FINAL-PROJECT
L3DAS21 – TASK1
3D SPEECH ENHANCEMENT

學生姓名:蔡承宏

學號:609415074

OUTLINE

1. Introduction
2. Datasets
3. Model
4. Demo
5. Conclusion
6. Reference

INTRODUCTION

1. 3D audio which field of application is incredibly wide and ranges from virtual and real conferencing to game development, music production, autonomous driving, surveillance and many more.

2. 3D audio formats permit to obtain an impressive performance in many machine learning-based tasks, usually bringing out a significant improvement over the single/dual-channel formats. Ex : Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features[1] 、Sound event detection using spatial features and convolutional recurrent neural network[2]

3. The L3DAS21 Challenge organized within the L3DAS(Learning 3D Audio Sources) is designed to encourage and foster research on machine learning for 3D audio signal processing.

4. 3D Speech Enhancement (SE) is relying on multiple-source and multiple-perspective (MSMP) Ambisonics recordings.

INTRODUCTION

Ambisonics recordings :

1. Ambisonics is a 360° surround sound format that, unlike conventional stereo and surround sound, captures the full directivity information for every soundwave that hits the microphone – including height information – and is therefore ideal for immersive audio applications.

2. First you need an ambisonic microphone with four capsules arranged in a tetrahedral array – such as the RØDE NT-SF1 – and a four-track recorder. The raw audio recorded when using the NT-SF1 is called A-Format, which in and of itself not especially useful. However, by using the SoundField by RØDE Plug-in, you can manipulate the signal to emulate any type of microphone pattern and polarity. These ‘virtual’ microphones can be pointed or 'steered' in any direction you choose

RØDE NT-SF1

INTRODUCTION

1. 3D SE aims at removing unwanted information from spurious spatial vocal recordings and further enhancing the speech intelligibility and clarity.

2. A widespread strategy to perform SE is to use deep neural networks (DNNs) to estimate a time-frequency mask in the Fourier domain that extracts clean speech signals from noisy spectra

3. Neural beamforming techniques as Filter and Sum Networks (FaSNet) provide state-of-the art results for Ambisonics-based SE and are usually suitable for low-latency scenarios.

4. Other techniques to perform SE include recurrent neural networks(RNNs), graph-based spectral subtraction, discriminative learning, dilated convolutions.

5. The objective of this task is the separation and enhancement of speech signals immersed in a noisy 3D environment.the models are expected to extract the monophonic voice signal from the 3D mixture that contains various background noises.

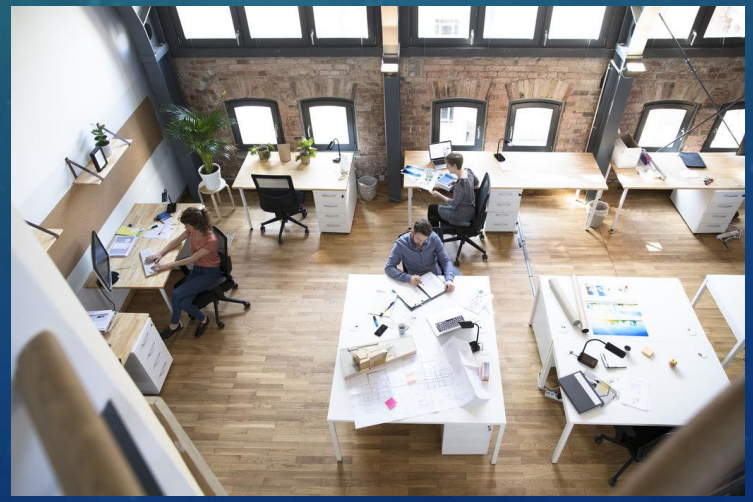
DATASETS

Environment Settings

1. The LEDAS21 dataset contains approximately 65 hours of MSMP Ambisonics audio recordings.

2. We sampledthe acoustic field of a large office room with the approximate dimensions of 6 m (length) by 5 m (width) by 3 m (height).

3. The room has typical office furniture: desks, chairs and wardrobe. The floor is made of wood parquet, while the walls and the ceiling are made of painted concrete.



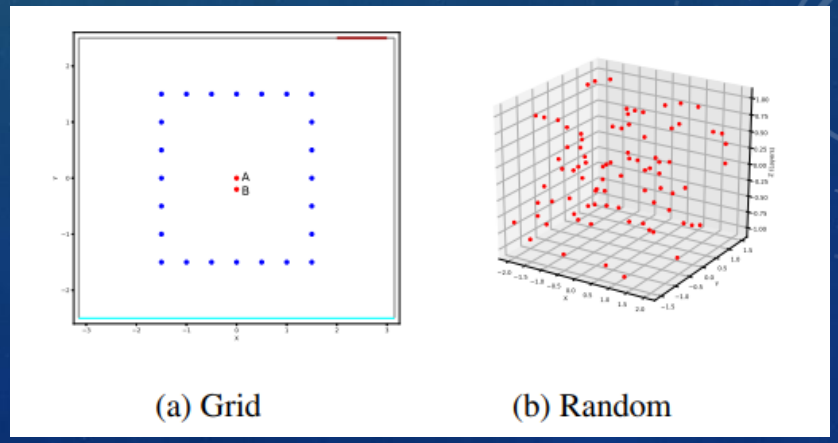
DATASETS

Microphones & Speakers Settings

1. We placed two Ambisonics microphones in the center of the room and we moved a speaker reproducing an analytic signal in 252 fixed spatial positions.

2. One microphone (mic A) lies in the exact center of the room, shown as a red dot in Fig. 1a, and the other (mic B) is 20 cm distant towards the width dimension. Both microphones are positioned at the same height of 1.3 m, which is the average ear height ear of a seated person. The capsules of both mics have the same orientation.

3. The analytic signal is a 24-bit exponential sinusoidal sweep that glides from 50 Hz to 16000 Hz in 20 seconds,reproduced at 90 dB SPL on average.



(a) Grid (b) Random

DATASETS

Noise Datasets

1. We used the Librispeech and FSD50K datasets.

2. We selected a total of 1440 noise sound files from FSD50K, divided into 14 transient noise classes: computer keyboard, drawer open/close, cupboard open/close, finger snapping, keys jangling, knock, laughter, scissors, telephone, writing, chink and clink, printer, female speech, male speech, and 4 continuous noise classes: alarm, crackle, mechanical fan, microwave oven. We collected 80 sounds for each noise class.

3. we extracted clean speech signals from Librispeech, taking only sound files up to 10 seconds.

DATASETS

SE(Speech Enhancement) Datasets

1. we created more than 30000 virtual 3D audio environments with a duration up to 10 seconds each, reaching a total duration of approximately 50 hours.

2. In each data point a speech signal is always present, mixed with various types of background noise. We extracted all sounds from the clean subset of Librispeech.

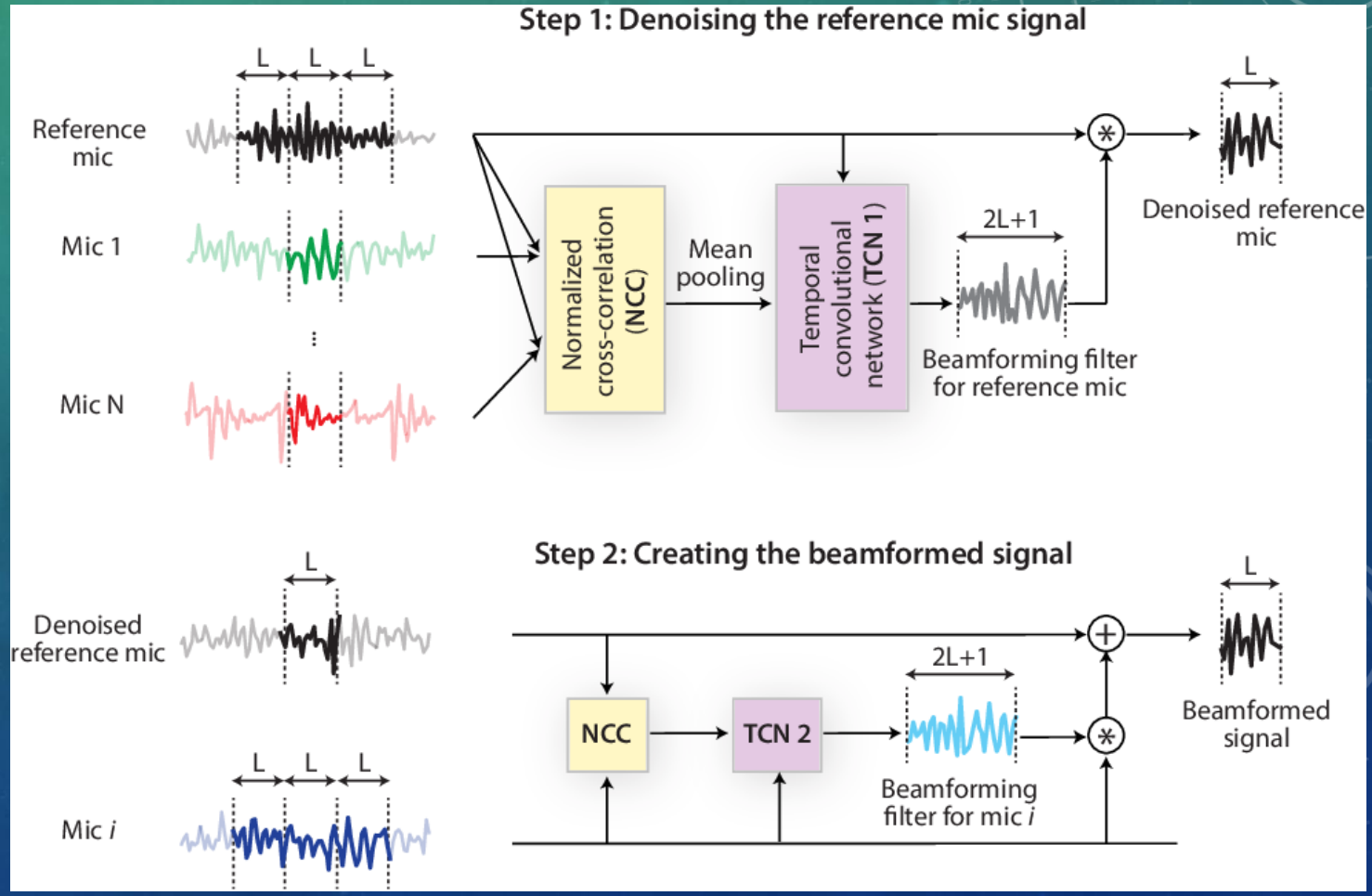
3. We add up to 3 non-speech background noises of the above-mentioned categories, extracting them from FSD50K.With a 25% chance, one of the background noises is a continuous noise. The signal-to-noise ratio ranges from 6 to 16 dB full scale (dBFS).

MODEL

FaSNet (Filter and Sum Network)

1. filter-and-sum network (FaSNet), a time-domain, filter-based beamforming approach suitable for low-latency scenarios.

2. FaSNet has a two-stage system design that first learns frame-level time-domain adaptive beamforming filters for a selected reference channel, and then calculate the filters for all remaining channels. The filtered outputs at all channels are summed to generate the final output.

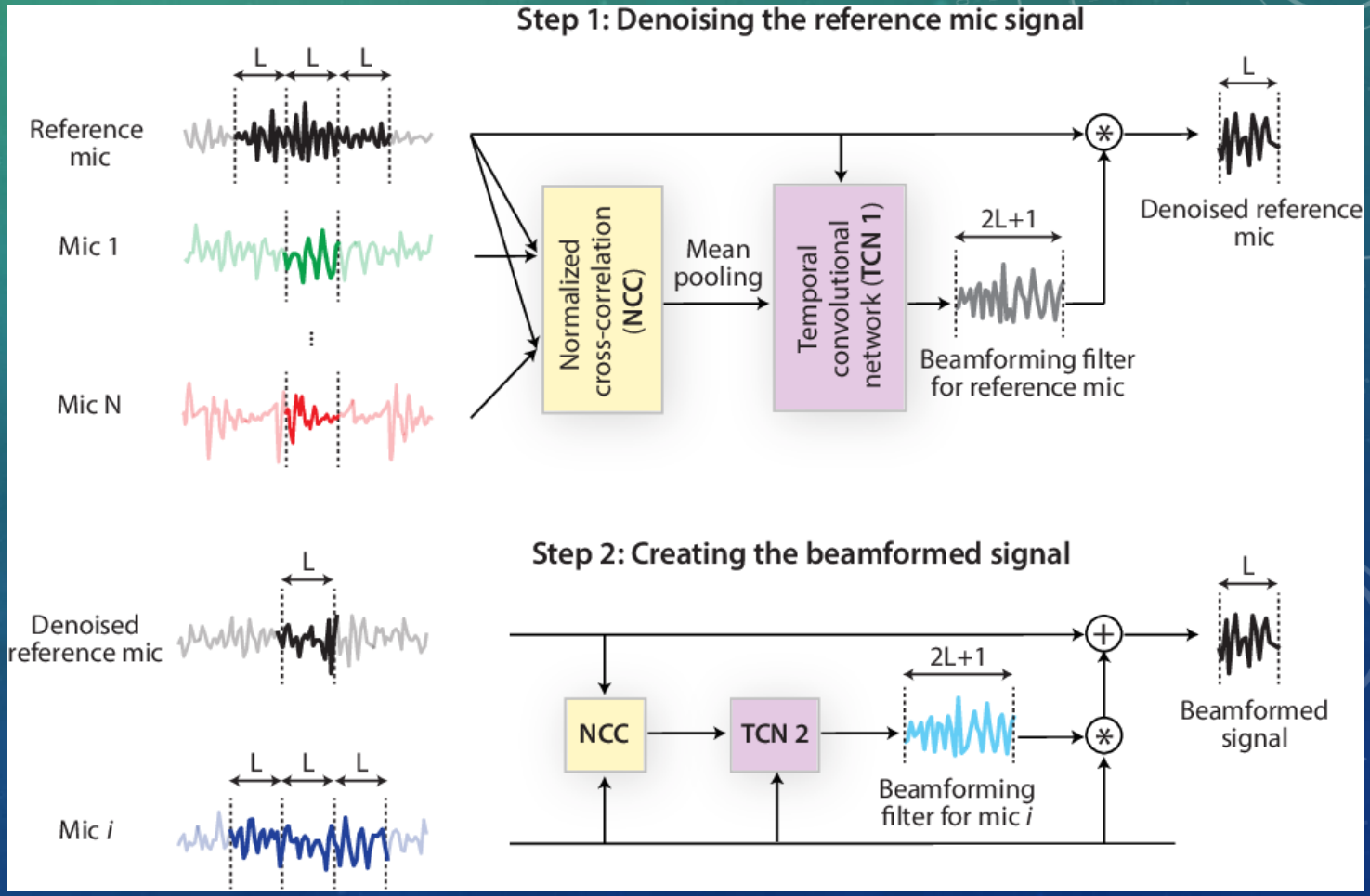


FaSNet system flowchart

MODEL

FaSNet (Filter and Sum Network)

3. The first stage estimates the frame-level beamforming filters for the reference microphone based on the normalized correlation correlation coefficient (NCC) feature, and the second stage uses the cleaned reference microphone signal to estimate the beamforming filters for all remaining microphones. Cosine similarity is used as the NCC feature, and the temporal convolutional network (TCN) is selected as the filter estimation module.



FaSNet system flowchart

MODEL

Evaluation Metrics

The evaluation metric for this task is the short-time objective intelligibility (STOI), which estimates the intelligibility of the output speech signal. Moreover, word error rate (WER).

- Short-Time Objective Intelligibility (STOI) :

1. It is one of the important metrics to measure the intelligibility of speech.
2. The value range of STOI is defined in 0 ~ 1.

- Word Error Rate (WER) :

1. Word error rate (WER) is a common metric of the performance of a speech recognition or machine translation system.

The final metric for this task is a combination of these two measures given by *(STOI + (1 − W ER))/2* .
This metric lies therefore in the 0-1 range and higher values are better.

DEMO

Environment Settings

Operation System	Windows 10	Package	Scipy 1.4.1
Programming language	Python 3.7	Package	Soundfile 0.10.3
Package	Pytorch 1.9	Package	Torch 1.0.1
Package	Jiwer 2.2	Package	Transformers 4.4.2
Package	Librosa 0.8	Package	Tqdm 4.36.1
Package	Numpy 1.18.1	Package	Wget 3.2
Package	Pandas 1.0.3	Package	Pystoi 0.3.3

DEMO

Environment Settings

Problem	Solution
NameError: name 'transformers' is not defined	pip install transformers[torch] (merge transformers & torch)
Can’t find FaSNet Model	Enter this Link to download FaSNet Model https://github.com/yluo42/TAC
MemoryError: Unable to allocate 14.9 GiB for an array with shape (15603, 4, 32000) and data type float64	Memory is not enough to training,so I decrease training‘s batch size from 20 to 8. It will increase training time.

DEMO

Pre-Processing

loading the raw audio waveforms and their correspondent metadata, apply custom pre-processing functions and save numpy arrays (.pkl files) containing the separate predictors and target matrices.

For Task1 the function returns 2 numpy arrays contatining:

- Input multichannel audio waveforms (3d noise+speech scenarios) - Shape: [n_data, n_channels, n_samples].
- Output monoaural audio waveforms (clean speech) - Shape [n_data, 1, n_samples].

DEMO

Training

	Epoch	Training Loss	Validation Loss	Test Loss
Our model	Pre-training+200	0.0034302054 (149 Epoch)	0.009131522 (149 Epoch)	0.008453996 (149 Epoch)

DEMO

Evaluate

Evaluate Metrics	Pre-training model Score	Our model Score
WER	0.4817644879932805	**0.4688954766711340 7**
STOI	0.722633032440839	**0.7412322822793246**
Task 1 metric	0.6204342722237781	**0.6361684028040964**

DEMO

Schedule

To-Do List	Complete
Download Datasets	☑ ok
Pre-Processing	☑ ok
Training	☑ ok
Evaluate	☑ ok
Report	☑ ok

CONCLUSION

- FaSNet is a time-domain adaptive beamforming method especially suitable for low-latency applications.
- FaSNet was designed as a two-stage system, where the first stage estimated the beamforming filter for a randomly selected reference microphone, and the second stage used the output of the first stage to calculate the filters for all the remaining microphones.
- FaSNet can also be concatenated with any other single-channel system for further performance improvement.

REFERENCE

1. Guizzo, Eric, et al. "L3DAS21 Challenge: Machine Learning for 3D Audio Signal Processing." arXiv preprint arXiv:2104.05499 (2021).

2. Luo, Y., Han, C., Mesgarani, N., Ceolini, E., & Liu, S. C. (2019, December). FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 260-267). IEEE.