The Journal of Digital Policy & Management 2013 Jun; 11(6): 169-174

음성 에너지 최대화와 묵음 특징 정규화를 이용한 잡음 환경에 강인한 음성 검출

안찬식*, 최기호*
광운대학교 컴퓨터공학과*

Voice Activity Detection in Noisy Environment using Speech Energy Maximization and Silence Feature Normalization

Chan-Shik Ahn*, Ki-ho Choi*
Dept. of Computer Engineering, The University of Kwangwoon*

요 약 음성 인식 성능 저하의 문제는 모델 훈련 환경과 인식 환경의 차이이다. 이러한 환경의 불일치를 줄이기 위한 방법으로 다양한 묵음 특징 정규화 방법을 사용하고 있다. 기존의 묵음 특징 정규화 방법은 낮은 신호 대 잡음비에서 묵음 구간의 에너지 레벨이 증가하여 음성과 비음성에 대한 분류의 정확도가 떨어짐으로 인해 인식 성능이 저하되는 문제점이 있다. 본 논문에서는 음성 에너지 최대화와 묵음 특징 정규화를 이용한 잡음 환경에 강인한 음성 검출 방법을 제안하였다. 제안한 방법은 높은 신호 대 잡음비에서는 음성 에너지를 최대화시켜 특징이 잡음의 영향을 적게 받는 특성을 이용하였고 낮은 신호 대 잡음비에서는 음성/비음성의 켑스트럼 특징 분포 특성을 이용하여 인식 성능을 향상시켰다. 인식 실험 결과 기존 방법에 비해 향상된 인식 성능을 확인할 수 있었다.

주제어 : 음성 인식, 음성 검출, 잡음 제거, 음성 에너지 최대화, 묵음 특징 정규화

Abstract Speech recognition, the problem of performance degradation is the difference between the model training and recognition environments. Silence features normalized using the method as a way to reduce the inconsistency of such an environment. Silence features normalized way of existing in the low signal-to-noise ratio. Increase the energy level of the silence interval for voice and non-voice classification accuracy due to the falling. There is a problem in the recognition performance is degraded. This paper proposed a robust speech detection method in noisy environments using a silence feature normalization and voice energy maximize. In the high signal-to-noise ratio for the proposed method was used to maximize the characteristics receive less characterized the effects of noise by the voice energy. Cepstral feature distribution of voice / non-voice characteristics in the low signal-to-noise ratio and improves the recognition performance. Result of the recognition experiment, recognition performance improved compared to the conventional method.

Key Words : Speech Recognition, Voice Detection, Noise Reduction, Speech Energy Maximization, Silence Feature Normalization

* 본 논문은 2013년 광운대학교 교내 학술연구비 지원에 의해 연구되었음.
Received 30 April 2013, Revised 27 May 2013
Accepted 20 June 2013
Corresponding Author: Chan-Shik Ahn(The University of Kwangwoon)
Email: absoluti@kw.ac.kr

ISSN: 1738-1916

1. 서론

음성 검출기는 모바일 음성 통신 환경에서 다양한 어플리케이션과 결합되어 음성 인식이나 잡음 제거 알고리즘과 같은 음성 처리 시스템에 적용되고 있으며 시스템 성능에 주요한 영향을 미치는 핵심부분으로 발전되고 있다. 음성 검출 알고리즘을 통해 잡음 신호를 추정하여 잡음을 제거하는 잡음 제거 알고리즘에서는 음성 검출 성능이 잡음 제거 알고리즘 성능에 전반적으로 영향을 주는 중요한 요소로 작용한다[1].

음성 검출 알고리즘은 음성과 비음성 신호를 판별하기 위한 특징 파라미터를 구하여 적절한 문턱(theshold) 값을 특징 파라미터에 적용하는 결정식(decision rule)의 형태로 음성과 비음성을 구분한다[2].

음성과 비음성 검출을 위하여 묵음 특징 정규화 방법이 사용되며 스펙트럼 에너지(spectral energy), ZCR(zero-crossing ratio), LPC(linear prediction coefficients), 통계적 모델에 기반한 likelihodd ratio(LR) 등이 다양한 형태로 특징 파라미터를 검출하여 사용된다[3]. 기존의 묵음 특징 정규화 방법은 낮은 신호 대 잡음비에서 묵음 구간의 에너지 레벨이 증가하여 음성과 비음성에 대한 분류의 정확도가 떨어짐으로 인해 인식 성능이 저하되는 문제점이 존재한다.

따라서 본 논문에서는 음성 에너지 최대화와 묵음 특징 정규화를 이용한 잡음 환경에 강인한 음성 검출 방법을 제안하였다. 제안한 방법은 높은 신호 대 잡음비에서는 음성 에너지를 최대화시켜 묵음 특징이 잡음의 영향을 적게 받는 특성을 이용하였고 낮은 신호 대 잡음비에서는 음성과 비음성의 켑스트럼 특징 분포 특성을 이용하여 인식 성능을 향상시켰다. 인식 실험 결과 기존 방법에 비해 향상된 인식 성능을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 언급하고 3장에서는 음성 에너지 최대화와 묵음 특징 정규화를 이용한 잡음 환경에 강인한 음성 검출 방법에 대해 설명하며, 4장에서는 시스템 평가를 수행하고 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

2.1 스펙트럼 에너지(spectral energy)

낮은 SNR 에너지 스펙트럼에서 음성 영역은 비음성 영역에 비해 상대적으로 높은 에너지 스펙트럼을 나타낸다. 따라서 음성 에너지 스펙트럼은 비음성 에너지 스펙트럼에 비해 상대적으로 높은 에너지 스펙트럼을 갖고 있다고 가정할 수 있다. 이와 같은 에너지 스펙트럼은 섀넌(Shannon)에 의해 소개된 정보 엔트로피와 유사하게 표현된다[5].

섀넌의 엔트로피는 다음과 같이 수식적으로 나타낼 수 있다[5].

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

N은 심볼의 전체를 나타내고 $s(i)$는 i의 심볼을 나타내며 $P(i)$는 심볼 i에 대한 사후 확률을 나타낸다. 엔트로피는 스펙트럴 에너지 영역에서 다음과 같이 수식적으로 나타낼 수 있다[6].

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

엔트로피의 계산을 위해 DTF(Discrete Fourier Transform)를 이용하여 이산 스펙트럴 파워를 계산한다. 스펙트럴 에너지 확률은 음과 같이 수식적으로 나타낼 수 있다.

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

k는 주파수 빈(Frequency bin) 인덱스를 나타내고 l은 프레임 인덱스를 나타낸다. 프레임 l에서 주파수 빈 k에 대한 스펙트럴 에너지 확률을 계산한다. 계산되어진 각 주파수 빈의 확률은 식 (2)에 의해 엔트로피로 계산되어진다[7].

2.2 크리티컬 밴드(critical band)

주파수 측에서 청취 가능 범위는 최저 가청 한계와 최대 가청 한계 사이에서 영역을 가지고 있으며 음의 세기에 대한 범위 또한 청취 가능 범위와 같은 영역을 가지고 있다. 절대 최소 가청 한계는 잡음이 없는 환경에서 사람이 소리를 감지할 수 있는 최소 가청 한계를 H. Fletcher에 의해 정의되어 다음과 같이 수식적으로 나타낼 수 있으며 $T_q(f)$는 주파수 영역에서 최소 가청 한계를 나타낸다[8].

$$T_q = 3.64(\frac{f}{1000})^{-0.8} - 6.5e^{-0.6(\frac{f}{1000-3.3})} + 10^{-3}(\frac{f}{1000})^4 \quad (3)$$

마스킹 효과는 큰 음에 의해 작은 음이 차단되어 들리지 않게 되는 현상을 말하며 큰 음을 마스커(masker), 작은 음을 마스키(maskee)라고 한다. 마스키는 마스커에 의해 계산된 마스킹 임계치(masking threshold) 이상의 음만 가청된다[9].

주파수 마스킹은 큰 음을 나타내는 마스커와 작은 음을 나타내는 마스키가 동시에 발생되었을 경우 마스키가 마스킹 되는 경우를 말하며 주파수 분석을 통해 마스키가 마스킹되는 정도를 계산 할 수 있다. 주파수 마스킹은 주파수 대역에 따라 마스킹 되는 음의 임계치가 달라지고 동일한 마스킹 임계치를 가지고 있는 주파수의 대역폭을 크리티컬 밴드(critical band)라고 부른다. 크리티컬 밴드의 폭은 1kHz 이하의 주파수에서는 100Hz의 폭으로 거의 일정하고 1kHz 이상에서는 주파수에 비례하여 증가되어는 특징을 가지고 있다[10].

크리티컬 밴드 대역은 다음과 같이 수식적으로 나타낼 수 있다.

$$BW_c(f) = 25 + 75(1 + 1.4(f/1000)^2)^{0.69} \quad (4)$$

크리티컬 밴드의 넓이를 바크(Bark)라고 하며, 주파수 단위 $z(f)$를 바크 단위로 변환하기 위하여 다음과 같이 수식적으로 나타낼 수 있다.

$$z(f) = 13\arctan(0.0076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \quad (5)$$

바크 스케일 주파수는 음성의 중요 요소가 있는 대역에 대한 표현이 우수하다[11].

3. 잡음 환경에 강인한 음성 검출

3.1 음성 에너지 최대화

음성은 모음에서 피치 주파수(pitch frequency)를 가지며 피치 주파수를 기본 주파수라고 한다. 기본 주파수는 음성 영역의 전 대역에 걸쳐 에너지가 가장 큰 특징을 가지고 있어 최대 에너지로 나타나며 최소 에너지는 음성 신호와 무관한 잡음 신호로 나타난다. 잡음 신호는 주어진 프레임에서 최소 에너지로 나타낸다. 음성 에너지의 특징을 이용하여 음성 에너지의 최대화를 위해 주어진 프레임에서 최대 에너지를 다음과 같이 수식적으로 나타낼 수 있다.

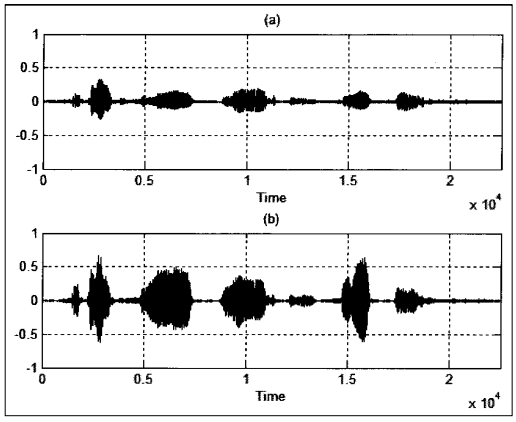
$$E_{\max}(l) = \min\{B(b_i, l), \ldots, B(b_{i-1}, l)\} \quad (6)$$

l은 해당 프레임 인덱스를 나타내고 b_i는 바크 스케일 인덱스를 나타낸다.

음성 에너지는 잡음에 비해 상대적으로 에너지가 크고 잡음은 에너지가 작기 때문에 음성 에너지와 잡음 에너지의 비율을 계산하고 출력에 대한 에너지 비율을 다음과 같이 수식적으로 나타낼 수 있다.

$$PSR(b_i, l) = \frac{B(b_i, l) - E_{\max}(l)}{\mu(l)} \quad (7)$$

음성 에너지와 잡음 에너지의 비율을 계산하여 비율을 나타낸다. [Fig. 1]은 음성 입력 신호와 음성 에너지 비율을 곱하여 음성 에너지는 잡음 에너지에 비해 상대적으로 신호가 증폭되어 진다.



[Fig. 1] (a) Input Speech Signal (b) The result of multiplying the input signal to the speech energy ratio.

3.2 묵음 특징 정규화

묵음 특징 정규화는 작은 로그 에너지 값을 갖는 묵음 구간을 찾아서 일정 값 이하로 줄여주는 방법이다. 로그 에너지 값을 다음 수식과 같이 무한 임펄스 응답

(IIR:Infinite Impulse Response) 필터를 통과시킨다.

$$\log\bar{E}(n)=\frac{1}{2}\left(\log E(n+1)-\log\bar{E}(n-1)\right) \quad (7)$$

$\log E(n)$은 n번째 프레임의 로그 에너지를 나타내고 $\log\bar{E}(n)$은 필터링의 출력 값을 나타낸다.

음성과 묵음에 대한 분류의 기준 값 T_0은 다음과 같이 수식적으로 나타낼 수 있다.

$$T_0=\frac{1}{N}\sum_{n=1}^{N}\log\bar{E}(n) \quad (8)$$

N은 각 음성 데이터의 프레임 수를 나타낸다. 정규화된 로그 에너지 $\log\bar{E}(n)$은 다음과 같이 수식적으로 나타낼 수 있다.

$$\log\hat{E}(n)=\begin{cases}\log E(n) & \text{if } \log\bar{E}(n)>T_0\\ \log(\epsilon)+\delta & \text{if } \log\bar{E}(n)\le T_0\end{cases} \quad (9)$$

식 (7)에서 계산된 $\log\bar{E}(n)$값이 기준 값 T_0보다 크면 음성으로 분류하여 로그 에너지 값을 원래의 값으로 유지하고, 작으면 묵음으로 분류하여 매우 작은 값 ϵ으로 정규화한다. ϵ은 상수 10^{-3}를 나타내고 δ는 평균 값 0, 분산 값 10^{-8}인 매우 작은 값을 나타낸다.

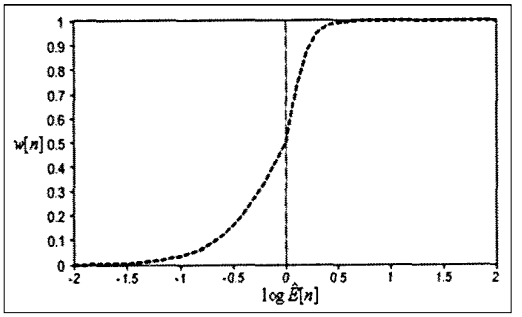
잡음에 오염된 음성 구간은 잡음만 존재하는 구간에 비해 더 넓은 주파수 대역폭을 가지고 있으며 큰 로그 에너지 값을 갖는 구간은 작은 로그 에너지 값을 갖는 구간에 비해 잡음의 영향을 상대적으로 적게 받기 때문에 가중 함수를 사용하였다. 가중 함수 $w(n)$은 다음과 같이 수식적으로 나타낼 수 있다.

$$w(n)=\begin{cases}1/(1+\exp(-(\log\bar{E}(n)-T_0)/\beta\sigma_1)) & \text{if } \log\bar{E}(n)>T_0\\ 1/(1+\exp(-(\log\bar{E}(n)-T_0)/\beta\sigma_2)) & \text{if } \log\bar{E}(n)\le T_0\end{cases} \quad (10)$$

출력 식 (9)에 가중 함수를 곱하여 음성 구간의 큰 값을 갖도록 가중치를 높이고 묵음 구간은 작은 값을 갖도록 가중치를 낮추게 되며 다음과 같이 수식적으로 나타낼 수 있다.

$$\log\tilde{E}(n)=w(n)\cdot\log\hat{E}(n) \quad (11)$$

수식에 의한 출력 결과는 [Fig. 2]와 같이 시그모이드 곡선의 형태로 나타난다.



[Fig. 2] Output Result of Weighting Function

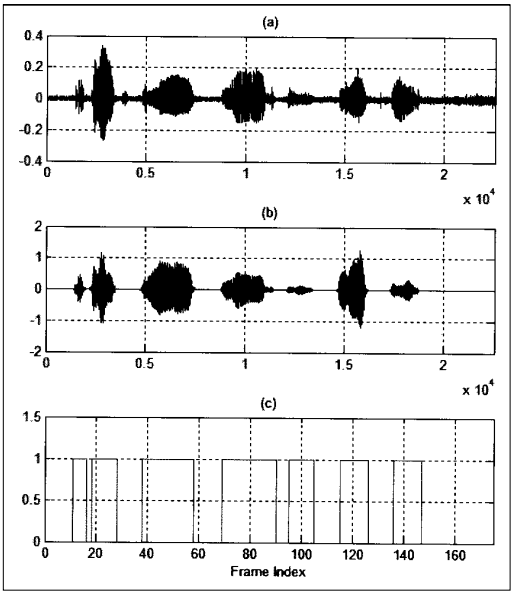
4. 실험 결과

본 논문에서 제안한 음성 에너지 최대화와 묵음 특징 정규화를 이용한 잡음 환경에 강인한 음성 검출 방법에 대해 인식 성능을 실험하였다.

평가를 위해 Aurora 2.0 데이터베이스를 사용하였다. Aurora 2.0에는 잡음 환경과 각각의 잡음 레벨로 구성되어 white gaussian noise, babble noise 등을 포함하며 street, airport, car noise 등 잡음 환경별로 구분되어 음성 향상 알고리듬의 성능 검증용으로 사용된다[11]. SNR 변화에 대한 성능을 검증하기 위해 잡음 환경(15dB, 10dB, 5dB, 0dB)을 구분하여 실험하였고, 끝점검출 성능 평가를 위해 비음성 구간에 대한 비음성 구간 적중률(PHR: Pause Hit Ratio)과 음성 구간에 대한 비음성 구간 실패율(FAR: False Alarm Ratio)을 사용하였다. 음원은 8kHz sampling rate, 16bit를 사용하였으며 FFT 크기는 256 샘플, 1/2 오버레핑(overlapping) 구간을 이용하였고 해밍 윈도우(Hamming Window)를 사용하였다[12].

실험 결과 [Fig. 3]의 (b)에서 보여 주는 바와 같이 SNR 15dB에서 음성 에너지부분이 증가한 결과를 확인할 수 있다.

잡음 구간에 대해서는 에너지를 0으로 주어 음성 구간과 비음성 구간이 분리되는 것을 확인할 수 있다. 표 1은 잡음 환경에 따른 음성 검출 성능을 평가한 결과이다.



[Fig. 3] (a) Input Signal of SNR 15dB (b) Results of Speech Energy Maximization and Silence Feature Normalization (c) Results of Speech Detection

〈Table 1〉 PHR and FAR for the SNR

Noise	SNR	VAD Result (%)
		PHR
Car	0	98
	5	98
	10	100
	15	100
Airport	0	86
	5	89
	10	93
	15	93
Street	0	89
	5	89
	10	95
	15	95

성능의 척도로 사용되어진 비음성 구간 적중률은 car 잡음 환경의 낮은 SNR 에서는 98%의 정확도를 보였으며 높은 SNR에서는 100%의 정확도를 보였다. Airport 잡음 환경의 낮은 SNR 5dB와 0dB에서는 각각 86%, 89%의 정확도를 보였으며 높은 SNR에서는 93%의 정확도를 보였다. Street 잡음 환경의 낮은 SNR에서는 89%의 정확도를 보였으며 높은 SNR에서는 95%의 정확도를 보였다.

5. 결 론

본 논문은 음성 에너지 최대화와 묵음 특징 정규화를 이용한 잡음 환경에 강인한 음성 검출과 인식 성능을 평가하였다. 다양한 환경적인 잡음이 존재하는 실제 잡음 환경이나 신호 대 잡음비가 낮은 음성 신호에 대해서는 피쳐 파라미터들이 잡음 신호에 민감하기 때문에 음성 검출의 성능이 저하되는 원인이 된다.

따라서 음성 에너지 최대화와 묵음 특징 정규화를 이용한 잡음 환경에 강인한 음성 검출 방법을 제안하였다. 제안한 방법은 높은 신호 대 잡음비에서는 음성 에너지를 최대화시켜 묵음 특징이 잡음의 영향을 적게 받는 특성을 이용하였고 낮은 신호 대 잡음비에서는 음성과 비음성의 켑스트럼 특징 분포 특성을 이용하여 인식 성능을 성능의 척도로 사용되어진 비음성 구간 적중률은 car 잡음 환경의 낮은 SNR 에서는 98%의 정확도를 보였으며 높은 SNR에서는 100%의 정확도를 보였다. Airport 잡음 환경의 낮은 SNR 5dB와 0dB에서는 각각 86%, 89%의 정확도를 보였으며 높은 SNR에서는 93%의 정확도를 보였다. Street 잡음 환경의 낮은 SNR에서는 89%의 정확도를 보였으며 높은 SNR에서는 95%의 정확도를 보였다. 인식 실험 결과 제안한 음성 에너지 최대화와 묵음 특징 정규화를 통해 음성 신호와 잡음 신호의 명확한 구분으로 인해 잡음을 제거율이 향상되어 기존 방법에 비해 잡음 환경에서 향상된 인식 성능을 확인할 수 있었다.

ACKNOWLEDGMENTS

This work was supported by the Kwangwoon University research fund of 2013.(KWU-과제번호)

REFERENCES

[1] Guanghu Shen, Hyun-Yeol Chung. Cepstral Distance and Log-Energy Based Silence Feature Normalization for Robust Speech Recognition. The Journal of the Acoustical Society of Korea. Vol 29, No. 4, pp. 278-285, 2010.

[2] Chan-Shik Ahn, Sang-Yeob Oh. Echo Noise Robust HMM Learning Model using Average Estimator LMS Algorithm. The Journal of Digital Policy and Management. Vol. 10, No. 10, pp. 277-282, 2012.
[3] Chan-Shik Ahn, Sang-Yeob Oh. Gaussian Model Optimization using Configuration Thread Control In CHMM Vocabulary Recognition. The Journal of Digital Policy and Management. Vol. 10, No. 7, pp. 167-172, 2012.
[4] Chan-Shik Ahn, Sang-Yeob Oh. CHMM Modeling using LMS Algorithm for Continuous Speech Recognition Improvement. The Journal of Digital Policy and Management. Vol. 10, No. 11, pp. 377-382, 2012.
[5] K. C. Wang, Y. H. Tsai. Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy. Second International Symposium on Universal Communication. pp. 423-428, 2008.
[6] Rix, A. W, Beerends, J. G, Hollier, M. P, Hekstra, A. P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. No. 2, pp. 749-752, 2001.
[7] Yi Hu, P. C. Loizou. Evaluation of objective quality measures for speech enhancement, IEEE Trans. ASLP, No. 16, pp. 229 - 238, 2008.
[8] K. S. Yao, E. Visser, O. W. Kwon, T. W. Lee. A Speech Processing Front-End with Eigenspace Normalization for Robust Speech Recognition in Noisy Automobile Environments, Proc. Eurospeech, pp. 9-12, 2003.
[9] C. F. Tai, J. W. Hung. Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments, Proc. ICSLP, pp. 2558-2561. 2006.
[10] Tuske Zoltan, Mihajlik Peter, Tobler Zoltan, Fegyo Tibor. Robust voice activity detection based on the entropy of noise-suppressed spectrum, in Proc. of INTER-SPEECH, pp. 245-248, 2005.
[11] S. Rangachari, P. C. Loizou. A noise-estimation algo-rithm for highly non-stationary errvironments, Speech Communication, Vol. 48, No. 2, pp. 220-231, 2006.
[12] Gab-Keun Choi, Soon-Hyob Kim. Voice Activity Detection Method Using Psycho-Acoustic Model Based on Speech Energy Maximization in Noisy Environments. The Journal of the Acoustical Society of Korea. Vol. 28, No. 5, pp. 447-453, 2009.

안 찬 식(Ahn, Chan Shik)

- 2002년 2월 : 광운대학교 컴퓨터공학과(공학석사)
- 2004년 8월 : 광운대학교 컴퓨터공학과(공학박사 수료)
- 관심분야 : 음성인식, 음성/음향 신호처리
- E-Mail : absoluti@kw.ac.kr

최 기 호(Choi, Ki Ho)

- 1973년 2월 : 한양대학교 전자공학과(공학사)
- 1977년 2월 : 한양대학교 전자공학과(공학석사)
- 1987년 2월 : 한양대학교 전자공학과(공학박사)
- 1977년 3월 ~ 1979년 2월 : 한국과학기술연구원(KIST) 전자공학부 연구원
- 2005년 1월 ~ 2005년 12월 : 한국ITS학회 회장
- 2006년 1월 ~ 2006년 12월 : 한국멀티미디어학회 회장
- 1979년 3월 ~ 현재 : 광운대학교 컴퓨터공학과 교수
- 관심분야 : 멀티미디어 응용, ITS, 상황인식
- E-Mail : khchoi@kw.ac.kr