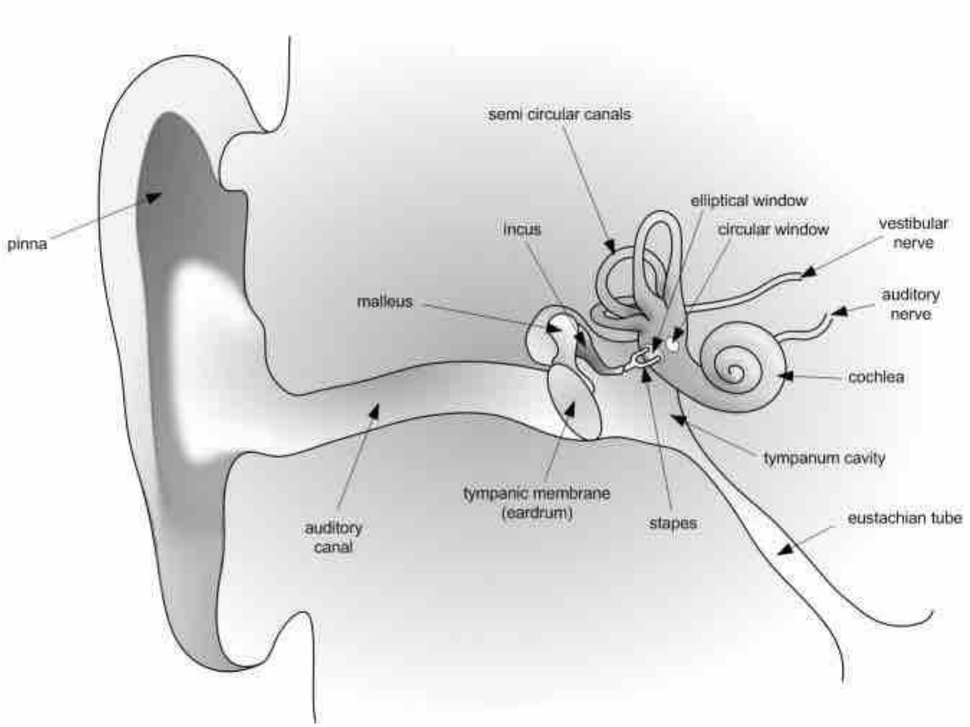
Scalable methods for general audio understanding

丁翰林

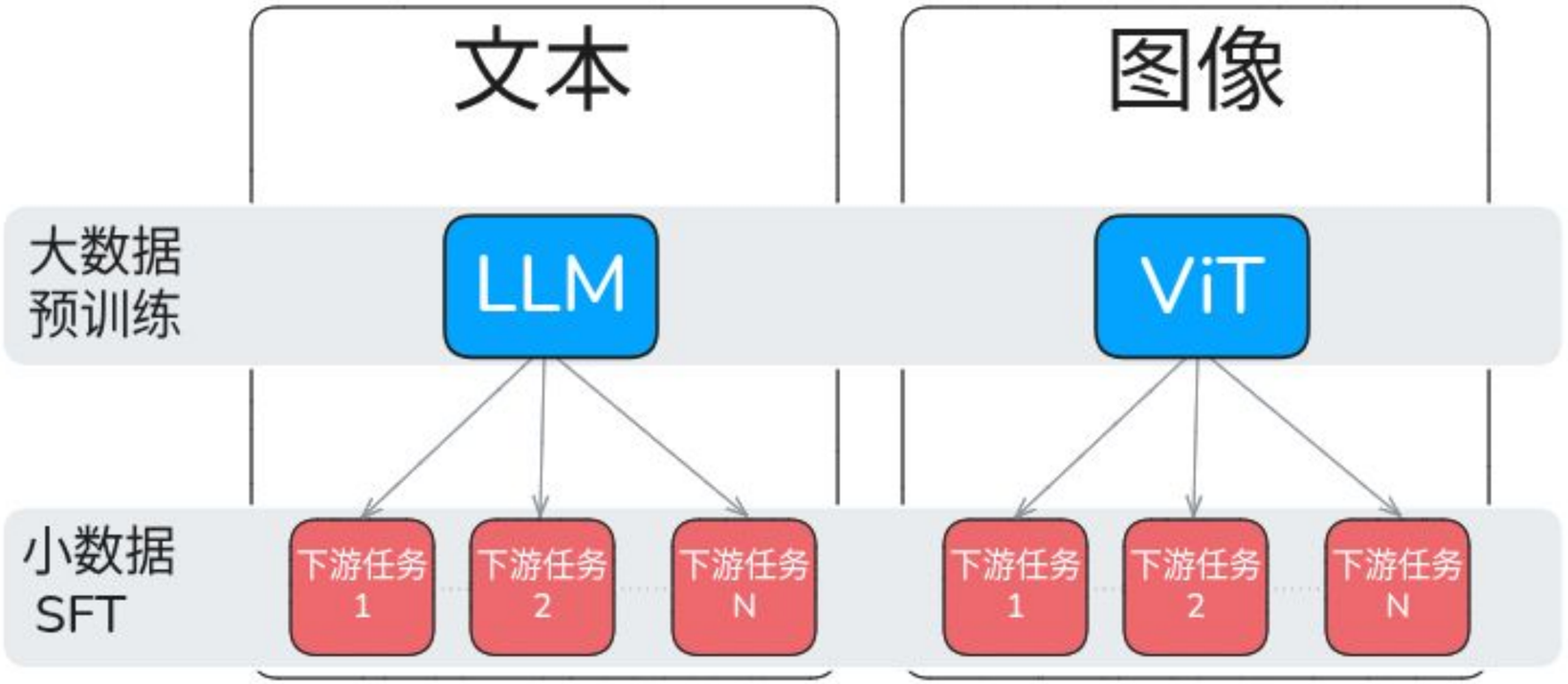
小米集团技术委员会
Xiaomi Technical Committee

Outline

- Motivation for General Audio Understanding
- Audio Encoder design and Challanges
- MiDashengLM - Efficient Audio Understanding via general audio captions



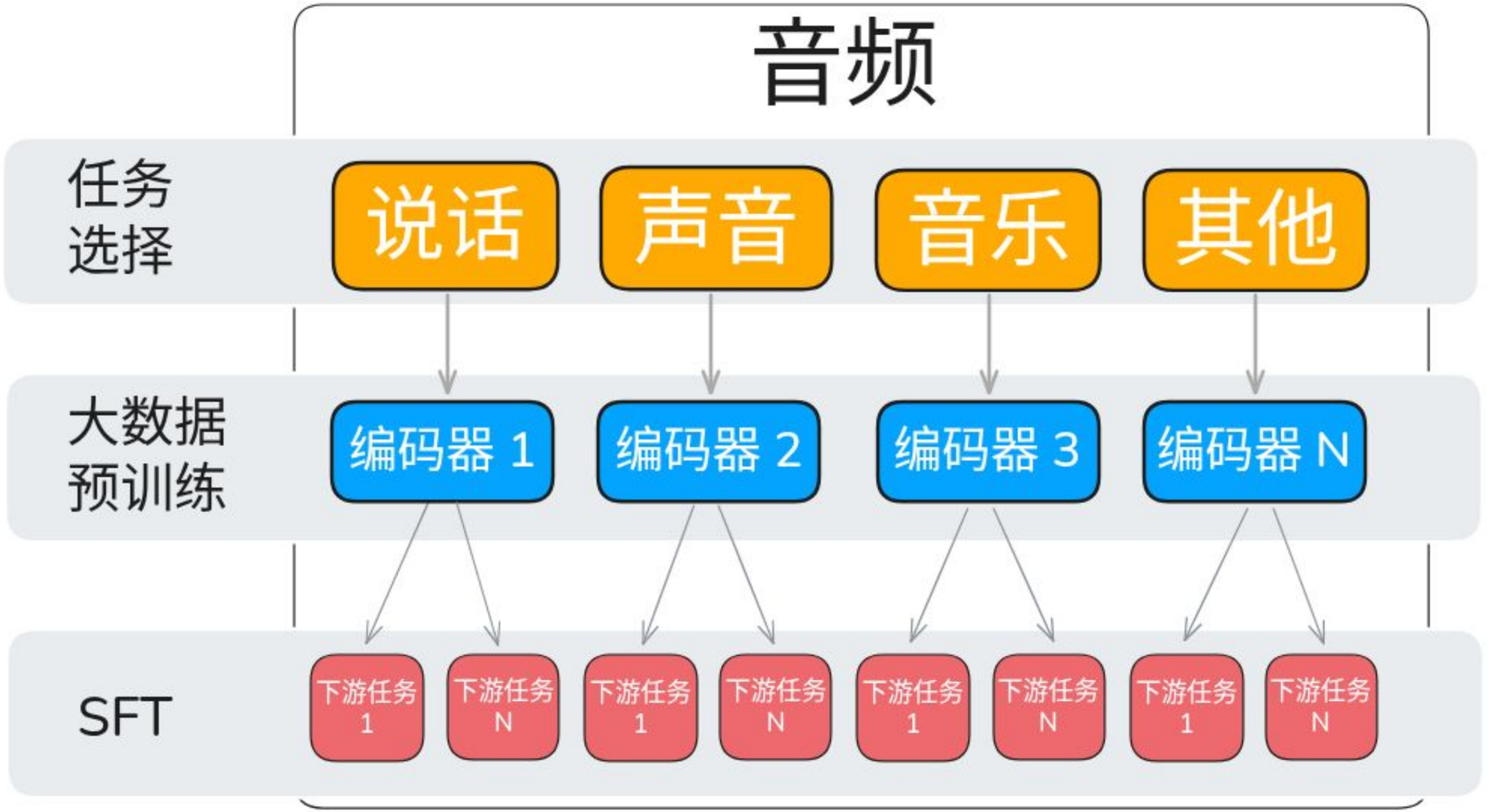
Motivation - Generalizable Features (1)



Motivation - Generalizable Features (2)

- Pretrain + SFT = Simple + Easy
- Works for most tasks
- Simple methods = Scalable
- Big Data = Scalable
- Big Data + Simple = Self-supervised (SSL)
- (Vision) Audio + Text ~= SSL

Audio


音频
任务选择
说话
声音
音乐
其他
大数据预训练
编码器 1
编码器 2
编码器 3
编码器 N
SFT
下游任务1
下游任务N
下游任务1
下游任务N
下游任务1
下游任务N
下游任务1
下游任务N

Audio

- Task specific
- Audio = Music + Speech + Sounds
- Lag behind NLP + Vision
- Very few works focus on general features (- 2025)
- Shinji did a talk in Interspeech2024 - So far little impact

OpenBEATs: A Fully Open-Source General-Purpose Audio Encoder

Shikhar Bharadwaj[1], *Samuele Cornell*[1], *Kwanghee Choi*[1], *Satoru Fukayama*[2],
Hye-jin Shim[1], *Soham Deshmukh*[1], *Shinji Watanabe*[1]

USAD: Universal Speech and Audio Representation via Distillation

Heng-Jui Chang, Saurabhchand Bhati, James Glass, Alexander H. Liu
MIT CSAIL
Cambridge, MA, USA
hengjui@mit.edu

Self-supervised Audio Teacher-Student Transformer for Both Clip-level and Frame-level Tasks

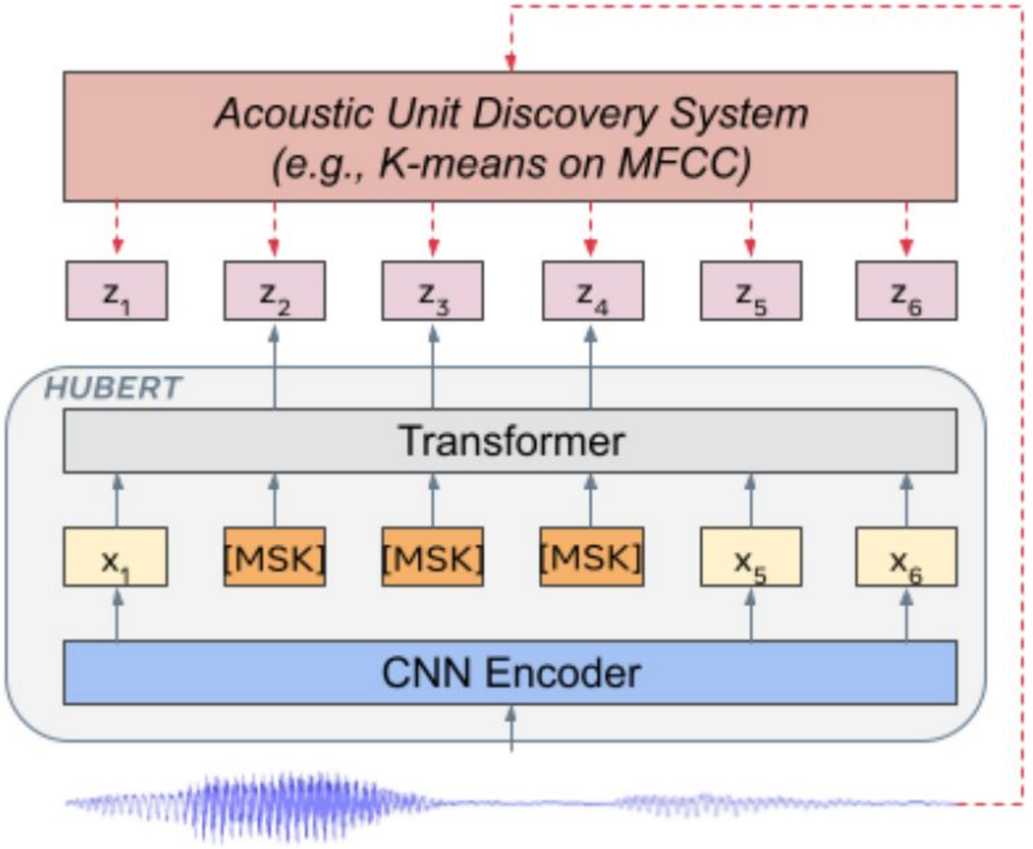
Xian Li, Nian Shao, and Xiaofei Li*

General Audio Encoders

- Wav2Vec2
- HuBERT
- WavLM (HuBERT + Denoise)
- AudioMAE
- Data2Vec2
- AST
- EAT 1/2
- Best-RQ
- BEATs
- Whisper (Supervised)

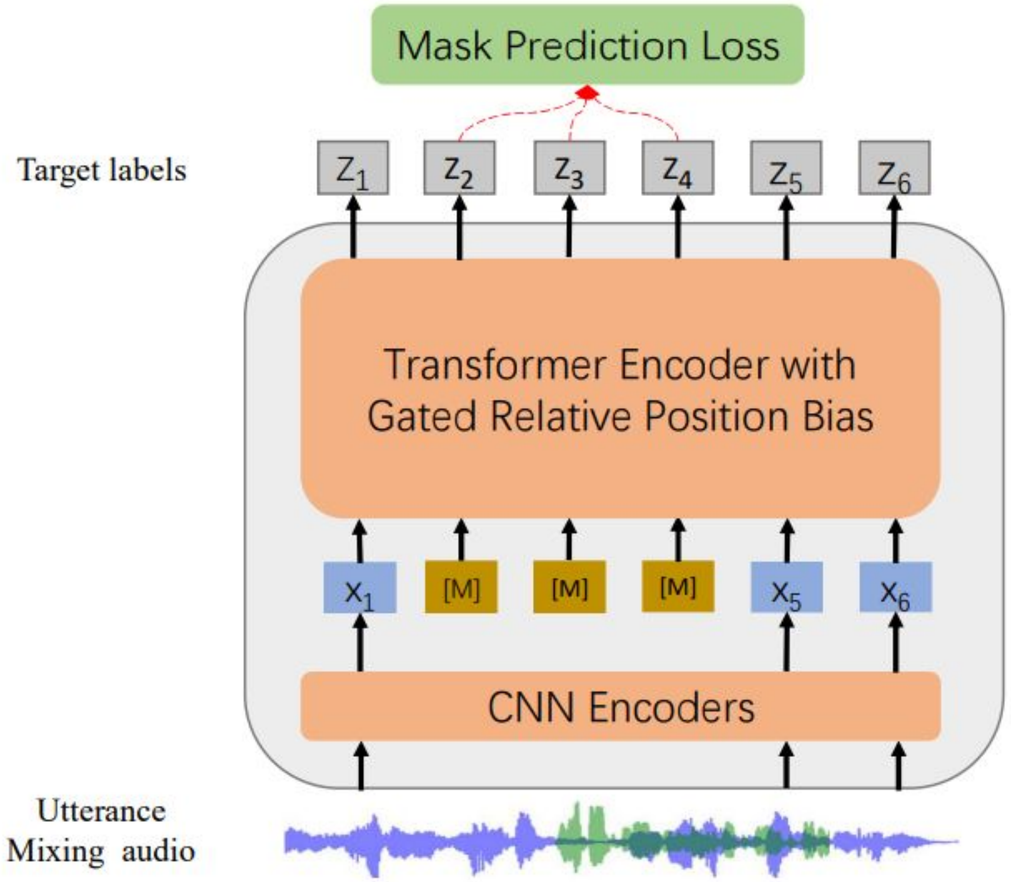
SSL - HuBERT

- Quantized Targets from Features (MFCC, Pretrained)
- Raw wave input
- Cross Entropy
- Slow, masked units need to be forwarded



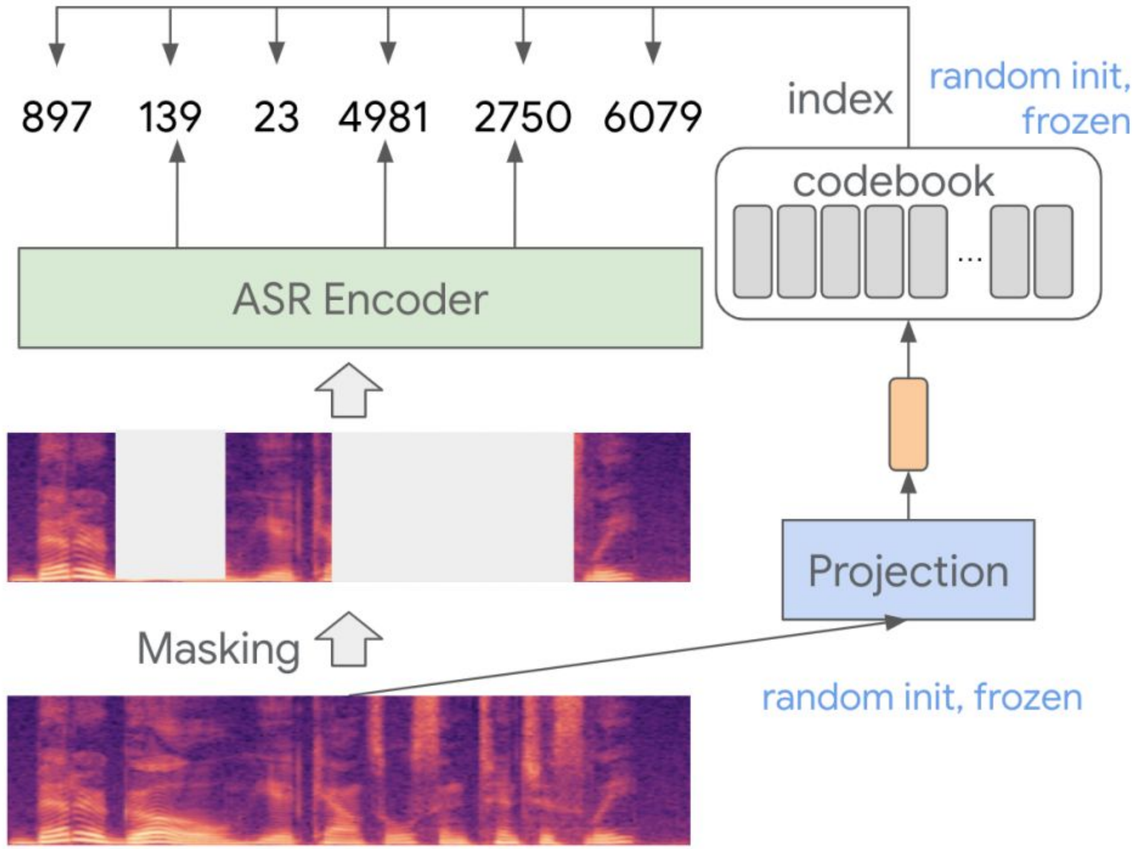
SSL - WavLM

- Quantized Targets from Features (MFCC, Pretrained), but speaker dependent
- Raw wave input
- Cross Entropy



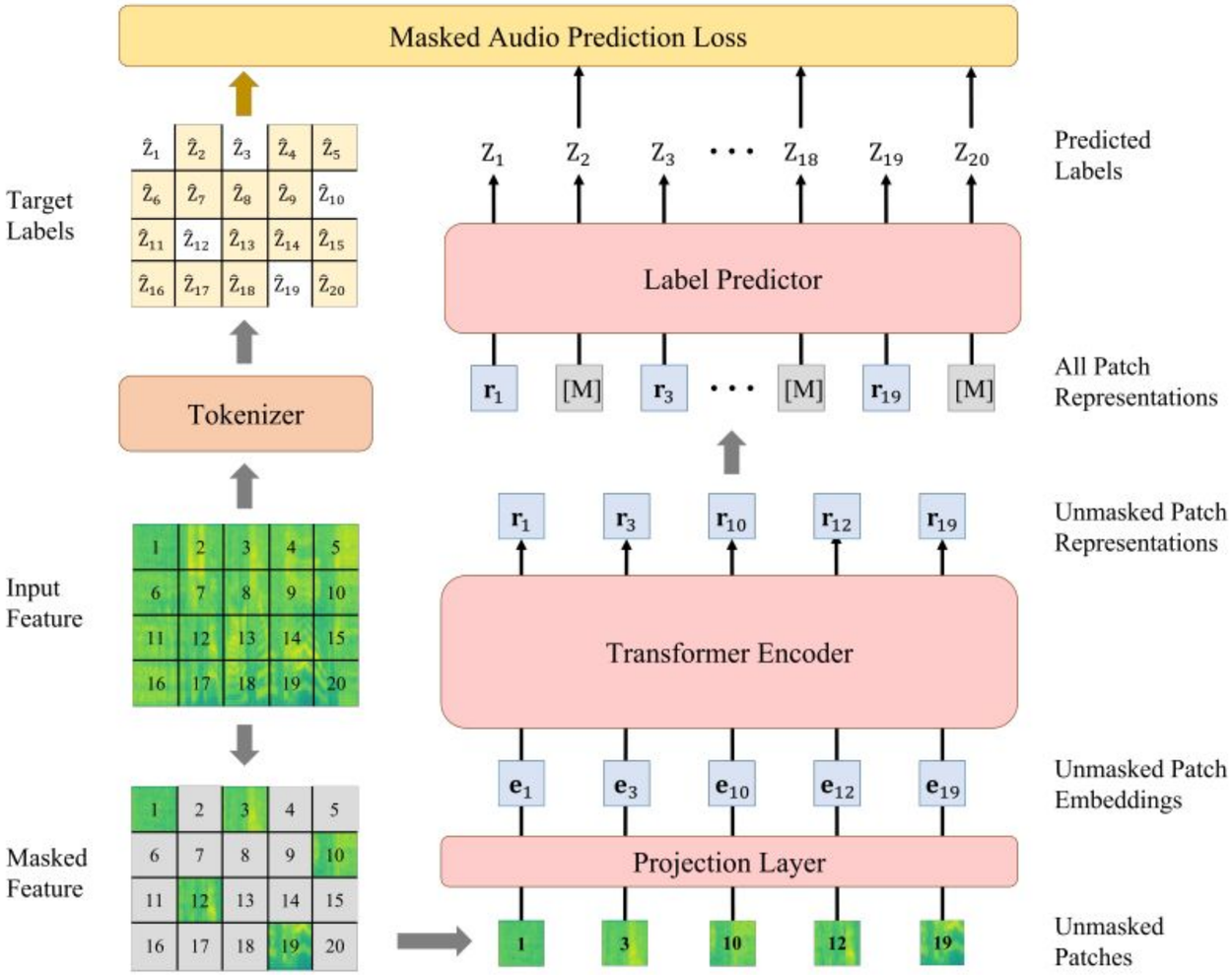
SSL - Best-RQ

- Quantized Targets from Features (MFCC, Pretrained), but from random projection
- Mel Input (Faster, better)
- Cross Entropy



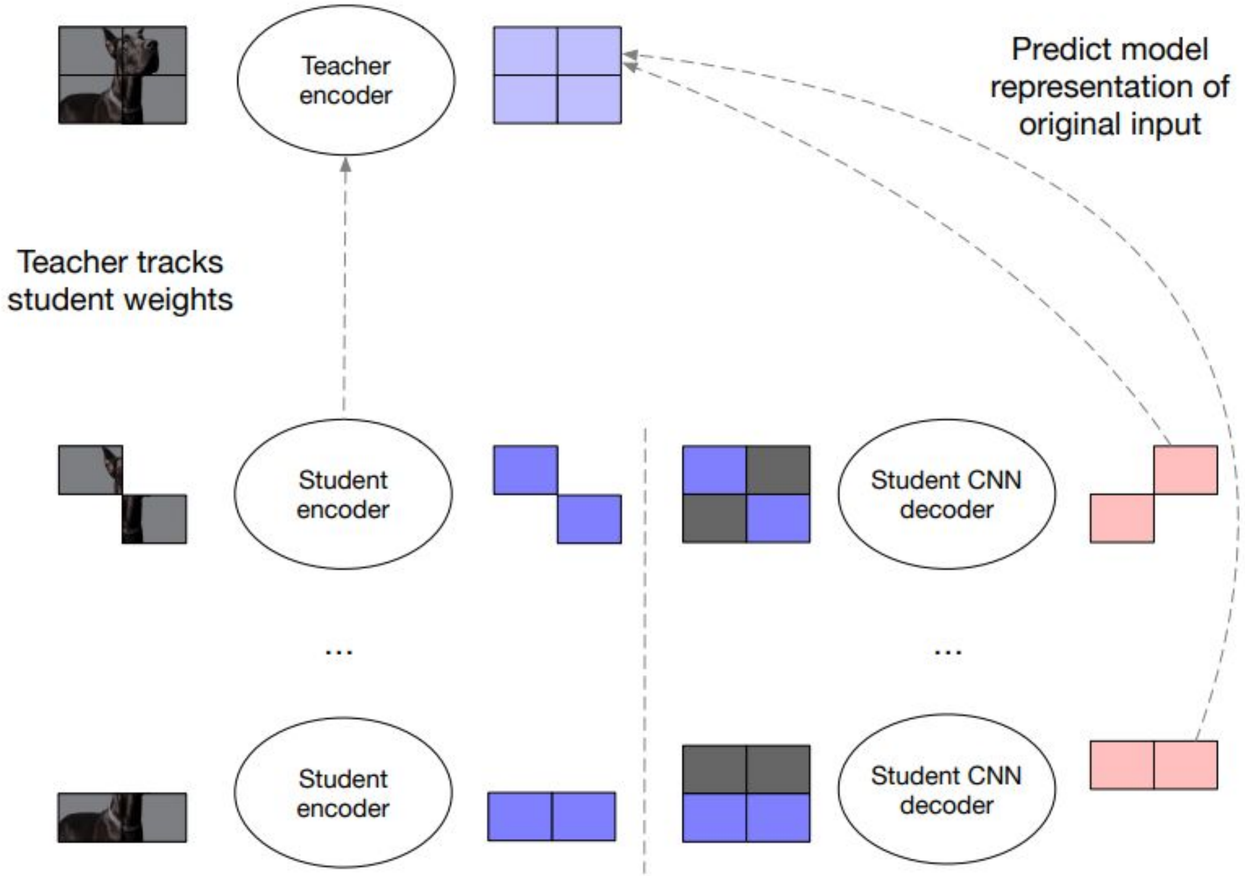
SSL - BEATs

- Same as Best-RQ but random projection on Mel-patches not frames
- Mel Input (Faster, better)
- Cross Entropy



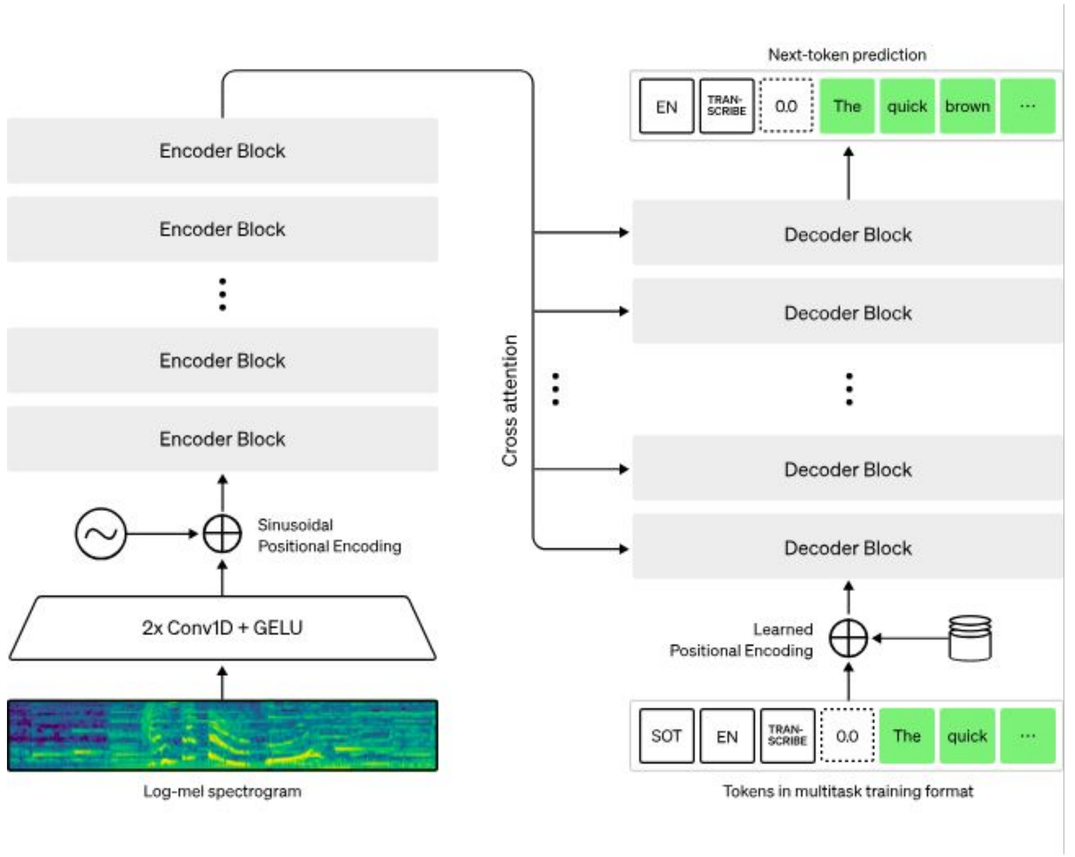
SSL - Data2Vec2

- Teacher-Student
  - 2x Parameters
- Student predicts hidden representations of Teacher
- Drops masked inputs
- MSE/L1 Loss



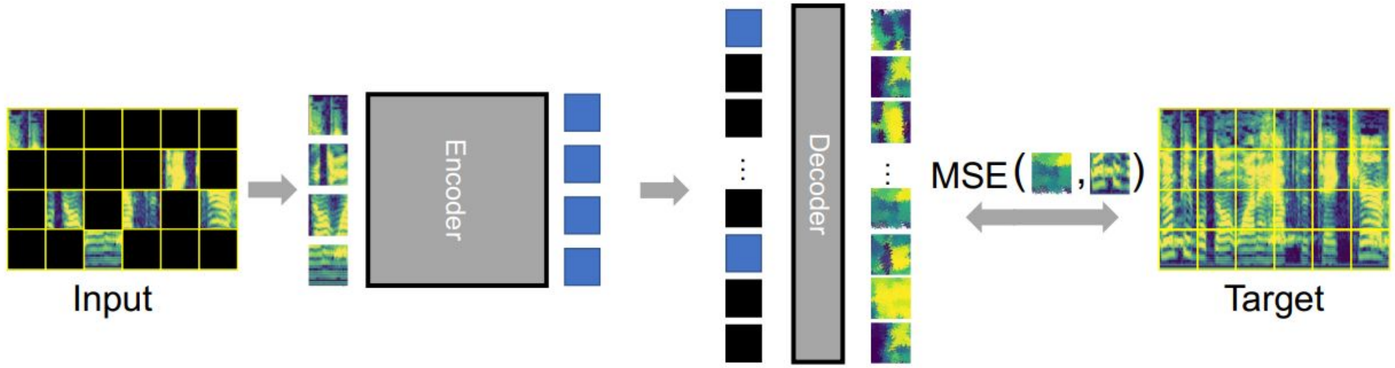
Supervised - Whisper

- Large scale pretraining (5 million hours)
- Weakly supervised data - ASR pipeline + Youtube Subtitles
- 680 M Encoder + 200 M Decoder
- Capable for multilingual ASR
- Unknown data, Unknown training, no reproducibility



SSL - AudioMAE

- Async Patch-level AutoEncoder
- Mel (Fast, Good)
- Large Encoder, Tiny Decoder
- Drops masked inputs (75 % - 85 %)
- MSE/L1 Loss
- Reconstruct input = Keeps information



Personal viewpoint (1)

- Classification targets for (Audio) SSL
  - SSL = Learn anything from data
  - Quantized representations = Keep important information
  - But what is important? How does a quantizer know?
  - Removing information = Bad
- SSL is not supervised learning
  - Low loss in supervised learning = good
  - Low loss in SSL = bad
- Raw-Wave inputs
  - Best case scenario: Model learns Mel spectrogram.
  - Average scenario: Compute wasted to learn Mel (Great!)
  - Worse case scenario: Harder to train, slower to converge, maybe collapse.

Personal viewpoint (2)

- Targets for SSL are important
  - Predicting hidden representations = Can only be used for classification
  - Predicting contextualized representations = Can only be used for classification
  - Predicting input = Can be used for classification + regression (separation).
- Data + parameter scale + simple > Complicated
  - What is better, given 1 week of compute ?
    - 10 M hours method A, Model = 7B
    - 1 M hours method B, Model = 100M
  - I think A > B

Personal viewpoint (3) - Quantized representations

SELF-SUPERVISED LEARNING FOR SPEECH ENHANCEMENT THROUGH SYNTHESIS

Bryce Irvin[1,2], *Marko Stamenovic*[1], *Mikolaj Kegler*[1], *Li-Chia Yang*[1]

[1] Bose Corporation, USA, [2] Georgia Institute of Technology, USA

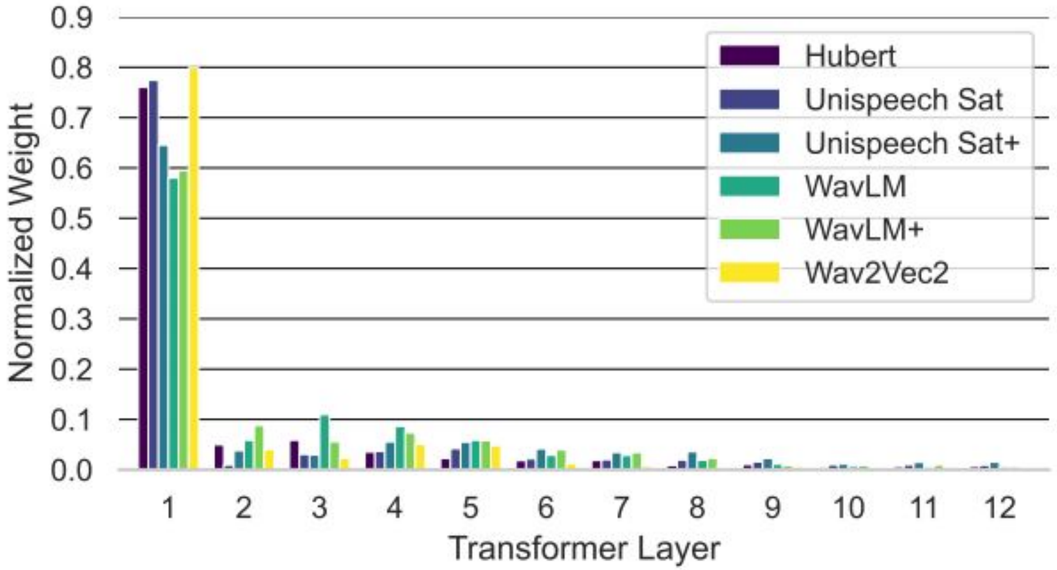
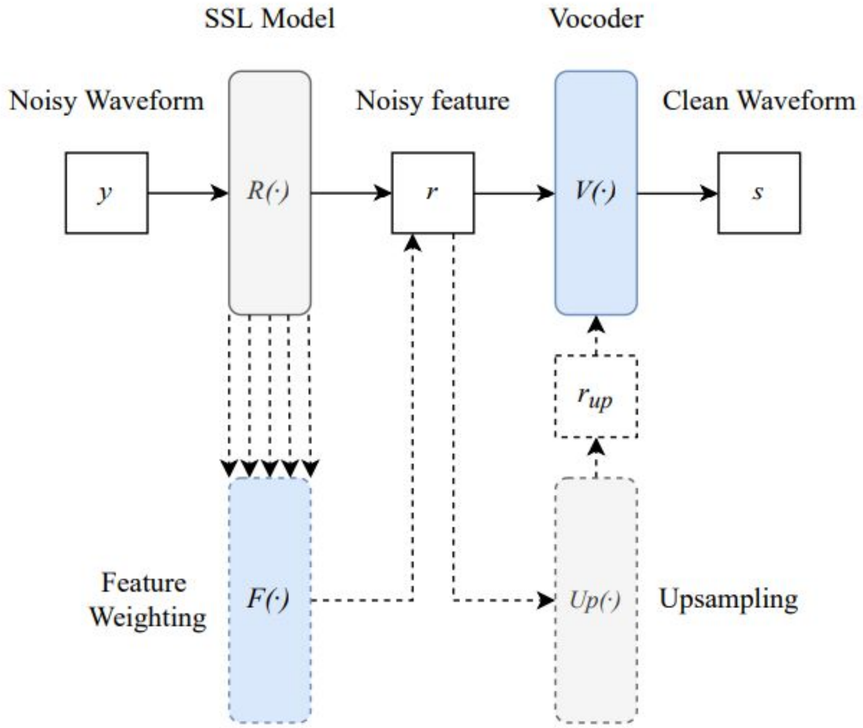


Fig. 2. SSL transformer layer weighting. The weights have been unit-normalized for each model.



Quantized SSL model = removed information = Needs to use lower features

Challenges

- HuBERT, BEST-RQ, BEATs drop information (CE-Loss + Quantization)
- Data2Vec2 uses 2x parameters = not scalable (1-100B)
- Data2Vec2 only useful for classification, not regression
- Whisper is supervised and drops information (speaker …)
- AudioMAE seems simplest and best for scale
  - Mainly for sound tasks
  - Scaling up has not been investigated

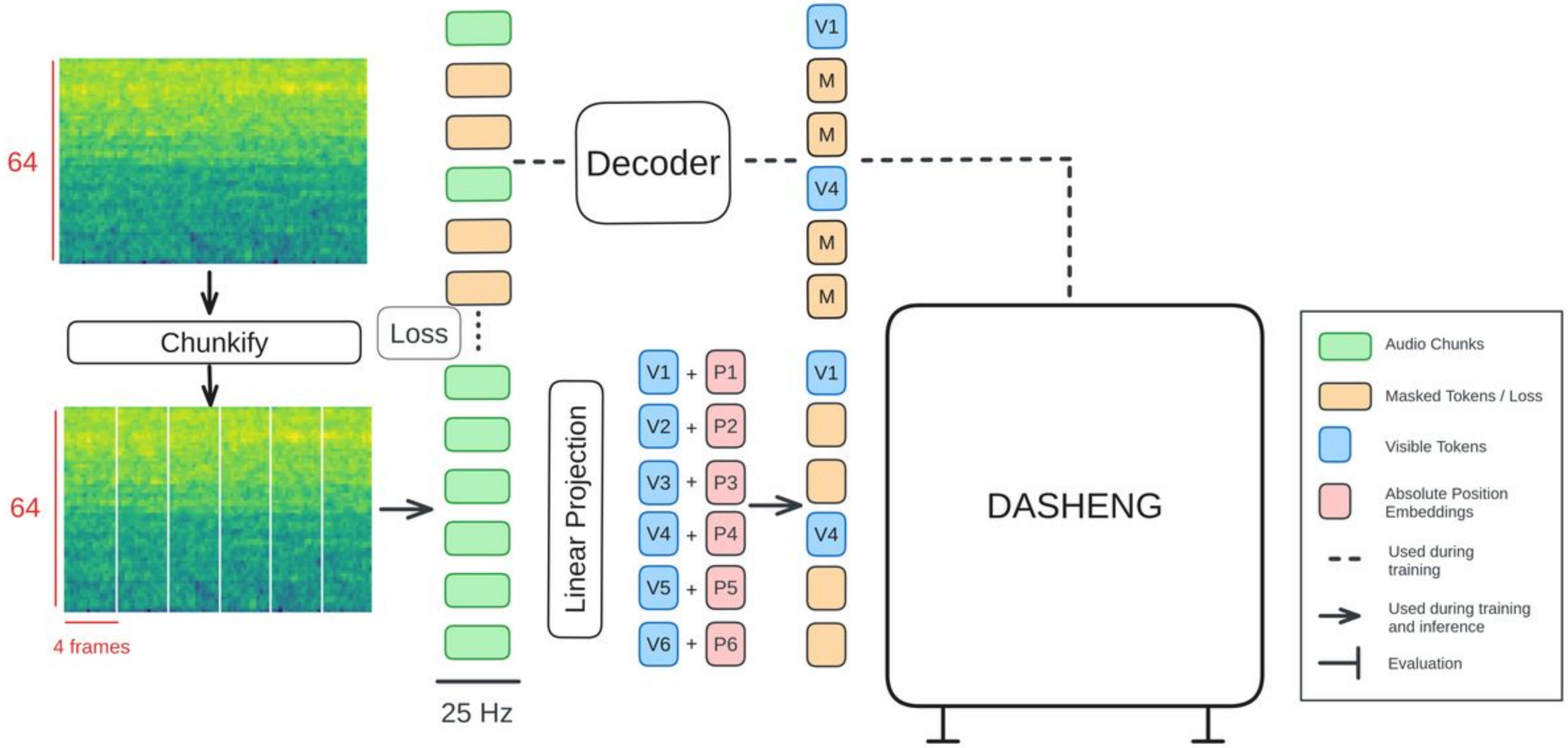
General Audio Features

- Data: Most is Speech
- Finding non-speech data is hard
- Filtered data via Audio-Visual correspondence might be good
- ACAV100M, 12 million hours filtered to 26k
- Not much other public data that I know of

Dataset	# Samples	Duration (h)	Type
ACAV100M	94,934,272	263,000	General
AudioSet	1,904,747	5,100	General
VGGSound	176,819	488	General
MTG-Jamendo	55,701	3,768	Music
All	97,071,539	272,356	General

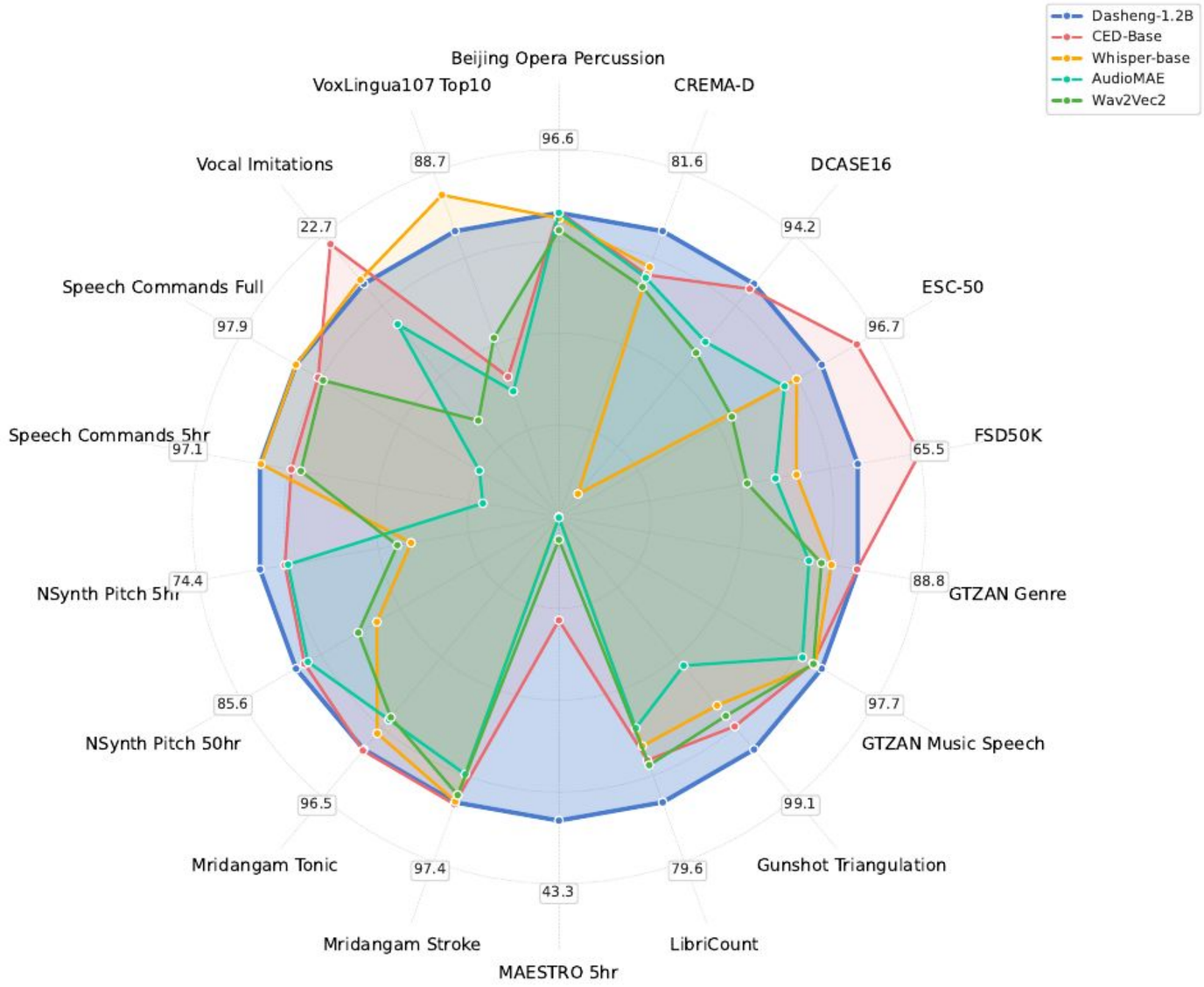
Dasheng

Dasheng = Scaled frame-level AudioMAE



HEAR Result

- MLP ontop of frozen embeddings
- 21 Tasks (19 unique)



HEAR Problems

- Evaluation mostly utterance level + sounds
- Random tasks
  - Beehive (5 min, 500 samples)
  - Gunshot (88 samples)
  - DCASE16 (72 samples)
- Speech related tasks too few
  - No speaker
  - No ASR
- “Speech” Tasks:
  - Speech Commands (Utterance-level , 5h + full)
  - Crema-D (Emotion)
  - VoxLingua (LangID, top 10 langs, mostly Indoeuropean)
- I/O implementation: GPU 1%, CPU 1000000000 %
  - Saves 1-20k npz files on disk, kkthxbye

ICME 2025 Audio Encoder Capability Challenge

- MLP (Linear) + kNN (for search) evaluation
- Requirement: No access to training data
- More diverse tasks
- Improved Speech: ASR + Speaker + LangID
- Baseline using Dasheng
- Better I/O using Webdataset
- Reasonable Code (I have never got SuperB to run)

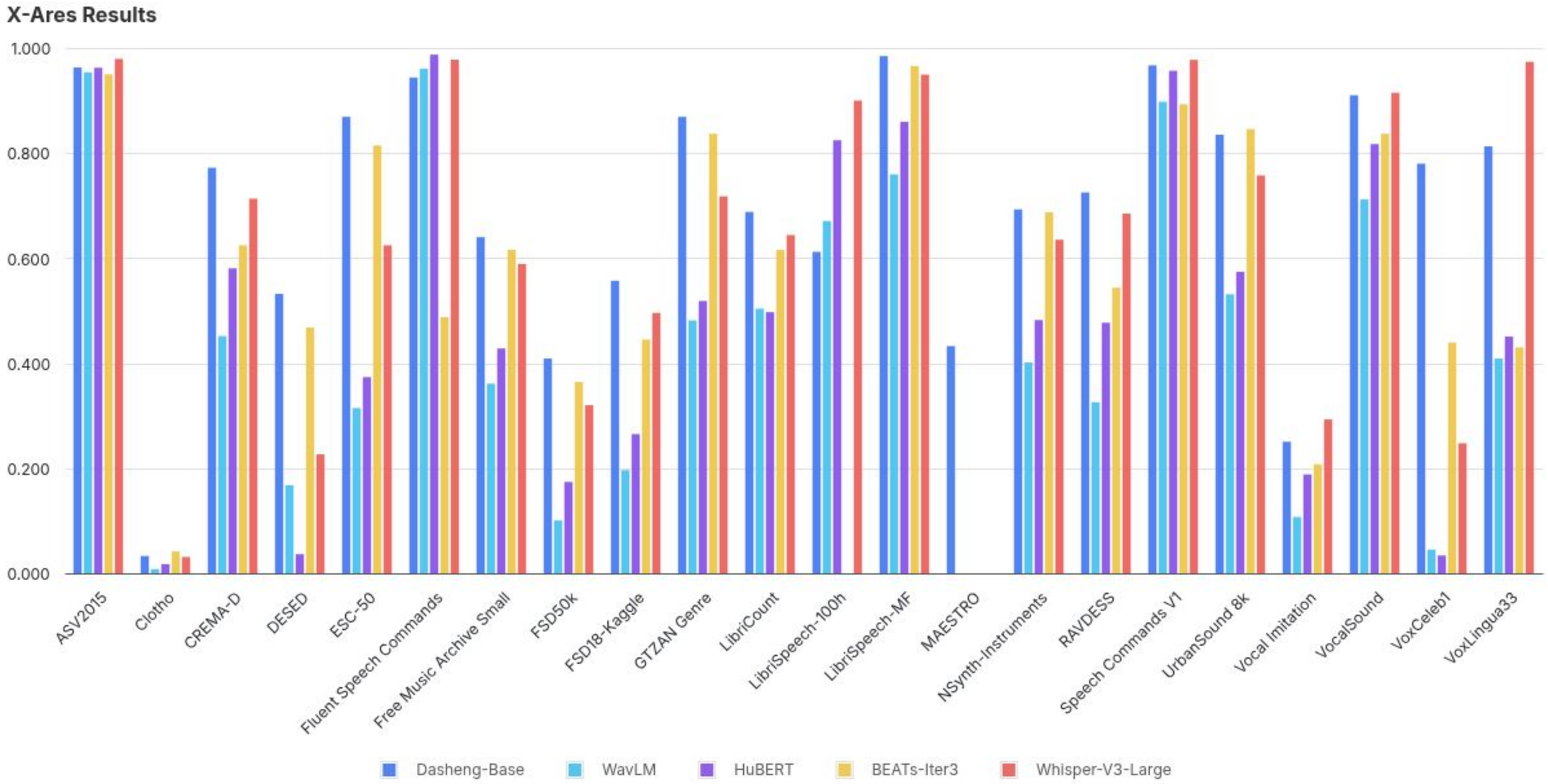
```python
class DashengEncoder(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.sampling_rate = 16000
        self.output_dim = 768
        self.hop_size_in_ms = 40
        self.max_length = int(10 * self.sampling_rate)
        self.model = dasheng_base()

    def forward(self, audio: torch.Tensor):
        assert isinstance(audio, torch.Tensor)
        if audio.ndim == 1:
            audio = audio.unsqueeze(0)

        self.model.eval()
        with torch.inference_mode():
            if audio.shape[-1] > self.max_length:
                output = []
                for chunk in audio.split(self.max_length, dim=-1):
                    if chunk.shape[-1] < self.sampling_rate:
                        chunk = torch.nn.functional.pad(chunk, (0, self.sampling_rate - chunk.shape[-1]))

                    tmp_output = self.model(chunk)
                    output.append(tmp_output)
                output = torch.cat(output, dim=1)
            else:
                output = self.model(audio)
        return output
```

```
1 pip install xares
2 python -m xares.run --max-jobs 8 example/dasheng/dasheng_encoder.py src/tasks/*.py
3
```

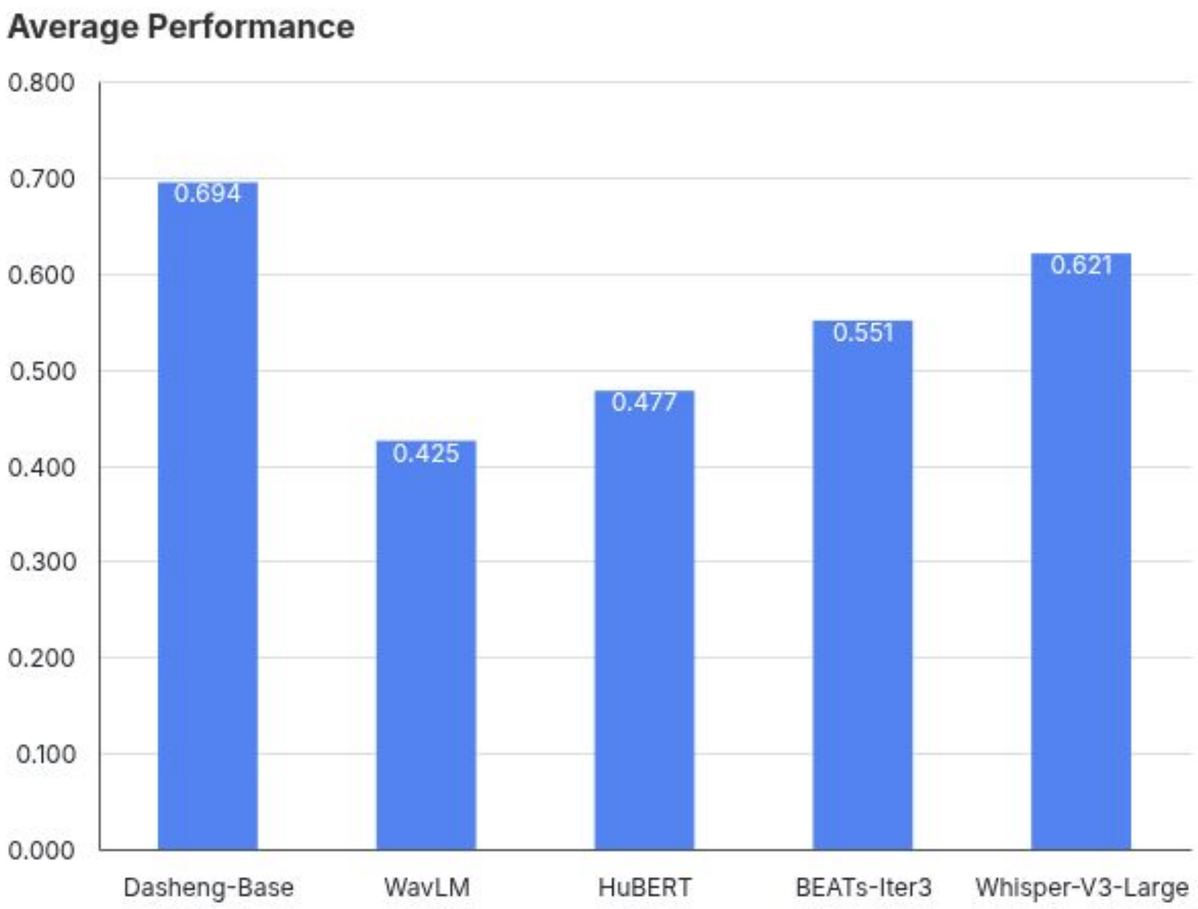
ICME 2025 Audio Encoder Capability Challenge

Domain	Dataset	Task Type	Metric	#	Track B
Speech	Speech Commands [16]	Keyword spotting	Acc	30	✓
	LibriCount [17]	Speaker counting	Acc	11	✓
	VoxLingua107 [18]	Language identification	Acc	33	✓
	VoxCeleb1 [19]	Speaker identification	Acc	1251	✓
	LibriSpeech [20]	Gender classification	Acc	2	✓
	Fluent Speech Commands [21]	Intent classification	Acc	248	✓
	VocalSound [22]	Non-speech sounds	Acc	6	✓
	CREMA-D [23]	Emotion recognition	Acc	5	✓
	speechocean762 [24]	Phoneme pronunciation	MSE	3	✗
	ASV2015 [25]	Spoofing detection	EER	2	✓
Sound	ESC-50 [26]	Environment classification	Acc	50	✓
	FSD50k [27]	Sound event detection	mAP	200	✗
	UrbanSound 8k [28]	Urban sound classification	Acc	10	✓
	DESED [29]	Sound event detection	Segment-F1	10	✓
	FSD18-Kaggle [30]	Sound event detection	mAP	41	✗
	Clotho [31]	Sound retrieval	Recall@1	-	✗
	Inside/outside car†	Sound event detection	Acc	2	✓
	Finger snap sound†	Sound event detection	Acc	2	✓
	Key scratching car†	Sound event detection	Acc	2	✓
	Subway broadcast†	Sound event detection	Acc	2	✓
	LiveEnv sounds†	Sound event detection	mAP	18	✗
Music	MAESTRO [32]	Note classification	Acc	88	✓
	GTZAN Genre [33]	Genre classification	Acc	10	✓
	NSynth-Instruments [34]	Instruments Classification	Acc	11	✓
	NSynth-Pitch [34]	Pitches Classification	Acc	128	✓
	Free Music Archive Small [35]	Music genre classification	Acc	8	✓

ICME 2025 Audio Encoder Capability Challenge Base



ICME 2025 Audio Encoder Capability Challenge Base



WavLab submission

- Patch-level BEATs reimplementation
- Useless for ASR
- Needs fusion

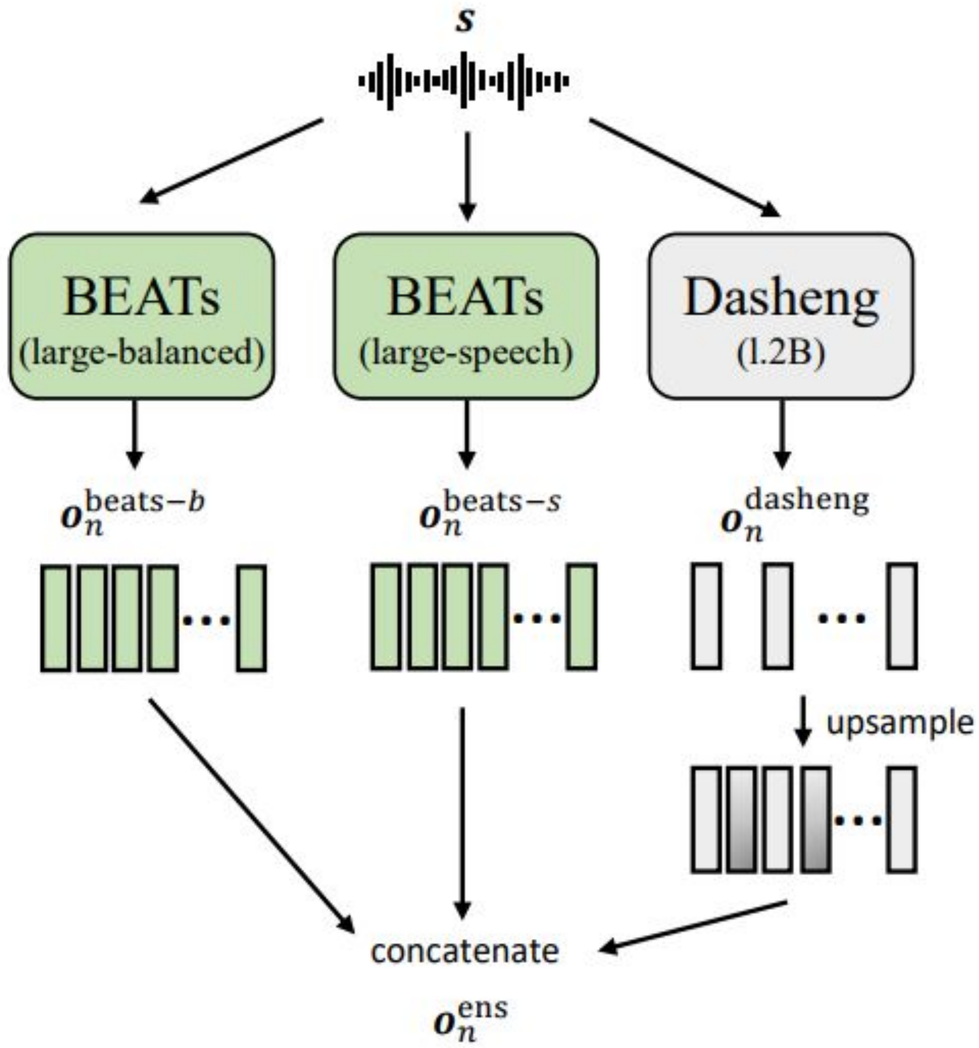


TABLE II
TRAINING DATA STATISTICS

Datasource	Domain	Hours
AudioSet [13]	Sound	5,000
Freesound	Sound	4,648
BBC Soundeffects	Sound	1,000
VGGSound [14]	Sound	548
Cochlscene [15]	Sound	169
EpicKitchen [16]	Sound	157
FMA [17]	Music	7,824
MTG-Jamendo [18]	Music	3,701
YODAS [9]	Speech	34,759
Commonvoice [19]	Speech	16,304
EARS [20]	Speech	77
Total		**74,187**

WavLab results

TABLE I

MLP TRACK RESULTS. WE USED X-ARES TOOLKIT [8] FOR EVALUATIONS. CHALLENGE BASELINE DENOTES THE BEST CHALLENGE BASELINE AMONG DASHENG-BASE, DATA2VEC AND WHISPER.

Task	Domain	Challenge Baseline	BEATs (90M) iter3 [1]	Dasheng 1.2B [2]	BEATs (300M) Balanced	BEATs (300M) Speech	Ensemble (Submission)
FSD50k	Sound	0.408	0.217	0.455	0.380	0.432	**0.463**
Vocal Imitation	Sound	0.238	0.212	0.293	0.214	0.223	**0.295**
FSD18-Kaggle	Sound	0.557	0.545	0.627	0.689	0.612	**0.764**
DESED	Sound	0.532	0.560	0.563	0.552	0.551	**0.566**
ESC-50	Sound	0.869	0.835	0.891	0.868	0.857	**0.904**
Clotho	Sound	0.033	0.042	0.036	0.040	**0.041**	0.038
UrbanSound 8k	Sound	0.835	0.853	0.846	**0.863**	0.857	0.862
NSynth-Instruments	Music	0.693	0.579	0.660	0.589	0.550	**0.729**
GTZAN Genre	Music	0.869	0.836	0.886	0.859	0.845	**0.898**
Free Music Archive Small	Music	0.640	0.614	**0.647**	0.624	0.616	0.637
LibriCount	Speech	0.688	0.665	0.728	0.699	0.705	**0.747**
CREMA-D	Speech	0.772	0.642	0.790	0.659	0.670	**0.815**
RAVDESS	Speech	0.725	0.564	**0.793**	0.630	0.655	0.792
Fluent Speech Commands	Speech	0.962	0.545	**0.973**	0.585	0.700	0.956
LibriSpeech-MF	Speech	0.985	0.970	0.975	0.973	**0.986**	0.985
Speech Commands V1	Speech	0.967	0.910	**0.973**	0.944	0.958	0.972
VoxLingua33	Speech	0.855	0.398	**0.860**	0.480	0.615	0.817
VocalSound	Speech	0.910	0.865	**0.925**	0.877	0.879	0.909

WavLab results

Baseline = Dasheng - Base (86M). Dasheng significantly outperforms proposed with less params + ASR.

TABLE I

MLP TRACK RESULTS. WE USED X-ARES TOOLKIT [8] FOR EVALUATIONS. CHALLENGE BASELINE DENOTES THE BEST CHALLENGE BASELINE AMONG DASHENG-BASE, DATA2VEC AND WHISPER.

Task	Domain	Challenge Baseline	BEATs (90M) iter3 [1]	Dasheng 1.2B [2]	BEATs (300M) Balanced	BEATs (300M) Speech	Ensemble (Submission)
FSD50k	Sound	0.408	0.217	0.455	0.380	0.432	**0.463**
Vocal Imitation	Sound	0.238	0.212	0.293	0.214	0.223	**0.295**
FSD18-Kaggle	Sound	0.557	0.545	0.627	0.689	0.612	**0.764**
DESED	Sound	0.532	0.560	0.563	0.552	0.551	**0.566**
ESC-50	Sound	0.869	0.835	0.891	0.868	0.857	**0.904**
Clotho	Sound	0.033	0.042	0.036	0.040	**0.041**	0.038
UrbanSound 8k	Sound	0.835	0.853	0.846	**0.863**	0.857	0.862
NSynth-Instruments	Music	0.693	0.579	0.660	0.589	0.550	**0.729**
GTZAN Genre	Music	0.869	0.836	0.886	0.859	0.845	**0.898**
Free Music Archive Small	Music	0.640	0.614	**0.647**	0.624	0.616	0.637
LibriCount	Speech	0.688	0.665	0.728	0.699	0.705	**0.747**
CREMA-D	Speech	0.772	0.642	0.790	0.659	0.670	**0.815**
RAVDESS	Speech	0.725	0.564	**0.793**	0.630	0.655	0.792
Fluent Speech Commands	Speech	0.962	0.545	**0.973**	0.585	0.700	0.956
LibriSpeech-MF	Speech	0.985	0.970	0.975	0.973	**0.986**	0.985
Speech Commands V1	Speech	0.967	0.910	**0.973**	0.944	0.958	0.972
VoxLingua33	Speech	0.855	0.398	**0.860**	0.480	0.615	0.817
VocalSound	Speech	0.910	0.865	**0.925**	0.877	0.879	0.909

WavLab conclusion

- Patch-level model that underperforms vs frame-level
- Slow training
- Unreasonable SSL objective, reconstruction >> classification
- OpenBEATs paper further demonstrates performance difference:

Model	Param	DSD	US8k	F50k	ESC	FKgl	Clotho	Avg
Dasheng B [25]	90M	0.53	0.84	0.41	0.87	0.56	0.03	0.54
Data2Vec [48]	90M	0.14	0.44	0.08	0.25	0.20	0.01	0.19
Whisper [49]	74M	0.13	0.72	0.26	0.61	0.48	0.03	0.37
EAT B [50]	88M	0.36	0.79	0.36	0.66	0.38	0.04	0.43
BEATs (iter3) [1]	90M	0.56	0.85	0.22	0.84	0.55	0.04	0.51
EAT L [50]	309M	0.44	0.82	0.44	0.73	0.58	**0.06**	0.51
Dasheng-0.6B [25]	600M	0.54	0.85	0.45	0.88	0.58	0.04	0.56
Dasheng-1.2B [25]	1.2B	0.56	0.85	**0.46**	**0.89**	**0.63**	0.04	0.57
OpenBEATs L	300M	**0.57**	**0.87**	0.43	0.86	0.54	0.05	0.55

Bytedance submission

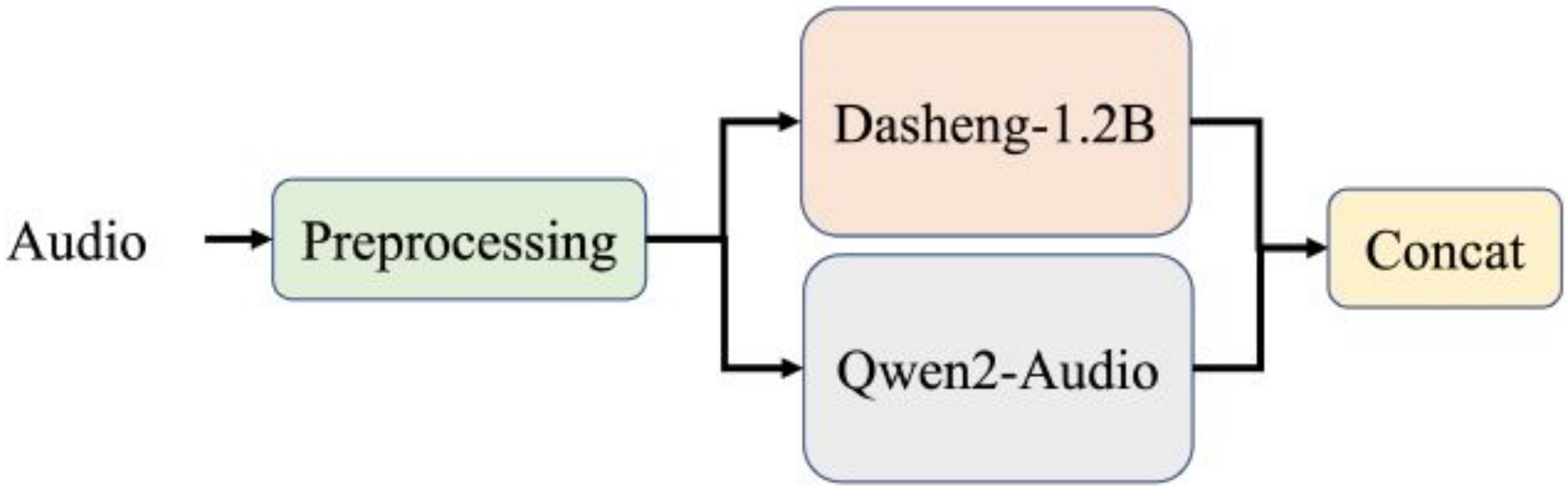


Figure 1: The DQF audio encoder.

Bytedance result

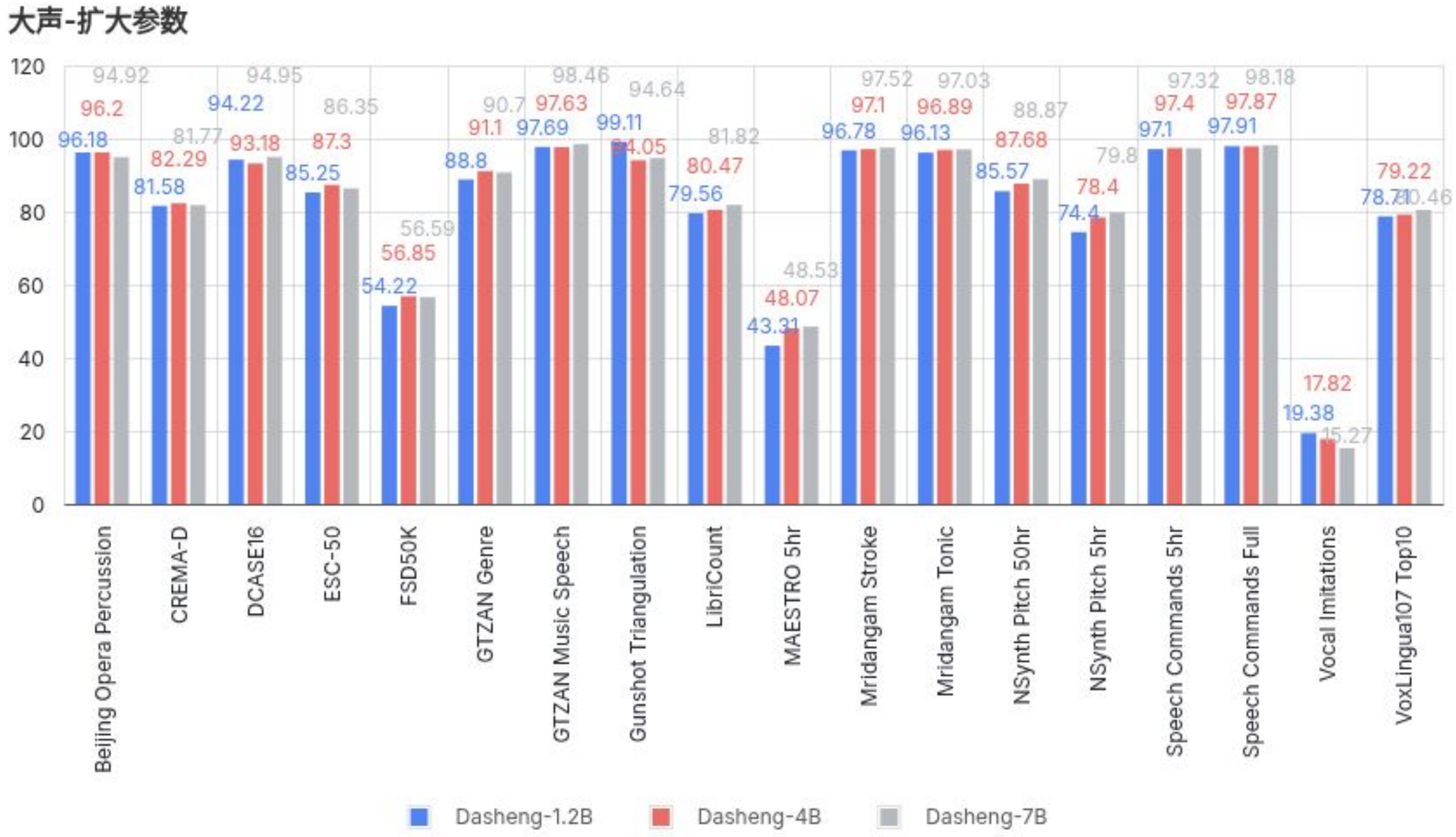
- Technically “cheated”: Qwen-Audio encoder saw training data
- But Qwen did not state in report …
- Again ensemble ….

	MLP		KNN	
Task	Dasheng	DQF	Dasheng	DQF
Clotho	0.033	**0.053**	\	\
CREMA-D	0.809	**0.871**	0.490	0.866
DESED	0.571	**0.572**	\	\
ESC-50	0.887	**0.967**	0.663	**0.917**
Fluent_Speech_Commands	0.973	**0.993**	0.445	**0.993**
Free_Music_Archive_Small	0.647	**0.697**	0.603	**0.637**
GTZAN_Genre	0.891	**0.933**	0.785	**0.861**
LibriCount	**0.727**	0.675	**0.408**	0.312
LibriSpeech-MF	**0.987**	0.982	**0.873**	0.859
NSynth-Instruments	0.749	**0.767**	0.433	**0.629**
RAVDESS	0.794	**0.922**	0.535	**0.834**
Speech_Commands_V1	0.971	**0.980**	0.935	**0.980**
UrbanSound_8k	0.846	**0.862**	0.683	**0.810**
Vocal_Imitation	0.297	**0.343**	0.112	**0.187**
weighted_average	0.731	**0.759**	0.560	**0.726**

Larger Dasheng

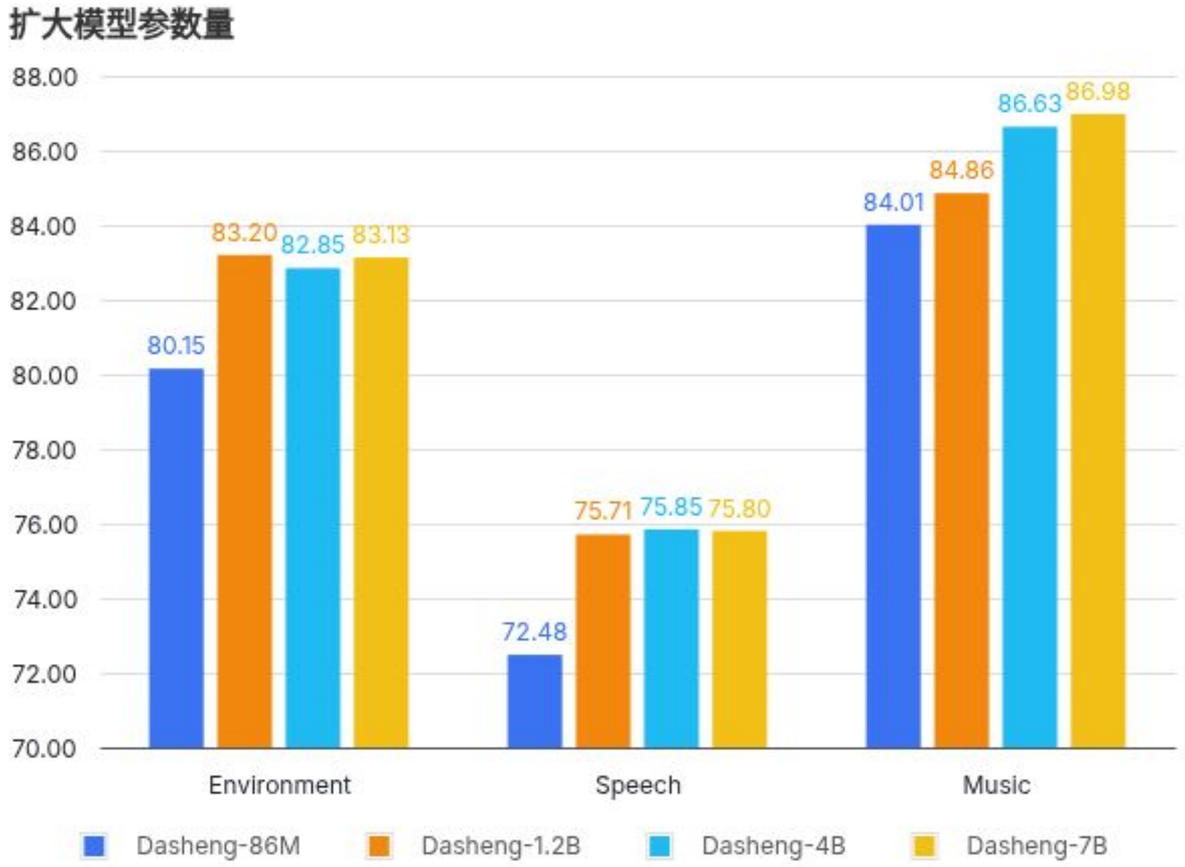
Training 4 days (8GPU) -> 7 Days

HEAR results not promising.



Larger Dasheng

Mostly same as 1.2B

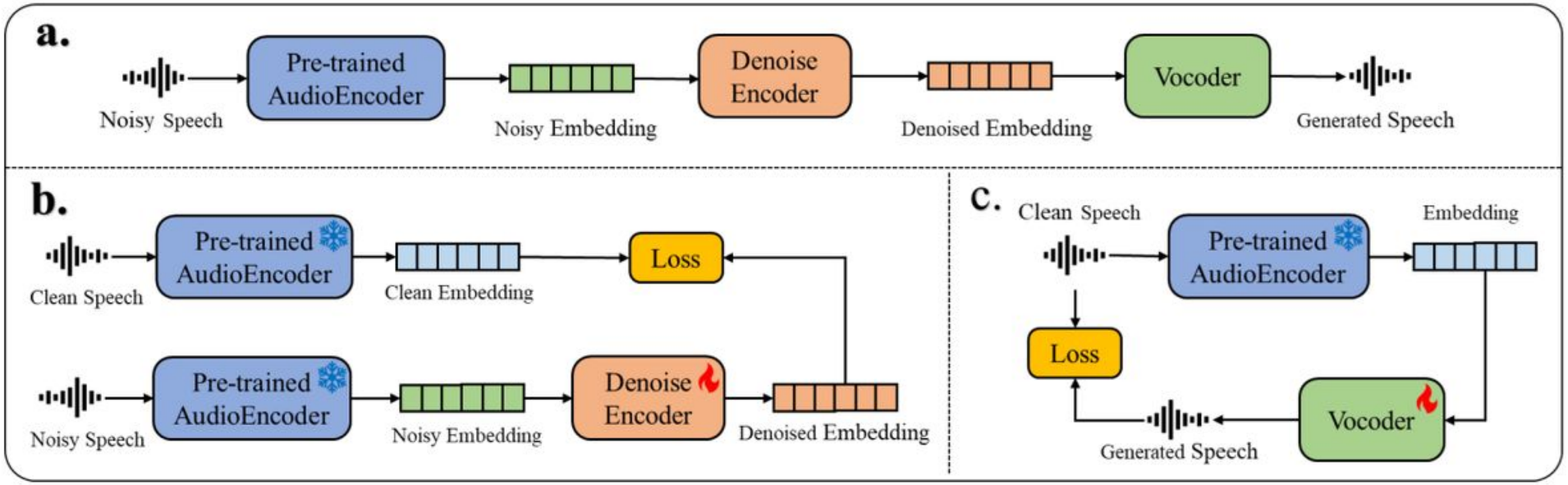


Audio Encoder Conclusion

- Depending on Loss, many audio encoders loose information
- AutoEncoders are good audio representation learners
- Proposed frame-level autoencoder (Dasheng) performs well across benchmarks
- Further scaling parameters does not work well
- Data scaling might also skew performance to Speech
- Obtaining more data is still problematic, need filtering

Applications of General Audio Features

Denoising



Noise

VoiceFixer

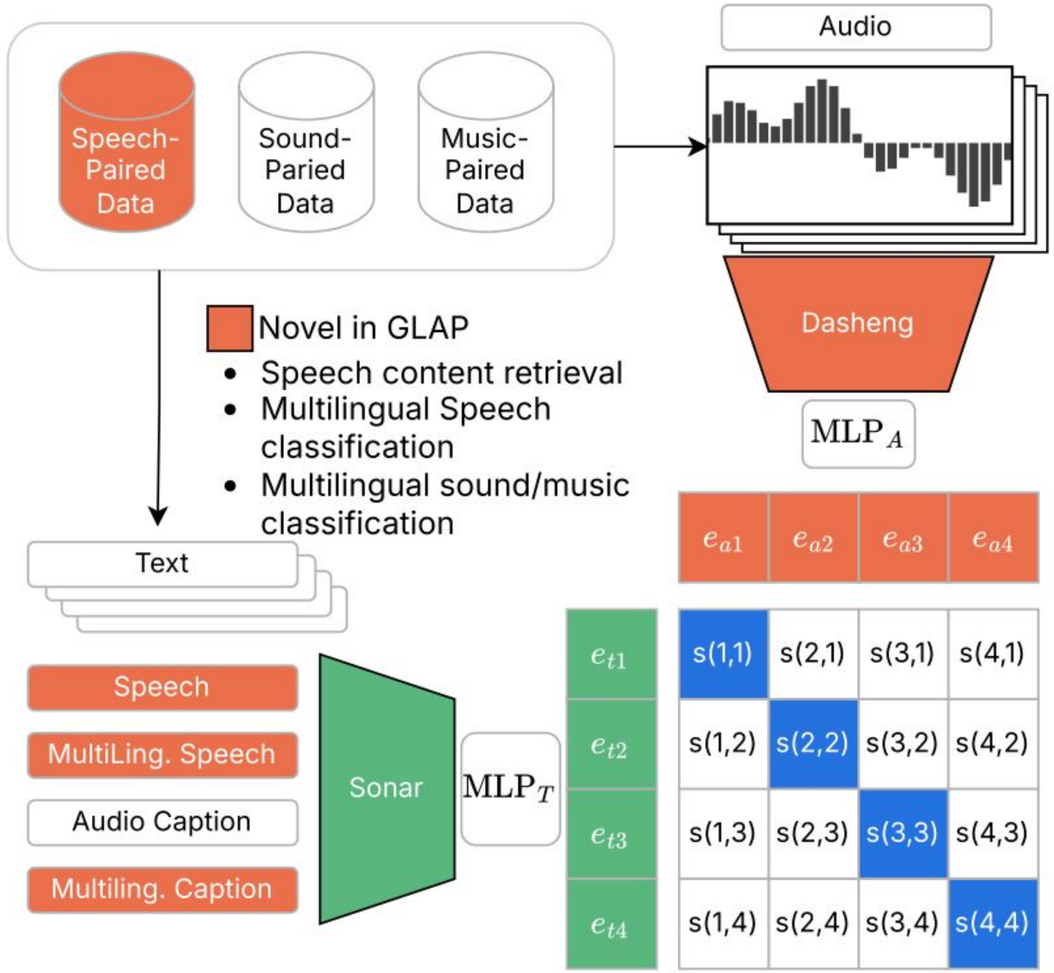
Ours

Large scale data = Large unusable data

How to filter data?

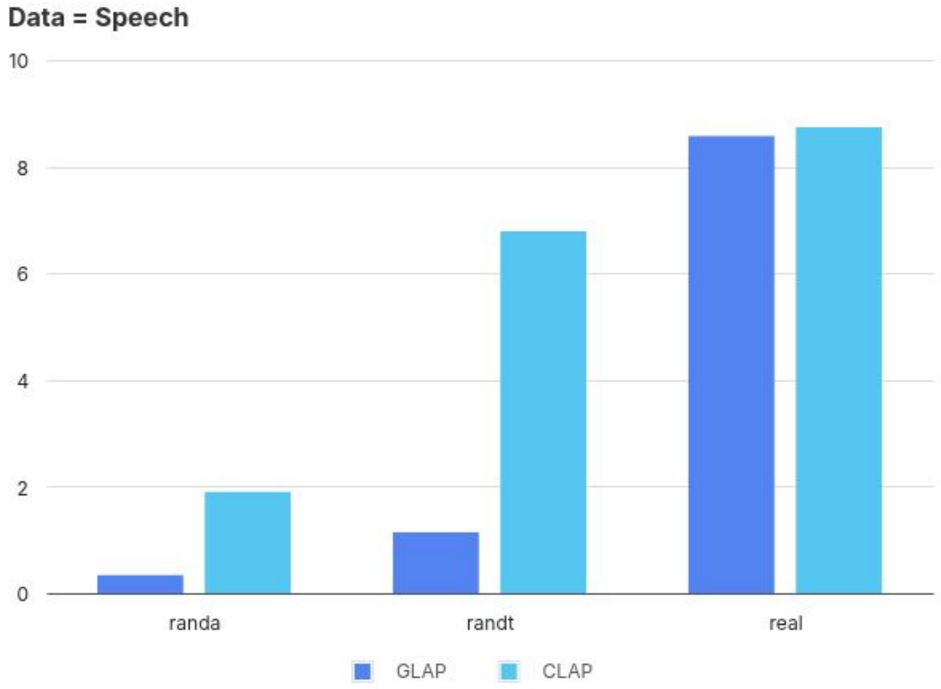
GLAP

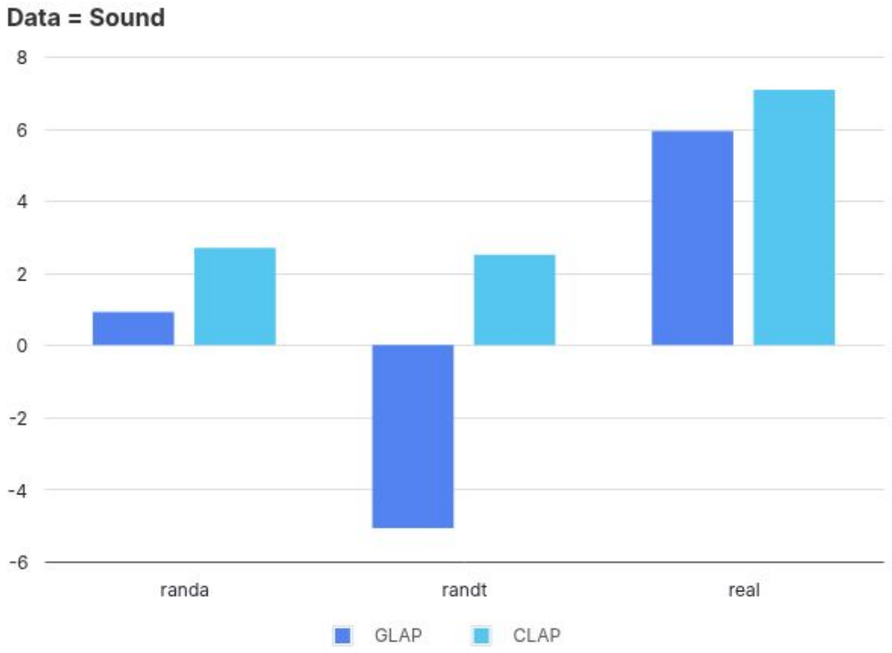
- CLAP has no spoken content capabilities
- Most CLAP's only work for English
- GLAP can be used as a general audio-text filter.
- Support for speech in 8 - 100 languages
- Support for music/sound in 8 languages.
- Filter audio via audio-text cosine distance.

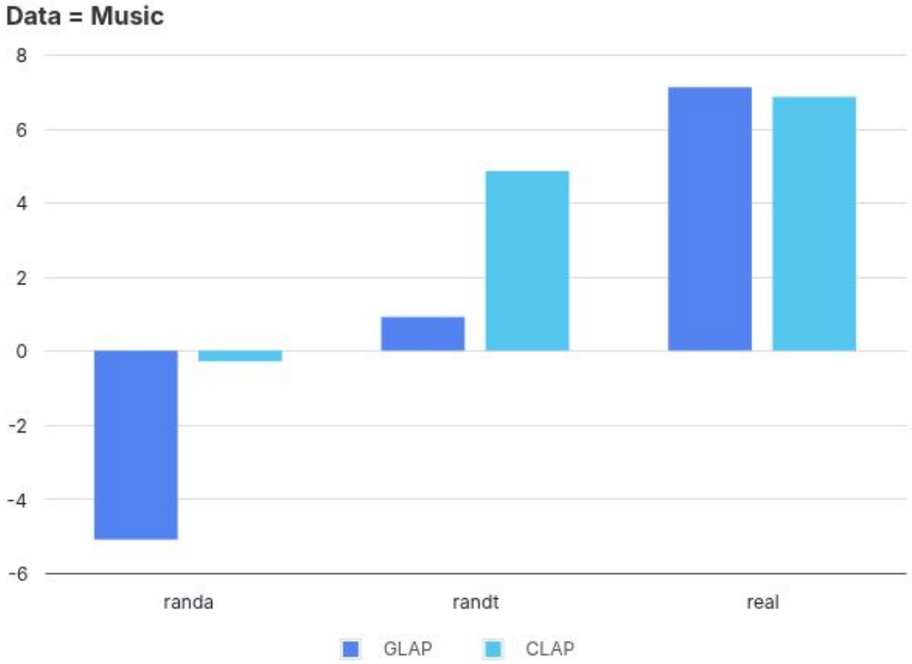


GLAP vs. CLAP

- randa = random audio
- randt = random text
- real = audio - text pair
- CLAP unreliable
- GLAP scores are related



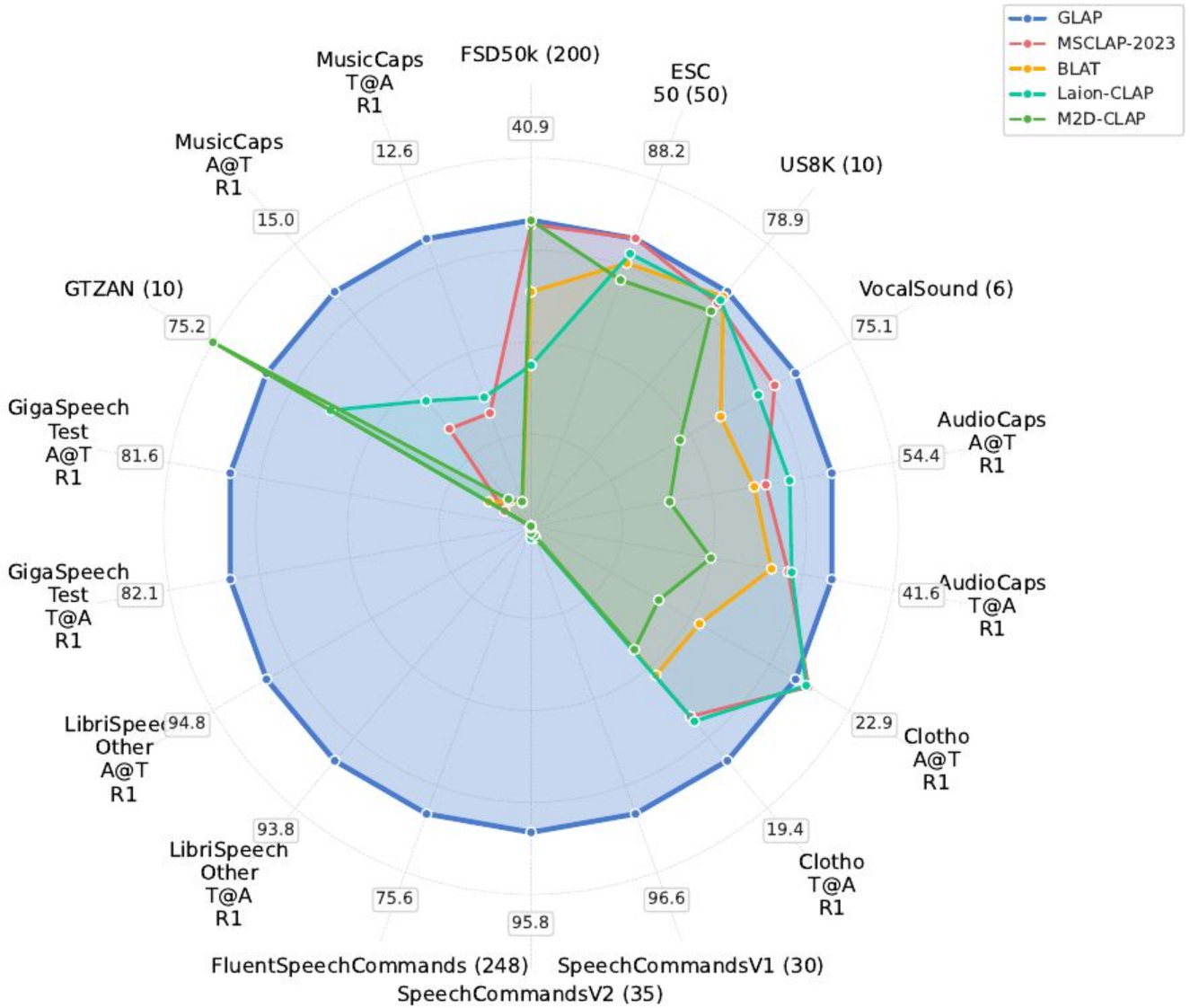




GLAP - Audio Encoder importance

	Encoder	Sound		Music	Speech	
声音	CED-Base	58.6	62.0	25.1	87.8	70.6
	Microsoft	55.1	64.3	23.9	91.8	44.0
说话	Openai-Whisper	46.5	52.9	15.8	98.9	99.4
	Microsoft-WavLM	36.1	47.5	14.8	99.9	96.3
	Dasheng	55.8	60.1	20.3	94.8	99.0

GLAP Results



GLAP Application

Prompt:
语速很快，声音干净，男声，充满激情

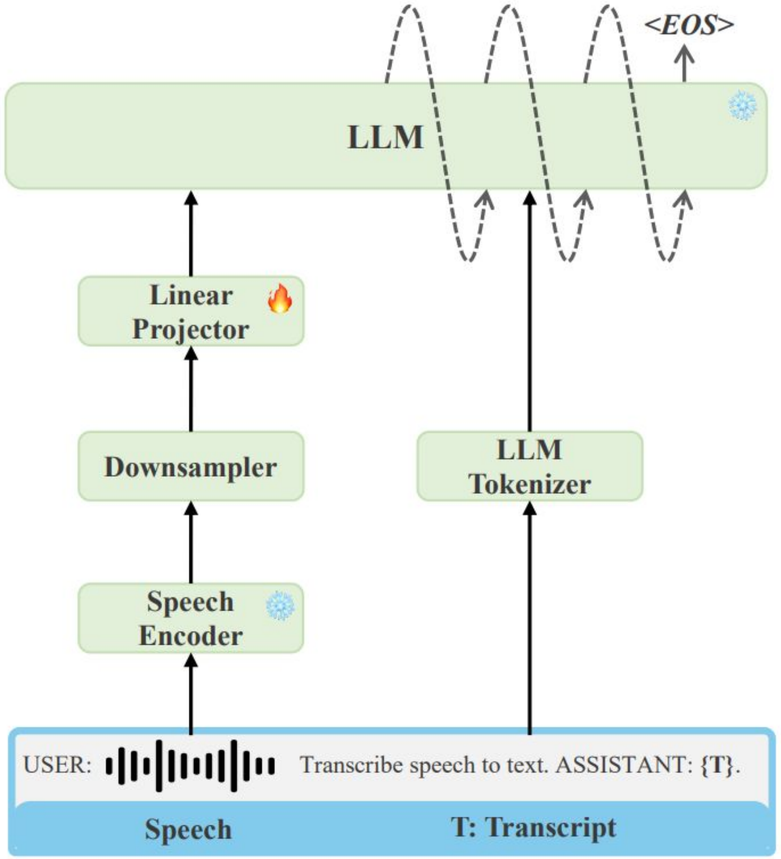
Result (2200 Sample Dataset)

GLAP - Application

MiDashengLM - Scalable Audio LLM

Audio LLM

- Input audio + text prompt
- Alignment problem of audio -> text
- Training identical to text-only.



General Audio Language LLM - Current State Audio Encoders

Supports **ASR** + General Audio

Qwen-Audio 1 & 2 & Omni
Kimi Audio
Audio Flamingo 3
SALMONN
Listen Think and Understand (LTU)

Whisper

Partial Support - Only ASR or Caption or Music

Audio Flamingo 1 & 2
GAMA
Pengi
DeSTA
SLAM-LLM

No-Whisper

General Audio Language LLM - Current State

Pretrain:

- 1: Obtain and filter data, generate text with ASR
- 2: Train with ASR data next token objective

Finetune:

- ASR + Sound + Music

General Audio Language LLM - Shortfalls

- ASR pretraining:
  - Local relationship, no global context needed
  - Simple objective (left - to - right alignment)
  - Heavy data waste (non ASR)
- Pretrained models:
  - Reliance on Whisper as encoder
  - Whisper = ASR encoder -> Loose of information (Speaker, Environment)
  - High frame-rate (50 Hz) -> High Compute
  - Unreasonable 30s padding for Whisper -> High Compute

The quick brownie points out that the

	The	quick	brown	ie	points	out	that	the
The	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
quick	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
brown	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ie	0.0000	0.6357	0.3643	0.0000	0.0000	0.0000	0.0000	0.0000
points	0.0000	0.3232	0.3457	0.3312	0.0000	0.0000	0.0000	0.0000
out	0.0000	0.1915	0.2085	0.3366	0.2634	0.0000	0.0000	0.0000
that	0.0000	0.1183	0.1298	0.1605	0.3424	0.2490	0.0000	0.0000
the	0.0000	0.0827	0.1104	0.1250	0.1788	0.2466	0.2564	0.0000

MiDashengLM - Motivation

- AudioLLM for all - reproducible public data + model + code
- Audio-text alignment without ASR
- Using general audio captions - learn local/global context
- Efficient training + inference - Very low frame rate audio features

MiDashengLM - Motivation

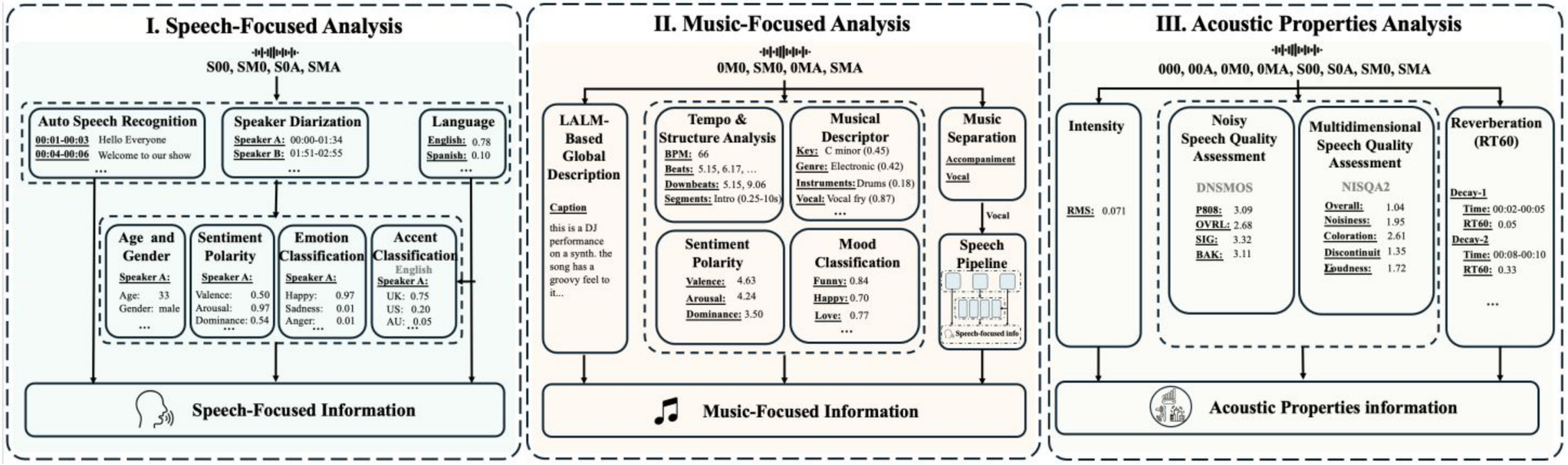
Why not use previous audio caption datasets?

- Overused Audioset (one audio - many text)
- All speech -> “Somebody is talking”
- English only

Dataset	Labeling	#Vocab	Avg. Sent	Source
ClothoV2 [11]	Manual	4366	11.32	Freesound
AudioCaps [13]		4844	8.70	Audioset
MusicCaps [42]		3730	47.17	Audioset
Songdescriber [43]		1811	26.31	MTG-Jamendo
LPMusicCaps-MTT [44]	LLM	4045	25.04	MagnaTagATune
LPMusicCaps-MSD [44]		14049	37.06	MillionSoundDatabase
SoundVECaps [17]		58401	31.48	Audioset
AutoACD [16]		20491	18.47	Audioset
AudiosetCaps [18]		21783	28.13	Audioset + VGGSound
WavCaps [15]		24592	7.84	Audioset + BBC + FreeSound + SoundBible
LAION-Audio-300M [45]		451927	37.55	?
Ours‡	Reasoning-LLM	644407	22.18	ACAV100M

MiDashengLM - General audio captions

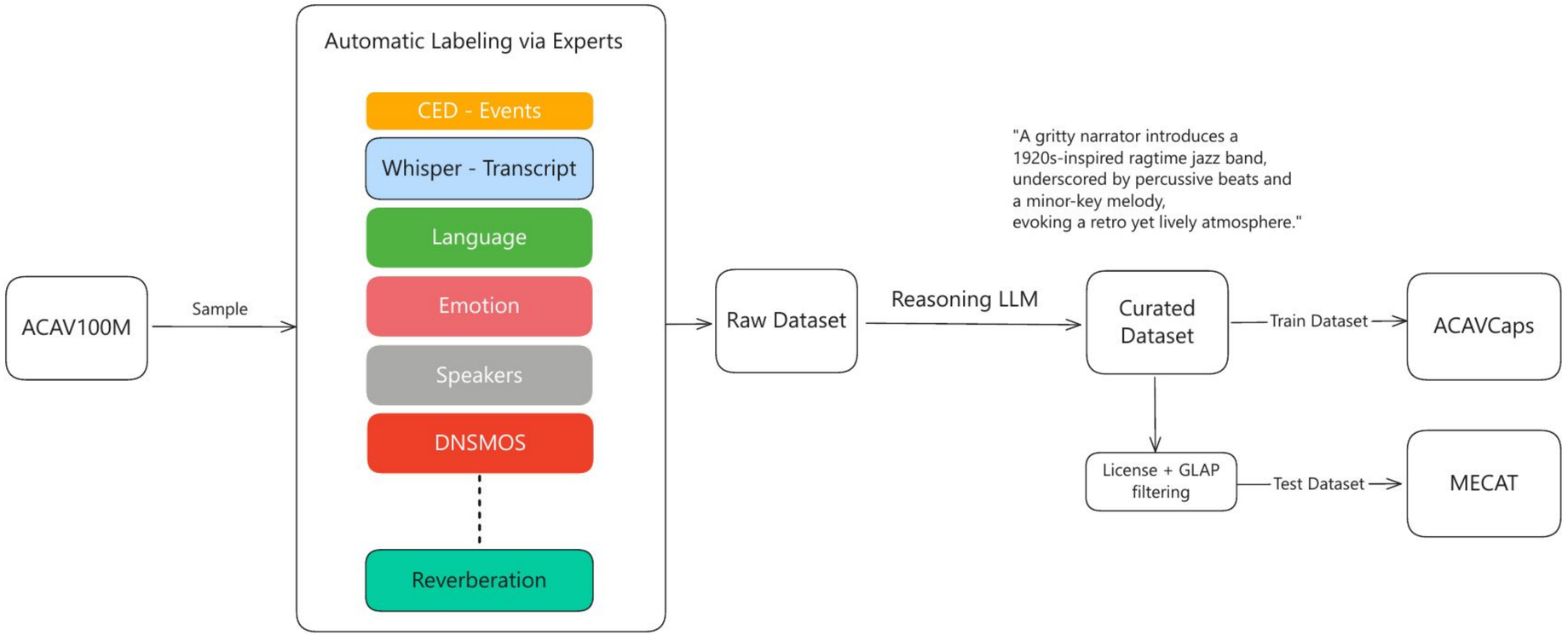
- Novel Pipeline that labels ACAV100M
- Music / Acoustic (Sound) / Speech
- Audio -> Text with experts
- Text + Reasoning LLM -> General caption



MiDashengLM - General audio captions

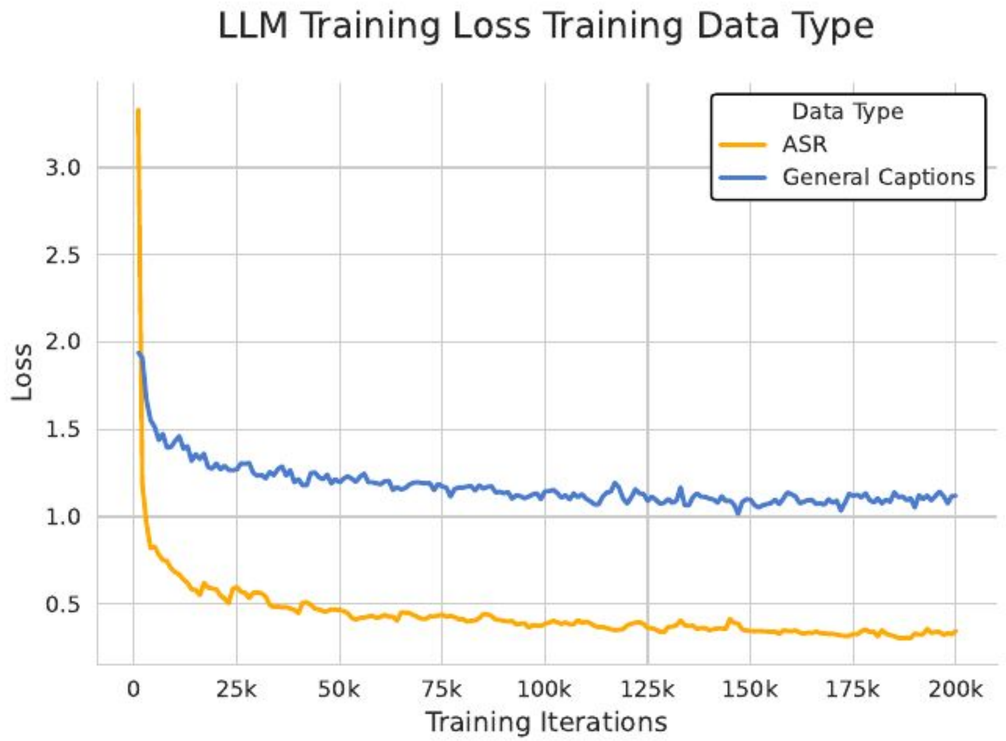
Category	Caption
Pure Speech	A female voice narrates a historical team competition (1966–1971) based on basketball rules, with intermittent synthetic speech modulation and variable acoustic reverberation.
Pure Sound	An outdoor scene with wind blowing, birds chirping, and a duck quacking, accompanied by significant background noise and low audio quality.
Pure Music	*"If I were a zombie, I'd want your heart, not your brain"* — A quirky electronic-pop anthem with gritty vocals, pulsing beats, and a dash of dark romance.
Mixed Music	The audio features a crowd cheering and clapping alongside electronic music with a synthesizer-driven, dark, and energetic soundscape.
Mixed Speech	A Russian voice demonstrates a synthesizer's capabilities over an experimental electronic backdrop, explaining its sound design and value in a gritty, vocal-fry tone.
Mixed Sound	A man speaks in English about entering a city and village, accompanied by the sounds of a running vehicle.

MiDashengLM - General audio captions - Train/Test



MiDashengLM - Alignment

- ASR vs. General caption losses
- ASR Data: 90 Languages, 55k hours
- General Caption: 1 Language, 38k hours



MiDashengLM - Audio Encoder

	Whisper-Large v3	Ours
Parameters	637.7M	630.3M
Pretraining data size	5M	270k
Training Objective	ASR	General captions
Context	30s	10s
Known pretraining data?	✗	✓ [37]
Open train code?	✗	✓ [47]
Open weight?	✓	✓

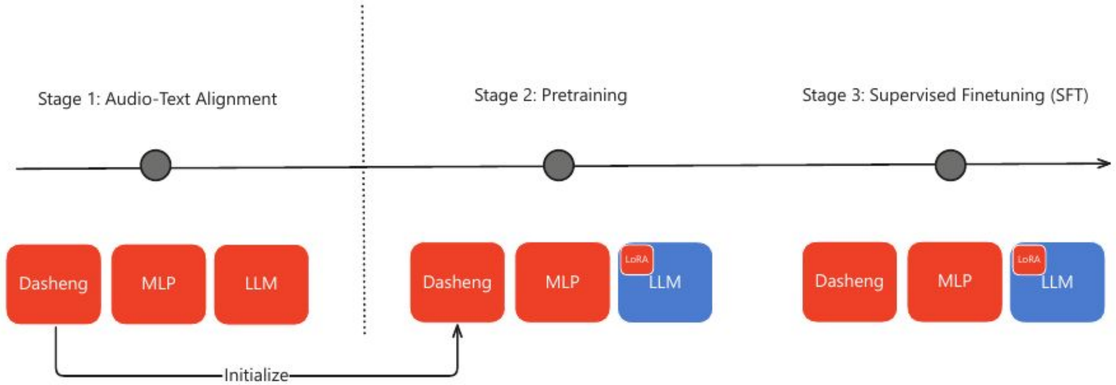
MiDashengLM - Alignment - Results

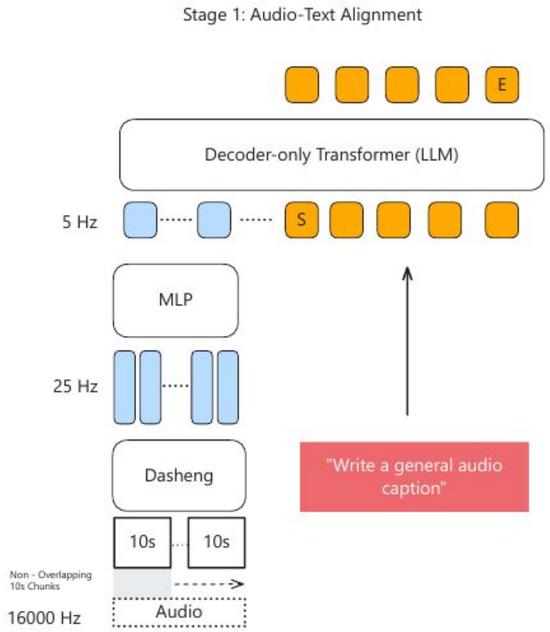
- Alignment much better than Whisper for most tasks
- Only for ASR (English, LibriSpeech) better performance
- VoxLingua33 is unfair: Whisper trained supervised
- Recall: Ours uses 5% of data, much cheaper training
- Significant gains for speaker recognition, sound events and music

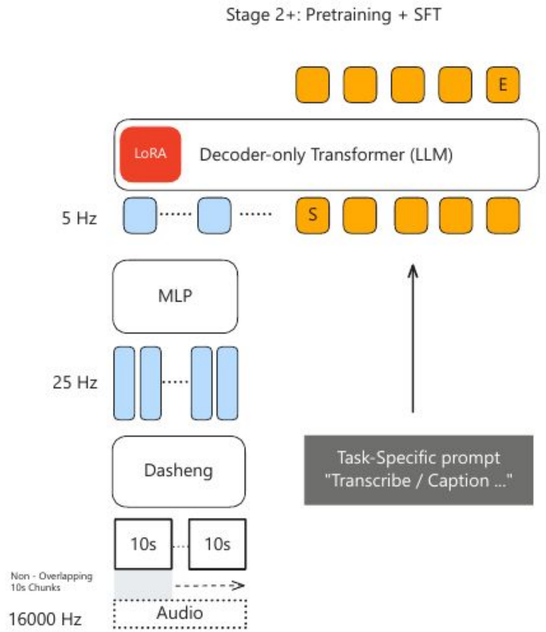
Domain	Dataset	Ours	Whisper	Ours vs. Whisper
Speech	LibriCount	61.9	**64.4**	-3.9
	LibriSpeech-100h	85.4	**90.0**	-5.1
	LibriSpeech-MF	**98.5**	94.9	+3.8
	VoxLingua33	92.3	**97.4**	-5.2
	Speech Commands V1	97.4	**97.7**	-0.3
	CREMA-D	**77.0**	71.3	+8.0
	Fluent Speech Commands	**98.1**	97.8	+0.3
	RAVDESS	**76.1**	68.5	+11.1
	Vocal Imitation	**31.2**	29.3	+6.5
	VocalSound	**93.2**	91.5	+1.9
	VoxCeleb1	**73.3**	24.8	+195.6
Sound	ASV2015	**99.3**	97.9	+1.4
	Clotho	**5.8**	3.1	+87.1
	DESED	**53.7**	22.6	+137.6
	ESC-50	**94.3**	62.5	+50.9
	FSD50k	**55.5**	32.0	+73.4
	FSD18-Kaggle	**82.2**	49.6	+65.7
	UrbanSound 8k	**87.9**	75.7	+16.1
Music	Free Music Archive Small	**67.2**	58.9	+14.1
	GTZAN Genre	**88.6**	71.8	+23.4
	MAESTRO	**54.5**	0.0	$+\infty$
	NSynth-Instruments	**72.2**	63.5	+13.7

MiDashengLM - Framework

- Dasheng 0.6B as pretrained AudioEncoder
- Audio-text alignment with general audio captions
- Standard Pretraining + SFT training regime
- 5 Hz features
- LoRA finetuning







Parameter	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
Encoder	Dasheng-based	Whisper-based	Whisper-based
Decoder Parameters ↓	**7B**	7B	7B
Audio-token framerate ↓	**5 Hz**	25 Hz	12.5 Hz
Audio-text alignment	General caption	ASR	ASR
Capable of ASR ?	✓	✓	✓
Known pretraining data ?	✓	✗	✗

MiDashengLM - Results Caption

Domain	Dataset	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
Music	MusicCaps	**47.15**	43.71	35.43
	Songdescriber	**49.97**	45.31	44.63
Sound	AudioCaps	**63.81**	60.79	49.00
	ClothoV2	**51.12**	47.55	48.01
	AutoACD	**67.08**	55.93	44.76

MiDashengLM - Results MECAT

Task	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
Content Long	**56.60**	48.34	40.83
Content Short	**58.40**	45.29	45.72
Pure Speech	**40.79**	37.27	25.57
Pure Sound	**53.47**	46.60	35.75
Pure Music	**59.29**	50.68	39.54
Mixed Speech	**42.19**	37.43	27.12
Mixed Sound	**38.59**	34.07	19.44
Mixed Music	**48.01**	34.71	16.18
Environment	**54.32**	47.84	16.66
Overall	**52.00**	43.80	36.32

MiDashengLM - Results Paralinguistic

Dataset	Metric	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
VoxCeleb1	ACC ↑	**97.05**	59.71	82.72
VoxLingua107		**92.79**	51.03	73.65
VoxCeleb-Gender		97.58	99.82	**99.69**
VGGSound		**53.46**	0.97	2.20
Cochlscene		**78.31**	23.88	18.34
NSynth		**82.32**	60.45	38.09
FMA		66.07	**66.77**	27.91
FSDKaggle2018	mAP ↑	**77.31**	31.38	24.75
AudioSet		**7.50**	6.48	3.47
FSD50K		**39.02**	23.87	27.23

MiDashengLM - Results ASR

Dataset	Language	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
LibriSpeech test-clean	English	5.6	1.7	**1.3**
LibriSpeech test-other		8.2	3.4	**2.4**
People's Speech		25.2	28.6	**22.3**
AISHELL2 Mic	Chinese	3.5	**2.5**	2.7
AISHELL2 iOS		3.5	**2.6**	**2.6**
AISHELL2 Android		3.6	2.7	**2.6**
GigaSpeech 2	Indonesian	**19.8**	21.2	>100
	Thai	**36.8**	53.8	>100
	Viet	**17.1**	18.6	>100

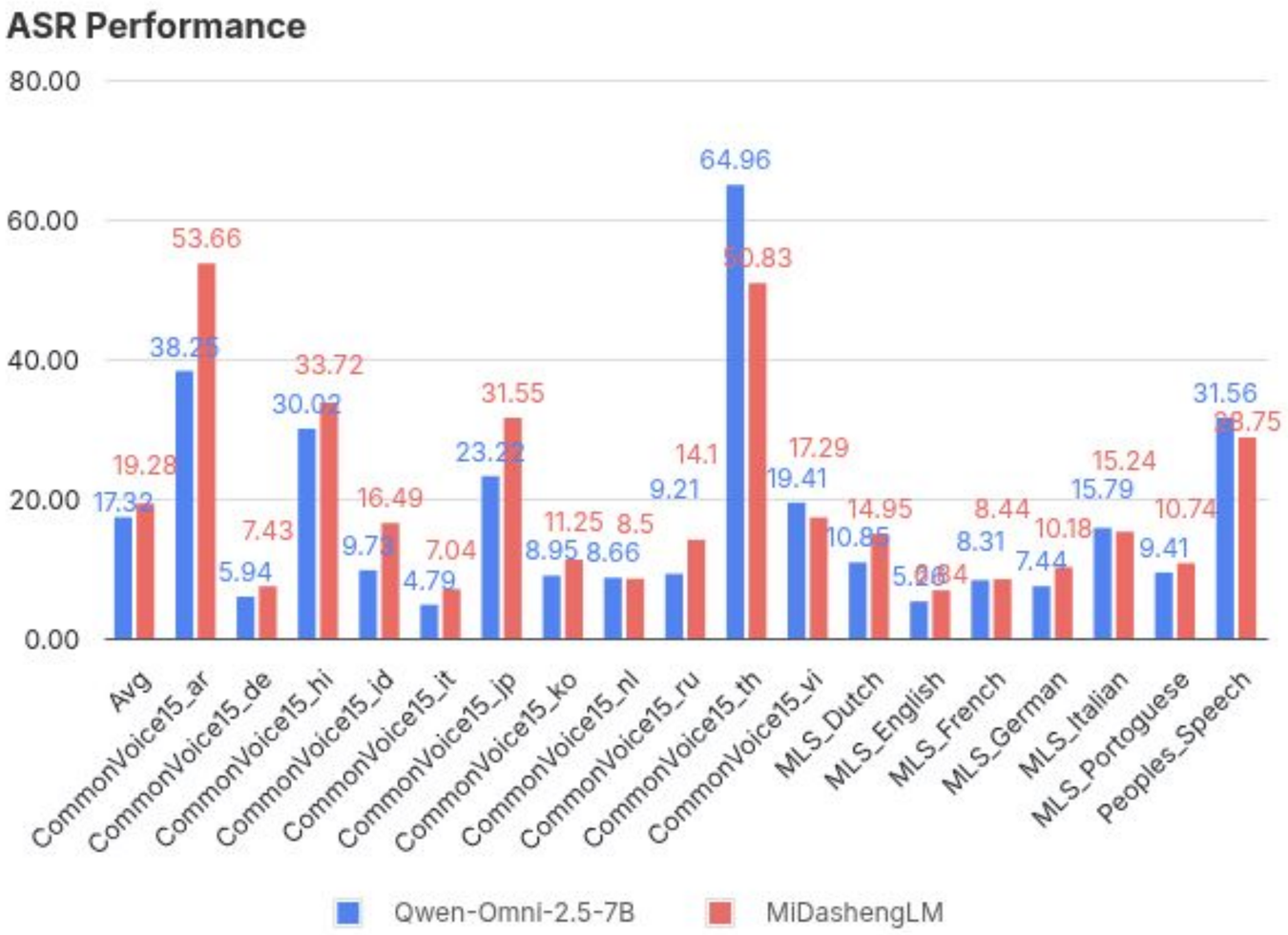
MiDashengLM - Results QA

Dataset	Subset	Metric	MiDashengLM 7B	Qwen2.5-Omni 7B	Kimi-Audio-Instruct 7B
MuChoMusic [55]		ACC ↑	**69.84**	64.79	67.40
MMAU [54]	Sound	ACC ↑	70.57	67.87	**74.17**
	Music		**70.96**	69.16	61.08
	Speech		59.46	**59.76**	57.66
	Average		**67.00**	65.60	64.30
MusicQA [56]		FENSE ↑	**61.18**	60.60	40.00
AudioCaps-QA [57]			**55.21**	53.28	47.34

MiDashengLM - Speedup

Batch Size	MiDashengLM 7B	Qwen2.5-Omni 7B	Speedup
1	0.65	0.45	1.4×
4	2.42	1.21	2.0×
8	4.67	1.44	3.2×
16	8.93		6.2×
32	14.36		10.0×
64	19.54	OOM	13.6×
128	24.26		16.8×
512	29.04		20.2×

MiDashengLM - ML ASR



MiDashengLM - Surprising Results

- Audio - text alignment with general audio captions greatly outperforms ASR (encoder)
- 5 Hz Audio features are workable
- 6.25 / 12.5 / 25 Hz features are much worse in performance
- 5 Hz also works for fast speech
- Much better performance can be achieved by fully finetuning

你看他非让我点确定你说他贱不见啊我说没见过这么贱的人啊没见过这么贱的游戏他强迫你玩别的啊他强迫你玩这个你不玩还不行

Material

X-Ares (ICME) https://github.com/jimbozhang/xares

Dasheng-Denoiser https://github.com/xiaomi-research/dasheng-denoiser

GLAP https://github.com/xiaomi-research/dasheng-glap

Dasheng https://github.com/XiaoMi/dasheng

MECAT https://github.com/xiaomi-research/mecat

MiDashengLM https://github.com/xiaomi-research/dasheng-lm


小米集团技术委员会
MITC
XIAOMI TECHNICAL COMMITTEE