CVSSP
Centre for Vision, Speech and Signal Processing

UNIVERSITY OF SURREY

Audio-Text Cross Modal Generation

Wenwu Wang

Centre for Vision, Speech and Signal Processing (CVSSP)
& Institute for People Centred Artificial Intelligence

University of Surrey

United Kingdom

Email: w.wang@surrey.ac.uk
Web: http://personal.ee.surrey.ac.uk/Personal/W.Wang/

Perspective Talk on ICASSP 2023, 6 June, Rhodes, Greece

Many thanks to...

- **Xinhao Mei**
- **Haohe Liu**
- **Xubo Liu**
- **Zehua Chen**
- **Qiuqiang Kong**
- **Mark Plumbley**
- **Danilo Mandic**
- Yi Yuan
- Jianyuan Sun
- Jian Guan
- Feiyang Xiao
- Yong Xu
- Yin Cao
- Turab Iqbal
- Yuxuan Wang
- Volkan Kilic
- Qiaoxi Zhu
- Philip Jackson
- Qiang Huang
- David Frohlich
- Emily Corrigan-Kavanagh
- Marc Green
- Andres Fernandes
- Christian Kroos
- Arshdeep Singh

Funding from:
UK Engineering and Physical Sciences Research Council (EPSRC) & British Council Newton Institutional Links Award

Outline

Sound Recognition

Audio-to-Text

- Problem description and existing methods
- Several open problems & our methods

Text-to-Audio

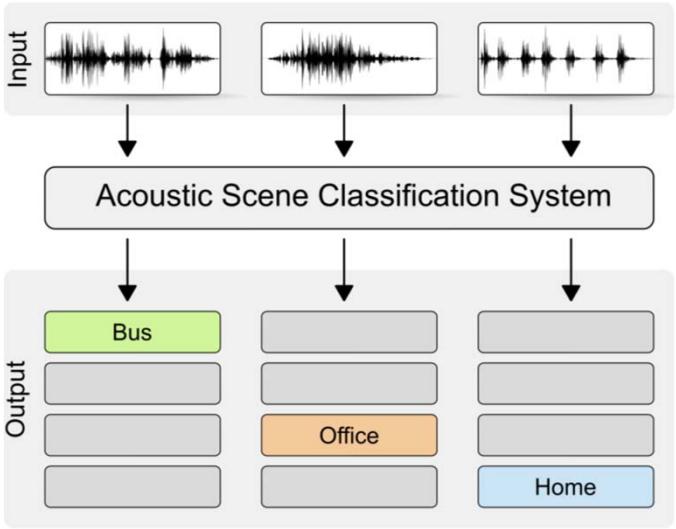
- Problem description and existing methods
- Our method AudioLDM & sound demos

Perspectives on Audio-Language Learning

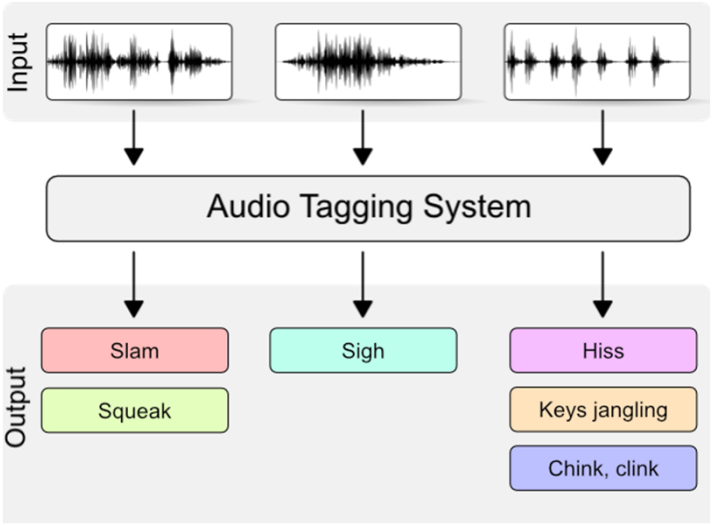
Conclusion & Future Work

Sound Recognition

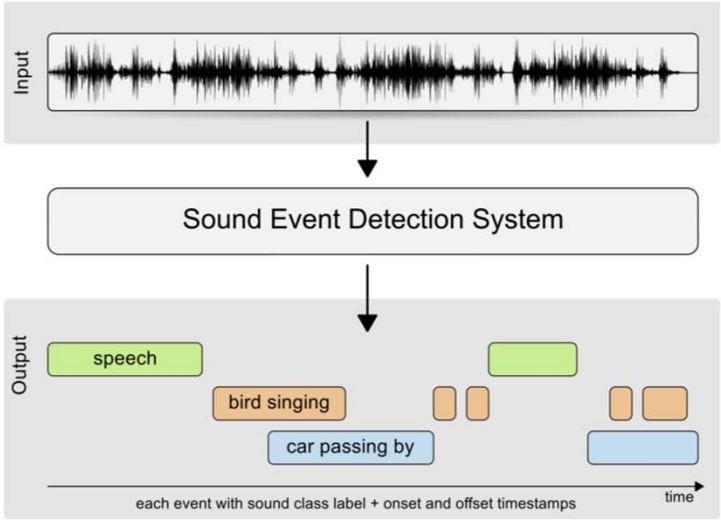
Scene Classification



Audio Tagging



Sound Event Detection



Applications:

- Security (glass break, smoke alarms)
- Medical sounds (heart, lung)
- Traffic (autonomous driving)
- Assisted living (human activity)
- Environmental (birds, insects, noise)
- Multimedia (video+sound search)

AudioSet (Google) (2017)

[Gemmeke, et al. ICASSP2017]

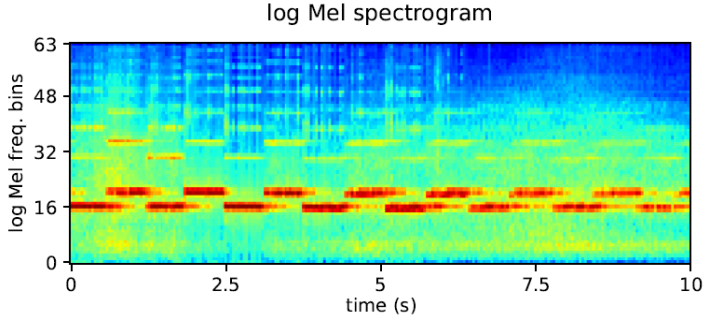
- From Google YouTube videos:
  2.1 million x 10-sec segments
- Labelled from 632 audio event categories (hierarchical tags)
- Multiple labels (average 2.7 per clip)
- Maximum depth: 6
- Example:
  Sound of things -> Vehicle
  -> Motor vehicle -> Emergency vehicle
  -> Siren -> Ambulance siren

Top-level Categories

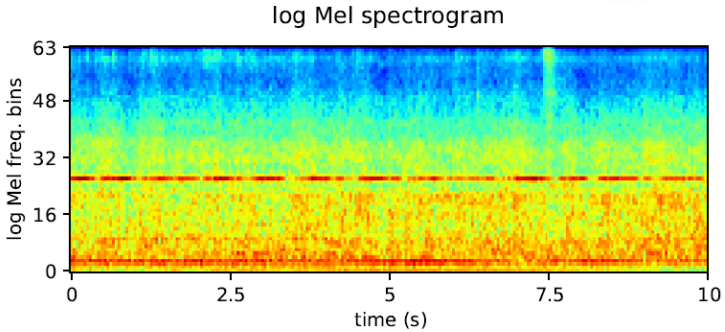
- **Human sounds**
  - *Human voice*
  - *Whistling*
  - *Respiratory sounds*
  - *Human locomotion*
  - *Digestive*
  - *Hands*
  - *Heart sounds, heartbeat*
  - *Otoacoustic emission*
  - *Human group actions*
- **Animal sounds**
  - *Domestic animals, pets*
  - *Livestock, farm animals, working animals*
  - *Wild animals*
- **Natural sounds**
  - *Wind*
  - *Thunderstorm*
  - *Water*
  - *Fire*
- **Music**
  - *Musical instrument*
  - *Music genre*
  - *Musical concepts*
  - *Music role*
  - *Music mood*
- **Sounds of things**
  - *Vehicle*
  - *Engine*
  - *Domestic sounds, home sounds*
  - *Bell*
  - *Alarm*
  - *Mechanisms*
  - *Tools*
  - *Explosion*
  - *Wood*
  - *Glass*
  - *Liquid*
  - *Miscellaneous sources*
  - *Specific impact sounds*
- **Source-ambiguous sounds**
  - *Generic impact sounds*
  - *Surface contact*
  - *Deformable shell*
  - *Onomatopoeia*
  - *Silence*
  - *Other sourceless*
- **Channel, environment and background**
  - *Acoustic environment*
  - *Noise*
  - *Sound reproduction*

Audio Tagging with Weakly Labelled Data

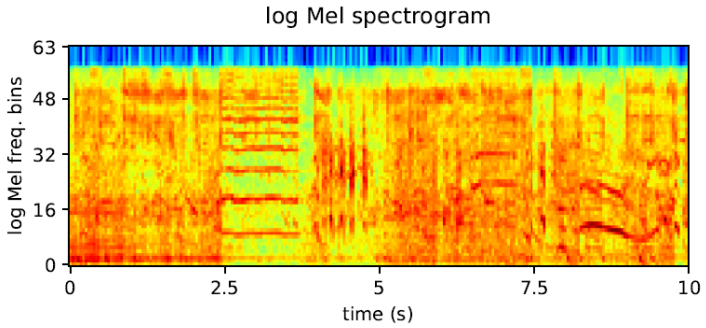
Ambulance (siren)



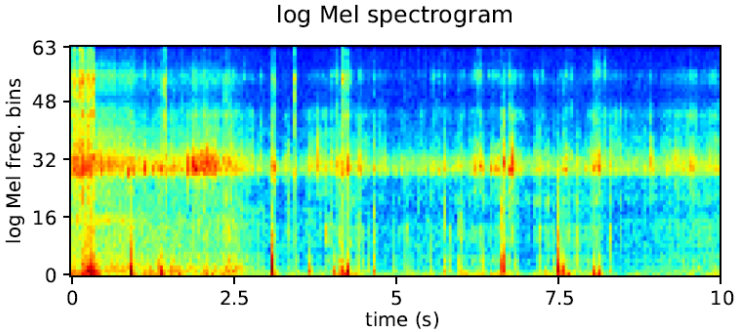
Reverse beep



Air horn, truck horn



Bicycle



Frequency

Time

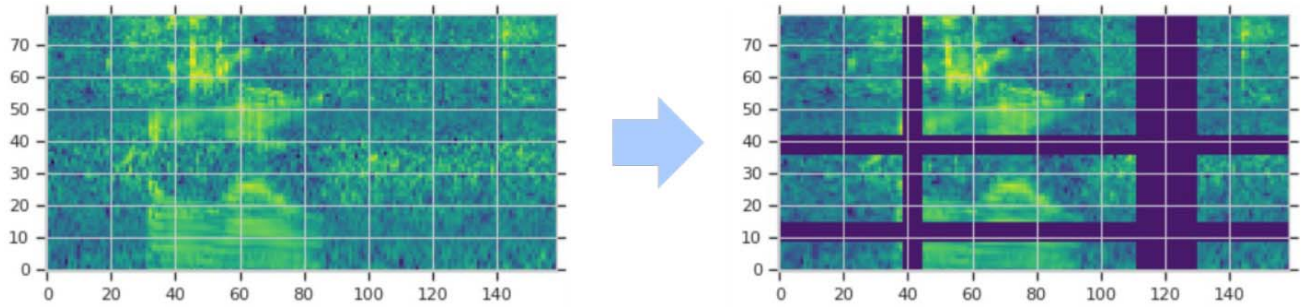
PANNs: Large-Scale Pre-trained Audio Neural Networks

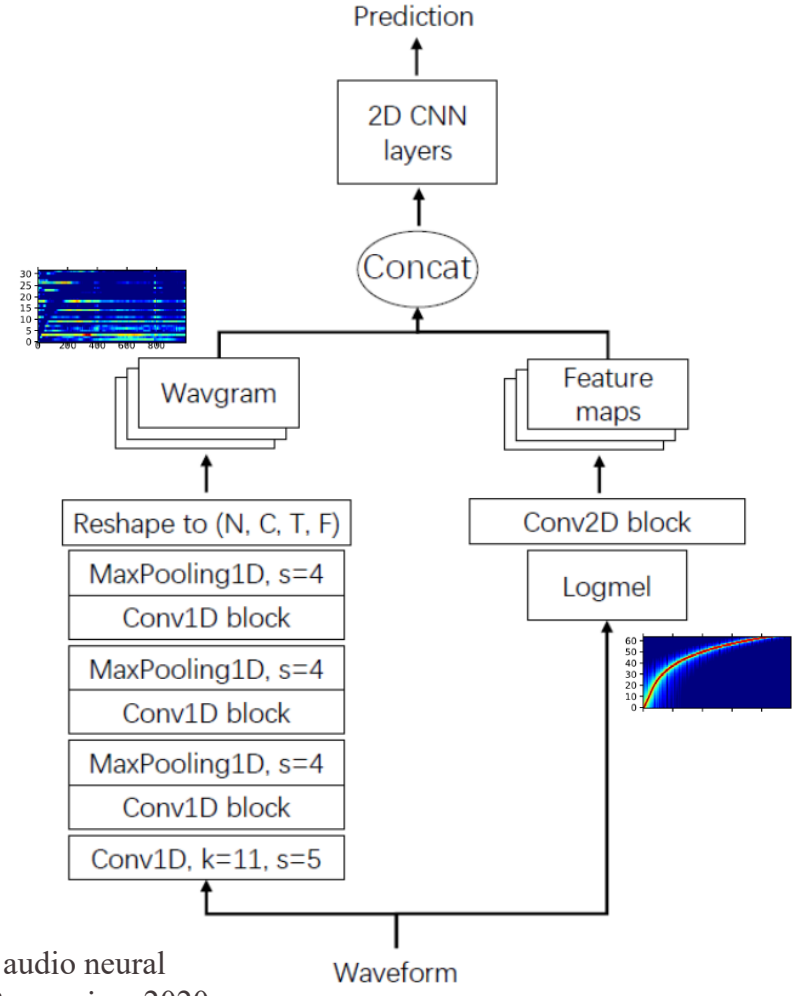
UNIVERSITY OF SURREY

Wavegram-Logmel-CNN for AudioSet tagging

- Time-domain ("Wavegram"), plus
- Log mel spectrogram

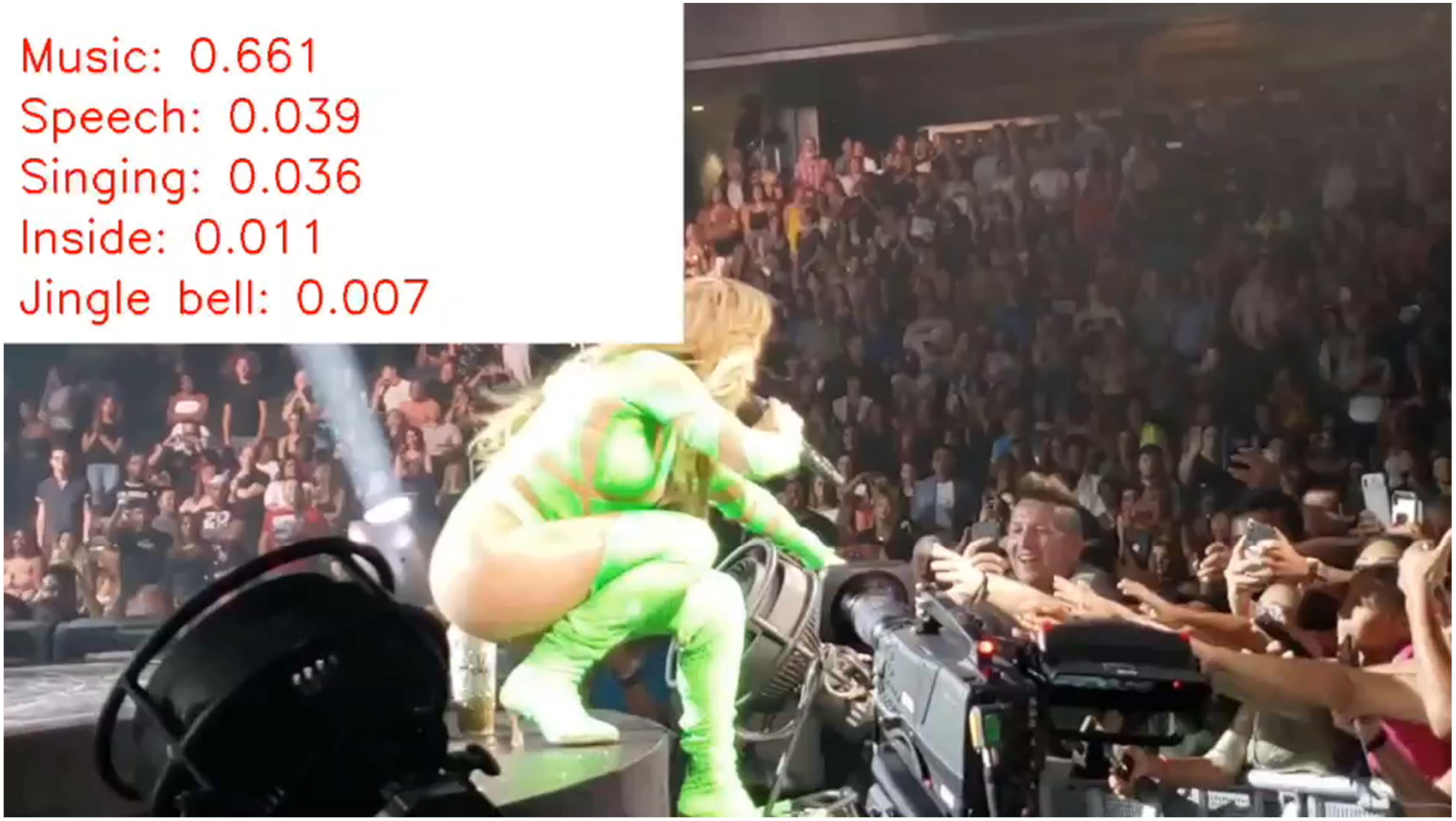
Data augmentation, e.g. use SpecAugment: randomly mask time and frequency stripes of log mel spectrogram





Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: large-scale pretrained audio neural networks for audio pattern recognition", *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2020.

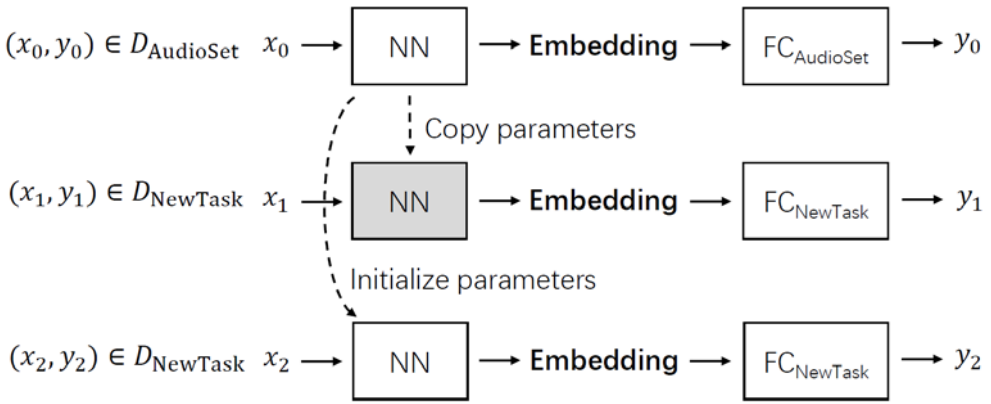
PANNs: Demo


Music: 0.661
Speech: 0.039
Singing: 0.036
Inside: 0.011
Jingle bell: 0.007

Impact of PANNs

Transfer to other datasets, e.g.:

- ESC-50 (sound events),
- GTZAN dataset (music)
- RAVDESS (human speech emotion)



Impact of PANNs

- Most cited article in IEEE/ACM Trans ASLP since 2020.
- IEEE SPS Young Author Best Paper Award (Kong, Iqbal) 2022.
- Baseline for e.g. HEAR challenge 2021, DCASE challenge 2021, 2022.
- Backbone of winner in DCASE 2021 & 2022 audio caption task.

Recent Developments: General Purpose Audio Representation Learning

Audio2Vec (Tagliasacchi et al, 2020)
CLAR (Al-Tahan and Mohsenzadeh, 2021)
COLA (Saeed et al, 2021)
BOYL-A (Niizumi et al, 2021)
AST (Gong et al, 2021)
ATST (Li and Li, 2022)
MAE-AST (Baade et al, 2022)
SSAST (Gong et al, 2022)
Audio-MAE (Huang et al, 2022)
BEATs (Chen et al, 2022)
ASiT (Atito et al, 2023)

mean Average Precision (mAP): **31.4** (2017) -> **50.6** (2022)

Audio to Text Generation

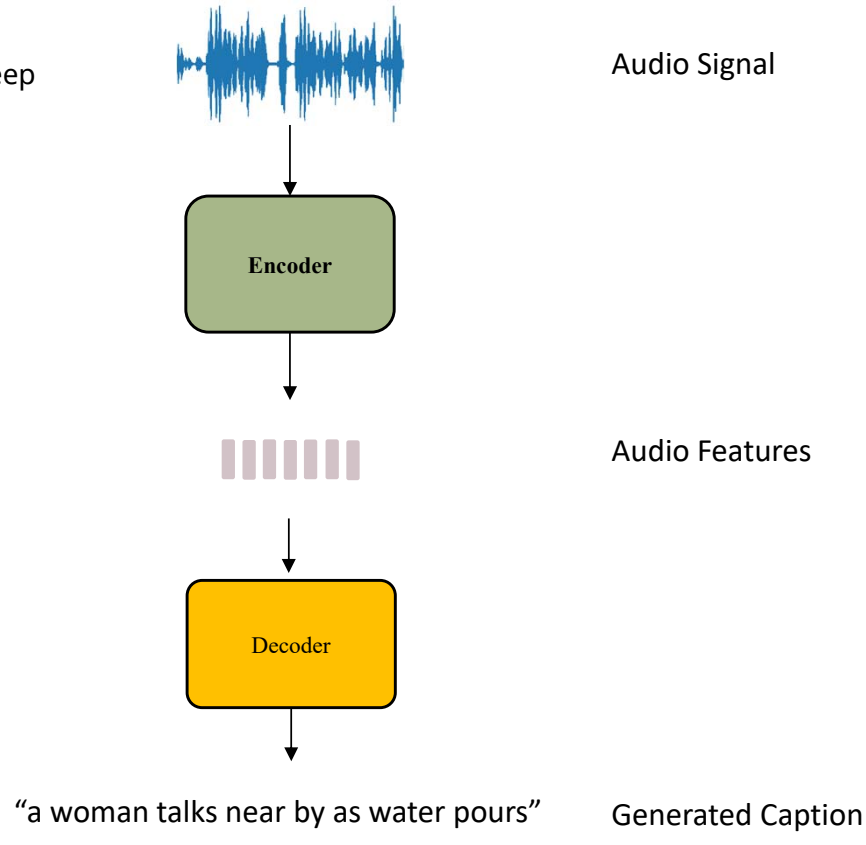
- **Automated audio captioning** (AAC) is a cross-modal translation task which aims at generating a natural language description given an audio clip.
- This task requires detecting the audio events and their spatial-temporal relationships and describing these information using natural language.
- Applications
  - Audio retrieval
  - Assist hearing-impared to understand environmental sounds
  - Subtitle for sounds in TV programs
- AAC started in 2017, and has received increasing attention in recent three years with freely available datasets released and being held as a task in DCASE Challenges 2020-2022.

AAC system

"a woman talks nearby as water pours"

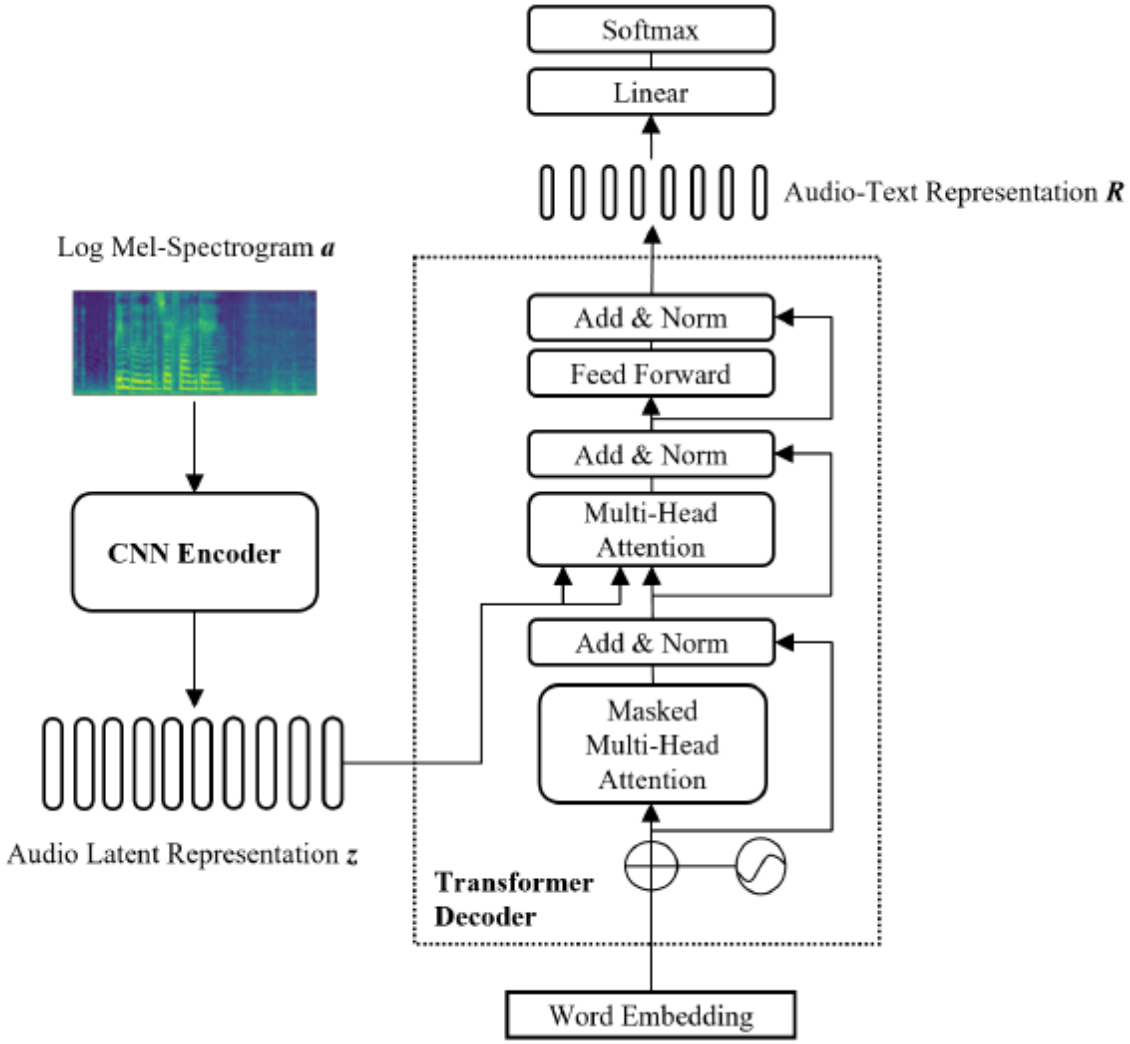
Existing Methods

- Model Architecture
  - All proposed methods are based on the encoder-decoder paradigm ultilizing deep learning techniques
    - RNN-RNN (2017)
    - CNN-RNN (2019)
    - CRNN-RNN(2020)
    - CNN-Transformer(2020)
    - Transformer-Transformer(2021)
    - Pre-trained models (2021)
    - Large language models (2022 & 2023)
    - Contrastive language-audio pretraining (2022 & 2023)
- Training
  - Cross-entropy training with maximum likelihood estimation
  - Reinforcement learning
- Auxiliary Information
  - Keywords or tag information
  - Sentence information
  - Pre-trained models
  - Data augmentations



X. Mei, X. Liu, M.D. Plumbley, and W. Wang, "Automated audio captioning: an overview of recent progress and new challenges," *EURASIP Journal on Audio Speech and Music Processing*, 2022.

An Example: CNN-Transformer Encoder-Decoder



Datasets

- Clotho: an audio captioning dataset whose audio clips are all collected from the Freesound archive, each audio clip has **5** reference captions and is between 15-30 seconds long. There are in total **3839**, **1045** and **1045** audio clips in training, validation and evaluation sets. Clotho is used as the official ranking dataset in DCASE challenge.

- AudioCaps: the largest audio captioning dataset which contains more than 50k 10-second audio clips sourced from AudioSet. Training set contains **49274** audio clips with **one** reference caption, validation and test sets contain **494** and **957** audio clips respectively with **5** reference captions.

- MACS: consists audio clips from the development set of TAU UrbanAcoustic Scenes 2019 dataset. The audio clips are all 10-second long recorded from three acoustic scenes (airport, public square and park) and are annotated by students. MACS contains **3930** audio clips without being split into subsets. The number of captions per audio clip **varies** in the dataset.

- AudioCaption: a domain-specific **Mandarin-annotated** audio captioning dataset. Two scene-specific sets has been published: one for hospital scene and another for in-car scene.

Performance Metrics

- Metrics
  - As a text generation problem, most metrics are directly borrowed from NLP tasks such as Machine Translation, Summarization. These metrics are all based on n-gram or sub-sequence matches.
    - BLEU: calculated as a weighted geometric mean of modified precision of n-grams.
    - METEOR: measures a harmonic mean of precision and recall based on word level matches between the candidate sentence and references
    - ROUGE: calculated as F-measures based on the longest common subsequence.
  - Metrics borrowed from Image Captioning
    - CIDEr: applies term frequency inverse document frequency (TF-IDF) weights to n-grams and calculates the cosine similarity between them.
    - SPICE: transforms captions into scene graphs and calculates F-score based on tuples in them.
    - SPIDEr: a linear combination of CIDEr and SPICE, where SPICE score ensures captions are semantically faithful to the audio clip, while CIDEr score ensures captions are syntactically fluent.
  - Other metrics
    - BERTScore, sentenceBERT, and FENSE, mBLEU, div-n, etc.

X. Mei, Q. Huang, X. Liu, G. Chen, J. Wu, Y. Wu, J. Zhao, S. Li, T. Ko, H.L. Tang, X. Shao, M.D. Plumbley, and W. Wang, "An encoder-decoder based audio captioning system with transfer and reinforcement learning", in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop* (DCASE 2021).
X. Mei, X. Liu, M.D. Plumbley, and W. Wang, “Automated audio captioning: an overview of recent progress and new challenges", *EURASIP Journal on Audio Speech and Music Processing*, 2022.

Several Open Problems

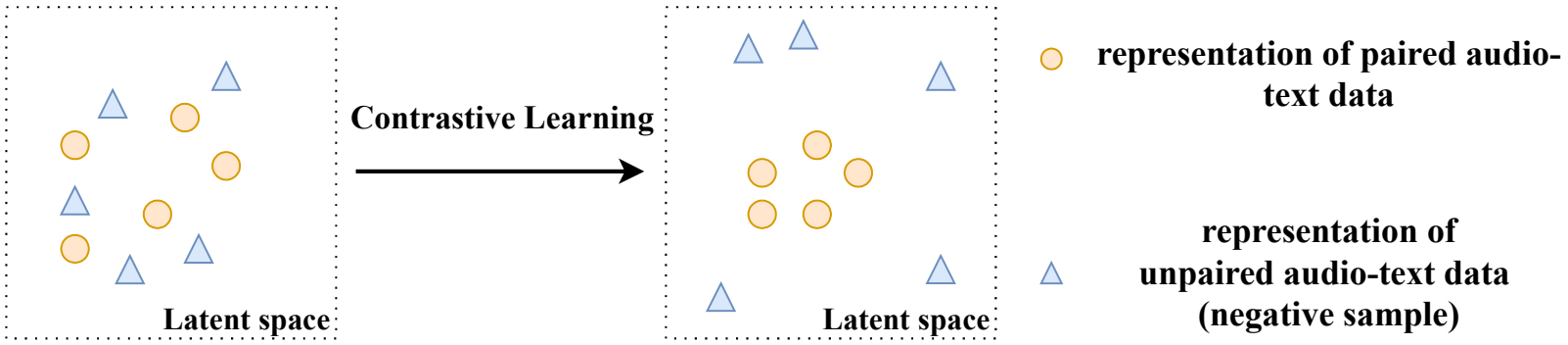
- Data scarcity
- Representations of audio, text and joint audio-text
- Diversity of captions
- Large language models + audio models
- Interactions with other modalities (e.g. vision)
- ....

CL4AC - Contrastive Learning for Audio Captioning

Example	paired caption C	unpaired caption $C_{negative}$
audio a	Something goes round that is playing its song	The Air is blowing some what fast outside
	At the fair, music is playing near a carousel through the speaker	A hand held sander was used as various speeds
	Chiming of bells, whistles and horns at a performance	A hard gravel ground is walked on by someone
	Fair kind music is being played at the circus grounds	A person using a hard object to tap and scrape glasses
	Polka or fair kind of music is being played	The wind is blowing and the waves are flowing

Table 1: Examples of paired audio-text training data $x = (a, C)$ and negative training sample $x_{negative} = (a, C_{negative})$. Examples are selected from the Clotho dataset, where each audio data has five corresponding captions.

- Contrastive Loss (CL) objective is designed to maximize the difference between the representation of the matched audio-caption pair and those from the negative pairs

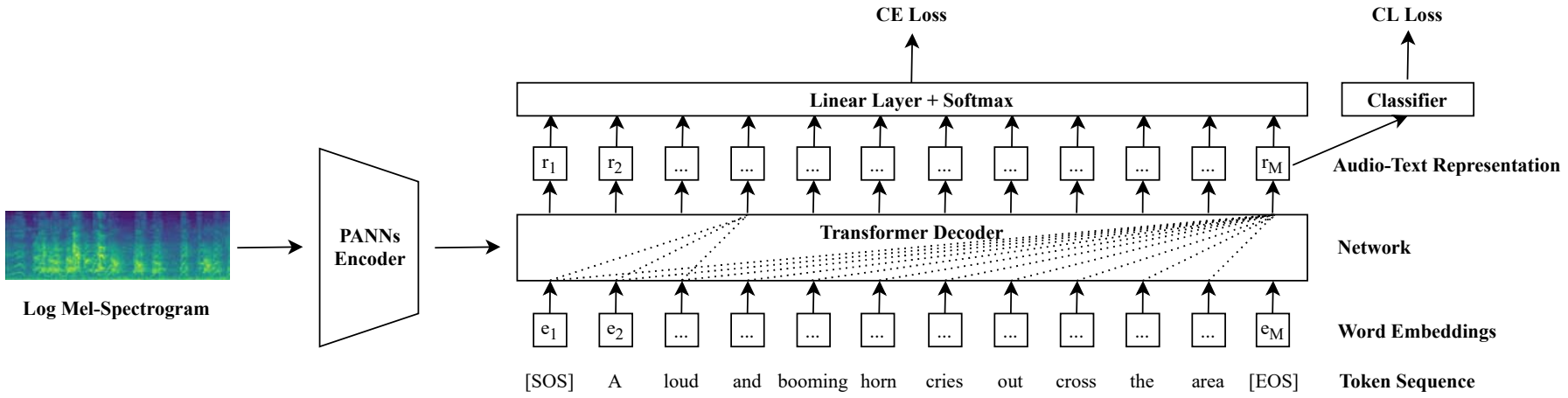


CL4AC: System Architecture

- Baseline system – CNN (PANNs) encoder and Transformer decoder
- Use the last vector in the audio-text representation to predict whether the input audio and caption are paired (y=0) or not (y=1)
- Training objective: $\text{Loss}_{\text{training}} = (1 - y)\,\text{Loss}_{\text{CE}} + \text{Loss}_{\text{CL}}$

$$\text{Loss}_{\text{CE}} = -\mathbb{E}_{(a,C)\sim D} \log p(w_m|z, w_1, ..., w_{m-1}).$$

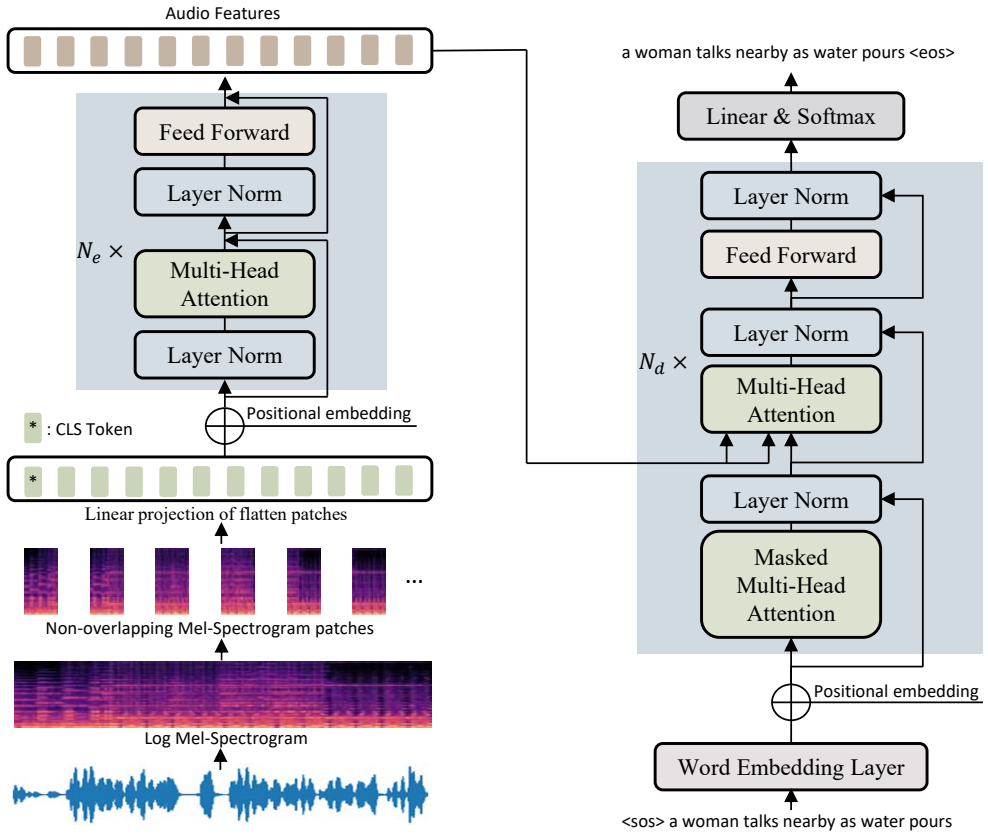
$$\text{Loss}_{\text{CL}} = -\mathbb{E}_{x'\sim D'} \log p(y|f(x')),$$



X. Liu, Q. Huang, X. Mei, T. Ko, H. Tang, M.D. Plumbley, and W. Wang, "CL4AC: A Contrastive Loss for Audio Captioning", in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop* (DCASE 2021).

Audio Captioning Transformer

- Motivation
  - CNN encoders can be limited at modeling temporal information for time series signals (audio)
  - RNN encoders can be limited at modeling long-range dependencies between time frames
  - Transformer shows powerful ability at modeling sequence data in NLP tasks
- We propose **Audio Captioning Transformer** (ACT), a full Transformer network to overcome the problems in existing methods.
- ACT encoder has a better ability to model the global information within an audio signal as well as capture temporal relationships between audio events.
- ACT encoder follows the settings in ViT containing 12 encoder blocks and 12 heads with an embedding dimension of 768.
- We perform experiments with 3 different decoder settings (# of layers, heads).



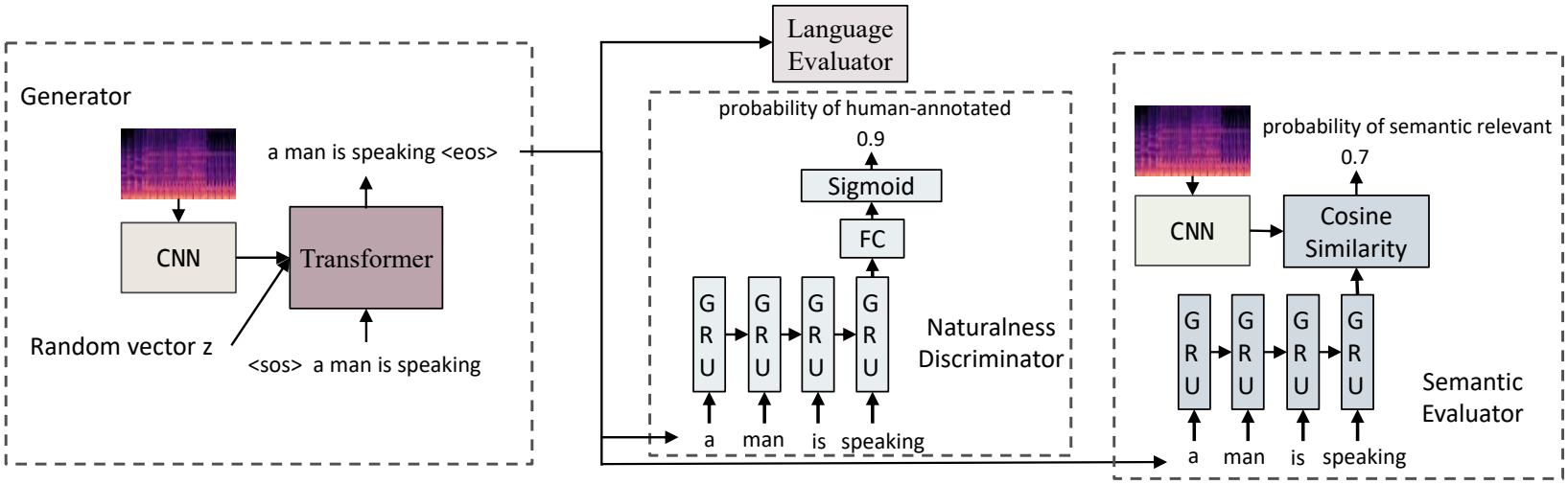
Diverse Audio Captioning

- Motivation
  - Existing models tend to generate deterministic (i.e. generating a fixed caption for a given audio clip), generic (i.e. generating same captions for similar audio clips), and simple (i.e. using simple and common words) captions.
  - Different people may describe an audio clip from different aspects using distinct words, phrases and grammars.
  - We argue that an audio captioning system should have the ability to generate diverse captions for a fixed audio clip and across similar audio clips.
- Each audio clip in Clotho dataset has five diverse reference human-annotated captions.
- Models trained on Clotho with MLE encourage the use of high frequent n-grams occurred in references, which leads to generic and simple captions.
- Examples

	Ground Truth	CNN-Transformer trained with MLE (Beam search with size 5)
Cap_1	A drill starts and stops four times with brief stops in between.	a jackhammer is being used to start a stop
Cap_2	A jackhammer digs into concrete, taking occasional short breaks.	a jackhammer is being used to get a stop
Cap_3	A jackhammer is being operated on a concrete surface.	a jackhammer is being used to start up and down
Cap_4	A small jackhammer rattles as it works at another hard object.	a jackhammer is being used to start up
Cap_5	Outdoors manually operated jack hammer or a tool	a jackhammer is being used to start a halt

Diverse Audio Captioning

- Methods
  - We propose an adversarial training framework for audio captioning based on a conditional generative adversarial network (C-GAN).
  - After training with MLE, the model is trained in an adversarial training manner to encourage diversity.
  - This framework is agnostic to the architecture of the caption generator.
  - We select CNN-Transformer as our generator.



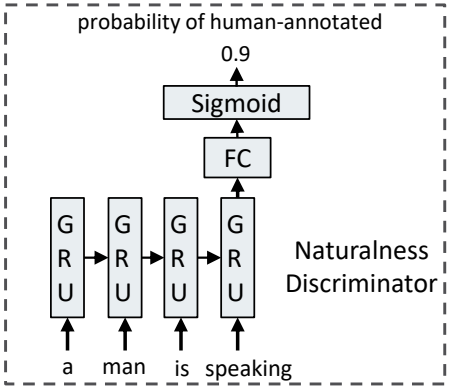
X. Mei, X. Liu, J. Sun, M. Plumbley, W. Wang, "Diverse audio captioning via adversarial training", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2022), Singapore, 22-27 May, 2022.

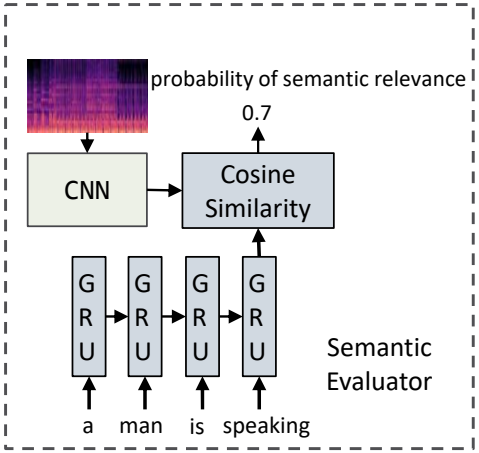
X. Mei, X. Liu, J. Sun, M. Plumbley, W. Wang, "Towards generating diverse audio captioning via adversarial training", arXiv:2212.02033, 2023.

Diverse Audio Captioning

- Naturalness discriminator
  - ND takes a caption as input and outputs a probability that indicates whether the given caption is human-annotated or machine-generated.
  - Generator and ND make up a conditional generative adversarial network (C-GAN) and are trained alternatively to compete with each other.
  - The aim of ND is to improve the naturalness of the generated captions.
- Semantic evaluator
  - SE is pre-trained using ground-truth captions and audio clips and frozen during the adversarial training stage.
  - Its aim is to ensure the semantic relevance of the generated caption with the given audio clip.
- Language evaluator
  - Evaluate CIDEr metrics of generated captions.
  - Its aim is to ensure the captions achieve high scores under the objective evaluation metrics.
- Final reward:
  - The final reward is a combination of these three scores

$$r(w) = \lambda\,(n + s) + (1 - \lambda) \cdot c$$





Diverse Audio Captioning

Ablation Study

	$BLEU_4$ (↑)	$CIDE_r$ (↑)	$SPIDE_r$ (↑)	vocab size (↑)	$mBLEU_4$ (↓)	div-1 (↑)	div-2 (↑)
C-GAN_ND	0.047	0.071	0.054	642	0.156	0.591	0.824
C-GAN_SE	0.000	0.012	0.025	224	0.715	0.096	0.269
C-GAN_LE	0.155	0.362	0.236	218	0.792	0.225	0.345

- When only using the naturalness discriminator, the model would generate random captions which are irrelevant to the given audio clip, thus these captions would be diverse but were not semantically faithful to the audio clip.
- When only using the semantic evaluator, the system cannot even generate reasonable sentences, but instead some disordered and repetitive words.
- When only optimizing CIDEr, the n-gram based evaluation metrics drops slightly, compared to the MLE model, as a random vector is appended to encourage the diversity of generated captions.

Example

	C-GAN	CNN-Transformer trained with MLE (Beam search with size 5)
Cap_1	a loud buzzing occurs and then starts again	a jackhammer is being used to start a stop
Cap_2	a jackhammer is being started and then slows down	a jackhammer is being used to get a stop
Cap_3	a loud drilling occurs and then slows down	a jackhammer is being used to start up and down
Cap_4	an old engine is running and then loud squealing	a jackhammer is being used to start up
Cap_5	a loud drilling occurs and then starts again	a jackhammer is being used to start a halt

Audio Captioning Demos

Text to Audio Generation

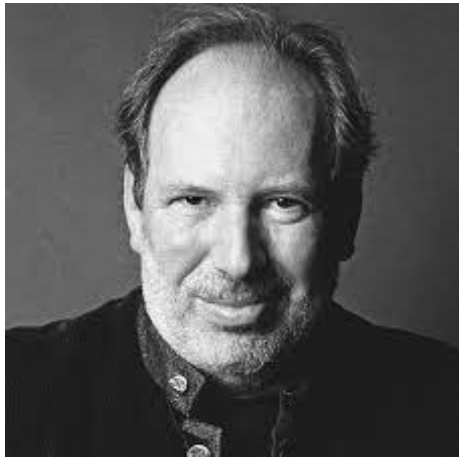
Computational “foley artist”: (e.g., https://www.thefoleybarn.com)

- *Game developer: e.g., A ghost is haunting a house.*
- *Audio producer: e.g., high heels hitting metal ground.*
- *Movie producer: e.g., the laser sound from a laser gun.*
- …

Automatic content creation (> 60 startups[1])

- Endless music
- Audiobook with ambient noises
- White noise for meditation
- …

Data Augmentations



Sound is often the unsung hero of the movie world
- Hans Zimmer

[1]https://github.com/csteinmetz1/ai-audio-startups

Related Works

Label-to-Audio Generation

- Acoustic Scene (Kong et al., 2019), Sound event (Liu et al., 2019), FootStep (Comunit et al. 2019), …

Text-to-Audio Generation

- DiffSound (Yang et al., 2022), AudioGen (Kreuk et al., 2022), Make-an-Audio (Huang et al., 2023)

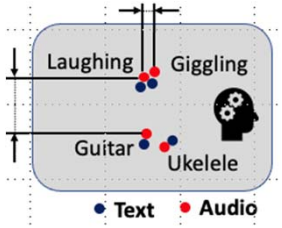
Text-to-Music Generation

- MusicLM (Andrea et al., 2023)
- Moûsai (Flavio et al., 2023)
- Noise2Music (Huang et al., 2023)

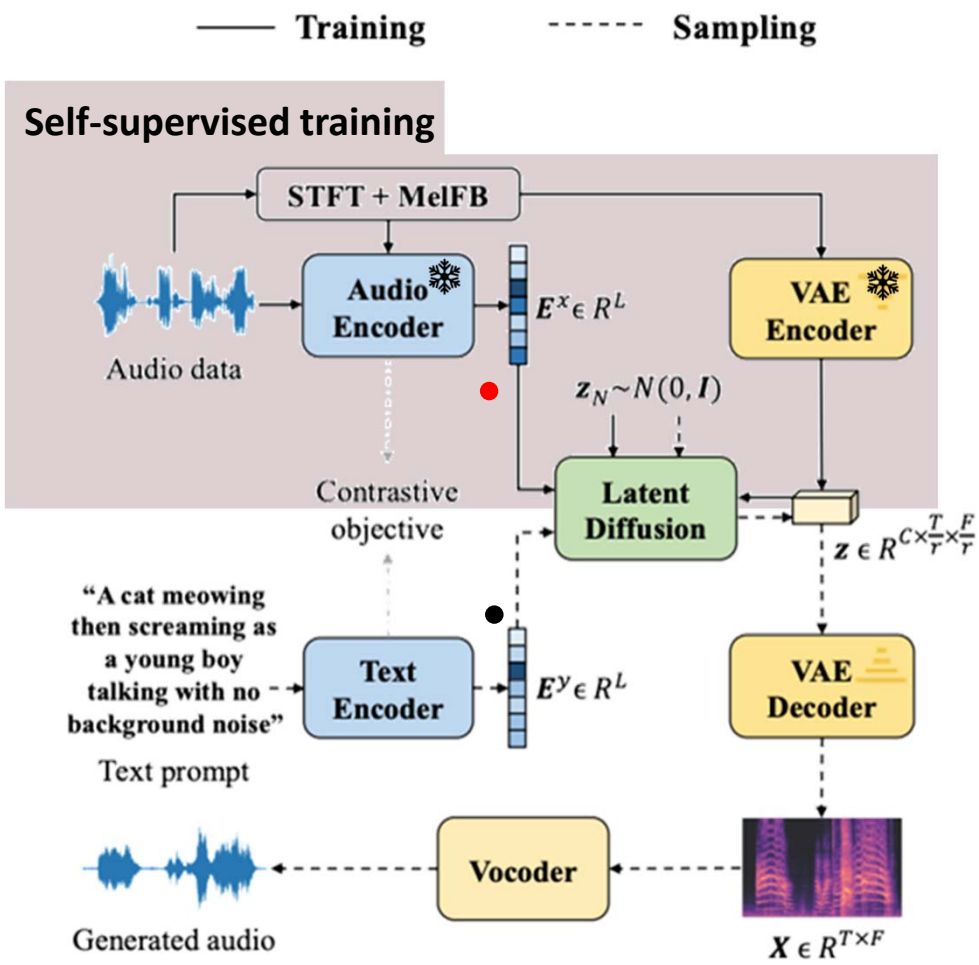
Others

- JukeBox (Dhariwal et al., 2020), AudioLM (Borsos et al., 2022), SingSong (Donahue et al., 2023),…

AudioLDM



1. **Contrastive Language-Audio Learning (CLAP) Encoders**
   - Align audio and text in one space.
2. **Latent Diffusion Models**
   - Learn to generate VAE latent conditioned on CLAP embedding
3. **Mel-spectrogram Autoencoder**
   - Learn latent representations.
4. **Mel-to-Waveform Vocoder**
   - Reverse Mel back to waveform



H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, M. D. Plumbley, "AudioLDM: text-to-audio generation with latent diffusion models," in *Proc. IEEE International Conference on Machine Learning* (ICML 2023), Hawaii, USA, 23-29 July, 2023.

Overall Advantages

Less computation cost

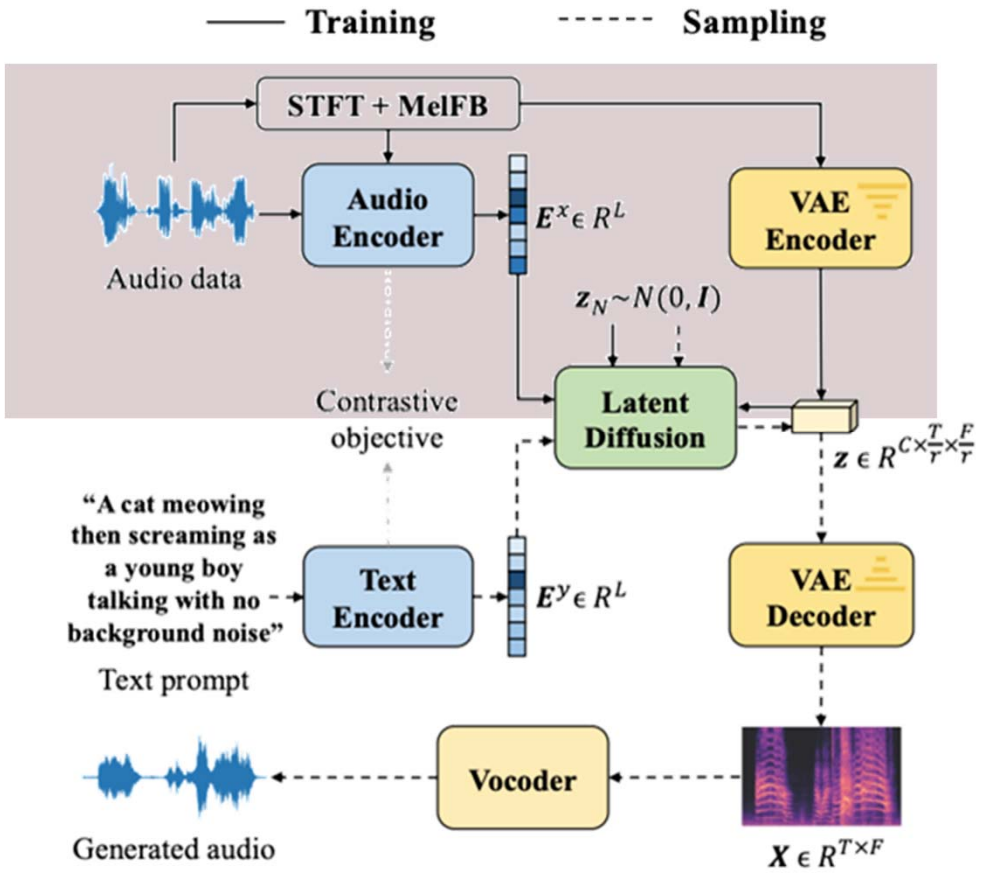
- Latent Diffusion Models

Less dependency on audio-text pairs

- Train LDMs by self supervision

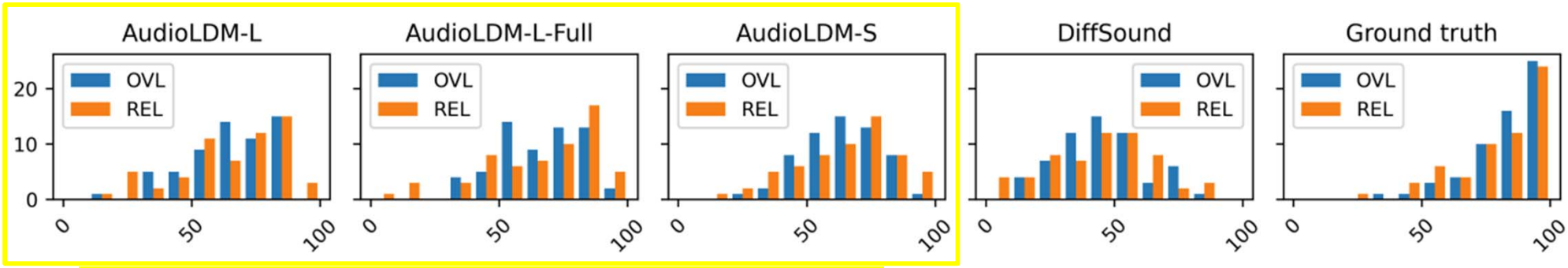
Continueous latent space

- Zero-shot audio style transfer
- Zero-shot audio super-resolution
- Zero-shot audio inpainting
- ...



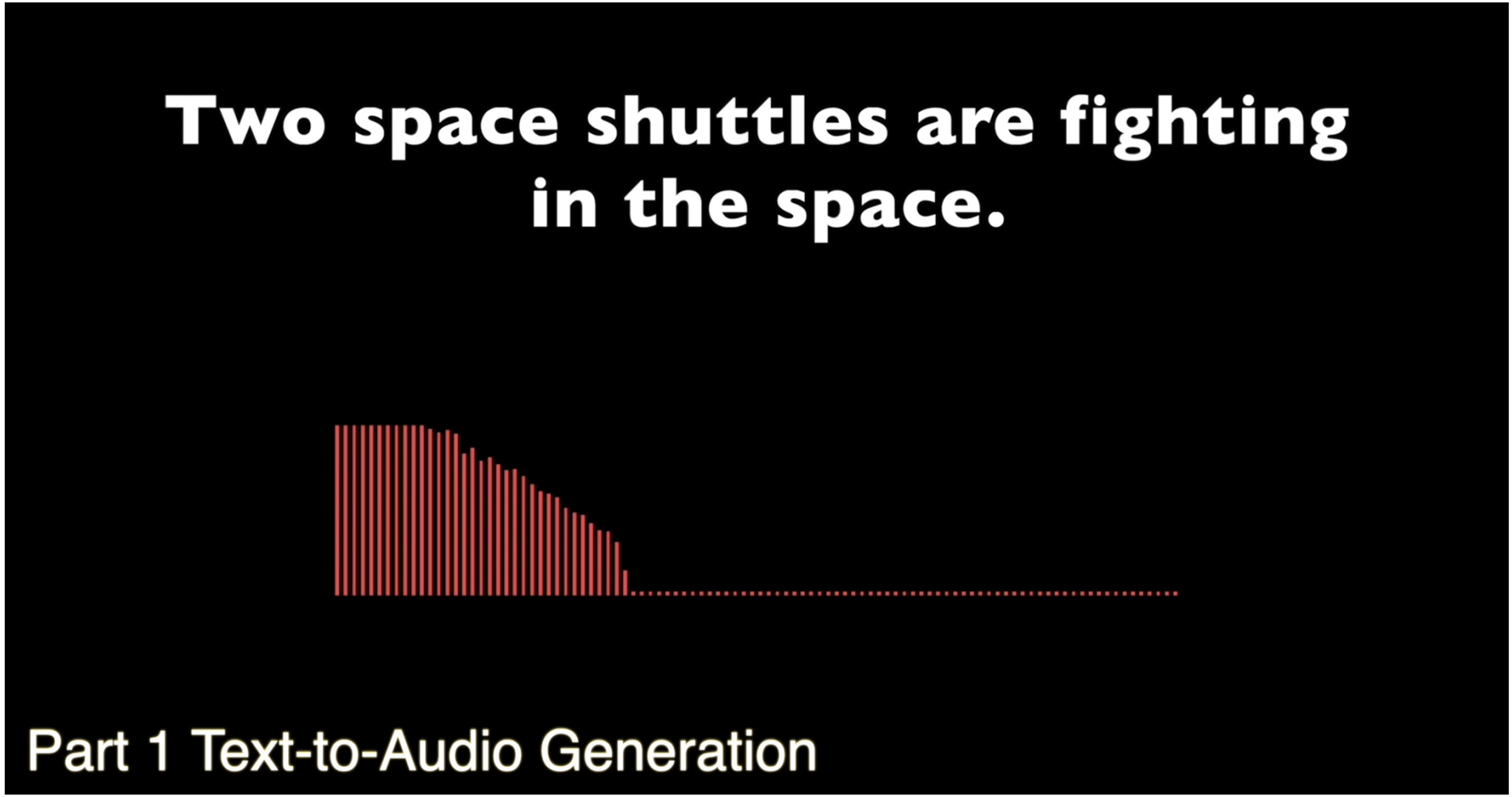
Result – SOTA Comparison

Model	Datasets	Text	Params	FD ↓	IS ↑	KL ↓	FAD ↓	OVL ↑	REL ↑
Ground truth	-	-	-	-	-	-	-	83.61	80.11
DiffSound† (Yang et al., 2022)	AS+AC	✓	400M	47.68	4.01	2.52	7.75	45.00	43.83
AudioGen† (Kreuk et al., 2022)	AS+AC+8 others	✓	285M	-	-	2.09	3.13	-	-
AudioLDM-S	AC	✗	181M	29.48	6.90	1.97	2.43	63.41	64.83
AudioLDM-L	AC	✗	739M	27.12	7.51	1.86	2.08	64.30	64.72
AudioLDM-L-Full	AS+AC+2 others	✗	739M	**23.31**	**8.13**	**1.59**	**1.96**	**65.91**	**65.97**



Trained on a single 3090 or A100 GPU!

AudioLDM: Demos



Impact of AudioLDM

- Accepted by the top machine learning conference ICML 2023
- One of the **25 most-liked machine learning apps** on Hugging Face Spaces among 25000+ apps
- Received more than **1400 stars** on Haohe's Github page.
- **Widely used** by people for creating albums, games, 3D animations
- **Widely reported and shared** on social media & official news channels (Youtube, Twitter, LinkedIn, Note, MarkTechPost, MachineHeart, Reddit, Zenodo, Replicate, University Press release, and many more)
- Search "AudioLDM" will pop out more than 10 pages of entries

Perspectives on Audio-Text Learning

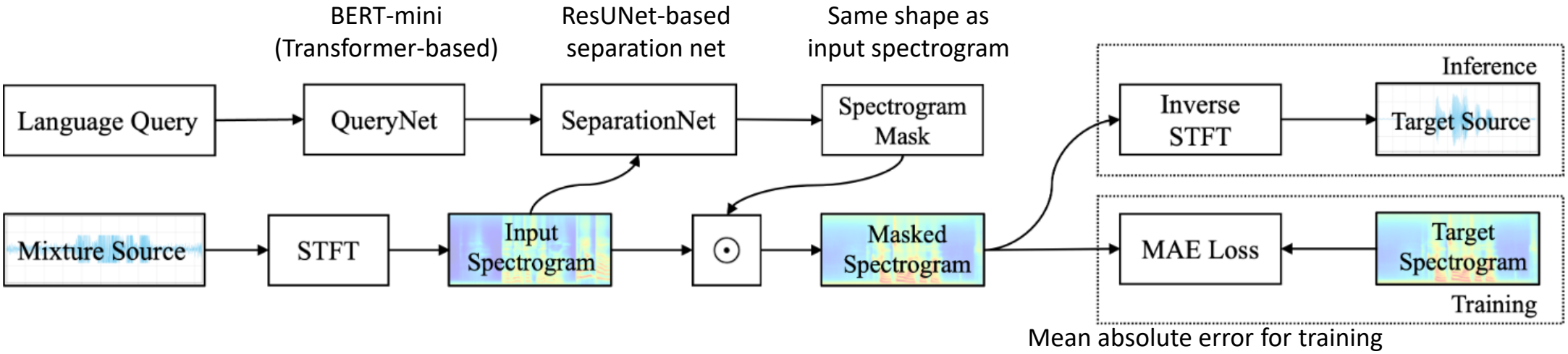
- Audio models (perception/classification) + language models (reasoning)
  - LTU (Gong et al., 2023): AST+LLaMA
- Unified models for open-ended and closed ended tasks
  - Pingi (Deshmukh et al, 2023): a unified model for audio classification + captioning+ question answering
- General representation models towards unlimited modalities
  - ONE-PEACE (Wang et al, 2023): vision, audio and language modalities
- Other applications: e.g. language-queried audio source separation, retrieval, etc.
  - Separate What You Describe (Liu et al, 2022 & 2023), CAPTURE (Okamoto, et al, 2023)

Perspectives on Audio-Text Learning (cont.)

- Interactions with large language models (LLMs)
  - Interacting with LLMs, e.g. GPT-3, ChatGPT, GPT-4, LaMDA, LLaMa, PaLM, Galactica, using techniques, e.g. LoRA (Hu et al 2021), Frozen (Tsimpoukelli et al, 2021)
  - Instruction tuning, e.g. Vicuna (Chiang et al, 2023)
- Creating new datasets for audio-language multimodal learning:
  - WavCaps (Mei et al, 2023), OpenAQA-5M (Gong et al, 2023)
  - Repurposing existing datasets, e.g. AudioSet, Freesound, ESC-50, FSD50K, VGGSound, LAION-Audio-630K, AudioCaps, Clotho, BBC Sound Effects, Sound Bible, ClothoAQA, WavText5K, SoundDescs, FindSound, MACS, CochlScene MOSEI, MELD, Nsynth, FMA, etc.
- Other audio/speech/music, & image/video related work:
  - CLAP (Elizalde et al, 2022); SpeechLM (Zhang et al, 2022); AudioGPT (Huang et al, 2023), SpeechGPT (Zhang et al, 2023); PandaGPT (Su et al, 2023); Audioclip (Guzhov et al, 2022)
  - ImageBind (Girdhar et al, 2023): one embedding space to bind images, text, audio, depth, thermal, and IMU data
  - Vision-language models: VisualBERT (Li et al, 2019); SimVLM (Wang et al, 2021); Flamingo (Alayrac et al, 2022)

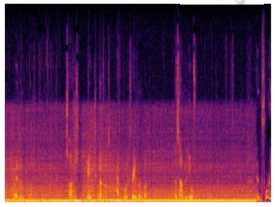
An Example: Language-Queried Audio Source Separation

Use language query to extract target source

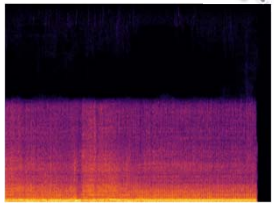


Human query: “The engine sound of a vehicle”

Mix

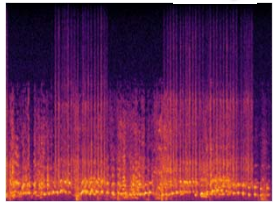


Separated

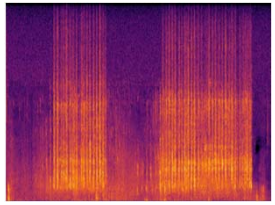


Human query: “The sound of hitting the keyboard”

Mix



Separated



X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M.D. Plumbley, and W. Wang,"Separate What You Describe: Language-Queried Audio Source Separation," in *Proc. 23rd Interspeech Conference* (INTERSPEECH 2022), 18-22 September, 2022, Incheon, Korea.

WavCaps

UNIVERSITY OF SURREY

WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong,
Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, Wenwu Wang

***Abstract*—The advancement of audio-language (AL) multimodal learning tasks has been significant in recent years. However, researchers face challenges due to the costly and time-consuming collection process of existing audio-language datasets, which are limited in size. To address this data scarcity issue, we introduce *WavCaps*, the first large-scale weakly-labelled audio captioning dataset, comprising approximately 400k audio clips with paired captions. We sourced audio clips and their raw descriptions from web sources and a sound event detection dataset.**

Research on AL takes a stride in the direction of empowering machines to comprehend audio signals at a human-like level.

While AL research is relatively young, vision-language (VL) multimodal learning [20], the counterpart of AL multimodal learning, has been studied for decades and has contributed to the success of many vision-language applications, such as cross-modal search [21], [22], image and video captioning [23], [24], text-to-image generation [25], [26], and

Paper: https://arxiv.org/abs/2303.17395

Code: https://github.com/xinhaomei/wavcaps

WavCaps

UNIVERSITY OF SURREY

TABLE I
EXAMPLE PROMPTS TO CHATGPT FOR FREESOUND AND AUDIOSET STRONGLY-LABELLED SUBSET. TRANSFORMATION EXAMPLES ARE ARE IGNORED. 'SL' REFERS TO 'STRONGLY-LABELLED'.

Data sources	Prompts
FreeSound BBC Sound Effects SoundBible	I will give you a list of descriptions of sounds. Process each individually. Extract the type of the sound and generate an audio caption describing the sound events. The audio caption should be less than 20 words. Delete the author of the sound. Delete locations, city names, country names. Delete the time. Delete device names. Delete the proper noun modifiers, number modifiers, and unit modifiers. Summarize each output into one sentence. Replace all named entities with their hypernyms. Replace people names with "someone". Do not write introductions or explanations. Only describe the sound events and do not use "heard", "recorded". Start each output sentence with its index. Make sure you are using grammatical subject-verb-object sentences. Output "Failure." if the description is not related to sound.
AudioSet SL	I will give you a number of lists containing sound events occurred sequentially in time. Process each individually. Write an one-sentence audio caption to describe these sounds. Make sure you are using grammatical subject-verb-object sentences. Directly describe the sounds and avoid using the word "heard". The caption should be less than 20 words.

TABLE II
EXAMPLES OF RAW DESCRIPTIONS AND CHATGPT'S OUTPUTS.

Data sources	raw descriptions	ChatGPT's output
FreeSound	This sound is of a book falling down the staircase in the Stanford University library west stacks. After 2h and 15min. Excerpt.	A book is falling down a staircase. Failure.
BBC Sound Effects	Timber & Wood - Rip saw, carpenters' workshop. Motor Car: Ford Ecsort 1300 (Automatic) - Ford Escort 1300, exterior, passing with horn. (1300cc engine, automatic transmission.)	Someone is using a rip saw in a carpenter's workshop. A car is passing with its horn.
SoundBible	Tasmanian Devil growling screaming hissing. Warning sounds from a Tasmanian Devil in Zoo. Large Tibetan Bells ringing in a temple. Could also use for Monastery or Monks.	An animal is growling, screaming, and hissing. Bells are ringing.
AudioSet SL	['Accelerating, revving, vroom', 'Race car, auto racing'] ['Female speech, woman speaking', 'Whoosh, swoosh, swish']	A race car is accelerating and revving. A woman is speaking while something whooshes.

TABLE IV
COMPARATIVE OVERVIEW OF MAIN AUDIO-LANGUAGE DATASETS BETWEEN OUR PROPOSED WAVCAPS DATASET.

Dataset	Num. audios	Duration (h)	Text source
AudioCaps [38]	52904	144.94	Human
Clotho [43]	5929	37.00	Human
MACS [44]	3537	9.83	Human
WavText5K [50]	4072	23.20	Online raw-data
SoundDescs [8]	32979	1060.4	Online raw-data
LAION-Audio-630K [51]	633526	4325.39	Online raw-data
WavCaps	403050	7567.92	ChatGPT

Conclusion & Future Works

- Summary
  - We have discussed some recent works on audio to text (e.g. diverse audio captioning) and text to audio generation (e.g. AudioLDM)
  - Using the proposed adversarial training methods, we can improve the diversity while maintaining the quality of the generated captions
  - AudioLDM offers state of the art performance in text-to-audio generation
  - We also discussed several perspectives about audio-language learning and some trending works
- Future Works
  - Further improving captioning performance, e.g. using pre-trained language models
  - Improving contrastive language-audio pre-training models
  - Improving the diversity and quality of the generated captions
  - Link to other applications such as audio retrieval and source separation with language query
  - Leveraging concept learning/reasoning with large language models & recognition/perception ability of audio models
  - Towards controllable & fine-grained sound generation
  - Physics based model + data driven model
  - Unified models for audio to text & audio to text generation

Paper, Codes, Demos, and More, …

Code avaliable at:
https://github.com/XinhaoMei/DCASE2021_task6_v2
https://github.com/XinhaoMei/ACT
https://github.com/liuxubo717/cl4ac
https://haoheliu.github.io/

Paper (https://arxiv.org/abs/2301.12503):
- AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

Hugging Face Space:
- https://huggingface.co/spaces/haoheliu/audioldm-text-to-audio-generation

Project Page: https://audioldm.github.io/

Github:
- Pretrained model: https://github.com/haoheliu/AudioLDM
- Evaluation tools: https://github.com/haoheliu/audioldm_eval

YouTube: https://www.youtube.com/watch?v=_0VTltNYhao

WavCaps: https://arxiv.org/abs/2303.17395

Code: https://github.com/xinhaomei/wavcaps

Thank you for listening!

This work was sponsored by a Newton Institutional Links Award from the British Council, titled “Automated Captioning of Image and Audio for Visually and Hearing Impaired” & EPSRC projects titled “Making Sense of Sounds” and “AI for Sounds”.