한국음향학회지 제36권 제6호 pp. 401~406 (2017)
The Journal of the Acoustical Society of Korea Vol.36, No.6 (2017)
https://doi.org/10.7776/ASK.2017.36.6.401

pISSN : 1225-4428
eISSN : 2287-3775

양서류 울음 소리 식별을 위한 특징 벡터 및 인식 알고리즘 성능 분석

Performance assessments of feature vectors and classification algorithms for amphibian sound classification

박상욱,[1] 고경득,[1] 고한석[1†]

(Sangwook Park,[1] Kyungdeuk Ko,[1] and Hanseok Ko[1†])

[1]고려대학교 전기전자공학과

(Received September 12, 2017; revised September 28, 2017; accepted November 29, 2017)

초 록: 본 논문에서는 양서류 울음소리를 통한 종 인식 시스템 개발을 위해, 음향 신호 분석에서 활용되는 주요 알고리즘의 인식 성능을 평가했다. 먼저, 멸종위기 종을 포함하여 총 9 종의 양서류를 선정하여, 각 종별 울음소리를 야생에서 녹음하여 실험 데이터를 구축했다. 성능평가를 위해, MFCC(Mel Frequency Cepstral Coefficient), RCGCC(Robust Compressive Gammachirp filterbank Cepstral Coefficient), SPCC(Subspace Projection Cepstral Coefficient)의 세 특징벡터와 GMM(Gaussian Mixture Model), SVM(Support Vector Machine), DBN-DNN(Deep Belief Network - Deep Neural Network)의 세 인식기가 고려됐다. 추가적으로, 화자 인식에 널리 사용되는 i-vector를 이용한 인식 실험도 수행했다. 인식 실험 결과, SPCC-SVM의 경우 98.81 %로 가장 높은 인식률을 확인 할 수 있었으며, 다른 알고리즘에서도 90 %에 가까운 인식률을 확인했다.

핵심용어: 음향 이벤트 인식, 환경음 인식, i-vector, MFCC (Mel Frequency Cepstral Coefficient)

ABSTRACT: This paper presents the performance assessment of several key algorithms conducted for amphibian species sound classification. Firstly, 9 target species including endangered species are defined and a database of their sounds is built. For performance assessment, three feature vectors such as MFCC (Mel Frequency Cepstral Coefficient), RCGCC (Robust Compressive Gammachirp filterbank Cepstral Coefficient), and SPCC (Subspace Projection Cepstral Coefficient), and three classifiers such as GMM(Gaussian Mixture Model), SVM(Support Vector Machine), DBN-DNN(Deep Belief Network - Deep Neural Network) are considered. In addition, i-vector based classification system which is widely used for speaker recognition, is used to assess for this task. Experimental results indicate that, SPCC-SVM achieved the best performance with 98.81 % while other methods also attained good performance with above 90 %.

Keywords: Acoustic event recognition, Environmental sound classification, i-vector, MFCC (Mel Frequency Cepstral Coefficient)

PACS numbers: 43.60.Bf, 43.60.Uv

I. 서 론

도시 계획, 사업구역 선정 등 환경에 영향을 줄 수 있는 사업을 시행하기 전 수행되는 환경 영향 평가의 주요 목적은 생태계를 보존하는 것이다. 이를 위해, 사업 예정 지역에 서식하는 야생 동물의 종류와 개체수를 파악하는 것이 중요하지만, 이러한 일을 수행할 수 있는 전문가가 많지 않다. 이 문제를 해결하기 위해, 최근 동물 울음소리로 종을 인식할 수 있는 인공지능 시스템에 관한 연구가 시작되었다. 본 논문에서는 양서류 울음소리를 분석하여 세부 종

†**Corresponding author:** Hanseok Ko (hsko@korea.ac.kr)
Engineering Building Room 419, Department of Electronics and Computer Engineering, Korea University Anam Campus, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea
(Tel: 82-2-3290-3239, Fax: 82-2-3291-2450)

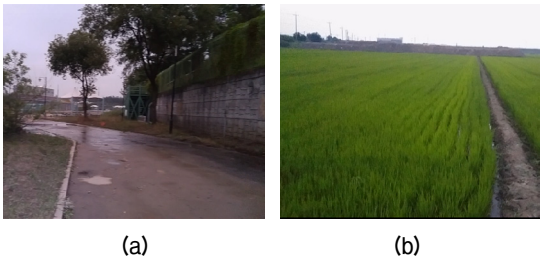
을 인식할 수 있는 인식 시스템 개발을 위해, 여러 특징 추출과 인식 알고리즘의 성능을 비교·분석했다.

특징 추출 알고리즘으로써, MFCCs(Mel Frequency Cepstral Coefficients)는 주파수 별 에너지 분포를 나타낸 스펙트로그램으로부터 추출되며, 환경이 다를 경우 인식 성능을 보장할 수 없다.[1] 반면, RCGCCs (Robust Compressive Gammachirp filterbank Cepstral Coefficients)는 스펙트로그램에서 추정된 잡음을 제거하기 때문에, 환경 잡음에 강인하다.[2] SPCCs(Subspace Projection Cepstral Coefficients)는 신호의 서브공간을 추정하여 투영함으로써 잡음 효과를 완화시킬 수 있을 뿐만 아니라, 클래스 간 서브공간의 차이를 반영하기 때문에 높은 인식 성능을 기대할 수 있다.[3] 이들을 분류하기 위한 인식 알고리즘으로는 GMM(Gaussian Mixture Model), SVM(Support Vector Machine), DBN-DNN(Deep Belief Network - Deep Neural Network)[4]를 고려했다.

추가적으로, 사람 목소리 중 구체적인 개인을 구분하는 화자 인식이 양서류 울음소리 중 세부 종을 인식하는 목표가 상당히 유사하기 때문에, 화자 인식에 널리 사용되어 우수한 성능을 보이는 i-vector[5]기반 인식기의 성능 평가도 수행했다. i-vector는 GMM의 평균 벡터가 연결되어 생성된 고차원의 슈퍼벡터 공간에서 확률기반 요소분석을 수행하여 차원을 축소하여 추출된 특징으로, 간단한 벡터 내적으로 인식 결과를 도출할 수 있다. 이들의 성능 평가를 위해, 데이터를 수집하고 동일한 조건에서 인식 실험을 수행했다. 이후 논문은 II 양서류 울음소리 데이터구축, III 울음소리 인식 알고리즘, IV 실험, V 결론으로 구성된다.

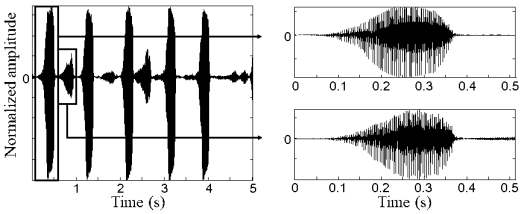
II. 양서류 울음소리 데이터 구축

본 논문에서는 국내에 서식하고 있는 양서류 중 멸종위기 종을 포함하여 총 9종을 선정하고 전문가와 함께 개별 서식지(Fig. 1)에서 울음 소리를 녹음했다 (44.1 kHz, mono, 16 bit-resolution). 녹음된 데이터 중 환경 잡음과 다른 종의 울음소리가 거의 없는 구간을 선정하여 데이터 구축에 활용했다. 이 과정에서 종별 데이터 수에 차이가 발생했다. Fig. 2는 맹꽁이(narrow mouth frog)의 울음소리를 녹음한 데이터



(a) (b)

Fig. 1. Places for data collection of (a) narrow mouth frog and (b) suweon tree frog.



(a) (b)

Fig. 2. Narrow mouth frog sound collected at the place described in Fig. 1: (a) sample for 5 seconds and (b) results of energy based end-point detection.

Table 1. Target classes for amphibian sound classification and the number of data.

Species	Abbreviation	# of data
Brown frog	BroFrog	342
Golden frog*	GolFrog	195
Green frog	GreFrog	260
Leopard frog	LeoFrog	367
Narrow mouth frog*	NarFrog	1,050
Rana rugosa	RanRugo	439
Red-bellied frog	RedFrog	152
Suweon tree frog*	SuwFrog	1,877
Water toad	WatToad	74

* endangered species

중 선정된 구간과 실험 데이터베이스로 구축된 파형을 보여준다. 환경 잡음이 포함되어 있지만, 맹꽁이 울음소리가 분명하다. 여기에 끝점검출 알고리즘[6]을 적용하고 16 kHz로 다운 샘플링 하여 Table 1에 요약된 바와 같이 데이터베이스를 구축했다.

III. 울음소리 인식 알고리즘

3.1 특징 추출 알고리즘

Fig. 3은 본 논문에서 고려한 세 특징벡터의 추출

과정을 보여준다. 입력신호의 고주파 감쇠를 보상하기 위해 증폭을 수행하고, 주파수 변환(STFT, Short Time Fourier Transform)을 수행한다. MFCC는 주파수 변환 결과의 크기를 멜 필터를 이용하여 필터링하고, 로그변환, 이산코사인변환(Discrete Cosine Transform, DCT)을 순차적으로 수행하여 추출된다. RCGCC는 주파수 변환 결과의 크기를 바탕으로 신호를 개선하고, 비선형변환과 이산코사인변환을 거쳐 추출된다. 이들과 달리, SPCC의 경우, 주파수로 변환된 입력 신호가 인식 대상에 대한 서브공간에 투영된다. 투영 벡터 크기에 스무딩과 이산코사인변환을 차례로 적용하면 SPCC가 추출된다. 끝으로, 입력 센서의 채널효과를 완화하기 위해, 추출된 세 특징은 정규화하여 인식기에 입력된다. Fig. 4는 60차원 공간에 분포하는 세 특징벡터를 t-SNE 알고리즘[7]을 통해 2차원 공간에 나타낸 분포를 보여준다. 세 특징 모두 서로 다른 종에 대한 경계를 확인할 수 있으며, 특히 SPCC의 경우, 그 경계가 분명한 것을 확인할 수 있다.

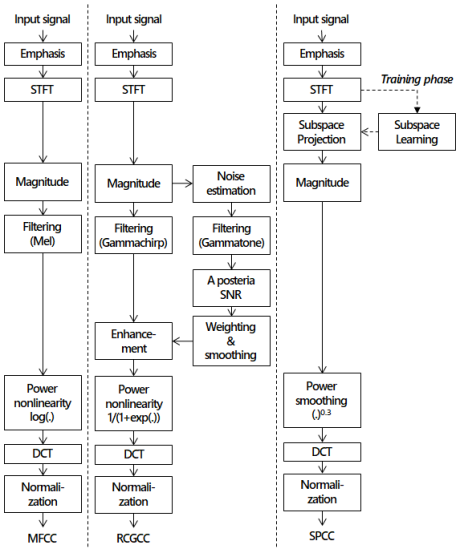


Fig. 3. Block diagrams for extracting features of MFCC, RCGCC, and SPCC, respectively.

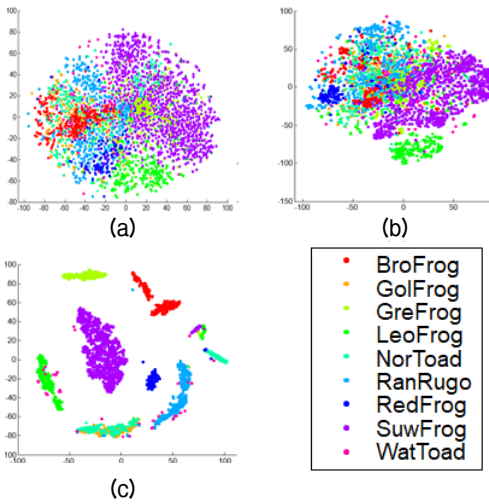


Fig. 4. Distributions of feature vectors: (a) MFCC (b) RCGCC (c) SPCC.

3.2 특징 분류를 위한 인식 알고리즘

GMM은 인식하려는 대상 클래스의 모델을 학습한 뒤, Eq. (1)과 같이 테스트 데이터를 인식한다.

$$\hat{c} = \arg max_c \left[\sum_{m=1}^{M} \log P(x_m | G_c) \right], \quad (1)$$

여기서 c와 m 은 각각 클래스와 프레임 인덱스이다. M은 총 프레임 수이고, G_c는 c에 대한 GMM 모델을 의미한다. $\log P(x_m|G_c)$는 c 클래스에 대한 m 번째 순간 특징 벡터의 로그우도(log-likelihood)로서, 매 순간순간 추출된 특징 벡터의 로그우도가 모두 동등한 비율로 반영되어 최종 결과를 도출한다.

SVM을 다중 클래스 분류에 적용하기 위해, '1 대 나머지' 방식으로 총 9개의 분류함수를 훈련했다. 이때, 분류함수의 수렴을 보장하기 위해, 프레임 기반 특징 대신, x의 평균과 표준편차를 연결하여 생성된 벡터 z를 적용했으며, Eq. (2)에 근거하여 테스트 데이터를 인식한다.

$$\hat{c} = \arg max_c \left[D_c(z) \right], \quad (2)$$

여기서 D_c는 클래스 c와 나머지 모든 클래스에 대해 학습된 분류함수이다. 다음으로, DBN-DNN을 이용한 실험에서는 벡터 z를 이용하여 네트워크를 훈련하고,[8] Eq. (3)과 같이 테스트 데이터를 인식했다.

$$\hat{c} = \arg max_c \left[O_c(z) \right], \quad (3)$$

여기서 O_c는 DBN-DNN의 출력 노드를 나타낸다.

3.3 i-vector 기반 인식 알고리즘

Fig. 5는 i-vector의 개요와 추출과정을 간략히 보여준다. 3.1절에 설명한 바와 같이 특징을 추출하면, 매 프레임 별 특징벡터가 캡스트럴 벡터공간에 분포한다. 이때, 모든 클래스에 대한 캡스트럴특징을 이용하여 전반적인 배경모델(GMM-UBM, Gaussian Mixture Model－Universal Background Model)을 훈련한다. GMM-UBM에서 혼합 요소의 평균벡터를 연결하면 전반적인 슈퍼벡터가 생성된다.

다음으로 GMM-UBM을 각 클래스에 대한 캡스트럴 특징으로 적응하면, 특정 클래스의 GMM을 훈련할 수 있고, 이때 혼합 요소의 평균벡터를 연결하여 클래스 슈퍼벡터를 슈퍼벡터 공간에 나타낼 수 있다. 슈퍼벡터 공간의 벡터는 고차원 연산으로 연산량 증가뿐만 아니라, 역변환이 가능한 공분산 행렬을 얻기 위해 무수히 많은 양의 데이터를 필요로 한다. 이러한 문제를 해결하기 위해, 전반적인 슈퍼벡터와 클래스 슈퍼벡터의 차이를 나타내는 벡터 Tx를 고려할 수 있고, 이때 x를 i-vector라 한다.[9]

캡스트럴 특징이 고차원 공간으로 확장되면서, 슈퍼벡터 공간에는 원래 특징이 가진 정보에 대한 서브공간 T가 존재한다. i-vector를 얻기 위해, 최대 우도 추정(maximum likelihood estimation) 기반 확률론적 주성분 분석(Probabilistic Principal Component Analysis, PPCA)[10]을 수행하여 T를 훈련하고, x를 추출한다. 추출된 i-vector는 Eq. (4)와 같이 테스트 i-vector x 와 클래스 c에 대한 대표 i-vector x_c의 코사인 점수를 산출하여 인식한다.

$$\hat{c} = \arg max_c \left[\frac{x^T x_c}{|x||x_c|} \right], \tag{4}$$

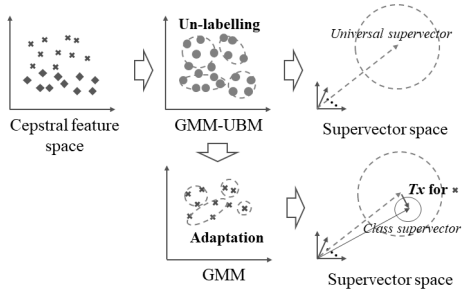


Fig. 5. The concept of i-vector.

IV. 실 험

4.1 실험 설정

MFCC, RCGCC, SPCC 추출을 위해, 프레임 길이를 2048로 설정하고, 50%중첩하여 다음 프레임을 정의했고, 해밍 윈도우를 적용하여 푸리에변환을 수행했다. MFCC와 RCGCC 추출에 사용되는 멜필터와 감마첩필터의 수는 40개로 설정하였고, 20차원의 캡스트럴 계수를 추출하고 델타와 델타-델타 계수를 연결하여 총 60차원 캡스트럴 특징 벡터를 추출했다. SPCC의 경우, 각 클래스별 주성분분석(Principal Component Analysis, PCA)을 통해 서브공간을 훈련했고, 이산코사인변환 이후, 델타와 델타-델타 계수를 연결하여 총 60차원 캡스트럴 특징 벡터를 추출했다.

i-vector 추출을 위한 캡스트럴 특징은 이전 실험을 위해 추출한 특징을 사용했다. 캡스트럴 공간에서 전반적인 배경 모델을 훈련했고, 슈퍼벡터 공간에서 200차원의 서브공간을 훈련하여, i-vector를 추출했다. 이후, i-vector의 인식률 향상을 위해 선형차별성분석(Linear Discriminant Learning, LDA)을 적용했다. 이때, 전반적인 배경 모델의 가우시안 혼합 모델 개수와 서브공간 차원은 실험을 통해 결정했다.

교차평가을 위해, Table 1에 명시된 데이터는 5개 집단으로 나뉘고, 훈련 데이터와 테스트 데이터 비율을 4:1로 설정했다. 인식기가 어느 한 클래스로 편중되는 것을 예방하기 위해, 훈련데이터는 클래스당 동일한 수의 데이터를 선정했다. 인식기에 필요한 변수는 교차평가를 통해 가장 높은 인식률을 보이는, 16 혼합 GMM과 선형기저함수(linear kernel function) 기반 SVM이 실험에 사용됐다. DBN-DNN의 구조는 과학습(overfitting)이 발생하지 않도록 실험적으로 결정했다. 그 결과, 200개 노드를 갖는 3개 은닉층으로 구성됐다(learning rate = 0.0001, epoch = 200). i-vector를 이용한 인식 실험에서는, 클래스 당 50개 i-vector의 평균으로 대푯값을 선정했다.

4.2 인식 실험 결과

앞서 소개한 특징 추출과 인식기의 조합으로 구성된 총 9가지 인식 시스템의 인식률을 Table 2에 요약

Table 2. Experiment results of amphibian classification using classifical recognition systems.

	Feature	Avg. Acc	BroFrog	GolFrog	GreFrog	LeoFrog	NarFrog	RanRugo	RedFrog	SuwFrog	WatToad
GMM	MFCC	96.86	99.42	95.90	91.54	98.64	98.57	97.04	92.11	98.51	100.00
	RCGCC	95.54	94.44	96.41	91.54	96.46	97.81	95.90	93.42	97.92	95.95
	SPCC	**98.42**	100.00	98.46	100.00	98.91	96.57	100.00	98.68	99.95	93.24
SVM	MFCC	71.74	75.15	54.36	83.08	82.02	90.48	59.68	86.84	74.91	39.19
	RCGCC	69.08	80.70	54.87	90.77	57.22	89.52	38.95	86.84	79.60	43.24
	SPCC	**98.81**	100.00	97.95	99.62	98.64	97.24	98.41	99.34	99.47	98.65
DBN-DNN	MFCC	86.50	97.95	84.10	88.46	89.92	97.81	93.85	82.24	98.19	45.95
	RCGCC	85.07	90.64	75.38	93.85	85.29	97.90	83.83	88.82	98.56	51.35
	SPCC	**98.17**	100.00	97.95	99.62	97.82	98.76	98.86	100.00	99.95	90.54

Table 3. Experiment results of amphibian classification using i-vector framework.

Cepstral feature	Avg. Acc	BroFrog	GolFrog	GreFrog	LeoFrog	NarFrog	RanRugo	RedFrog	SuwFrog	WatToad
MFCC	**88.04**	97.08	92.82	95.00	63.76	86.95	92.94	85.53	93.18	85.14
SPCC	87.12	100.00	82.56	99.23	54.77	99.05	92.48	98.68	95.10	62.16

했다. GMM 기반 인식 실험에서는 모든 양서류에 대해 90 % 이상의 인식률이 확인되었다. 특히 SPCC를 특징벡터로 적용한 경우, 세 종류의 양서류에서 100 % 인식률을 확인할 수 있었다. 반면, SVM과 DBN-DNN 인식 실험에서는 특징에 따라 성능 차이가 크다. 특히, 물두꺼비의 경우 금개구리와 수원청개구리로 오인식이 많이 발생했다.

MFCC와 RCGCC의 경우, GMM은 매 프레임에서 추출된 특징벡터를 혼합모델로 나타내어, 순간순간의 변화를 모델링할 수 있지만, SVM과 DBN-DNN에 적용된 두 특징은 프레임 평균과 표준편차로 변환되면서, 순간적인 특성이 손실된다. 반면, SPCC는 서브공간이 시간에 대한 변화가 없기 때문에, 세 인식기에서 모두 유사한 결과를 확인할 수 있다. 또한 MFCC와 RCGCC는 Fig. 4에서 보듯이 서로 다른 클래스 사이의 경계가 인접해 있어 복잡한 인식기를 요구하는 반면, SPCC는 그 경계가 분명하기 때문에 인식기 별 성능 차이가 미비하다.

Fig. 6은 i-vector 추출을 위한 혼합모델 수를 결정하기 위해 수행한 실험 결과를 보여준다. MFCC는 512 혼합모델을 제외하고, 128 혼합모델 이상에서 80 % 이상의 인식률을 확인할 수 있는 반면, SPCC는 혼합모델수가 적을수록 높은 인식률을 확인할 수 있다. 1024 혼합모델을 사용한 경우, 공분산 행렬이 비가역행렬인 가우시안 모델이 생성되면서 i-vector 추출

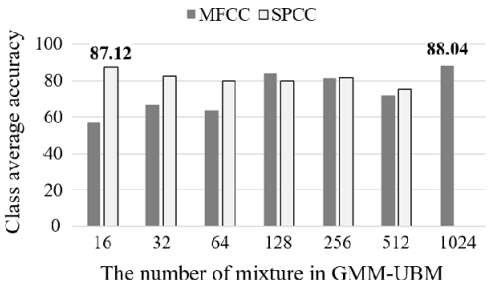


Fig. 6. Class average accuracies according to the number of mixture component in UBM.

에 실패했다. 이는 SPCC는 시간에 따른 변화가 적기 때문에 적은 수의 혼합 모델로도 특징 분포를 모델링하는데 충분하기 때문이다. 전반적인 배경 모델의 혼합 모델 수는 슈퍼벡터 공간 차원을 결정하므로, SPCC는 MFCC와 비교하여 훨씬 낮은 차원에서 인식에 필요한 차별성을 추출할 수 있다.

Table 3은 1024 혼합모델을 이용한 MFCC 기반 i-vector와 16 혼합모델을 이용한 SPCC 기반 i-vector의 인식 대상별 인식률을 보여준다. MFCC와 SPCC에 기반한 i-vector 기반 인식 결과, 산개구리(BroFog), 옴개구리(RanRugo), 수원청개구리(SuwFrog)에서 모두 90 % 이상의 인식률을 확인할 수 있었지만, 참개구리(LeoFrog), 물두꺼비(WatToad)에서 상대적으로 낮은 인식률을 확인할 수 있었다. 전반적인 배경모델은 i-vector의 인식 성능을 결정하는데 매우 중요하

다. 보통 전반적인 배경모델은 수많은 인식 대상에 대한 다량의 데이터로 학습되지만, 본 실험은 Table 1에 요약된 데이터 중 일부만 사용했기 때문에, Table 2와 비교하여 낮은 인식률을 확인할 수 있다.

V. 결 론

본 논문은 양서류 울음소리를 통해 세부 종을 인식할 수 있는 시스템 개발을 위해, 데이터를 수집하고 여러 인식 시스템에서 인식 성능 평가를 수행했다. 실험에서 고려한 특징 추출 알고리즘은 MFCC, RCGCC, SPCC이고, GMM, SVM, DBN-DNN이 인식기로 고려되었다. 이들 조합으로 구성된 9가지 시스템에 대한 성능을 평가했고, 추가적으로 i-vector 기반 인식 시스템에 대한 성능 평가도 수행했다. 인식 실험 결과 특징 벡터와 인식기 조합에 따라 성능에 차이는 있지만, 전반적인 실험 결과로부터 양서류 울음소리를 통한 세부 종 인식이 가능하다는 점을 확인했다. 향후에는 데이터를 추가 수집하고, 잡음과 중첩 상황 등 보다 현실적인 문제를 고려한 연구를 수행할 계획이다.

감사의 글

본 연구는 환경부의 환경산업 선진화 기술개발 사업에서 지원받았음.

References

1. S. Park, W. Choi, D. K. Han, and H. Ko “Acoustic event filterbank for enabling robust event recognition by cleaning robot,” IEEE Trans. Consu. Electro., **61**, 189-196 (2015).
2. M. J. Alam, P. Kenny, and D. O’Shaughnessy, “Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique,” Digital Signal Processing, **29**, 147-157 (2014).
3. S. Park, Y. Lee, D. K. Han, and H. Ko, “Subspace projection cepstral coefficients for noise robust acoustic event recognition,” Proc. ICASSP, 761-765 (2017).
4. G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” Neural Computation, **18**, 1527-1554 (2006).
5. N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans. Audio, Speech. and Lang. Proc. **19**, 788-798 (2011).
6. J. Park, W. Kim, D. K. Han, and H. Ko, “Voice activity detection in noisy environments based on double-combined fourier transform and line fitting,” The Scientific World J. 2014, 1-12 (2014).
7. L. J. P. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Machine Learning Research, **9**, 2579-2605 (2008).
8. Z. Kons and O. Toledo-Ronen, “Audio event classification using deep neural networks,” Proc. INTERSPEECH, 1482-1486 (2013).
9. P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modelling with sparse training data,” IEEE Trans. Speech and Audio Processing, **13**, 345-354 (2005).
10. M. E. Tipping and C. M. Bishop, “Mixtures of probabilistic principal component analyzers,” Neural Computation, **11**, 443-482 (1999).

저자 약력

▸ 박 상 욱 (Sangwook, Park)

2012년 2월: 중앙대학교 전자전기공학부 (전자전기공학사)
2017년 8월: 고려대학교 전기전자공학부 (공학박사)
2017년 9월 ~ 현재: 박사 후 연구원

▸ 고 경 득 (Kyungdeuk Ko)

2017년 8월: 연세대학교 의공학부(공학사)
2017년 9월 ~ 현재: 고려대학교 전기전자공학부 석사과정

▸ 고 한 석 (Hanseok Ko)

1982년 5월: 미국 카네기 멜론 대학교 전기공학 (공학사)
1986년 5월: 미국 메릴랜드 대학교 시스템 공학(공학석사)
1988년 5월: 미국 존스 홉킨스 대학교 전기공학 (공학석사)
1992년 5월: 미국 카톨릭 대학교 전기공학 (공학박사)
1995년 3월 ~ 현재:고려대학교 전기전자공학부 교수