 1과 2는 실제 음성에 대한 무잡음환경과 잡음환경에서의 음성구간추출에 대한 예를 보여주고 있다.

또한 본 논문에서는 입력되는 음성이 끝나는 시점 설정에서 음성이 아닌 신호가 연속적으로 15프레임이 입력되면 음성이 끝나는 것으로 간주하여 더 이상의 음성데이터를 입력받지 않았다.

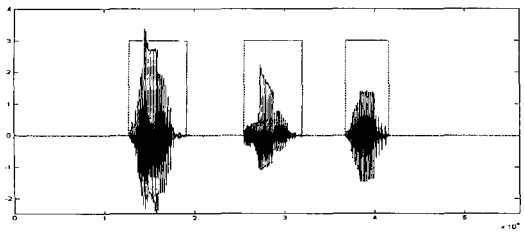

Fig. 1. 무 잡음환경에서의 음성구간검출(올챙이)

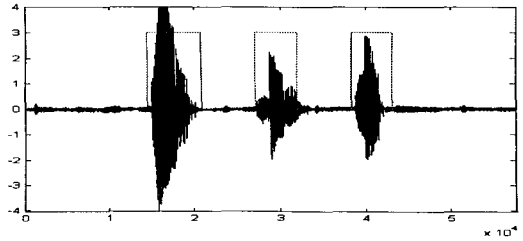

Fig. 2. 잡음환경에서의 음성구간검출(올챙이)

2.2 Short Time Energy

음성 신호의 진폭은 다소나마 시간에 따라 변화하는데, 무성음 분절(Segments)의 진폭은 일반적으로 유성음 분절의 진폭에 비해 훨씬 작으며 음성신호의 단구간 에너지는 이러한 진폭의 변화를 쉽게 나타내는데 다음과 같은 식으로 정의된다.[1-4]

$$E(m) = \sum_{m=-\infty}^{\infty} [x(n)w(n-m)]^2 \quad (4)$$

$$= \sum_{m=-\infty}^{\infty} x(n)^2 h(n-m) \quad (5)$$

여기서, $h(n) = w^2(n)$이 된다.

단구간 에너지는 무성음에서 작고 유성음에서는 크게 나타난다. 그러나 마찰음과 같은 일부 자음에서는 세기가 약한 모음보다 크게 나타날 수도 있다. 이 에너지는 유성음과 무성음의 구별에도 이용될 수 있으며 높은 음질의 신호에서는 묵음과 무성음의 구별에도 유용하게 쓰인다.

2.3 Zero Crossing Rate

스펙트럼에서 에너지가 집중되는 주파수를 찾는데 유용한 특징파라미터로 널리 사용되는 영교차율은 분석구간 프레임 내에서 신호 파형이 영점(Zero) 축과 교차하는 횟수를 말하며, 이산신호에서 연속 샘플링 값이 서로 다른 부호일 때 발생한다.

영교차율(ZCR)은 주어진 분석 구간 내에 음성 신호가 기준점인 영점 축을 교차하는 횟수를 말한다. 즉 이산 신호에서 연속 샘플링 값이 서로 다른 부호일 때 발생하는데, 이는 음성의 분할, 분석, 인식에 매우 유용하게 쓰인다. 영교차율은 화자의 성량에 적게 의존하며 음성 발생 시 성대에 의해 음성스펙트럼이 감쇄되므로 유성음은 낮은 주파수에 집중되므로 영교차율은 낮다. 이를 이용하여 유.무성음의 구별을 할 수 있다.

N개의 디지털 음성 신호로 구성되어 그에 따른 단구간 영교차율에 대한 일반식은 식 (6)과 같다. m은 분석 음성구간을 의미한다.[1-5]

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m)$$

$$w(n) = \begin{pmatrix} \frac{1}{2N} & 0 \le n \le N-1 \\ 0 & otherwise \end{pmatrix}$$

$$sgn[x(n)] = \begin{pmatrix} 1 & x(m) \ge 0 \\ 0 & x(m) < 0 \end{pmatrix} \quad (6)$$

sgn[x(n)] : 샘플링 값 w(n-m) : 창 함수(window function)

2.4 Pitch

시간축에서 크게 나타나는 피크들의 주파수로 음성신호 중에서 가장 기본이 되는 주파수이며, 성대의 주기적인 떨림에 의해서 생성된다.

피치는 인간의 청각에 매우 민감하게 반응하는 파라미터로서, 음성신호의 화자를 구분하는데 사용하며, 음성신호의 자연성(naturalness)에 큰 영향을 미친다. 따라서 정확한 피치 해석은 음성합성의 음질을 좌우하는 중요한 요인이며 음성코딩에 있어서도 피치의 정확한 추출과 복원은 음질에 결정적인 역할을 한다.[6,7]

2.5 Period Detecting method

음성의 피치는 전체 유성음에서 하나의 주기로써 나타내어진다. 따라서 본 논문에서는 유성음의 프레임마다 한 주기의 음성을 검출하여 피치를 검출하는 방법을 제안한다.

그림 3은 음성구간 검출과 피치 검출에 대한 전체 알고리즘을 나타낸 것이다.

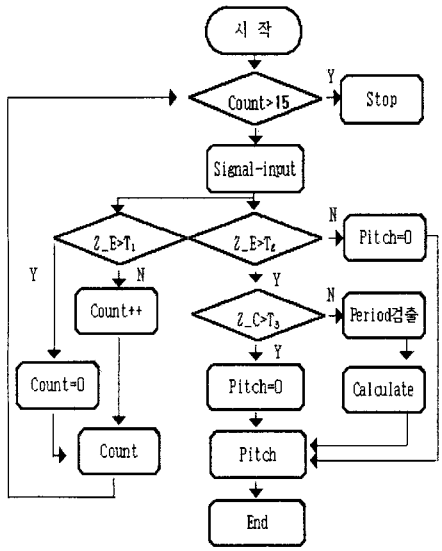

Fig. 3. 음성의 실시간 피치검출 전체 알고리즘

전체적으로 3개의 문턱값을 사용하여 음성구간과 유성음/무성음 구간을 검출하였다.

그림 4는 그림 3에서 피치 검출 부분을 상세히 나타낸 부분이며 그림 5는 신호파형에서 피치 검출의 예를 들고 있으며 유성음의 프레임이 입력되었을 때 한 주기의 검출은 가장 큰 최대점(maximum point)를 검출하고 최대점의 앞부분의 음성 중 가장 나중에 나오는 0점을 검출한다. 이 0점이 시작점이나 끝점이 된다.

찾은 0점에서 앞뒤 길이를 비교하여 길이가 긴 부분을 선택하여 다시 최대점을 찾고 0점을 찾는 과정을 되풀이하여 전체적인 한 주기를 검출한다.

여기서 프레임의 전체 크기는 400 표본점(sampling point) 이상이 되어야 한다.

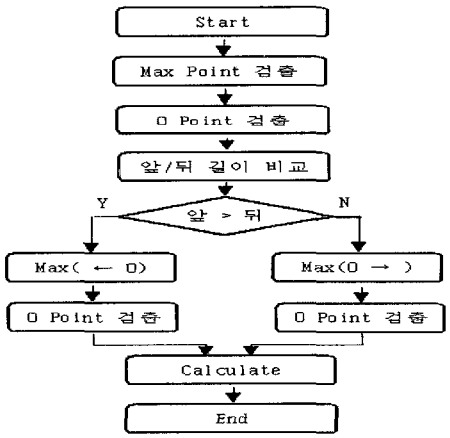

Fig. 4. 유성음의 한 주기 검출

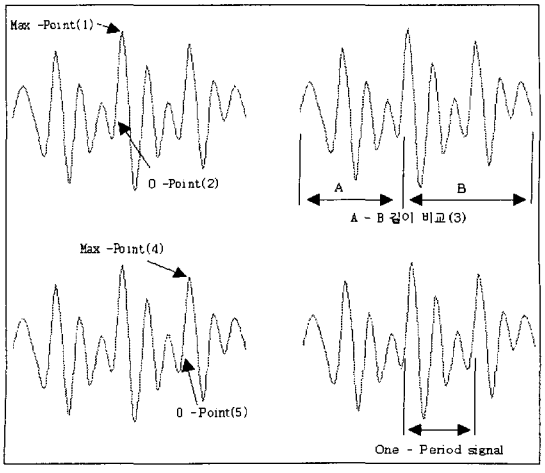

Fig. 5. Signal에서의 주기검출 요약도

예를 들어 16 [*kHz*]의 표본 주파수(sampling frequency)로 음성 신호를 입력받을 때 최소 프레임 단위는 0.03 [*s*]이상 되어야 한다.(16000×0.03=480)

그 이하이면 한 주기 이상이 프레임 내에 존재하지 않을 수 있기 때문이다.

한 주기가 검출되면 그 검출된 신호의 표본점의 수를 찾아 다음 식과 같이 계산함으로써 Pitch Frequency를 얻을 수 있다.

$$Pitch\ \ Freq = \frac{Sampling\ \ freq}{One\ \ Period}\ [\mathrm{Hz}] \qquad (7)$$

본 논문에서는 표본 주파수는 16 [*kHz*]를 사용하였고, 표본점을 480으로 하여 주기를 검출하였다. 그림 6, 그림 7, 그림 8은 연속음성에서 유성음 구간을 검출하여 피치를 계산한 결과를 나타낸 것이다.

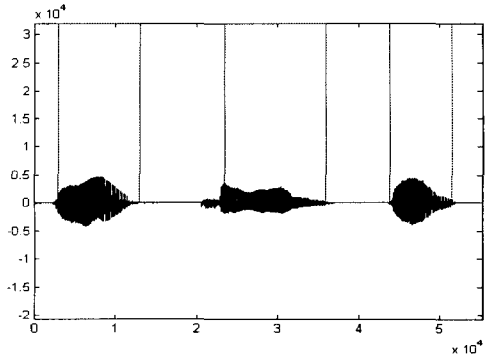

Fig. 6. 연속음성에서 유성음구간 검출

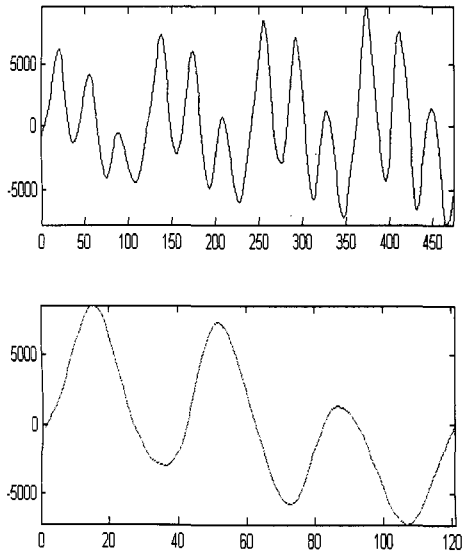

Fig. 7. '노' 프레임의 한 주기 검출(132.23[Hz])

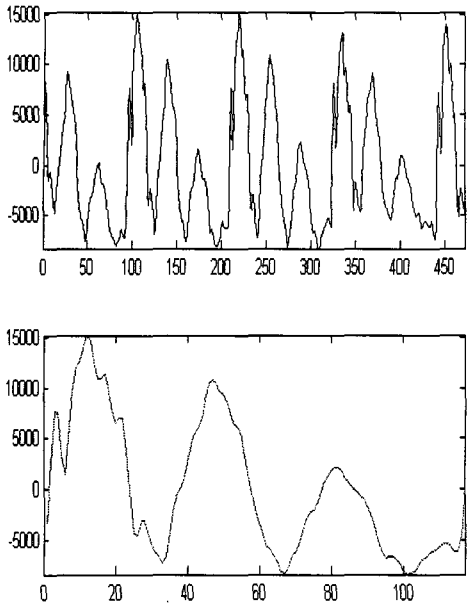

Fig. 8. 'H' 프레임의 한 주기 검출(136.75[Hz])

입력받은 음성의 프레임 단위 피치는 무잡음 환경과 잡음 환경에서의 변화를 그림 9와 10과 같이 얻을 수 있다.

연속으로 입력되는 숫자음성에 대한 음성구간 추출의 예를 그림 12와 13에서 보이고 있으며 음성구간만을 추출하여 단어인식 대상으로 사용한다면 좋은 인식결과를 나타낼 것으로 보인다.

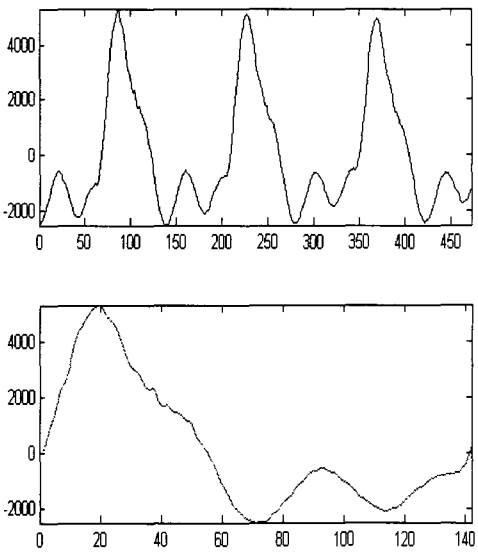

Fig. 9. 'I' 프레임의 한 주기 검출 (112.68[Hz])

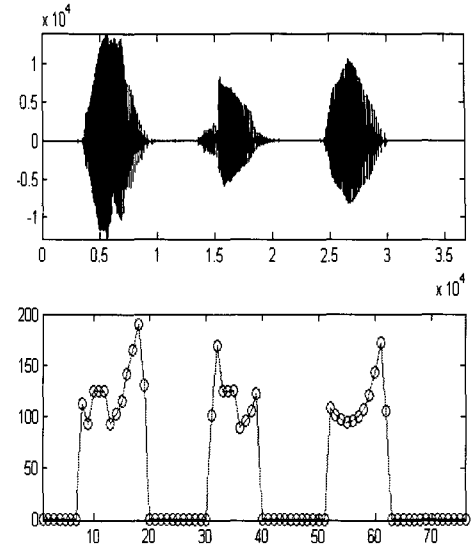

Fig. 10. 무 잡음 환경에서의 Pitch (올챙이)

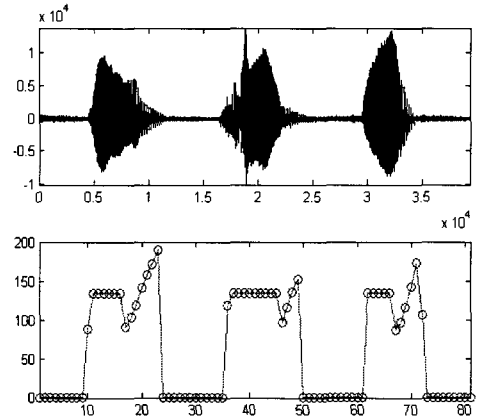

Fig. 11. 잡음 환경에서의 Pitch (올챙이)

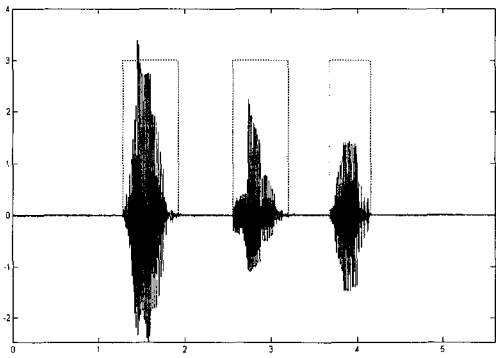

Fig. 12. 무잡음 환경에서의 연속 숫자음

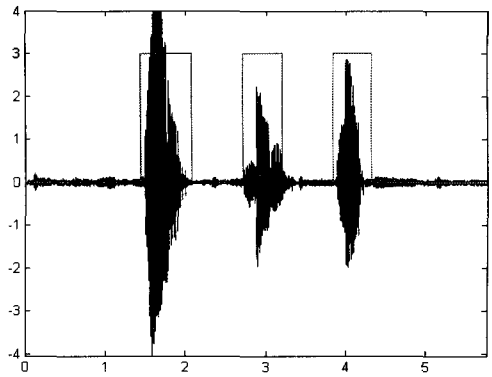

Fig. 13. 잡음환경에서의 연속숫자음

3. 결 론

본 논문에서는 연속적으로 입력되는 음성에 대해 실시간에 가깝게 유성음/무성음을 구분하였다. 본 논문에서 제안한 알고리즘은 간단하면서도 실제환경 잡음에 강인한 특성을 실험적으로 보였으며 처리속도가 빠른 특성을 보였다.

유성음의 구성은 거의 같은 주기의 연속이므로 한 주기를 검출하게되면 음소 인식에서 그 특징을 좀 더 정확하게 얻을 수 있는 방법이 될 것이다.

음성인식에서는 음성구간의 추출이 정확하게 이루어져야 하며, 특히 연속음성 인식에서는 본 논문에서 제안하는 방법을 사용하면 실시간에 가깝게 주기검출 및 유성음/무성음의 구분이 가능해질 것이다.

차후 음성인식부분과 연계하면 현재의 음성인식의 수준을 보다 향상시킬 수 있을 것이다.

본 논문에서의 주기검출 방법을 이용하여 연속음성 인식을 원활하게 하기 위해서는 부가적인 연구가 병행되어야 할 것이며 음성인식부분의 접목에 대한 연구가 진행 될 것이다.

참 고 문 헌

[1] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals,* 1978. Prentice Hall.1. C. Becchetti and L.P. Ricotti.

[2] *Speech Recognition(Thery and C++ Implementation),* 1999. John Wiley & Sons.

[3] D.J. Fucci and N.J. Lass, *Fundamentals of Speech Science,* 1999. Allyn and Bacon.

[4] B. Gold and N. Morgan, *Speech and Audio Signal Processing,* 2000. John Wiley & Sons.

[5] D. Jurafsky and J.H. Martin, *Speech and Language Processeing,* 2000. Prentice Hall.

[6] 유창동, "유성음/무성음 분리를 이용한 잡음처리", 한국음향학회지, 제21권, 제4호, pp. 374~379, 2002.

[7] 박정임, "잡음환경에서 우리말 연속음성의 무성자음 구간 추출 방법", 한국음향학회지, 제22권, 제4호, pp. 286~292, 2003.

김 태 석

1981년 경북대학교 전자공학과 졸업(공학사)
1989년 일본 KEIO대학 이공학부 계산기과학전공(공학 석사)
1993년 일본 KEIO대학 이공학부 계산기과학전공(공학 박사)
1993년 일본 국제전신전화연구소(KDD) 기술고문
1993년 일본 KEIO대학 이공학부 객원연구원
1994년~현재 동의대학교 소프트웨어공학과 교수
관심분야 : 정보시스템, 기계번역, 인터넷비즈니스

장 종 칠

1996년 2월 부경대학교 전자공학과(공학사)
1999년 2월 부경대학교 전자공학과(석사)
1999년 3월~현재 부경대학교 전자공학과 박사수료, 부경시스템 선임연구원, 동의대학교 소프트웨어공학과 겸임교수
관심분야 : 음성신호처리, 음성인식, 신경망