2021년도 박사후국내외연수사업
신규과제 연수계획서

과제명	국문	미지의 손상(unknown corruption)이 발생한 음성에 대한 딥 러닝 기반 자동 복원 모델 개발
	영문	Deep Learning-based Automatic Audio Repair against Unknown Corruptions

1. 연구개발과제의 필요성

1) 연구의 필요성

- 음성 발화문에서 소음, 잔향 등을 제거하는 음성강화 기술은 다양한 음성 애플리케이션의 기반 기술로도 활용됐으며, 비대면 회의가 보편화 되는 포스트 코로나 시대에는 더욱 기술적 수요가 높음. 또한, Youtube 등의 다양한 차세대 1인 매체에서는 비전문가도 사용 가능한 음성강화 기술의 수요가 높음

- 화상통화 (facebook[10], Google[11], Microsoft[14]), 음성 편집 (Adobe[9]), 1인 매체 (Bytedance [15]), 음성 하드웨어 (Bose[12], Qualcomm[13]) 등의 분야에서 음성강화 전문가 채용이 큰 폭으로 증가하고 있는 최근 석/박사 취업 시장 동향은 음성강화 기술에 대한 높은 수요를 방증함

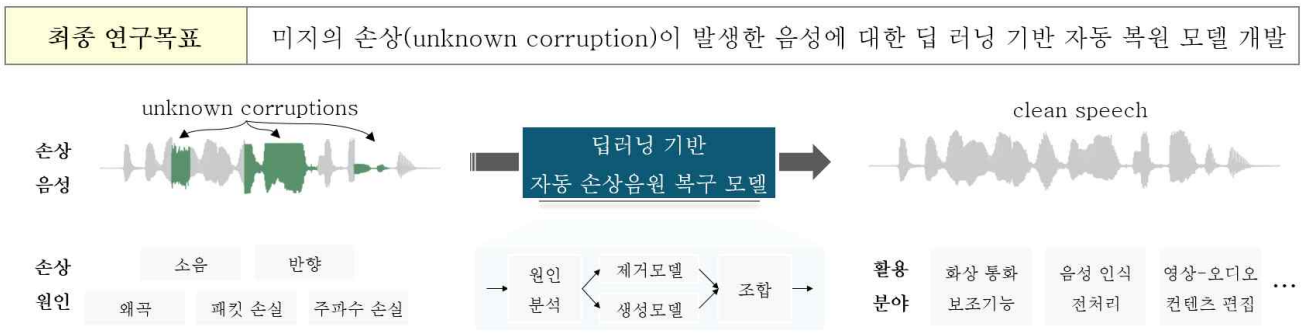
2) 연수의 목적

- 연수자의 박사학위 논문 주제인 기계학습 기반 오디오 음원 분리 기술을 확장하여 비전문가도 실생활에서 유용하게 활용할 수 있는 자동 손상 음원 복구 기술을 연구함으로써 연구의 지속가능성을 높임
- 기계학습 기반 신호처리 기술 연구를 선도하는 해외기관인 C4DM@QMUL[1]에서의 연수를 통해 연수자의 연구 능력을 질적으로 향상하며, 연수 이후에도 활용 가능한 글로벌 인적 네트워크를 구축함

2. 연구개발과제의 목표 및 내용

1) 연구개발과제의 최종 목표

- 본 연구는 [그림 1]과 같이 손상이 발생한 음원의 손상 원인을 파악하고, 이를 복원하여 향상된 음원을 반환하는 딥러닝 기반 기술개발을 최종 연구목표로 함



[그림 1] 최종 연구목표 요약

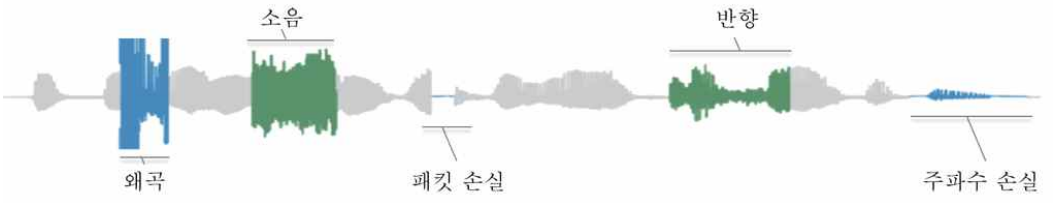
- 녹음된 음성 발화문은 소음, 잔향 효과, 패킷손실 등의 다양한 원인[8]으로 인해 손상됨. 특정 손실 원인에 종속적인 복구 기능을 제공하는 기존연구 및 시스템은 오디오 비전문가가 활용하기에 어려움

- 본 연구에서는 비전문가 또한 손상 음원을 손쉽게 복구할 수 있도록, 임의 손상이 가해진 음원을 분석하고 손상이 가해진 구간에 필요한 조처하는 딥러닝 모델 개발을 목표로 함

1) Queen Mary University in London의 the Centre for Digital Music (https://c4dm.eecs.qmul.ac.uk/)

○ **연구범위-1.** 본 연구에서 다루는 음원 손상의 종류

- 본 연구에서는 발화문에서 흔히 발생하는 다섯 가지[8]의 음원 손상을 중점적으로 다룸: 소음(noise), 잔향(reverb), 왜곡(clipped), 단기 패킷손실 (short-term packet loss), 주파수 대역 손실 (frequency cut-off)



[그림 2] 손상의 종류와 예시

✓ 소음: 기계적·환경적 요인 등으로 생성된 음성 발화문 이외의 모든 소리
✓ 잔향: 공간 구조물에 반사되어 단기간 지연되어 들리는 소리 (메아리보다 짧은 term의 반사)
✓ 왜곡: sample 값이 마이크/스피커 한계 입출력 범위 $[-1, 1]$ 를 초과하는 손상 (찢어지는 느낌의 소리)
✓ 단기 패킷손실: 네트워크 결함 등의 사유로 일부 구간이 손실되는 손상 (본 연구는 1초 미만 단기손실만 다룸)
✓ 주파수 대역 손실: 손실압축/해독 및 네트워크 결함 등의 사유로 특정 주파수 대역이 손실 (ex. 전화 목소리)

○ **연구범위-2.** 본 연구에서 가정한 환경

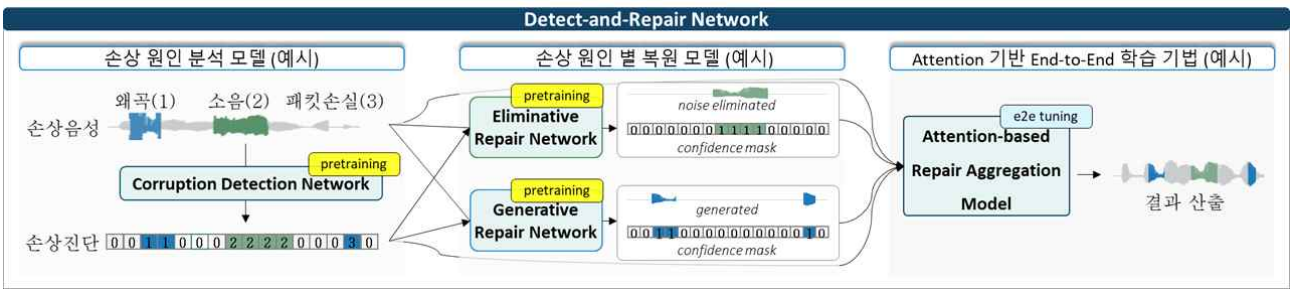
- [1차 목표 환경] [그림 2]와 같이, 단일화자의 발화문 녹음본의 특정 구간에서 발생한 음원 손상을 복원. 손상 원인이 다른 다수의 구간이 존재하며, 각 구간끼리는 겹치지 않음
- [확장환경] 1차 목표에서의 연구 목표가 조기 달성 시 다음과 같이 더욱 복잡한 환경에서 연구를 수행 (단일화자 => 복수화자, 구간끼리 겹치지 않음 => 구간끼리 겹칠 수 있음)

○ **최종목표의 도전성**

- [미지의 음원 손상] 임의의 구간에 임의의 손상이 발생한 음원을 복구해야 하는 본 연구의 환경은 손상이 사전에 알려지지 않았다는 점에서 기존 음성강화 기법으로 해결할 수 없는 도전적인 과제임
- [손상 원인에 따라 달라져야 하는 복구 방법] 본 연구자의 사전연구수행 결과[2], 손상의 성격이 다름에도 단일 복구 방법을 적용할 시[4] 복원 품질이 떨어지는 경향이 있었음. 본 연구에서 개발하고자 하는 모델은 다양한 손상을 고려해야 한다는 점에서 기존기법 적용 시 성능상의 제약이 있는 도전적 과제임

2) 연구개발과제의 내용

- 본 제안연구에서는 최종 연구목표 달성을 위해 [그림 3]과 같은 기계학습 모델을 제안함



[그림 3] 연구내용 요약도: Detect-and-Repair (DaR) Network

2) 본 연구자의 사전연구에서 개발한 텍스트 기반의 음원선별적 편집을 위한 U-Net 모델(데모)의 경우, eliminative 작업은 수월하게 모사하지만 generative한 작업 (왜곡, 단기 패킷손실, 주파수 대역손실에 대한 복구)에서는 성능이 좋지 않음

○ **문제해결 전략:** 제안 모델은 다음과 같은 전략으로 최종목표의 도전성을 해소하고자 함

- [미지의 음원 손상] 본 연구에서는 손상음성의 각 sample에 어떠한 손상이 가해졌는지 감지(detect)한 후, 감지 결과에 conditioned[21]된 복원(repair)을 수행함으로써 [미지의 음원 손상] 문제를 해소함
- [손상 원인에 따라 달라지는 복구 방법] 본 연구에서는 손상의 성격을 크게 두 부류로 분류하여, 불필요한 정보를 제거하는 역할인 eliminative repair model과 부족한 정보를 생성하는 generative repair model을 별도로 둔 후[3], 이들의 복구 결과를 attentive[22]하게 수합하는 방식의 모델을 제안함으로써 [손상 원인에 따라 달라지는 복구 방법]을 제안함

○ 목표 달성을 위해 다음과 같은 세부연구를 수행함 ([그림 3] 및 3장 참조)

<표 1> 세부 연구내용

세부연구	연구내용	성과물
세부연구 1	소음 데이터 및 DSP 기반 학습 프레임워크 LibriFix 구축	LibriFix
세부연구 2	손상 원인 분석 모델 연구/개발	Detect
세부연구 3	손상 원인별 복원 모델 연구/개발	Repair
세부연구 4	Attentive Aggregation을 통한 DaR-Net의 End-to-End 학습 기법 연구	DaR-Net

3) 연구개발과제의 창의성 및 혁신성 등

○ **연구의 창의성**: 본 연구는 기존 음성강화 기법은 해결하지 못했던 창의적인 문제를 다룸 (<표 1> 참조)

<표 2> 기존 기술과 제안 기술의 비교 (N=소음, R=잔향, C=왜곡, L=패킷손실, F=주파수 대역손실)

	기반기법	대상 음원손상					음성손상의 형태	미지의 손상에 따른 대처
		N	R	C	L	F		
[1]	algorithmic	○	·	○	·	○	알려진 손상	N/A
[2]	GAN	○	·	·	·	·		
[3]	Unet	○	○	·	·	·		
[5]	DNN	·	·	○	·	·		
[6]	Unet	·	·	·	·	○		
[7]	GAN	·	·	·	○	·		
[4]	GAN	○	○	·	·	○	미지의 손상 (전체구간)	단일 방법론(GAN)에 의존 => 성능저하
제안	DaR-Net	○	○	○	○	○	미지의 손상 (임의구간)	두 가지 이상의 방법론 사용 예정

○ **연구의 혁신성**: 본 연구는 다음과 같은 관점에서 음성강화 분야의 혁신을 가져올 수 있는 연구임

- izotope 사의 RX8 등의 기존 음성강화 기법은 다양한 음성 복원 기능을 제공하지만, 복원 이전에 음향 전문가의 진단이 필요하다는 불편함이 있음. 이를테면 왜곡(clipped)을 인지해야 관련 복원 기능인 declipping을 적용하는데, 음향 전문가가 아닌 일반인은 왜곡이 발생했는지조차 진단하기 어려우므로 복원 기능을 효과적으로 사용하는 것이 어려움
- 제안 연구의 경우, 기존기법들이 개별적으로 다루었던 다섯 가지 주요 손상에 대한 원클릭(one-click)에 복구 방법론을 제안함. 비전문가 또한 손쉽게 음성을 복원할 수 있도록, 복원에 필요한 진단을 모델이 스스로 내리는 기능까지 탑재한 모델을 개발함으로써 음성강화 분야의 혁신을 선도하고자 함.

3) 본 연구자의 사전연구 결과, 불필요한 정보를 제거하는 복구(소음, 잔향)에는 U-Net[19] 기법이, 부족한 정보를 생성해야 하는 복구(그 외)에는 GAN[20] 기법이 더 좋은 성능을 보였음

3. 연구개발과제의 추진전략 · 방법 및 추진체계

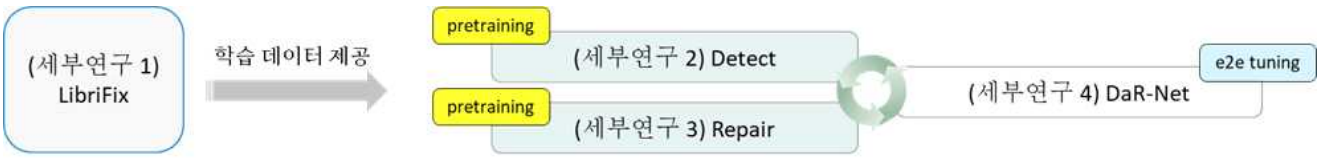
1) 연구개발과제의 추진전략·방법

○ 상기 연구 목표를 달성하기 위해 본 연구자는 세부 연구목표별로 다음과 같은 독창적 전략을 수립하였음

- [세부연구 1: LibriFix] 소음 데이터 및 DSP 기반 학습 프레임워크 LibriFix 구축
  음성데이터(LibriSpeech[17]), 소음데이터(Wham![18]), Digital Signal Processing(DSP)를 활용하여 다섯 가지 손상을 시뮬레이션하여 지도학습을 위한 훈련 데이터를 제공하는 프레임워크 LibriFix를 개발
  유사한 DSP 기반 기존연구인 LibriMix[16]를 참조 및 활용하여 개발 시간 단축
- [세부연구 2: Detect] 손상 원인 분석 모델 개발
  LibriFix를 기반으로, 손상음성을 입력받고 각 sample에 어떠한 손상(corruption)이 발생하였는지 감지하는 corruption detection network를 연구/개발함.
  입출력의 특성을 반영하여 Wavefrom-to-Waveform 모델을 우선적으로 구현함
- [세부연구 3: Repair] 손상 원인별 복원 모델 개발
  손상 원인 분석 결과에 conditioned된 복원 모델을 연구/개발함. 불필요한 정보를 제거하는 역할인 eliminative repair model과 부족한 정보를 생성하는 generative repair model을 별도로 훈련함. 사전연구 수행 결과 후자의 경우 Time-Frequency 기반보다는 Time domain 기반 기법이 더 유리하였음
- [세부연구 4: DaR-Net] Attentive Aggregation을 통한 DaR-Net의 End-to-End 학습 기법 연구
  사전훈련된 Detect 모델과 Repair 모델을 연동하여 손상 원인을 감지한 후 이를 기반으로 음성 복구를 하는 Detect-and-Repair Network (DaR-Net)을 구현하고, 이를 end-to-end로 fine-tuning 함.
  Attention 기반의 aggregation 모듈을 통해 각 모델의 복원 결과를 soft masking하여 수합함

2) 연구개발과제의 추진체계

- 본 연구자는 세부연구별 의존성을 고려하여 다음과 같은 추진체계를 수립하였음



[그림 4] 제안 연구의 세부연구의 의존성을 고려한 추친체계

4. 연구자의 연구 수행역량

- 본 연구자는 머신러닝 기반의 디지털 신호처리 분야에서 다음과 같은 연구를 수행해 옴. 특히 본 과제와 가장 연관성이 깊은 연구인 “다양한 음향신호처리에 범용적 적용이 가능한 신경망 기반 펙트로그램 변환블록 연구” 및 “Deep Attention Network 기반의 음원분리 기법 개발”을 통해 축적된 경험을 바탕으로 제안연구를 차질없이 수행할 계획임. 다음은 본 연구자의 최근 과제 참여 및 수행 내역임

<표 3> 과제 수행내역 (*표시된 박사과정생연구지원장려금지원사업 제외, 지도교수인 고려대 정순영교수가 연구책임자)

수행 기간	과제명	과제 기관
2020 – 2021	다양한 음향신호처리에 범용적 적용이 가능한 신경망 기반 펙트로그램 변환블록 연구	한국연구재단
2019 – 2021	Deep Attention Network 기반의 음원분리 기법 개발*	한국연구재단
2019 - 2020	인공신경망 기반 고차원 범위질의처리 기법 연구	한국연구재단
2016 – 2018	다차원 의미정보 기반의 예측적 이동성 집계 분석 기법 개발	한국연구재단

○ **[선행연구]** 음원분리 분야에서의 연구경험과 성과

- [가창음원분리 모델] Frequency Transformation을 도입하여 MUSDB18[24]에 대한 가창음원분리 부분에서 SOTA를 달성하였으며, 이를 권위 있는 음악검색 분야 학술대회인 ISMIR2020에서 발표 [21]
  ISMIR는 h5 인덱스 기준 음악 분야에서 rank #1, 멀티미디어 분야에서 rank #8인 top conference 임
  – 관련 링크: 인터렉티브 데모 [26] 링크, Github [27] 링크, 발표 포스터 [28] 링크

- [잠재 음원성분 분석 기반 조건부음원분리 모델] 가창음원분리 모델을 일반화한 조건부 음원분리 모델에 잠재 음원성분 분석 기법을 탑재하여 MUSDB18에 대해 보컬뿐만이 아닌 기타악기에서도 SOTA를 달성하였으며, 이를 권위 있는 신호처리 분햐 학술대회인 ICASSP에 투고하여 accept 됨 [23]
  ICASSP는 h5 인덱스 기준 신호처리 분야에서 rank #4 인 top conference 임
  – 관련 링크: 인터렉티브 데모 [29] 링크 , 데모 [30] 링크, Github [31] 링크, 발표 슬라이드 [32] 링크

- 이 외에도 데이터 셋 확보 등의 작업을 수행하여 차질없이 과제를 수행할 수 있도록 기반을 갖추었음

5. 연수기관 및 지도교수의 적합성

- 기계학습 기반 디지털 신호처리기법을 기반으로 하는 본 연구와 관련된 분야를 선도하는 정상 연구기관으로는 Queen Mary University in London의 Centre for Digital Music (C4DM), Pompeu Fabra University의 Music Teechnology Group, 서울대학교 Music and Audio Research Group, Georgia Tech Center for Music Technology의 Music Informatics Group 등이 있음

- 본 연구자는 이 중 C4DM의 소속인 Joshua D Reiss 교수를 본 제안연구를 가장 잘 지도해줄 수 있는 역량을 갖추었다고 판단하였음. 해당 교수는 기계학습과 디지털 신호처리를 접목한 다양한 기법[25, 33-36]을 ICML, ICASSP 등의 권위 있는 학술대회에서 발표한 이력이 있음. 또한, 본 연구자는 해당 교수의 박사과정 지도 학생이었던 현 Sony 소속 Marco Antonio Martinez Ramirez 박사와 공동연구를 수행 중으로 추후 3자 간의 협업을 기대할 수 있기에 해당 교수를 가장 적합한 지도교수라고 판단함

- 해당 교수에 문의 결과, 본 연구자를 박사후과정연구생으로서 적극적으로 추천한다는 회신을 받음

6. 연구개발성과의 기대효과

- 비전문가도 사용 가능한 음성강화 기술의 수요가 높으므로, 본 연구과제 결과물은 상업적인 가치가 뛰어남. 화상통화, 음성 하드웨어 등의 분야에서 음성강화 전문가 채용이 큰 폭으로 증가하고 있는 최근 석/박사 취업 시장 동향은 음성강화 기술에 대한 높은 수요를 방증함

- 제안 연구의 경우, 기존기법들이 개별적으로 다루었던 다섯 가지 주요 손상에 대한 원클릭(one-click) 에 복구 방법론을 제안함. 비전문가도 손쉽게 음성을 복원할 수 있도록, 복원에 필요한 진단을 모델이 스스로 내리는 기능까지 탑재한 모델을 개발함으로써 음성강화의 학술적 혁신 또한 선도할 수 있음

7. 기타 (참고문헌 포함)

해당사항 없음

[참고문헌(Reference)] (※ 작성분량에서 제외)

<표 2>에서 활용 [1~8]

[1] Laguna, Christopher. “REPAIR : A Web Application For Audio Quality Enhancement.” (2016).

[2] Pascual, Santiago, Antonio Bonafonte, and Joan Serra. “SEGAN: Speech Enhancement Generative Adversarial Network.“ Proc. Interspeech 2017 (2017): 3642-3646.

[3] Choi, Hyeong-Seok, et al. “Phase-aware single-stage speech denoising and dereverberation with u-net.“ arXiv preprint arXiv:2006.00687 (2020).

[4] Su, Jiaqi, Zeyu Jin, and Adam Finkelstein. “HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks.“ arXiv preprint arXiv:2006.05694 (2020). (separation + dereverberance + equalization)

[5] W. Mack and E. A. P. Habets, “Declipping speech using deep filtering,” in 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct. 2019, pp. 200?204.

[6] Birnbaum, Sawyer, et al. “Temporal FiLM: Capturing Long-Range Sequence Dependencies with Feature-Wise Modulations.“ arXiv e-prints (2019): arXiv-1909.

[7] Marafioti, Andres, et al. “GACELA--A generative adversarial context encoder for long audio inpainting.“ arXiv preprint arXiv:2005.05032 (2020).

[8] Barry, Dan, Alessandro Ragano, and Andrew Hines. “Audio Inpainting based on Self-similarity for Sound Source Separation Applications.“ 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2020.

1-1. 연구의 필요성에서 활용 [9-15]

[9] Adobe, “2021 Research Intern ? Audio.”, adobe.wd5.myworkdayjobs.com/en-US/external_university/job/Job-Title--2021-Research-Intern---Audio_R102056. (Accessed 2021-02-26)

[10] Facebook, “Research Scientist, Speech Enhancement“ https://www.facebook.com/careers/v2/jobs/480683763298425/. (Accessed 2021-02-26)

[11] Google, “Algorithm Engineer, Audio, Devices & Services“ https://careers.google.com/jobs/results/98751346741519046-algorithm-engineer-audio-devices-services/. (Accessed 2021-02-26)

[12] Bose, “Machine Learning Research Engineer“ https://boseallaboutme.wd1.myworkdayjobs.com/en-US/Bose_Careers/job/US-MA---Framingham/Machine-Learning-Research-Engineer_R19511-1. (Accessed 2021-02-26)

[13] Qualcomm, “Machine Learning Researcher ? Audio and Speech“ https://qualcomm.wd5.myworkdayjobs.com/en-US/External/job/Seoul/Machine-Learning-Researcher---Audio-and-Speech_1979539. (Accessed 2021-02-26)

[14] Microsoft, “Research Intern - Azure Cognitive Services: Speech“ https://careers.microsoft.com/us/en/job/926631/Research-Intern-Azure-Cognitive-Services-Speech.

(Accessed 2021-02-26)
[15] Bytedance, “Research Scientist in Speech & Audio (Intelligent Creation Lab) - 2021 StartApplied“ https://jobs.bytedance.com/en/position/6893626812188018957/detail. (Accessed 2021-02-26)

3장에서 활용 [16-18]

[16] Cosentino, Joris, et al. “Librimix: An open-source dataset for generalizable speech separation.“ arXiv preprint arXiv:2005.11262 (2020).
[17] Panayotov, Vassil, et al. “Librispeech: an asr corpus based on public domain audio books.“ 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015.
[18] Wichern, Gordon, et al. “WHAM!: Extending speech separation to noisy environments.“ arXiv preprint arXiv:1907.01160 (2019).

4페이지 주석, 본 연구자의 사전연구 [19-20]

[19] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.“ International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.
[20] Goodfellow, Ian J., et al. “Generative adversarial networks.“ arXiv preprint arXiv:1406.2661 (2014).

문제해결 전략 [21-22]

[21] Choi, Woosung., Kim, Minseok., Chung, Jaehwa., and Jung, Soonyoung. “LaSAFT: Latent Source Attentive Frequency Transformation for Conditioned Source Separation.” arXiv preprint arXiv:2010.11631 (2020). (accepted to ICASSP 2021)
[22] Vaswani, Ashish, et al. “Attention is all you need.“ arXiv preprint arXiv:1706.03762 (2017).

4. 연구수행역량 [23-32]

[23] Choi, Woosung., Kim, Minseok., Chung, Jaehwa., Lee, Daewon., and Jung, Soonyoung. “Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation.” 21th International Society for Music Information Retrieval Conference, ISMIR, 2020.
[24] Rafii, Zafar, et al. “MUSDB18-a corpus for music separation.“ (2017).
[25] Ramirez, Marco A. Martinez, Emmanouil Benetos, and Joshua D. Reiss. “Modeling plate and spring reverberation using a DSP-informed deep neural network.“ ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[26] Choi, Woosung, et al. https://github.com/ws-choi/ISMIR2020_U_Nets_SVS/tree/master/colab_demo
[27] Choi, Woosung, et al. https://github.com/ws-choi/ISMIR2020_U_Nets_SVS
[28] Choi, Woosung, et al. https://program.ismir2020.net/poster_2-04.html
[29] Choi, Woosung, et al. https://github.com/ws-choi/Conditioned-Source-Separation-LaSAFT/tree/main/colab_demo
[30] Choi, Woosung, et al. https://lasaft.github.io/

[31] Choi, Woosung, et al. https://github.com/ws-choi/Conditioned-Source-Separation-LaSAFT
[32] Choi, Woosung, et al. https://lightsaft.github.io/slide/gaudio/

5. 연수기관 및 지도교수의 적합성 [33~36]

[33] Ramirez, MA Martinez, Emmanouil Benetos, and Joshua D. Reiss. “Deep learning for black-box modeling of audio effects.“ Applied Sciences 10.2 (2020): 638.
[34] Ramirez, Marco A. Martinez, and Joshua D. Reiss. “Modeling nonlinear audio effects with end-to-end deep neural networks.“ ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[35] Wilkinson, William J., et al. “Unifying probabilistic models for time-frequency analysis.“ ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[36] Wilkinson, William J., et al. “End-to-End Probabilistic Inference for Nonstationary Audio Analysis.“ Thirty-sixth International Conference on Machine Learning. 2019.