Regular paper http://dx.doi.org/10.13067/JKIECS.2015.10.8.901

Bi-Level HMM을 이용한 효율적인 음성구간 검출 방법

장광우* · 정문호**

An Efficient Voice Activity Detection Method using Bi-Level HMM

Guang-Woo Jang* · Mun-Ho Jeong**

요 약

본 논문에서는 Bi-Level HMM을 이용한 음성구간 검출 방법을 제안하였다. 기존의 음성 구간 검출법은 짧은 상태변화 오류(Burst Clipping)를 제거하기 위하여 별도의 후처리 과정을 거치든가, 규칙 기반 지연 프레임을 설정해야만 한다. 이러한 문제에 대처하기 위하여 기존의 HMM 모델에 상태 계층을 추가한 Bi-Level HMM을 이용하여 음성구간 판정을 위해 음성상태의 사후 확률비를 이용하였다. 사람의 청각특성을 고려한 MFCC를 특징치로 하여, 다양한 SNR의 음성 데이터에 대한 평가지표를 활용한 실험을 수행하여 기존의 음성상태 분류법보다 우수한 결과를 얻을 수 있었다.

ABSTRACT

We presented a method for Vad(Voice Activity Detection) using Bi-level HMM. Conventional methods need to do an additional post processing or set rule-based delayed frames. To cope with the problem, we applied to VAD a Bi-level HMM that has an inserted state layer into a typical HMM. And we used posterior ratio of voice states to detect voice period. Considering MFCCs(: Mel-Frequency Cepstral Coefficients) as observation vectors, we performed some experiments with voice data of different SNRs and achieved satisfactory results compared with well-known methods.

키워드

Voice Activity Detection, Bi-Level HMM, Posterior Ratio, MFCC
음성 구간 검출, Bi-Level HMM, 사후 확률비, 멜주파수 캡스트럼 계수

Ⅰ. 서 론

최근 음성인식 기술이 스마트기기, 가전제품, 자동차의 음성 인터페이스 등 여러 분야에 적용됨에 따라 음성구간 검출(Voice Activity Detection, VAD) 기술의 중요성이 커지고 있다[1]. 음성구간 검출은 입력신호로부터 음성구간과 비음성구간을 구분하는 것을 방법으로서, 다양한 음성 기반 알고리즘의 전처리 기법으로 적용되어 음성분석, 음성인식, 제한된 대역폭에서의 효율적인 음성 부호화 등 음성신호처리 분야에 필수적인 것이다[1-5].

음성구간 검출이 어려운 것은 비정적인 잡음에 노출 된 여러 가지 복잡한 환경 속에서 음성구간의 시작점과 끝점을 찾아야 한다는 점 때문이다. 이러한 환경에서 최근의 통계적 모델 기반의 여러 가지 방법들은[6-10] 규칙기반의 초기 음성구간 검출 방법들에[11] 비해 우위를 가진다. 음성과 비음성의 통계적 모델을 이용하는 것에는 우도비(Likelihood Ratio)를 이

* 광운대학교 로봇학부(tieqima@naver.com)
** 교신저자 (corresponding author) : 광운대학교 로봇학부(mhjeong@kw.ac.kr)
접수일자 : 2015. 07. 08 심사(수정)일자 : 2015. 08. 13 게재확정일자 : 2015. 08. 23

용하는 것과[6,12], HMM(: Hidden Markov Model)[8-10], SVM(Support Vector Machine)[13]을 이용하는 것이 대표적이다. 그러나 문제가 되는 것은 실제 비음성 구간인데도 음성 구간으로 인식되는 짧은 구간(Burst Clipping) 혹은, 음성구간 중에 비음성 구간으로 간주되어지는 짧은 구간이 빈번히 발생한다는 것이다. 그래서 대개 통계적 모델로 음성구간을 검출한 후에 규칙 기반의 후처리 과정을 거쳐 이러한 짧은 구간의 오류를 제거한다. 하지만, 이러한 방법은 비정적 잡음에 대해 신뢰성을 주기 어렵고 실시간 처리를 요구하는 시스템에 적용하는데 무리가 있다.

실시간으로 이러한 문제를 다루는 전형적인 방법은 우도비를 사용하여 음성구간을 검출하고 음성(혹은 비음성) 상태에서 비음성(혹은 음성) 상태로의 전환에 특정 길이의 지연 프레임을 부여하는 것이다. Sohn[7]은 HMM 기반 Hang-over 기법을 고안했는데, HMM의 Forward 알고리즘에 의한 Posterior Probability의 비를 이용하는 것이다. 그러나 HMM의 전형적인 방법이 상태의 Duration을 나타내는데 단점이 있고, 제안된 방법은 여전히 이 한계를 가진다. Veisi[8]는 HMM 기반의 우도비로 음성상태를 판별한 다음, 특정 길이 미만의 상태변화를 무시하는 Heuristic Hang-over 기법을 도입했다. 좋은 성능을 나타냈으나, 여전히 잡음의 상태에 연관된 특정길이를 어떻게 정하는가 하는 문제가 있다. Tiaw[14]는 HMM 기반의 상태판별과 Hang-over 기법을 통합한 Bi-Level HMM을 고안하여 시각 기반 발화 검출에 적용했다. 이것은 짧은 오류 상태인식 구간을 제거하는 기존의 후처리 과정이나 규칙 기반의 방법들에 의존함이 없는 하나의 통합된 통계적 모델이라는데 의미를 가진다.

본 논문에서는 Bi-Level HMM을 실시간 음성구간 검출에 적용하였다. 특징치로는 주파수 특성의 다양한 연관관계와 청각특성을 고려한 MFCC를 사용하였다[15]. Tiaw[14]는 상태의 Posterior 확률을 이용했지만, 본 논문에서는 Posterior의 비를 사용함으로써 더욱더 평활한 음성구간을 검출할 수 있었다.

2장에서는 MFCC 특징치 추출에 대한 내용을 정리하였고 3장에서 음성구간 검출을 위한 Bi-Level HMM 적용을 다루었다. 4장에서는 본 제안한 방법의 실증을 위한 실험 내용 및 결과를 보여준다. 이에 따른 결론을 5장에서 설명한다.

II. MFCC 특징치 추출

2.1 Mel-Frequency Cepstral Coefficents

Mel-Frequency Cepstrum이란 음성 주파수의 비선형 Mel Scale의 로그 파워 스펙트럼을 기반으로 선형 변환하는 과정을 말한다. Cepstrum과 구분되는 점은 MFC의 Frequency Band가 Mel Scale상에서 균일하게 분할되어 일반적인 Cepstrum의 선형구간의 Frequency Band 보다 사람의 청각 시스템에 더 가깝다는데 있다[11]. 다음과 같은 과정을 거쳐 구한다.

① 고주파 강화 : 발성과정에서 성대와 입술의 영향에 의해 감쇠되는 고주파 성분의 보상

② 프레임 분할 : 샘플링 된 N개의 음성신호를 하나의 관측단위로 묶어 주는 것을 말함. 8 khz로 샘플링 하고 한 프레임을 25ms로 할 경우 200 개의 샘플이 한 프레임을 구성함. 프레임 사이의 이동은 10 ms로 하여 중첩이 일어나게 함.

③ Hamming Window 적용 : 프레임의 연속성을 향상시키기 위해 식 (1)과 같은 연산을 적용함.

$$S'(n) = S(n) * W(n),\ 0 \le n \le N-1, \tag{1}$$

여기서, $W(n) = 0.54 - 0.4\cos(2\pi n/(N-1))$, $S(n)$은 샘플된 음성신호임.

④ FFT 적용 : 음성신호를 시간영역에서 주파수 영역으로 변환함.

⑤ Mel 필터링 : FFT를 통해 얻은 스펙트럼에 대해 M(=20) 개의 삼각 대역 필터링을 적용하고 그것의 로그 스펙트럼을 구함.

⑥ DCT 적용 : Mel 필터링된 스펙트럼에 DCT(: Discrete Cosine Transform)을 적용하여 최종 13개의 계수를(MFCC) 구함.

2.2 MFCC 특징치 벡터 정의

DCT를 통해 구한 13개의 계수 중 첫 번째 계수는 프레임 에너지에 해당하는 것으로 잡음의 크기에 민감하므로 제외한다. 그리고 음성구간의 검출 정확도와 관계가 적은 열 번째부터 열세 번째 계수를 제외한 여덟 개의 계수를 특징치 벡터로 정의한다[15].

III. Bi-Level HMM 기반 음성신호 모델링

3.1 확률 기반 음성상태 정의

앞서 말한 바와 같이 입력된 음성신호는 제한된 대역폭에서의 효율적인 음성부호화, 음성인식 등을 위해 음성 구간과 비음성 구간으로 나눌 필요가 있다. 이 때, 음성 구간이라 할지라도 단어와 단어의 사이, 무성음 등에서는 비음성 구간의 특성을 가지고 있다. 또, 비음성 구간에서도 비정적 잡음에 의해 음성 구간의 특성이 나타날 수 있다. 본 논문에서는 이러한 점을 고려한 확률적 모델링을 위해 다음 두 가지를 가정한다. 첫째, 음성신호는 MFCC 특징치 벡터의 확률적 분포로 정의되는 활성과 비활성의 두 가지 상태로 나누어진다(식 (2) 참조). 둘째, 1차 마코프 모델 특성을 갖는 음성 상태와 비음성 상태가 존재하고, 각 상태는 활성 상태와 비활성 상태의 빈도수에 의해 결정된다(식 (3) 참조). 즉, 음성 상태는 활성 상태의 빈도수가 비활성 상태보다 많은 것이고, 비음성 상태는 비활성 상태의 빈도수가 상대적으로 많다는 것이다.

$$q_t^l:\ p(o_t|q_t^l),\ q_t^l \in \{active, \in active\} \qquad (2)$$
$$p(o_t|q_t^l) = \sum_{k=1}^{K} w_k G(o_t, \mu_k^l, \Sigma_k^l),\ \sum_{k=1}^{K} w_k = 1$$

$$q_t^c: P(q_t^l|q_t^c) = B_{q_t^l q_t^c},\ q_t^c \in \{speech, nospeech\} \qquad (3)$$

여기서, o_t는 8차원 MFCC 벡터, B의 행은 음성 상태와 비음성 상태를 나타내고 B의 열은 활성 상태와 비활성 상태를 나타낸다.

위의 두 번째 가정은 비정적 잡음에 의한 짧은 시간의 상태변화 오류를 없앨 수 있는 핵심적인 것이다. 이것에 의해 연속된 활성상태(혹은 비활성)에서 가끔 짧게 비활성(혹은 활성)으로 갔다가 다시 활성상태로 돌아오는 것을 음성 상태(혹은 비음성 상태)의 연속으로 유지시킬 수가 있기 때문이다.

3.2 Bi-Level HMM

은닉 마르코프 모델(Hidden Markov Model)은 관측 벡터의 시간열에 대한 시공간 확률분석 모델로서 서로 독립된 관측벡터 계층과 숨겨진 상태열의 두 계층으로 나누어진다. Bi-level HMM은 상태 계층과 관측벡터 계층 사이에 또 하나의 상태계층을 추가한 것으로 그림 1과 같다[14].

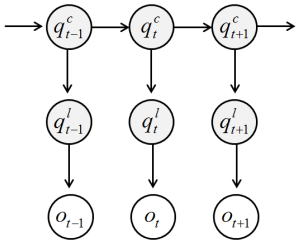


그림 1. Bi-Level HMM 확률 그래프 구조
Fig. 1 Probabilistic Graph Structure of Bi-Level HMM

확률 기반 음성상태를 Bi-Level HMM 구조에서 보면, 상위 상태 계층에서 음성 상태와 비음성 상태가 1차 마코프 모델의 관계를 가진다. 상위 각 상태에 정의된 빈도수(B)에 따라 활성 상태, 비활성 상태가 결정되고 하위 상태 계층을 이룬다. 관측벡터 계층은 활성, 비활성 상태에 따라 확률적으로 분포하는 독립적인 관측벡터로 구성된다.

3.3 음성 상태 추론

전방향 상태 추론은 현재까지 입력된 관측벡터로부터 상태를 재귀적 방법 추정하는 것으로 다음과 같다[14].

$$\alpha_t(j^c) = P(o_1 \cdots o_t, q_t^c = j^c|\Theta) \qquad (4)$$
$$= \sum_{j^l=1}^{L^l} \left[B_{j^c j^l} P_{j^l}(o_t)\right] \sum_{j^l=1}^{L^l} \left[A_{i^c j^c}\, \alpha_{t-1}(i^c)\right]$$

여기서, i^c, j^c는 각각 $t-1$, t에서의 상위 계층 상태를 말하며, j^l은 t에서의 하위 계층 상태(활성, 비활성)를 나타낸다. A는 상태천이 행렬로서, $A_{i^c j^c}$는 $P(j^c|i^c)$로 정의된다. Θ는 A와 B 및 활성/비활성 상태의 관측벡터 확률분포 파라미터로 구성된다(식 (2) 참조). 3.2절에서 말한 하위 상태 계층의 도입으로 인한 짧은 시간의 상태변화 오류 제거는 실제 식 (4)의 $B_{j^c j^l}$항에 기인한다.

시간 t에서의 사후확률(Posterior Probability)은, 식 (5)와 같이 전방향 상태추론에 비례한다. 이에 따라 음성 상태는 확률적으로 식 (6)과 같이 정할 수 있다.

$$P(q_t^c|o_1 \cdots o_t) = \frac{P(o_1 \cdots o_t, q_t^c)}{P(o_1 \cdots o_t)} \propto P(o_1 \cdots o_t, q_t^c) \quad (5)$$

$$q_t^c = speech \quad \text{if } \alpha_t(speech) > \alpha_t(nospeech) \quad (6)$$
$$nospeech \text{ otherwise}$$

한편, 상태변경에서 문턱을 두기위하여 우도비와 같이 전방향 상태비를 도입하면 식 (7)과 같이 음성 상태를 판정하게 되고 이와 같이 문턱을 두면 상태의 빈번한 변화를 방지하는 효과가 있다.

$$q_t^c = speech \quad \text{if } \Gamma(t) = \frac{\alpha_t(j^c = speech)}{\alpha_t(j^c = nospeech)} > \xi \quad (7)$$
$$nospeech \text{ otherwise}$$

Ⅳ. 실험결과

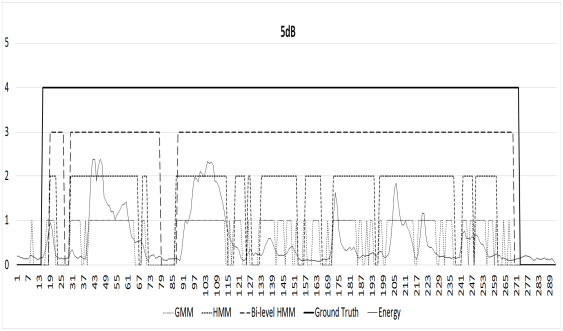
자동차 소음환경에서 다양한 SNR로 구분된 음성 데이터베이스부터[1]을 이용하여 MFCC 특징치를 추출하였다. 본 논문에서 제안하는 기법은 GMM에 의한 우도비 및 HMM 방법과 비교 실험되었는데, 타 두 가지 방식은 음성 상태의 짧은 상태변화 오류(Burst Clipping)를 제거하기 위한 별도의 후처리 과정을 거치지 않았다.

그림 2는 SNR에 따른 각 방법의 음성상태 추정결과를 나타내고 있다. 음성 데이터가 연속된 단어의 문장으로 이루어져 있어서 단어와 단어 사이를 비음성 구간으로 판정하는 경우가 많음을 알 수 있다. 이것은 음성인식, 전송을 위한 음성부호화 등의 성능저하에 연결될 수 있다. 제안한 방법은 다른 방법에 비해 월등히 뛰어난 결과를 보여준다.

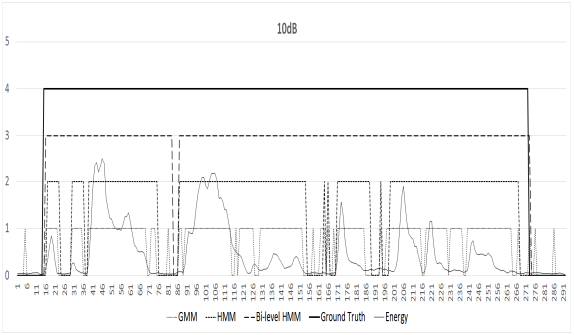
제안한 방법의 평가를 위해 음성구간 검출율(P_D), 짧은 상태변화 오류(P_{sc})를 지표로 활용하였다(식 (8-9) 참조). 이 지표를 이용한 ROC(: Receiver Operating Characteristic) 공간은 그림 3과 같다. 본 논문에서 제안한 방법은 P_{sc} 측면에서 기존 방법에 비해 현저히 우수함을 확인할 수 있다.

$$P_D = \frac{\text{음성구간으로 판정된 프레임 수}}{\text{참 음성구간의 프레임 수}} \quad (8)$$

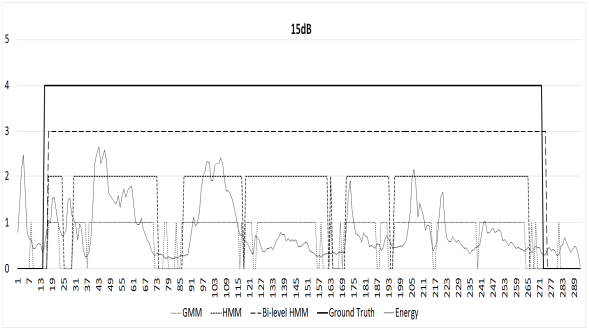
$$P_{SC} = \frac{1}{2}\left(\frac{\text{비음성 상태로 바뀐 회수}}{\text{참 음성구간의 프레임 수}} + \frac{\text{음성 상태로 바뀐 회수}}{\text{참 비음성구간의 프레임 수}}\right) \quad (9)$$



(a)



(b)



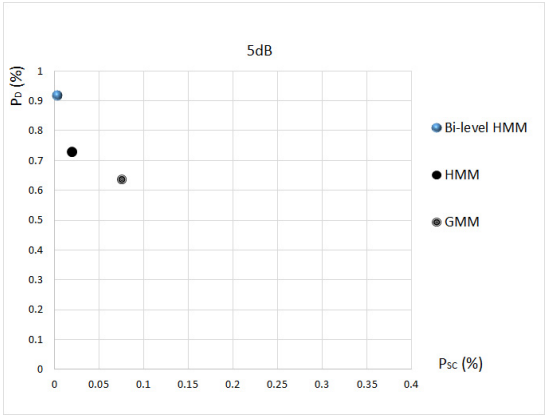
(c)

그림 2. 음성 구간 검출 결과 (a) SNR 5dB (b) SNR 10dB (c) SNR 15dB
Fig. 2 Result of Voice Activity Detection (a) SNR 5dB (b) SNR 10dB (c) SNR 15dB

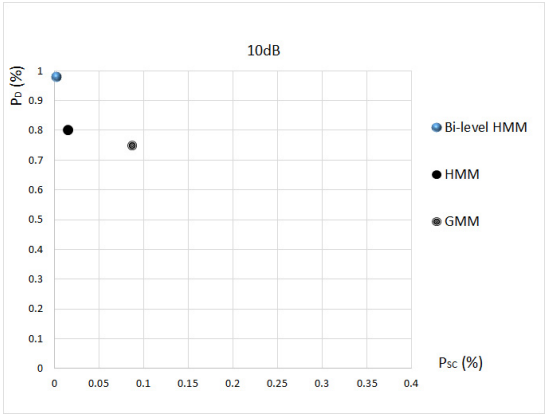
1) http://ecs.utdallas.edu/loizou/speech/noizeus/

Ⅴ. 결 론

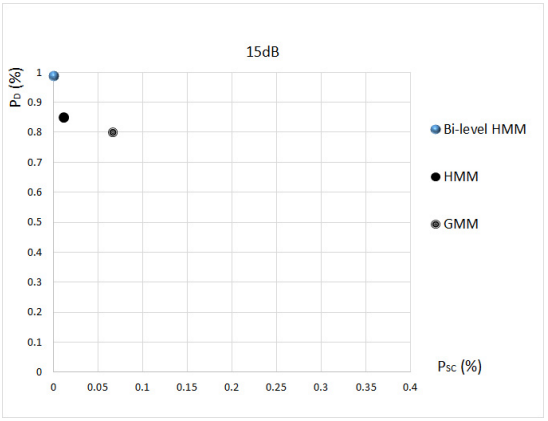
본 논문에서는 Bi-Level HMM을 적용한 음성구간 검출법을 제안하였다. 기존의 방법은 짧은 상태변화 오류(Burst Clipping)를 제거하기 위하여 별도의 후처리 과정을 거치든가, 규칙 기반 지연 프레임을 설정했다. 제안한 방법은 기존의 HMM 모델에 상태 계층을 추가한 Bi-Level HMM을 음성구간 검출에 적용하고 전방향 상태의 사후 확률비를 통한 음성상태 판정으로 통해 이 문제를 해결하였다. 다양한 SNR의 음성 데이터에 대한 실험을 수행하여 기존의 음성상태 분류법보다 우수한 결과를 얻을 수 있었다.



(a)



(b)



(c)

그림 3. 음성구간 검출 ROC (a) 노이즈가 5dB인 환경 (b) 노이즈가 10dB인 환경 (c) 노이즈가 15dB인 환경
Fig. 3 ROC of Voice Activity Detection (a) Noise of environment 5dB (b) Noise of environment 10dB (c) Noise of environment 15dB

References

[1] Y. Zhang, Z. Tang, Y. Li, and Y. Luo, "A hierarchical framework approach for voice activity detection and speech enhancement," *The Scientific World J.*, vol. 2014, 2014, pp. 1-8.

[2] J. Choi, "Speech and Noise Recognition System by Neural Network," *The J. of Korea Institute of Electronic Communication Science*, vol. 5, no. 4, 2010, pp. 357-362.

[3] J. Choi, "Subband Based Spectrum Subtraction Algorithm" *The J. of Korea Institute of Electronic Communication Science*, vol. 8, no. 4, 2013, pp. 555-560.

[4] J. Choi, "Voiced-Unvoiced-Silence Detection Algorithm using Perceptron Neural Network," *The J. of Korea Institute of Electronic Communication Science*, vol. 6, no 2, 2011, pp. 237-242.

[5] C. Lee and D. Kim, "Adaptive Noise Reduction of Speech Using Wavelet Transform," *The J. of Korea Institute of Electronic Communication Science*, vol. 4, no. 3, 2009, pp. 190-196.

[6] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical Voice Detection using a Multiple Observation Likelihood Ratio Test," *IEEE*

Signal Proc. Letters, vol. 12, no. 10, 2005, pp. 689-692.
[7] J. Sohn, N.-S. Kim, and W. Sung, "A statistical model-based voice activity detection[J]," *Signal Proc. Letters, IEEE*, vol. 6, no. 1, 1999, pp. 1-3.
[8] H. Veisi and H. Sameti, "Hidden Markov Model-based Voice Activity Detector with High Speech Detection Rate for Speech Enhancement," *IET Signal Proc.*, vol. 6, no. 3, 2010, pp. 54-63.
[9] H. Othman and T. Aboulnasr, "A Semi-Continuous State-Transition Probability HMM-Based Voice Activity Detector," *EURASIP J. on Audio, Speech, and Music Proc.*, vol. 2007, 2007, pp. 1-7.
[10] X. Liu, Y. Liang, Y. Lou, H. Li, and B. Shan, "Noise-Robust Voice Activity Detector Based on Hidden Semi-Markov Models," *Int. Conf. on Pattern Recognition, Istanbul, Turkey,* August 2010, pp. 81-84.
[11] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "ITU-T Recommendation G.729-Annex B. A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *IEEE Communication Mag.*, Sept. 1997, pp. 64-70.
[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, 2000, pp. 19-41.
[13] S. Chen, R. C. Guido, T. Truong, and Y. Chang, "Improbed Voice Activity Detection Algorithm using Wavelet and Support Vector Machine," *Computer Speech and Language*, vol. 24, no. 3, 2010, pp. 531-543.
[14] P. Tiawongsombat, M. Jeong, J. Yun, B. You, and S. Oh, "Robust visual speakingness detection using bi-level HMM," *Pattern Recognition,* vol. 45, no. 2, 2012, pp. 783-793.
[15] S. Skorik and F. Berthommier, "On a cepstrum-based speech detector robust to white noise," *Computing Research Repository*, vol. cs.CL/00100014, 2000, pp. 1-4.

저자 소개

장광우(Guang-Woo Jang)

2013년 연변대학교 컴퓨터과학기술학과 졸업(공학사)
2013년 ~현재 광운대학교 대학원 제어계측공학과 석사과정
※ 관심분야 : 로봇비전, Computer Graphics, 지능로봇

정문호(Mun-Ho Jeong)

1988년 KAIST 정밀공학과 졸업(공학사)
1994년 KAIST 대학원 자동화 및 설계공학과 졸업(공학석사)
2002년 오사카대학 전자제어기계공학(공학박사)
2002년 KIST 지능로봇연구센터 선임연구원
2010년 ~현재 광운대학교 로봇학부 교수
※ 관심분야 : 로봇비전, HRI, 지능로봇