논문 2010-47SP-1-20

# 엔트로피와 하모닉 검출을 이용한 잡음환경에 강인한 음성검출

## ( Robust Voice Activity Detection in Noisy Environment Using Entropy and Harmonics Detection )

최 갑 근*, 김 순 협**

( Gab-Keun Choi and Soon-Hyob Kim )

요 약

이 논문은 잡음환경에서 음성인식률 향상을 위한 끝점 검출 방법에 대해 소개한다. 제안된 방법은 엔트로피와 음성의 하모닉 검출을 이용해 음성 구간과 비음성 구간을 검출한다. 음성의 스펙트럴 에너지에 대한 엔트로피를 사용하여 끝점검출을 하게 되면 비교적 높은 SNR 환경(SNR 15dB)에서는 성능이 우수하나 잡음환경의 변화에 따라 음성과 비음성의 문턱값이 변화하여 낮은 SNR환경(SNR 0dB)에서는 정확한 끝점 검출이 어렵다. 본 논문은 낮은 SNR 환경(0dB)에서도 정확한 끝점을 검출할 수 있도록 음성의 스펙트럴 엔트로피와 하모닉 성분을 검출하여 끝점을 검출하는 방법을 제안한다. 실험결과 기존의 엔트로피만을 이용한 방법보다 개선된 성능을 보였다.

Abstract

This paper explains end-point detection method for better speech recognition rates. The proposed method determines speech and non-speech region with the entropy and the harmonic detection of speech. The end-point detection using entropy on the speech spectral energy has good performance at the high SNR(SNR 15dB) environments. At the low SNR environment(SNR 0dB), however, the threshold level of speech and noise varies, so the precise end-point detection is difficult. Therefore, this paper introduces the end-point detection methods which uses speech spectral entropy and harmonics. Experiment shows better performance than the conventional entropy methods.

**Keywords :** Voice Activity Detection, End-Point Detection, Speech Recognition

## Ⅰ. 서 론

잡음환경에서 음성검출은 음성통신, 음성코딩, 음성인식분야에서 성능향상을 위해 그 중요성이 나날이 커지고 있는 분야이다. 배경잡음으로 인한 잘못된 음성검출로 인해 발생하는 성능저하를 막기 위해서는 잡음환경의 변화에 둔감한 문턱값의 설정이 매우 중요하다. 이러한 문제점을 해결하기 위해 아래의 방법들에 대한 연구가 활발한 상황이다.[1, 3]

1. Energy + ZCR(Zero Crossing Rate) threshold
2. Pitch Detection + Periodicity measure
3. Spectrum analysis + Cepstral Analysis
4. Chi-Square test
5. Entropy

문턱값에 기반한 끝점검출 방법은 SNR(Signal to Noise Ratio)의 변화에 능동적으로 대처하기 어렵다. 전통적으로 시간영역에서 음성의 끝점을 검출하는 방법인 Energy와 ZCR(Zero Crossing Rate)은 연산량에서는

* 학생회원, ** 정회원, 광운대학교 컴퓨터공학과
(Computer Engineering Department, Kwangwoon University)
접수일자: 2009년6월15일, 수정완료일: 2009년12월28일

우수하나 낮은 SNR에서 그 성능이 매우 떨어진다. 따라서 잡음환경에 안정적인 성능을 보이는 음성검출을 위해 음성의 주기와 기본주파수를 검출하는 방법과 주파수 영역에 대한 분석방법, 통계적인 접근 방법등이 주로 사용되며 특히 잡음환경에 비교적 강인한 성능을 보이는 통계적 방법을 많은 연구자들이 선호하고 있다.[1~4, 6~7, 9]

근래 들어서는 정보이론에 기초한 엔트로피를 이용한 방법등이 소개되면서 이와 관련된 연구가 진행되고 있다.[8] 그러나 엔트로피 역시 매우 낮은 SNR환경(SNR 0dB)에서 음성을 검출할 때 최대 스펙트럴 피크 에너지와 최소 스펙트럴 피크 에너지의 차가 적어 음성검출 성능이 다소 불안하다. 본 논문에서는 낮은 잡음환경에서 음성과 비음성을 구분할 안정적 성능의 음성검출을 위해 음성의 스펙트럴 에너지에 대한 엔트로피와 음성에 포함되어 있는 하모닉 성분을 검출한다. 음성의 하모닉 성분은 [그림 2]와 같이 비음성구간과는 확연히 구분되는 스펙트럴 피크 트렉을 갖고 있다. 스펙트럴 파크 트렉은 인간의 기본주파수에 대한 배음구조를 갖고 있어 음성과 비음성을 구분할 수 있는 좋은 수단이 된다. 따라서 본 논문은 음성의 스펙트럴 에너지에 기반한 엔트로피의 불안한 음성검출 성능을 음성의 하모닉 성분으로 보완하여 전체적인 음성검출 성능을 향상시켰다.

## Ⅱ. 에너지 스펙트럼 엔트로피 (Entropy of Energy Spectrum)

낮은 SNR 에너지 스펙트럼에서 음성영역은 비음성 영역에 비해 상대적으로 높은 에너지 스펙트럼을 나타낸다. 따라서 음성에너지 스펙트럼은 비음성 에너지 스펙트럼에 비해 상대적으로 높은 에너지 스펙트럼을 갖고 있다고 가정 할 수 있다. 이와 같은 에너지 스펙트럼은 셰넌(Shannon)에 의해 소개된 정보 엔트로피와 유사하게 표현할 수 있다.[6]

섀넌의 엔트로피는 식 (1)과 같이 정의된다.

$$H(S) = -\sum_{i=1}^{N} P(s(i) \cdot \log_2(P(s(i))) \qquad (1)$$

여기서 $N$은 심볼의 수, $s(i)$는 심볼 $i$, 그리고 $P(i)$는 심볼 $i$에 대한 사후 확률이다. 엔트로피는 스펙트럴 에너지 영역에서 식 (2)와 같이 정의 할 수 있다.

$$H(|Y(k,l)|^2) = -\sum_{k=1}^{N/2} \{P(|Y(k,l)|^2) \cdot \log_2(P(|Y(k,l)|^2))\} \qquad (2)$$

엔트로피의 계산을 위해 먼저 DFT(Discrete Fourier Transform)를 이용하여 이산 스펙트럴 파워를 계산한다. 여기서 $k$는 주파수 빈(Frequency bin)인덱스 이고 $l$은 프레임 인덱스이다. 주어진 프레임 $l$에서 주파수빈 $k$에 대한 스펙트럴 에너지 확률은 식 (3)과 같이 계산한다.

$$P(|Y(k,l)|^2) = \frac{|Y(k,l)|^2}{\sum_{k=1}^{N/2} |Y(k,l)|^2} \qquad (3)$$

구해진 각 주파수 빈의 확률은 식(2)에 의해 엔트로피로 계산되어진다.

[그림 1]은 SNR 0dB~15dB에 대한 엔트로피를 나타낸다. 여기서 [그림 1]의 (a)는 입력신호의 파형을 나타내고 [그림 1]의 (b)는 그에 대한 엔트로피 계산결과를 나타낸다. [그림 1]에서 보여 지는 바와 같이 SNR 15dB에서는 끝점검출을 위한 문턱값 설정이 비교적 용이하나 낮은 SNR에서 문턱값 변화가 심하여 안정된 성능을 갖기가 어렵다.

본 논문은 엔트로피의 이러한 문제를 개선하기 위해 음성의 하모닉을 이용하여 잡음을 추정하고 추정된 잡음을 차감하여 음성과 비음성을 구분하는 안정적 성능을 갖는 문턱값을 구하는 방법을 제안한다.

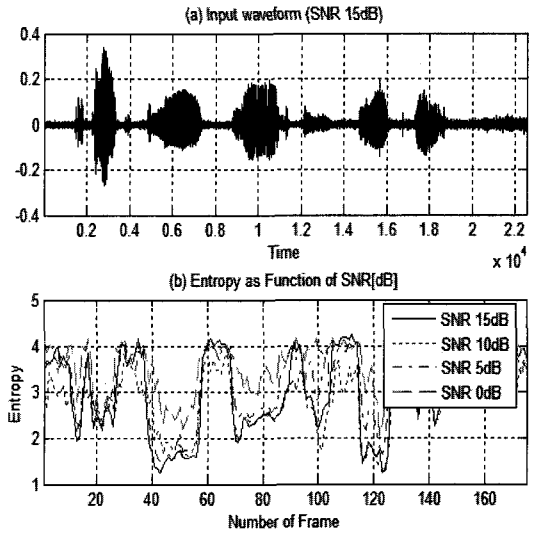


그림 1. SNR [15~0dB]에 대한 엔트로피
Fig. 1. Entropy for SNR 15~0dB.

## Ⅲ. 엔트로피와 하모닉을 이용한 음성 검출 (Voice Activity Detection Using Entropy and Harmonic)

음성의 하모닉(Harmonic)은 모음에서 주로 발견된다. 인간의 음성은 모음에서 성대의 공명을 통해 발생된 주기성이 강한 기본주파수(Fundamental Frequency)를 중심으로 정수배에 해당하는 주파수 영역에 하모닉 성분이 발견되며 이러한 특징은 음성과 비음성을 구분하는 중요한 요소이다. 또한 기본주파수는 피치(Pitch)라고도 한다.[8] 본 논문에서는 잡음과 음성의 구분이 뚜렷한 하모닉 성분과 잡음의 스펙트럴 에너지 특성과 음성의 스펙트럴 에너지 특성을 반영하는 스펙트럴 에너지 엔트로피를 이용하여 음성을 검출하는 EH-VAD(Entropy and Harmonic Voice Activity Detection) 알고리듬을 제안한다.

[그림 2]에서는 "한국"에 대한 스펙트로그램과 하모닉성분을 보여준다.

제안된 알고리듬을 처리하기 위해 입력신호에 대한 단구간 퓨리에 분석은 음성신호 $x(t)$를 식(4)와 같이 정의하고 주파수 영역에서 처리하기 위하여 식(6)을 이용하여 DFT(Discrete fourier Transform) 처리하여 식(5)와 같이 주파수 성분을 $[X_k]$로 표시한다.

$$x(t) = [x_k] = [x_0, x_1, x_2, \ldots, x_{N-1}] \tag{4}$$

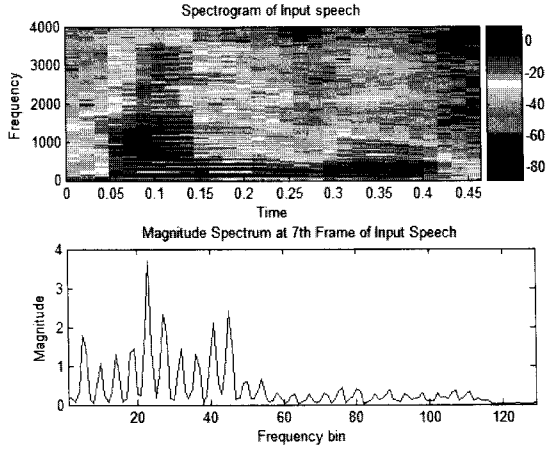


그림 2. 입력 음성에 대한 스펙트로그램과 입력음성의 7번째 프레임의 매그니튜드 스펙트럼

Fig. 2. Spectrogram of Input speech and Magnitude Spectrum at 7th Frame of Input Speech.

$$[X_k] = DFT[x_k] = [X_0, X_1, X_2, \ldots X_{N/2}] \tag{5}$$

$$X_k = \left| \sum_{n=0}^{N-1} x_n e^{-j(2\pi nk/N)} \right| \tag{6}$$

여기서 식(6)은 우측항이 복소수이므로 절대치를 $X_k$로 하며 $k$는 주파수 빈(frequency bin)의 인덱스(index)이다.

$$\begin{aligned} &\text{if} \quad X(k,l) > X(k-1,l),\ X(k,l) > X(k+1,l) \\ &then \quad H_{peak}(k,l) = X(k,l) \\ &else \quad H_{peak}(k,l) = 0 \end{aligned} \tag{7}$$

식(7)은 주어진 프레임 $l$에서 스펙트럴 피크를 계산하기 위한 것으로 각각의 주파수 빈에 대해 인접한 $k-1$ 주파수 빈과 $k+1$ 주파수 빈을 비교하여 $k$번째 주파수 빈이 크면 스펙트럴 피크로 본다. 그렇지 않은 경우에는 에너지를 0으로 주어 하모닉을 조사하기에 용이하도록 한다.

$$F_{freq} = F_k \times \frac{Fs/2}{N/2} (Hz) \tag{8}$$

식 (8)은 발견된 첫 번째 최대 에너지 피크 $F_k$를 기본 주파수로 보고 식(9)와 같이 하모닉 계산을 수행한다. 하모닉 성분은 배음구조이며 기본주파수의 정수배로 계산되어진다.

$$\overline{F} = \frac{Fs/2}{F_{freq}} \tag{9}$$

식 (10)은 주어진 프레임 내에서 하모닉 성분으로 선택된 스펙트럴 피크들의 평균을 계산한다.

$$H_\mu(l) = \frac{1}{N} \sum_{k=1}^{N-1} [\overline{F}(k+1,l) - \overline{F}(k,l)] \tag{10}$$

식(11)은 주어진 프레임 내에서 하모닉 성분으로 선택된 스펙트럴 피크들의 분산 값을 계산한다.

$$H_{\sigma^2}(l) = \frac{1}{N} \sum_{k=1}^{N-1} (\overline{F}(k,l) - H_\mu(l))^2 \tag{11}$$

알고리듬은 주어진 프레임에서 하모닉 성분으로 선택되어진 스펙트럴 피크들의 평균과 분산을 계산하고 식(12)를 이용하여 프레임 내에서 하모닉 성분들의 관

계를 조사한다.

$$H_d(l) = \frac{[\overline{F}(k+1,l) - \overline{F}(k,l)]}{H_{\sigma^2}(l)} \quad (12)$$

식(13)은 식(12)에 의해 조사된 스펙트럴 피크들의 거리의 합을 계산한다.

$$H_{har}(l) = \sum_{k=1}^{N-1} H_d(k,l) \quad (13)$$

식(13)에 의해 계산되어진 하모닉 가중치는 식(14)와 같이 식(2)에 의해 계산되어진 엔트로피에 곱해지게 된다. 이로서 엔트로피는 음성이 존재하고 있는 구간에서의 안정적인 끝점검출 문턱값을 설정할 수 있게 된다.

$$EH(l) = E(l) \times (H_{har}(l)) \quad (14)$$

[그림 5]에서 보여지는 바와 같이 식(14)의 계산결과는 잡음보다 음성이 상대적인 에너지가 크다고 가정된 상황에서 엔트로피는 비교적 우수한 성능을 보이나 임계점 설정구간의 변화에 의해 안정한 임계점 설정이 어렵다. 또한 낮은 SNR(0dB) 상황에서는 잡음과 음성의 에너지차가 적어 비음성 구간의 문턱값의 변동이 심하다. 음성의 하모닉 성분은 음성의 모음에서 발견되는 주요한 특징으로 이 하모닉 성분의 스펙트럴 특성을 계산한 것을 엔트로피에 곱하게 되면 음성구간으로 선택된 영역에서 매우 안정한 평탄도를 보이게 된다.

## Ⅳ. 실험 및 결과

엔트로피를 이용한 음성의 검출은 낮은 SNR환경에서 불안정한 검출성능을 보인다. 따라서 안정적인 검출성능을 보이기 위해 본 논문에서는 음성에 존재하는 하모닉 성분을 검출하여 엔트로피에 가중치를 주는 EH-VAD(Entropy and Harmonic Voice Activity Detection)방법을 제안했다. 제안된 알고리듬을 실험하기 위해 SNR(15dB, 10dB, 5dB, 0dB)로 구분된 NOIZEUS DB를 이용하였다.[5] 잡음의 종류로는 Car, babble, street등으로 구분된 신호를 사용하였으며 신호분석을 위해 8kHz 샘플링레이트, 16비트 양자화, 해밍윈도우를 사용하였고 실험을 위한 시스템 설계는 [그림 3]과 같다.

[그림 4]와 [그림 5]는 SNR 15dB입력신호에 대한 엔트로피와 엔트로피에 검출된 하모닉 성분을 곱한 결과

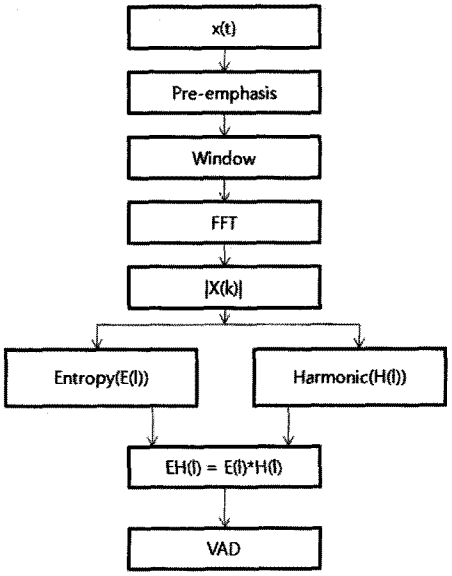


그림 3. EH-VAD 시스템
Fig. 3. Overview of EH-VAD(Entropy Harmonic Voice Activity Detection) System.

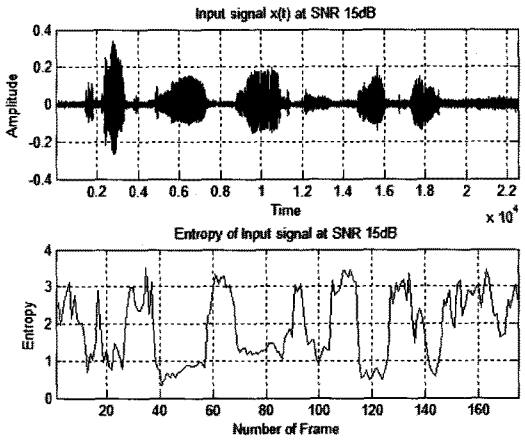


그림 4. SNR 15dB 입력 음성에 대한 웨이브폼과 엔트로피
Fig. 4. Wave form and Entropy of Input signal at SNR 15dB.

이다. [그림 5]에서 보여지는 바와 같이 엔트로피만을 사용했을 때보다 음성으로 검출될 영역에서 엔트로피 하모닉값이 안정한 특성을 볼 수 있다. 따라서 변이 폭이 매우 작은 구간을 임계지점으로 선택하게 되면 [그림 6]에서 보여지는 바와 같이 안정한 음성 검출이 가능하다.

[표 1]은 실험을 위해 사용한 고정 문턱값에서 SNR 변화에 따른 음성검출 성능을 보여준다. PHR(Pause Hit Ratio)는 비음성구간 적중률이고, FAR(False Alarm Rate)은 음성구간을 비음성 구간으로 오인한 오

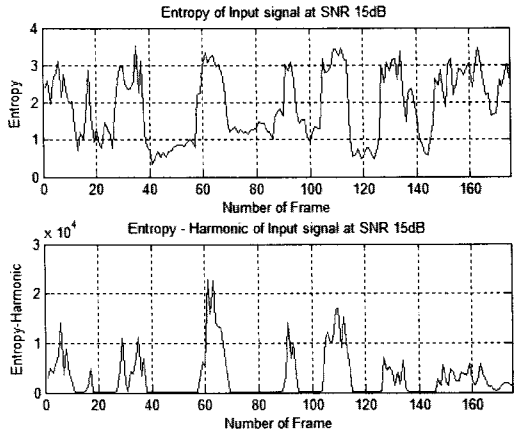


그림 5. SNR 15dB 입력 신호에 대한 엔트로피와 하모닉

Fig. 5. Entropy and Harmonic of Input signal at SNR 15dB.

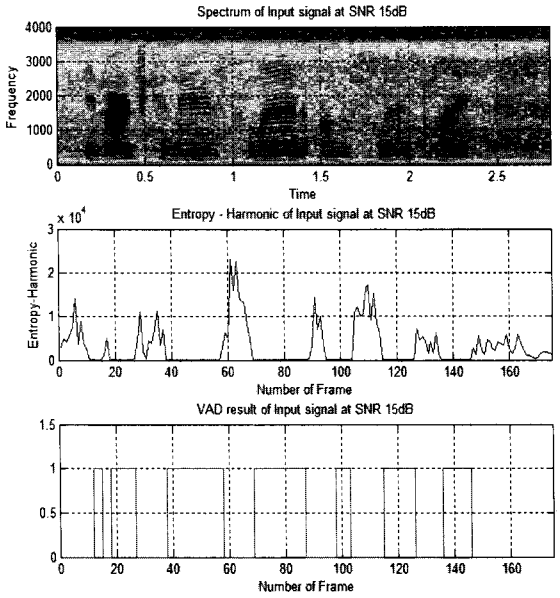


그림 6. SNR 15dB 입력 신호에 대한 엔트로피-하모닉과 음성신호 검출결과

Fig. 6. Entropy-Harmonic and VAD result of Input signal at SNR 15dB.

보율이다. 실험결과 제안된 EH-VAD 알고리듬이 기존의 엔트로피만을 이용했을 때보다 FAR이 개선된 것을 알 수 있다.

표 1. 제안된 알고리듬과 엔트로피를 적용한 알고리듬의 PHR과 FAR 분석결과

Table 1. Analysis of PHR and FAR result for the proposed VAD algorithm (EH-VAD) and conventional entropy technique.

| Noise | SNR(dB) | VAD Result(%) | | | |
|---|---|---|---|---|---|
| | | Entropy VAD | | EH-VAD | |
| | | PHR | FAR | PHR | FAR |
| Car | 0 | 74 | 30 | 89 | 13 |
| | 5 | 80 | 28 | 92 | 7 |
| | 10 | 88 | 23 | 96 | 5 |
| | 15 | 95 | 20 | 99 | 2 |
| Babble | 0 | 60 | 41 | 78 | 23 |
| | 5 | 77 | 31 | 88 | 10 |
| | 10 | 82 | 27 | 90 | 8 |
| | 15 | 89 | 23 | 91 | 8 |
| Street | 0 | 71 | 34 | 87 | 12 |
| | 5 | 78 | 30 | 90 | 11 |
| | 10 | 82 | 28 | 94 | 9 |
| | 15 | 92 | 20 | 95 | 7 |

## V. 결 론

본 논문에서는 낮은 SNR환경에서 엔트로피를 이용한 음성검출방법이 문턱값 설정이 어려운 것을 보완하기 위해 음성의 하모닉 성분을 가중치로 곱해주는 엔트로피-하모닉 음성검출 방법을 제안하였다. 제안된 방법을 Car, babble, street 잡음이 부가된 음성에 대해서 실험을 통해 PHR(Pause Hit Ratio)과 FAR(False Alarm Rate) 성능을 분석한 결과 기존의 엔트로피를 이용한 방법보다 우수함을 알 수 있었다.

## 참 고 문 헌

[1] 하동경, 조석제, 진강규, 신옥근, "엔트로피 차와 신호의 에너지에 기반한 잡음환경에서의 음성검출" *한국마린엔지니어링학회지*, 제32권 제5호, 768-774쪽, 2008년 7월

[2] Ahmed, B. Holmes, P.H., "A voice activity detector using the chi-square test", *Acoustics, Speech, and Signal Processing, 2004. Proceedings.*, pp. I-625-8, R. Melbourne Inst. of Technol., Vic., Australia, May 2004.

[3] L.R. Rabiner, M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", *The Bell System Technical Journal*, Vol. 54, No. 2, pp.297-315, 1975.

[4] Zoltan Tuske, Peter Mihajlik, Zoltan Tobler and

Tibor Fegyo, "Robust Voice Activity Detection Based on the Entropy of Noise Suppressed Spectrum" *Interspeech 2005*, pp. 245-248, Lisbon Portugal., september 2005.
[5] Yi Hu, Philip Loizou, "NOIZEUS Speech Corpus', http://www.utdallas.edu/~loizou/speech/noizeus/
[6] Abdallah I., Montresor S., Baudry M, "Robust speech/non-speech detection in adverse conditions using an entropy based estimator" *Digital Signal Processing Proceedings 1997*, pp. 752-760, Santorini Greece, Jul 1997.
[7] David Kozel, Constantin Apostoaia, "Colored Noise Reduction Using Bark Scale Spectral Subtraction, Statistics, and Multiple Time Frames" *IEEE EIT Proceedings 2007*, pp. 416-421, Chicago USA, May 2007.
[8] Ramalho, M.A. Mammone, R.J. "New speech enhancement techniques using the pitch mode modulation model" *Circuits and Systems, 1993 Proceedings of the 36th Midwest Symposium*, pp. 1531-1534, Detroit, USA, Aug 1993.
[9] 조규행, 박윤식, 장준혁, "Smoothed Global Soft Decision에 근거한 음성향상 기법" *전자공학회 논문지*, 제 44권, SP편 제 6호, pp. 734-739, 2007년 11월

## 저 자 소 개

최 갑 근(학생회원)
1999년 광운대학교 정보과학원 학사 졸업.
2002년 광운대학교 컴퓨터공학과 석사 졸업.
2006년 광운대학교 컴퓨터공학과 박사수료.
2008년~현재 에이펫 주식회사 IT사업부장
<주관심분야 : 음성인식, 오디오신호처리>

김 순 협(정회원)
1974년 울산대학교 전자공학과 학사 졸업.
1976년 연세대학교 전자공학과 석사 졸업.
1983년 연세대학교 전자공학과 박사 졸업.
1979년~현재 광운대학교 컴퓨터공학과 교수
<주관심분야 : 음성인식, 신호처리>