지능형 음향환경파라미터추정기술

이명인, 장준혁
한양대학교

요 약

인간은 소리를 통해 많은 정보를 얻을 수 있다. 누가 어떤 말을 하는지 뿐만 아니라, 상황에 따라서는 소리가 발화된 환경 또한 시각적인 정보 없이 유추할 수 있다. 이러한 판단을 내리기까지, 인간은 경험을 통해 스스로 학습하는 과정을 거친다. 이와 같은 학습 과정에 생물의 사고과정을 모방하여 복잡한 상관관계를 추론하는 인공지능형 알고리즘을 적용하면, 인간의 두뇌가 경험을 통해 학습하고 판단하던 역할을 기계적으로도 모방할 수 있게 된다. 본고에서는 음향이 발화된 환경의 정보를 나타낼 수 있는 파라미터들에 대해 알아보고, 그 파라미터들을 지능형 알고리즘을 이용해 도출해내는 과정과 기법들을 소개한다.

Ⅰ. 서 론

소리는 많은 정보를 담고 있다. 인간은 두 귀를 이용해 시각적인 정보 없이도 어디에서 어떤 소리가 나는지, 그 소리는 어떤 메시지를 담고있는지 등 많은 정보를 판단할 수 있다. 우리가 소리를 통해 판단할 수 있는 것은 그 안에 포함돼있는 언어 또는 메시지 그 자체일 수도 있지만, 상황에 따라서는 그 소리가 발화된 환경이기도하다. 예를 들어, 전화통화를 통해 대화하는 중에 상대방의 소리만을 듣고 누가 통화를 하고 있는지, 얼마나 소란한 환경에 있는지, 얼마나 울리는 공간에 있는지 등을 알아맞추는 일은 우리가 일상생활에서 청각에 의존해 정보를 얻는 것의 일부분이라 할 수 있다. 우리는 이러한 판단을 내리기 까지 청각적 경험을 통해 스스로 학습하는 과정을 거친다. 이러한 정보를 비단 인간의 두뇌를 통해 얻어내는 것이 아니라 마이크와 알고리즘을 이용해 도출할 수 있다면, 음성, 음향 신호처리 분야에서 성능을 현저하게 저하시키는 요소들을 정량화 하여 알아냄으로써 음성인식, 방향추정, 위치추정 등의 알고리즘에 적응적으로 활용할 수 있다. 공간의 음향적 특성은 공간의 구조, 표면의 반사율, 그리고 발화점과 관찰점의 위치 등의 요소에 의해 결정되는데, 그 특성을 나타내는 파라미터에는 대표적으로 Early Decay Time (EDT), Deutlichkeit (D50), Clarity index (C50), Direct-to-reverberant ratio (DRR), 잔향시간 (T60) 등이 있다. 이와 같은 파라미터들을 추정하기 위해 기존의 방법들은 환경에 대한 가정을 기반으로 한 물리적 모델링을 이용해왔지만, 이론적인 시뮬레이션 환경 외에서 낮은 정확도를 보이는 경우가 있다. 이에 반해 생물의 사고과정을 모방하여 충분히 많은 데이터를 기반으로 학습을 통해 복잡한 상관관계를 추론하는 인공지능형 알고리즘을 적용하면, 인간의 두뇌가 경험을 통해 학습하고 판단하던 역할을 기계적으로도 모방할 수 있게 된다. 이와 같은 인공지능의 활용은 음성인식, 문자 인식 등의 최신 패턴 인식 기술에서 분류, 판별에 중요한 역할을 하고 있으며, 그 성능을 다양한 분야에서 입증 받고 있다. 본고에서는 공간의 특성을 나타내는 다양한 음향환경파라미터들을 소개하고, 그 특성에 따른 분석을 통한 지능형 추정기술과 그 가능성을 알아본다.

Ⅱ. 본 론

1. 인공지능

인공지능은 넓은 의미에서 인간이 가지는 문제 해결 능력을 컴퓨터가 갖도록 하는 것이다. 이러한 개념에 대한 제안은 오래전 제시되었지만, 관심에 비해 당시 기술의 한계로 인해 침체기와 발전기를 거듭해 왔다. 그 결과, 기존 인공신경망 모델의 단점을 극복하는 모델이 제시되고, 하드웨어가 비약적으로 발전하면서, 몇 주 이상 걸리던 복잡한 행렬과 벡터의 연산을 단 시간으로 줄여낼 수 있게 되었다[1]. 그 결과 인간이 판단을 위해 학습하는 과정을 기계적으로 구현할 수 있게 되었고, 환경에 대한 가정을 기반으로 한 물리적 모델링을 이용했던 기존의 방법들을 대용하는데 적극적으로 활용되기 시작했다[1]-[4].

딥 러닝은 비선형적인 변환기법들의 조합을 이용해 다량의 데이터 속에서 핵심적인 내용, 또는 복잡한 상관관계

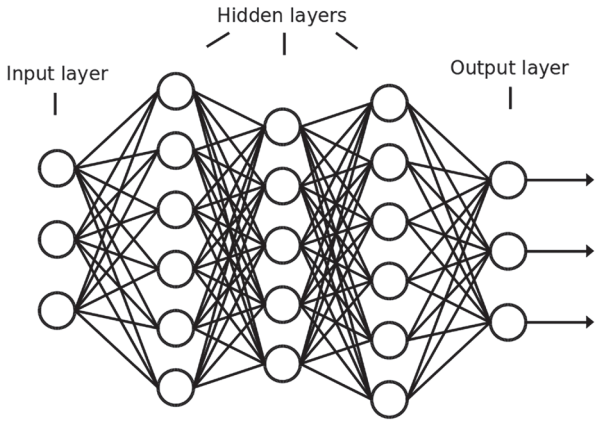

그림 1. DNN 알고리즘의 구조

를 도출해내는 기계학습 알고리즘의 집합이다. 여기에는 심층 신경망 (Deep Neural Network, DNN), 합성곱 신경망 (Convolutional Neural Network, CNN), 순환 신경망 (Recurrent Neural Network, RNN), 제한 볼츠만 머신 (Restricted Boltzmann Machine, RBM), 심층 신뢰 신경망 (Deep Belief Network, DBN) 등의 알고리즘이 있는데 이들은 각각 적용하고자 하는 데이터의 집합에 따라 적절한 것을 선택하여 쓰는 것이 바람직하다. 이와 같은 알고리즘들은 그 성능을 인정받아 컴퓨터비전, 음석인식, 자연어처리, 음성신호처리 등의 다양한 분야에 적용되어 최신 기술 동향을 선도하고 있다.

그중 심층 신경망 (Deep Neural Network, DNN)은 입력 계층 (input layer)과 출력 계층 (output layer) 사이에 복수개의 은닉 계층 (hidden layer)들로 이루어진 인공신경망으로서, 많은 데이터 속에서 복잡한 비선형 관계를 모델링하는데 좋은 성능을 보이는 것으로 알려져 있다. 합성곱 신경망 (Convolutional Neural Network, CNN)은 하나 또는 복수의 합성곱 계층과 그 위에 인공 신경망 계층들이 쌓여있는 구조인데, DNN에 비해 적은 전처리를 거친 자료 속에서도 비선형 관계를 잘 찾아내는 것으로 알려져 있다. 이러한 기법들은 연속적인 음성신호 속에서 마치 훈련된 인간이 듣고 판단을 내릴 수 있듯, 음향환경파라미터들을 추정하는데 활용될 수 있다.

2. 음향환경파라미터

현실 속의 대부분의 경우에서 신호는 공간을 통해 전파되고, 그 과정에서 청취자, 또는 마이크에 최단거리를 통해 전달될 뿐만 아니라 공간 구조와 그 특성에 따른 반사적인 경로를 거쳐 잔향을 포함한 신호가 도달하게 된다. 이러한 현상은 음성인식, 방향추정, 위치추정을 포함한 음성, 음향 신호처리 기술들의 정

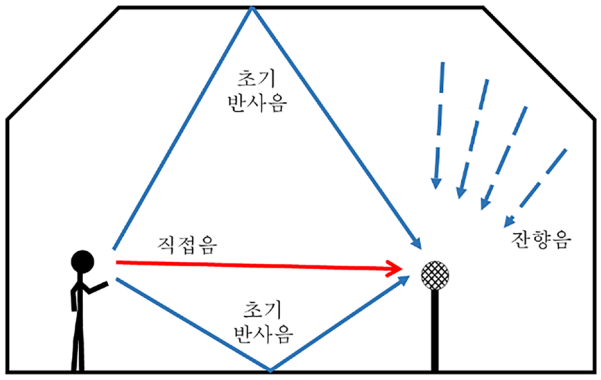

그림 2. 잔향환경에서의 소리 구성

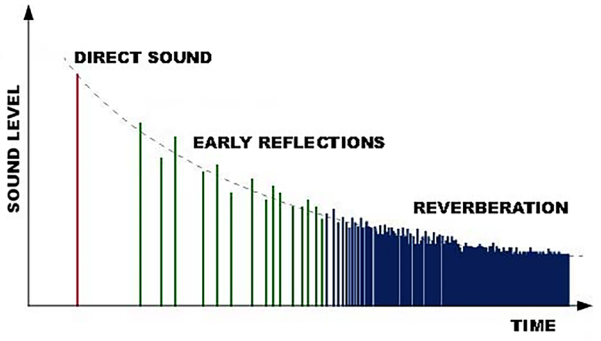

그림 3. 잔향환경에서의Impulse response구성

확도를 현저하게 저하시키기 때문에, 그 정도와 특성을 적절한 음향환경파라미터로 표현하는 것은 해당 분야에서 늘 중요하게 여겨져 왔다. 따라서 그 정의와 용도에 따라 다양한 음향환경파라미터들이 제안되어 왔으며, 추정 방법 또한 제안되어왔다.

가. Early Decay Time (EDT)

EDT는 신호의 전파와 반사 과정에서 일어나는 감쇄를 통해 첫 10 dB 만큼의 감쇄가 일어나는데 걸리는 시간을 뜻한다. 음성의 명료도 관점에서는 EDT 값이 작을수록 직접음에 간섭하는 반사음의 비율이 낮아져서 명료한 소리를 가지게 된다.

나. Deutlichkeit (D50)

직접음이 처음 도달한 후 50 ms까지의 에너지와 전체 에너지의 비를 나타낸 수치로서 Deutlichkeit, 또는 Definition으로 정의된다.

$$D_{50} = 10\ log_{10}\left(\frac{\sum_{m=0}^{50} h^2(m)}{\sum_{m=0}^{M-1} h^2(m)}\right) dB$$

50 ms까지의 초기 반사음 초기 반사음은 분리된 소리로 인지되지 않으며, 하스효과 (Hass effect) 에 의해 오히려 직접음을

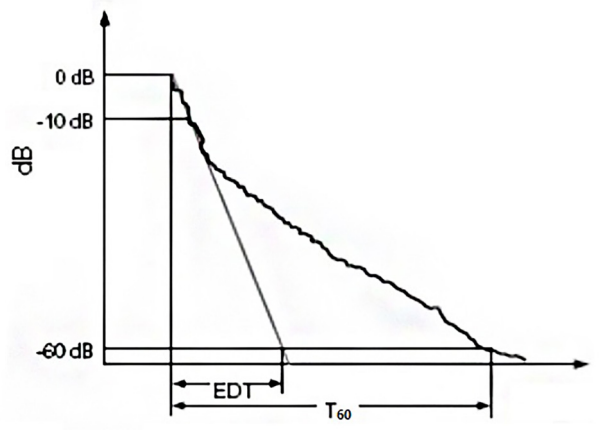

그림 4. Impulse response에서 관찰할 수 있는 EDT의 예

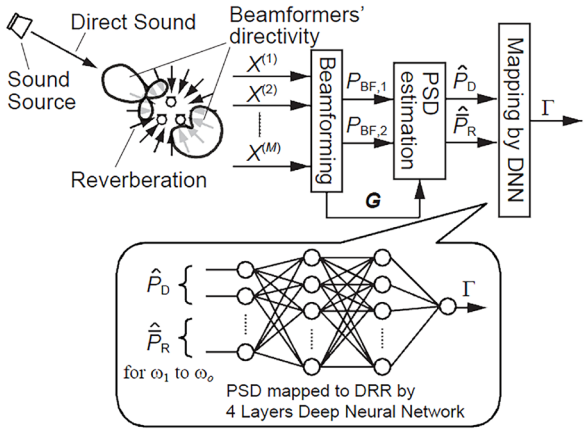

그림 5. 지능형 DRR 추정 기술 알고리즘[6]

보강하고, 명료도를 높여 주는 효과가 있다. 50 ms 보다 뒤에 도달하는 반사음은 듣는 이로 하여금 공간감을 느끼게 한다. 공간감은 발생으로부터 반사경로를 통해 도달하는 시간이 늦을수록 큰 공간감을 형성하는데, 따라서 D50는 클수록 잔향음에 비해 직접음과 초기 반사음의 비율이 높아 명료도가 높으며, 작으면 울림의 양이 많아 명료도가 낮다.

다. Klarheitsmass, Clarity (C50)

C50는 직접음을 포함한 50 ms까지의 에너지와 그 50 ms 이후의 에너지의 비로 정의되며, Klarheitsmass, 또는 Clarity를 뜻한다. 건축음향, 및 음악의 명료도 관점에서는 C50뿐만 아니라 용도에 따라 C80 등의 기준치를 변경한 파라미터를 사용하기도 하는데, 공통적으로는 직접음과 하스효과 내에 있는 소리의 에너지와 그 외의 에너지를 비교한다는 점에서 음소 관점의 명료도를 가장 잘 나타내는 파라미터로 알려져 있다[5].

$$C_{50} = 10\,log_{10}\left(\frac{\sum_{m=0}^{50} h^2(m)}{\sum_{m=50+1}^{M-1} h^2(m)}\right) dB$$

따라서 C50를 추정하는 방법은 다양하게 고안되어왔는데, 최근에는 잔향으로 인해 발생하는 주파수 변조와 분사스펙트럼을 데이터 드리븐 (data-driven)에 적용시킨 추정방법이 제시된 바 있다. 이와 같은 추정방법들은 잔향의 정도를 정량적으로 측정하기에 적합한 특징벡터를 신호로부터 도출하는 것으로부터 시작되는데, 잔향환경에서 반사과정을 통해 직접음은 파형의 크기, 에너지가 줄어들 뿐만 아니라, 주파수 특성 또한 변화하여 변조가 일어난다. 이러한 중요 특징 벡터를 관찰하는 것은 공간 정보 추정에서 매우 중요한 역할을 한다.

라. Direct-to-reverberant ratio (DRR, D/R)

D50 는 최초로 마이크에 도착하는 직접음뿐만 아니라 50 ms 내의 소리를 함께 인지관점에서 직접음으로 간주하고 에너지를 도출하는 반면, DRR은 최초에 도달하는 직접음의 에너지와 반사된 음의 에너지의 비로 정의된다. 직접음 만이 소스의 위치와 방향에 대한 정보를 가지고 있다는 점에서, 인식과 명료도 관점에서 직접음뿐만 아니라 하스효과를 고려한 다른 공간 파라미터들과는 다른 의미를 가진다. 정확한 DRR의 측정을 위해서는 공간의 임펄스 특성 (Room Impulse Response, RIR)을 알아내는 과정이 필요하다. 하지만, RIR을 정확히 측정하는 과정은 특수한 장비와 다른 소프트웨어를 필요로하며, 이것은 모바일, 또는 대부분의 실용적인 마이크 환경에 현실적으로 적용이 불가능하다. 그러므로 DRR을 RIR 또는 공간의 크기, 반사계수 등의 정보 없이 도출하는 방법들이 연구되어왔다.

이에 따라 심층 신경망을 이용한 DRR 추정방식이 최근 제안된 바 있는데, 다채널 마이크를 이용해 직접음과 반사음을 서로 다른 방향에서 오는 소리를 분리해내는 빔포밍 (beamforming)을 이용해 분류하고, 그 결과의 파워스펙트럼 밀도(Power Spectrum Density, PSD)를 심층 신경망의 입력으로 주어 DRR 값을 도출할 수 있도록 학습시켰다. 이와 같은 추정방식은 DRR의 물리적인 특성을 고려했을 때, 모델링을 사용했다면 직접음의 크기를 도출해야하므로 모델링 과정에서 큰 오차가 발생했겠지만, 별도의 모델링 없이 파라미터를 나타낼 수 있는 입력을 심층 신경망에 부여하여, 알고리즘 스스로 복잡한 상관관계를 찾아내도록 한 사례이다.

마. 잔향시간 (Reverberation time, RT60, T60)

잔향시간은 RT60, 또는 T60로 표현되며, Sabine에 의해 1920년대에 정의된 가장 기본적이면서도 거시적인 실내음향의 특성이다. 잔향시간의 정의는 음압이 초기의 크기로부터 반사를 통해 감쇄하면서, 그 레벨이 60dB 감쇠되는데 소요되는 시간이며, 이것은 음압 관점에서는 100만분의 1에 해당하는 것으

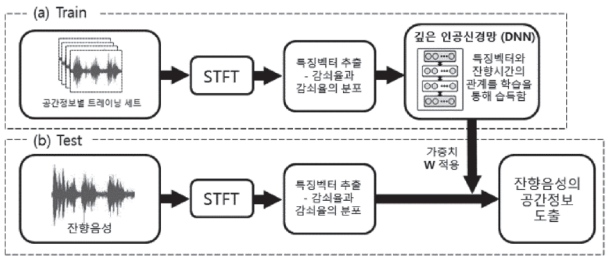

그림 6. 지능형 잔향시간 추정 기술 알고리즘

로, 음원의 생성으로부터 반사 경로를 거쳐 거의 소멸에 해당하는 시간에 대한 정보를 준다. 고전적인 측정방법으로는 임펄스를 생성시켜 60dB만큼 감쇠되는 시간을 직접 측정하는 방법이 있는데, 이와 같은 방법은 RIR을 도출하는 것만큼이나 시간적 비용적 측면에서 비효율적이다. 그러므로 공간에 대한 정보 없이 음성신호에서 도출한 특징 벡터만을 사용해 잔향시간을 추정하는 시도가 이어져왔다. 그 중에서 잔향은 주파수 변조를 일으킬 뿐만 아니라[7], 감쇠율의 분포와도 연관이 있음이 연구를 통해 증명되었으며, 감쇠율의 분포를 2차 다항식 회귀(polynomial regression)의 입력으로 이용해 잔향시간을 추정하는 방식이 최근 널리 사용되었다[8]. 하지만, 단일한 특징 벡터에 의존해 잔향 추정하는 방식은 잡음에 강인하지 않은 특성을 보였으며, 환경에 따라 다른 다항식을 적용해주어야하는 단점이 있었다. 따라서, 감쇠율의 분포뿐만 아니라, 주파수 변조 또한 함께 추정에 함께 활용될 수 있다면, 보다 상호보완적이고 정확한 추정 알고리즘을 구현할 수 있다.

제안된 알고리즘에서는 STFT를 이용해 도출한 시간, 주파수 정보를 이용해 주파수 대역별로 포락선의 감쇠율과 그 분포를 도출하고, 그것을 모두 심층 신경망의 특징 벡터로 입력하였다. 심층 신경망은 많은 데이터 속에서 복잡한 비선형 관계를 모델링하는데 좋은 성능을 보이는 것으로 알려져 있으므로, 이미 알려져있는 감쇠율의 분포와 잔향시간의 관계뿐만 아니라, 주파수 밴드별 감쇠율 간의 관계를 이용해 주파수 변조와 잔향시간의 관계를 동시에 도출해낼 수 있다. 뿐만 아니라, 감쇠율의 분포와 변조 간에도 기존에 수학적 모델링으로 제시되지 못했던 관계를 도출하고, 그 결과 기존 방법들을 상호보완하는 결과를 기대할 수 있다. 그 결과 기존의 추정방식들에 비해 지능형 추정기술은 높은 정확도와 잡음에 강인한 특성을 가지게 되었다. 이와 같은 결과는 인간의 잔향에 대한 인지 관점에서도 의미를 찾을 수 있는데, 인간은 전화통화를 할 때, 또는 시각적 정보가 없는 상황에서 서로 대화를 할 때, 상대방이 소리가 울리는 환경에서 있는지 아닌지를 직감적으로 알 수 있다. 그것은 우리가 경험적으로 인간의 목소리가 어느 정도의 특성을 가지고 있는지, 그리고 잔향환경에서는 그 특성들이 정도에 따라 어떻게 변화하는지를 의식하지 않은 상태에서도 학습을 통해 습득해왔기 때문이다.

하지만 기계에게 그 특성을 바로 알아채고 빠른 시간 안에 효율적으로 습득하기만을 마냥 기대하는 것은 경우에 따라 바람직하지 않을 수 있다. 그러므로 중요한 특징벡터들을 신경망의 입력으로 주는 방식은 지능형 추정기술에서 중요하다. 딥 러닝 기술과 그 성능이 나날이 개선되고 있지만, 가공되지 않은 데이터 자체만을 입력으로 주고 원하는 추정 결과를 얻어내기에는 부족한 경우가 있으며, 대신 가공된 형태의 특징벡터 또는 높은 차수의 특징벡터를 입력으로 주었을 때, 딥 러닝 알고리즘은 그 속의 복잡한 관계를 보다 쉽게 발견해내는 것을 확인할 수 있었다. 이러한 현상은 깊은 신경망뿐만 아니라 보다 데이터를 특성들을 잘 도출해내는 것으로 알려진 신경망 알고리즘에서도 발견되고는 하는데, 향후 신경망 알고리즘이 더욱 정교해지고 더욱 깊은 특성을 충분히 발견할 수 있게 되면 가공되지 않은 데이터만을 가지고도 알고리즘을 학습시킬 수 있겠지만, 그 정도에 타협점을 찾을 필요가 있다. 이러한 추정방법은 비단 잔향시간 추정에 국한될 뿐만 아니라, 앞서 소개한 다양한 음향환경파라미터의 추정에도 충분히 활용될 가능성을 가지고 있다.

Ⅲ. 결 론

본고에서는 공간의 특성을 나타내는 다양한 음향환경파라미터들을 소개하고, 그 특성에 따른 분석을 통한 지능형 추정기술과 그 가능성을 알아봤다. 추정기술에 활용가능한 대표적인 딥 러닝 알고리즘에 대해서도 소개했으며, 정의와 용도에 따라 다양한 음향파라미터가 있음을 소개했다. 그 중 대표적으로 활용도가 높은 파라미터들에 대한 추정방법을 소개했으며, 딥러닝을 이용한 잔향시간 추정 알고리즘을 소개했다. 특징벡터 선택의 중요성을 확인해볼 수 있었으며, 그 결과 기존의 추정방식들에 비해 높은 정확도와 잡음에 강인한 알고리즘을 도출할 수 있다. 인간이 청취를 통해 도출할 수 있는 음향환경의 특성에 대해 고찰함으로서, 지능형 알고리즘을 이용하면 높은 정확도로 그 것을 정량화할 수 있음을 소개했다.

지능형 추정기술은 잔향시간 추정에만 활용될 수 있을 뿐만 아니라, 앞서 소개한 다양한 음향환경파라미터의 추정에도 활용될 수 있다. 파라미터마다 음향환경에 대해 서로 다른 특성을 가지고 있기 때문에, 그 특성에 대해 이해하고 기계가 학습할 수 있는 형태로 딥 러닝에 적용하면, 모델링을 기반으로한 추정방법에 비해 다양한 환경에 적응적인 성능을 보이는 알고리즘

을 도출할 수 있을 것이다.

참 고 문 헌

[1] Hinton G, and Salakhutdinov R. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504-507.

[2] Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition. in Proc. NIPS, 2009.

[3] Mohamed A, Dahl G, Hinton G. Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech & Language Processing, 2012;20(1):14-22.

[4] Hinton G, Deng L, Dahl G, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process Magazine, 2012;29(6):82-97.

[5] P. A. Naylor and N. D. Gaubitch, Speech Dereverberation, Springer, 2010.

[6] Y. Hioka, K. Niwa, Estimating direct-to-reverberant ratio mapped from power spectral density using deep neural network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20-25 March 2016; pp. 149-152.

[7] T. Falk, C. Zheng, and W.-Y. Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Transactions on Audio, Speech & Language Processing, vol. 18, no. 7, pp. 1766-1774, Sept. 2010.

[8] N. D. Gaubitch, H. W. Lollmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes. Performance comparison of algorithms for blind reverberation time estimation from speech. International Workshop on Acoustic Echo and Noise Control, Aachen, Germany, Sept. 2012.

약 력

이 명 인

2008년~2015년 한양대학교 공학사
2015년~현재 한양대학교 음성음향오디오신호처리연구실 석사과정
관심분야: 음성음향신호처리, 오디오신호처리, 딥 러닝

장 준 혁

1992년~1998년 경북대학교 공학사
1998년~2000년 서울대학교 공학석사
2000년~2004년 서울대학교 공학박사
2000년~2005년 Netdus Corporation, Chief Engineer
2004년~2005년 University of California, Santa Barbara, Unisted States, Postdoctoral Fellow
2005년~2005년 KIST, Research scientist
2005년~2011년 인하대학교 전자공학과 조교수
2011년~현재 한양대학교 융합전자공학부 부교수
관심분야: 딥 러닝, 음성인식 및 통신 전처리기술, 오디오신호처리, 바이오신호처리