Inteligencia Artificial Generativa

Let's talk about hype stuff

Valentin Barriere // Clemente Henriquez

Universidad de Chile – DCC

Diplomado de Postítulo en Inteligencia Artificial, Primavera 2025

Audio Models

Audio Specificities

- Speech inputs have a variable number of lexical units per sequence.
- Speech is a long sequence that doesn't have segment boundaries.
- Speech is continuous without a predefined dictionary of units to explicitly model in the self-supervised setting.
- Speech processing tasks might require orthogonal information, e.g., ASR and Speaker ID.

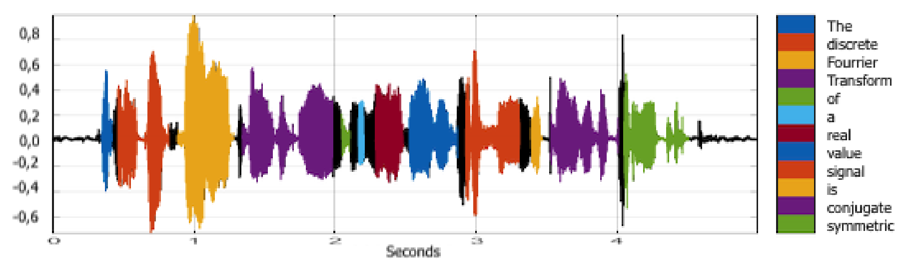
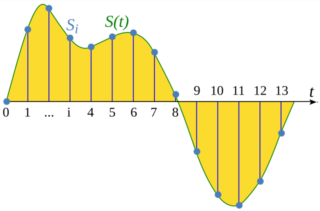


Figure 1: Speech is continuous while text is discrete

Audio Specificities

What is Audio Data?

- Sound is a continuous wave — computers store it as a series of numbers (samples).
- **Sampling rate** defines how many times per second we capture the signal.
- The resulting array of values forms a **waveform**.
- Each point represents amplitude — how "loud" the sound is at a given instant.
- This digital representation allows AI models to **analyze, generate, or understand** sound.



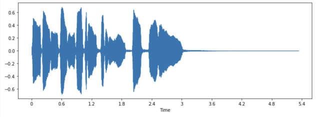
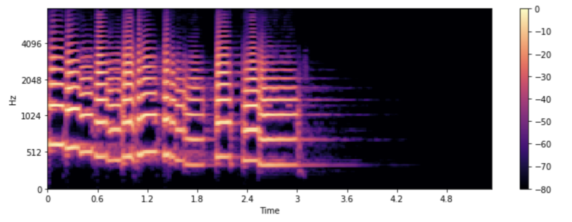


Figure 3: Waveform: time vs amplitude

Spectrogram: Classical way to understand Sounds

Why process audio?

- Models can't interpret raw sound directly — we convert it to **features**.
- The most common view: the **spectrogram** — time on one axis, frequency on the other.
- A **mel spectrogram** reshapes frequencies to match human hearing.
- These representations make speech, music, and environmental sounds measurable and learnable.



Classical Audio Tasks

Task	Input	Output	Description
Audio Classification	Audio	Label	Classify sounds, music genres, environmental sounds
Audio Speech Recognition	Speech	Text	Convert spoken language to written text
Speaker Diarization	Audio	Segments + IDs	"Who spoke when?" - identify speakers over time
Speaker Identification	Audio	Speaker ID	Identify which person is speaking
Text-To-Speech	Text	Speech	Convert written text to spoken audio
Voice Conversion	Audio + Target	Audio	Change voice characteristics (speaker, emotion)
Music Generation	Text/Audio	Audio	Generate music from prompts or continuations
Audio Enhancement	Noisy audio	Clean audio	Remove noise, improve quality

Two Main Paradigms

Understanding (Audio $\rightarrow$ Information): Classification, ASR, Diarization, Identification
Generation (Information $\rightarrow$ Audio): TTS, Music Generation, Voice Conversion

Model Input Approaches: Raw Audio vs Spectrograms

Raw Audio Input

- Direct waveform processing
- 1D temporal signal
- Sample rate: 16kHz - 48kHz
- Learning features directly from raw signal

Examples: HuBERT [13], wav2vec2 [3], wavLM [5], EnCodec [**?**], ...

Spectrogram Input

- Pre-computed frequency representation
- 2D time-frequency image
- Mel-scale or linear scale
- Leverages image processing techniques

Examples: Whisper [20], AST [11], CLAP [23], BYOL-A [12], ...

Trend

Modern models increasingly use **raw audio** for end-to-end learning, but spectrograms remain effective for many tasks

Outline : Representations

Raw Audio Models: Learning from the Waveform

Key Characteristics

- **Input**: Raw waveform (1D signal)
- **Processing**: CNN or 1D convolutions
- **Architecture**: Encoder learns representations directly

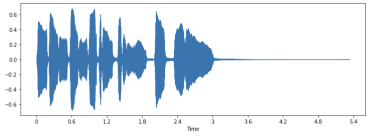


Figure 4: Raw audio waveform

Advantages:

Raw Audio Models: Learning from the Waveform

Key Characteristics

- **Input**: Raw waveform (1D signal)
- **Processing**: CNN or 1D convolutions
- **Architecture**: Encoder learns representations directly

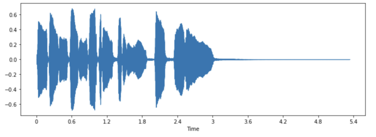


Figure 4: Raw audio waveform

Advantages:

- End-to-end learning

Raw Audio Models: Learning from the Waveform

Key Characteristics

- **Input**: Raw waveform (1D signal)
- **Processing**: CNN or 1D convolutions
- **Architecture**: Encoder learns representations directly

Advantages:

- End-to-end learning
- No information loss from feature extraction

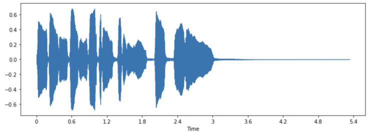


Figure 4: Raw audio waveform

Raw Audio Models: Learning from the Waveform

Key Characteristics

- **Input**: Raw waveform (1D signal)
- **Processing**: CNN or 1D convolutions
- **Architecture**: Encoder learns representations directly

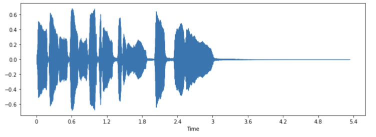


Figure 4: Raw audio waveform

Advantages:

- End-to-end learning
- No information loss from feature extraction
- Learns optimal features for the task

Raw Audio Models: Learning from the Waveform

Key Characteristics

- **Input**: Raw waveform (1D signal)
- **Processing**: CNN or 1D convolutions
- **Architecture**: Encoder learns representations directly

Advantages:

- End-to-end learning
- No information loss from feature extraction
- Learns optimal features for the task
- Works across different sampling rates

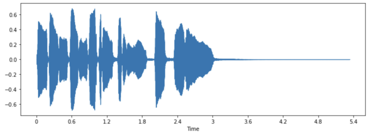


Figure 4: Raw audio waveform

Raw Audio Models: Learning from the Waveform

Key Characteristics

- **Input**: Raw waveform (1D signal)
- **Processing**: CNN or 1D convolutions
- **Architecture**: Encoder learns representations directly

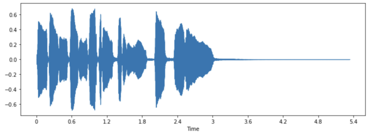


Figure 4: Raw audio waveform

Advantages:

- End-to-end learning
- No information loss from feature extraction
- Learns optimal features for the task
- Works across different sampling rates

Example: wav2vec 2.0

- 1D CNN feature encoder
- Converts 16kHz audio to latent representations
- Transformer processes these representations

Spectrogram-based Models: Visual Audio Representation

Key Characteristics

- **Input**: Mel-spectrogram (2D image)
- **Processing**: 2D CNN or Vision Transformers
- **Architecture**: Treats audio as an image

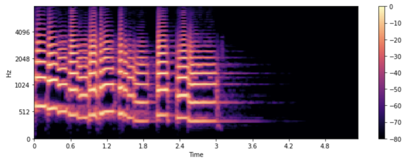


Figure 5: Mel-spectrogram

Advantages:

Spectrogram-based Models: Visual Audio Representation

Key Characteristics

- **Input**: Mel-spectrogram (2D image)
- **Processing**: 2D CNN or Vision Transformers
- **Architecture**: Treats audio as an image

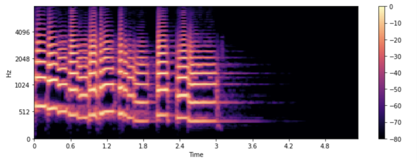


Figure 5: Mel-spectrogram

Advantages:

- Leverages CVis techniques

Spectrogram-based Models: Visual Audio Representation

Key Characteristics

- **Input**: Mel-spectrogram (2D image)
- **Processing**: 2D CNN or Vision Transformers
- **Architecture**: Treats audio as an image

Advantages:

- Leverages CVis techniques
- Interpretable time-frequency patterns

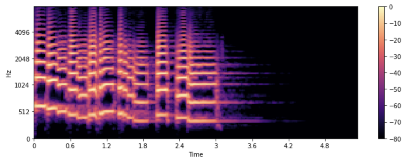


Figure 5: Mel-spectrogram

Spectrogram-based Models: Visual Audio Representation

Key Characteristics

- **Input**: Mel-spectrogram (2D image)
- **Processing**: 2D CNN or Vision Transformers
- **Architecture**: Treats audio as an image

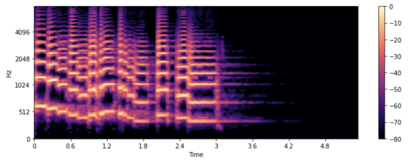


Figure 5: Mel-spectrogram

Advantages:

- Leverages CVis techniques
- Interpretable time-frequency patterns
- Efficient processing (compressed representation)

Spectrogram-based Models: Visual Audio Representation

Key Characteristics

- **Input**: Mel-spectrogram (2D image)
- **Processing**: 2D CNN or Vision Transformers
- **Architecture**: Treats audio as an image

Advantages:

- Leverages CVis techniques
- Interpretable time-frequency patterns
- Efficient processing (compressed representation)
- Transfer learning from vision

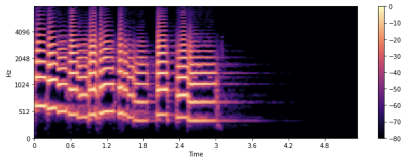


Figure 5: Mel-spectrogram

Spectrogram-based Models: Visual Audio Representation

Key Characteristics

- **Input**: Mel-spectrogram (2D image)
- **Processing**: 2D CNN or Vision Transformers
- **Architecture**: Treats audio as an image

Advantages:

- Leverages CVis techniques
- Interpretable time-frequency patterns
- Efficient processing (compressed representation)
- Transfer learning from vision

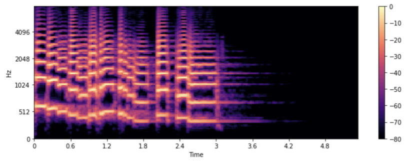


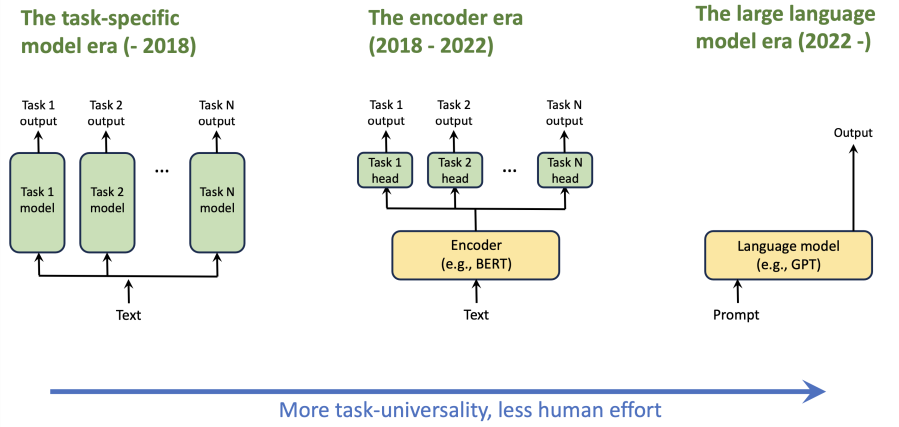
Figure 5: Mel-spectrogram

Example: Whisper

- Log-mel spectrogram (80 channels)
- 2D convolutions for feature extraction
- Encoder-decoder Transformer architecture

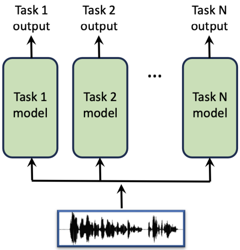
Outline : Speech Models

Evolution of text and speech foundation models

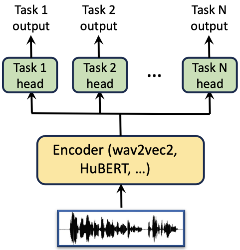


Evolution of text and speech foundation models

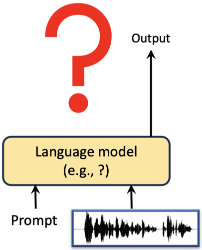
The task-specific model era (- 2020)



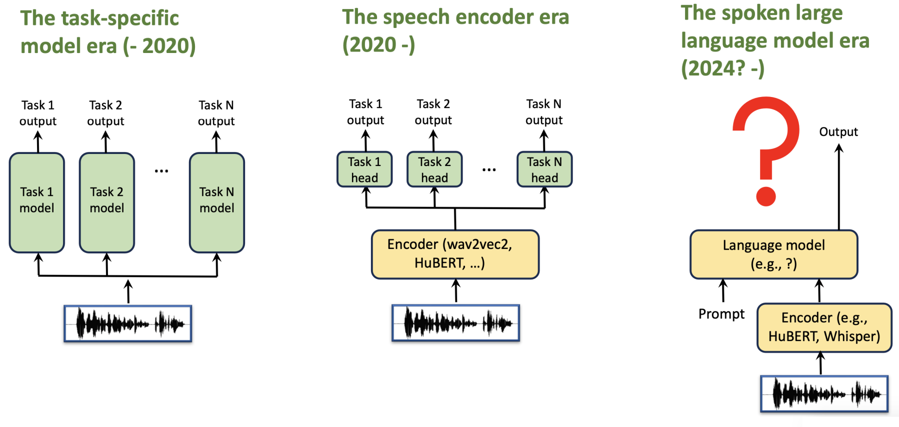
The speech encoder era (2020 -)



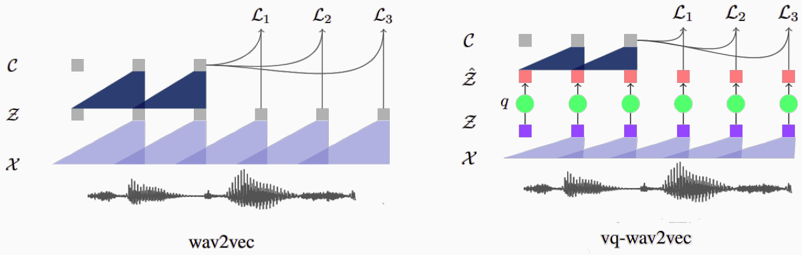
The spoken large language model era (2024? -)



Evolution of text and speech foundation models



vq-wav2vec [2]



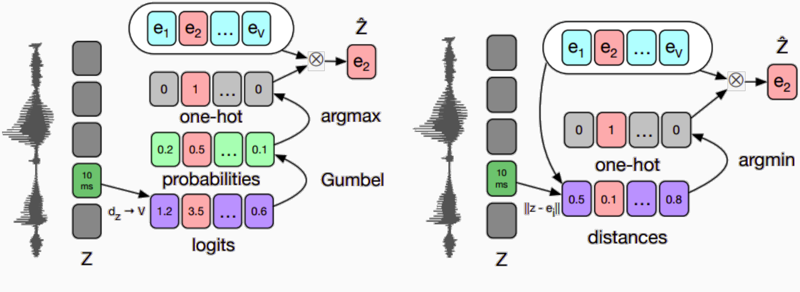
- Same as wav2vec, but processing a quantization of the hidden state.
- Using the InfoNCE loss (Contrastive Predicting Coding) such as wav2vec and word2vec

vq-wav2vec [2] Hidden states quantization

The quantization module replaces the original continuous representation $\mathcal{Z}$ by a fixed-size discrete representation $\hat{\mathcal{Z}} = e_i$ where **code-book** $e \in \mathbb{R}^{V \times d}$ contains V representations of size d.

vq-wav2vec [2] Hidden states quantization

The quantization module replaces the original continuous representation $\mathcal{Z}$ by a fixed-size discrete representation $\hat{\mathcal{Z}} = e_i$ where **code-book** $e \in \mathbb{R}^{V \times d}$ contains V representations of size d.

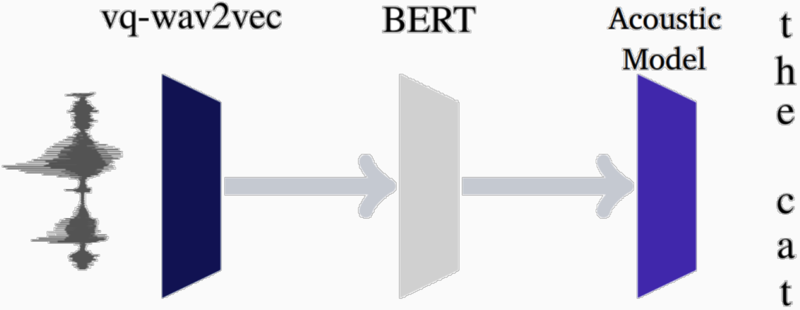


(a) Gumbel-Softmax (b) K-means clustering.

Two techniques to pass from dense to quantized vectors:

- **Gumbel-Softmax**: a differentiable approximation of the arg max for computing one-hot representations
- **K-means**: Similar as VQ-VAE [21]: $||sg(z)\hat{z}||^2 + \gamma * ||zsg(\hat{z})||^2$

vq-wav2vec [2]



Quantization **makes it possible to pre-train transformer using a BERT-like architecture** and MLM objective using the quantized values.

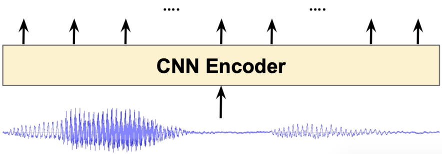
More info [in this blogpost](#).

wav2vec 2.0: [3]

- Predict masked speech frames

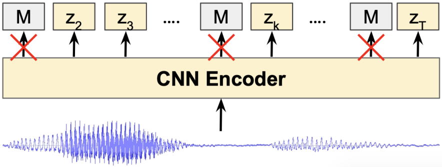
wav2vec 2.0: [3]

- Predict masked speech frames



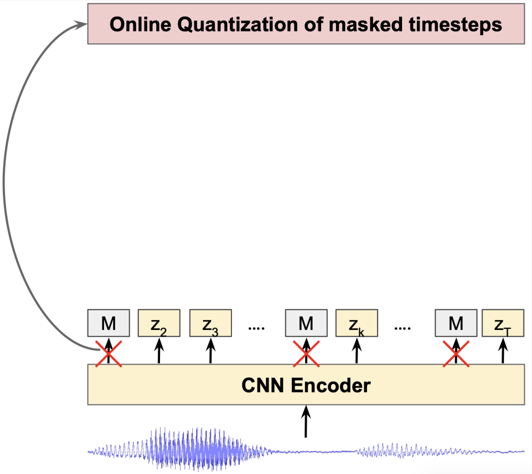
wav2vec 2.0: [3]

- Predict masked speech frames



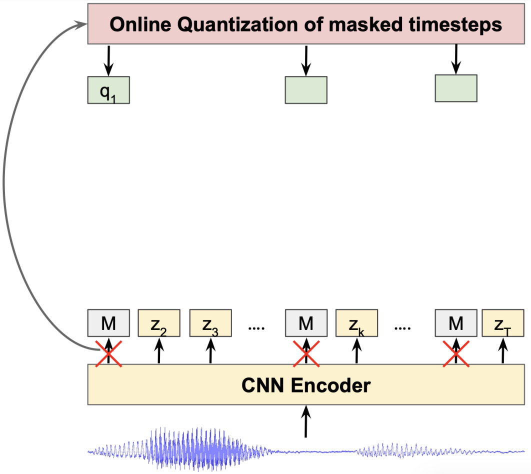
wav2vec 2.0: [3]

- Predict masked speech frames



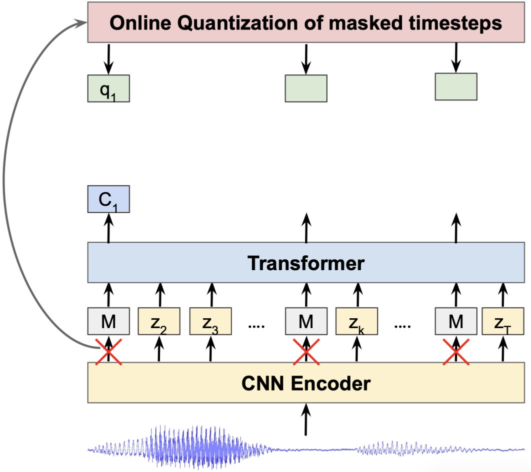
wav2vec 2.0: [3]

- Predict masked speech frames



wav2vec 2.0: [3]

- Predict masked speech frames
- **Contrastive Loss**: Predicted frame representations should be similar to quantized input features at the same frame
- ...and different from inputs at different frames

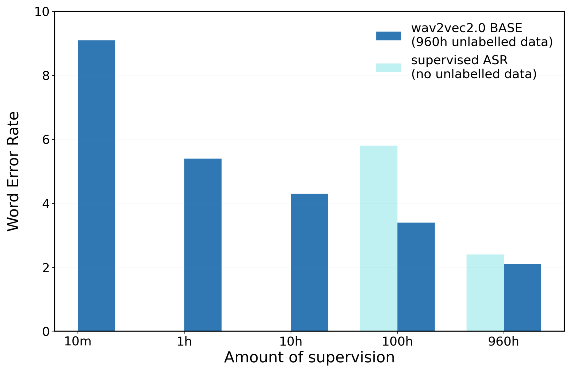


$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\mathbf{q}\sim\mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \mathbf{q})/\kappa)}$$

wav2vec 2.0: Results

First major improvements on ASR using self- supervised learning

- Improved performance and labeled data efficiency on the LibriSpeech benchmark
- Matches a supervised model using only 1% of the labeled data (100 hours $\rightarrow$ 1 hour)

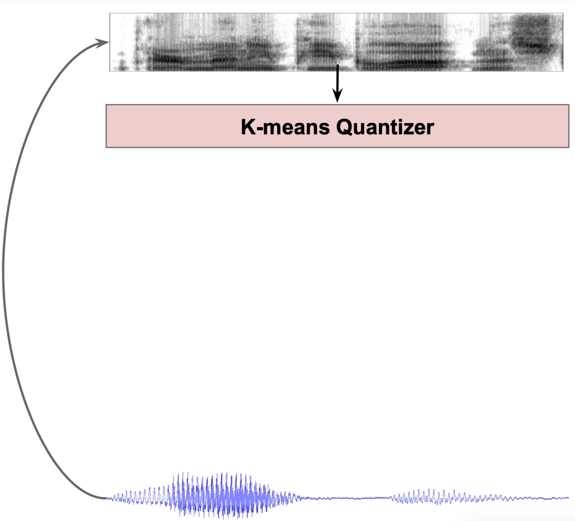


HuBERT: Hidden-unit BERT [13]

- A simple method to apply BERT style representation learning for speech.
- Matched or beat the SOTA on ASR while being the best for many speech tasks.
- With its high-quality discrete units, HuBERT facilitated Textless NLP research.

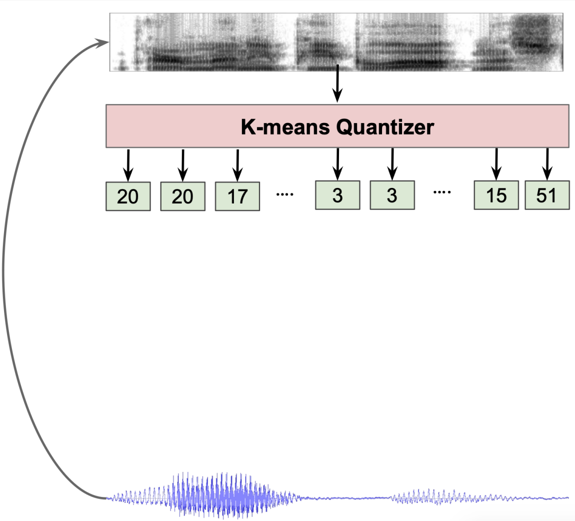
HuBERT: Hidden-unit BERT [13]

HuBERT: Hidden-unit BERT [13]



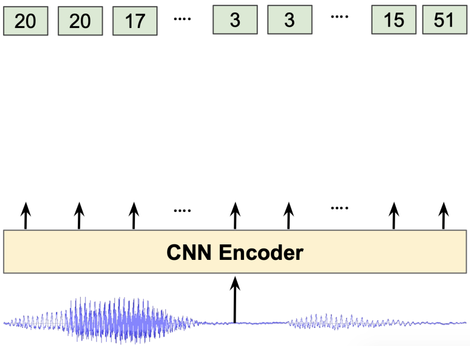
HuBERT: Hidden-unit BERT [13]

- The K-means quantizer produces frame-level labels.



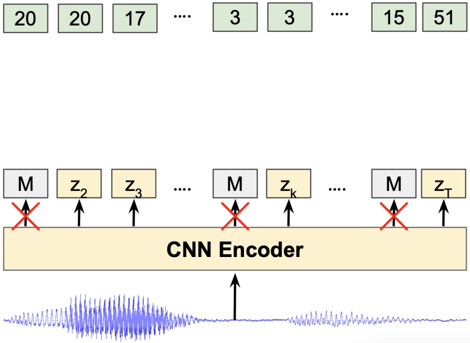
HuBERT: Hidden-unit BERT [13]

- The K-means quantizer produces frame-level labels.



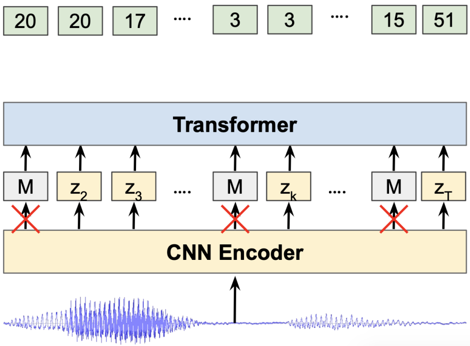
HuBERT: Hidden-unit BERT [13]

- The K-means quantizer produces frame-level labels.



HuBERT: Hidden-unit BERT [13]

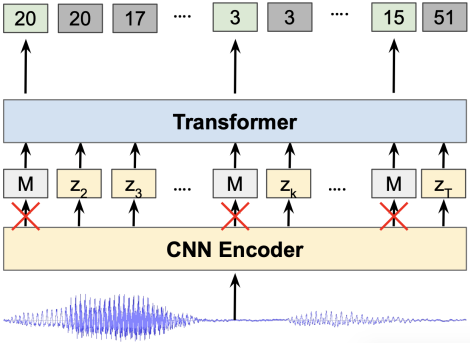
- The K-means quantizer produces frame-level labels.



HuBERT: Hidden-unit BERT [13]

- The K-means quantizer produces frame-level labels.
- BERT-like masked prediction loss:

$$\mathcal{L}_m = \sum_{t \in M} - \log p(y_t | X)$$

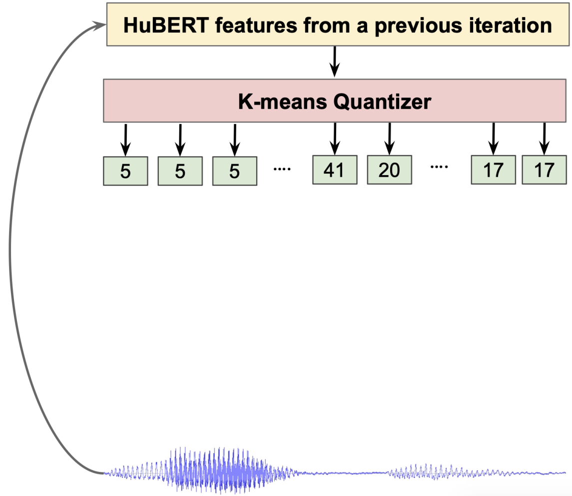


HuBERT: Hidden-unit BERT [13]

- The K-means quantizer produces frame-level labels.
- BERT-like masked prediction loss:

$\mathcal{L}_m = \sum_{t \in M} - \log p(y_t | X)$

- First iteration uses quantized spectrogram, then HuBERT features from previous iteration

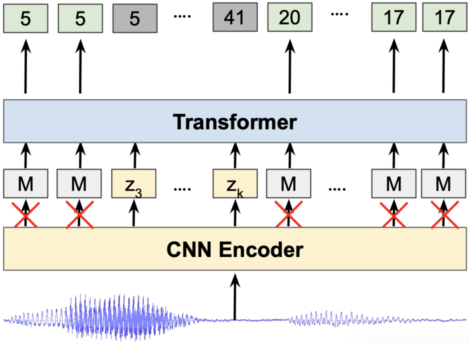


HuBERT: Hidden-unit BERT [13]

- The K-means quantizer produces frame-level labels.
- BERT-like masked prediction loss:

$$\mathcal{L}_m = \sum_{t \in M} - \log p(y_t | X)$$

- First iteration uses quantized spectrogram, then HuBERT features from previous iteration

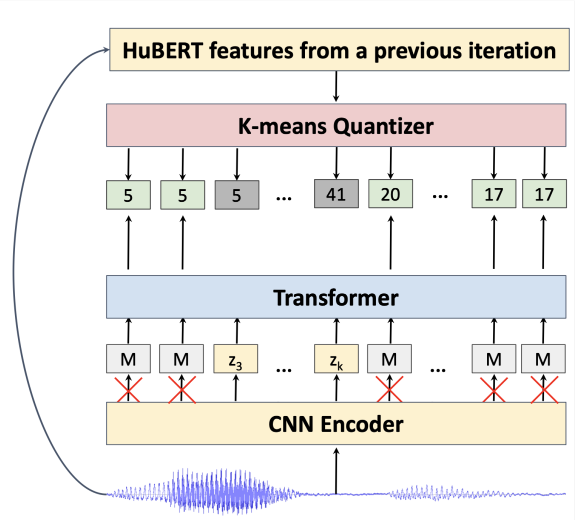


HuBERT: Hidden-unit BERT [13]

- The K-means quantizer produces frame-level labels.
- BERT-like masked prediction loss:

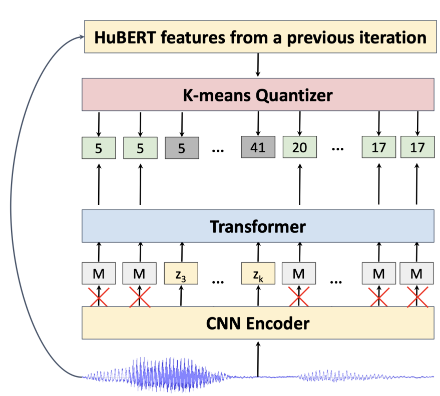
$\mathcal{L}_m = \sum_{t \in M} - \log p(y_t | X)$

- First iteration uses quantized spectrogram, then HuBERT features from previous iteration



HuBERT: Hidden-unit BERT [13]

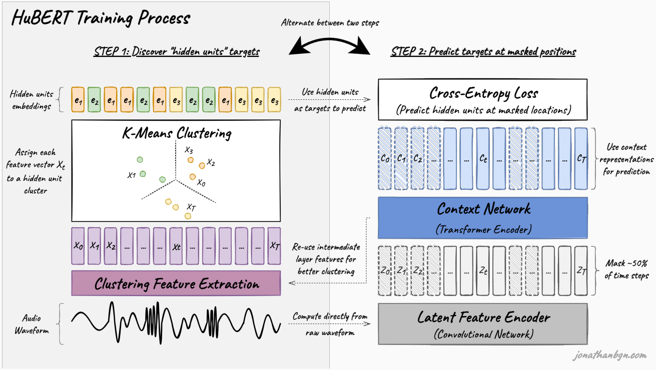
- A small codebook size, e.g., 50, 100, is used for the initial training iteration to focus on phonetic differences rather than speaker and style.
- Layer 6 for iteration 1, layer 9 for iteration 2 used for the clustering steps. They found empirically to contain higher quality features over many speech tasks.



Results

- Matched or beat the SOTA on ASR
- Best representations for multiple downstream tasks: ASR, Speaker Diarization, Keyword Detection, etc...

HuBERT: A Visual Explanation



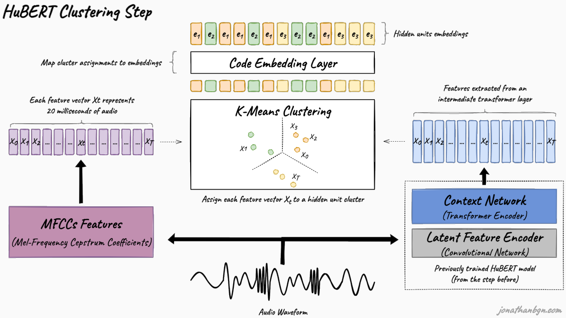
Key Innovation

Learning meaningful speech representations **without labeled data**

- **Problem**: Speech is continuous, not discrete like text
- **Solution**: Create discrete units through clustering
- **Training**: Use BERT-style masked prediction
- **Result**: Rich representations for any speech task

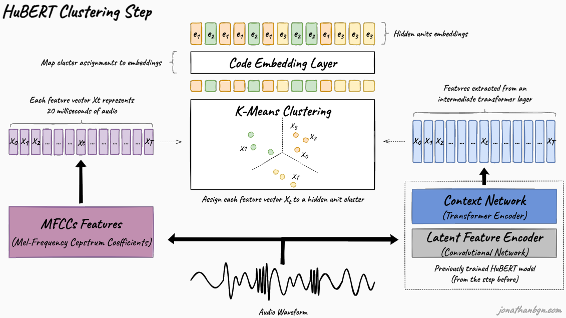
HuBERT Step 1: Clustering Speech Segments

- Audio divided into **25ms segments**



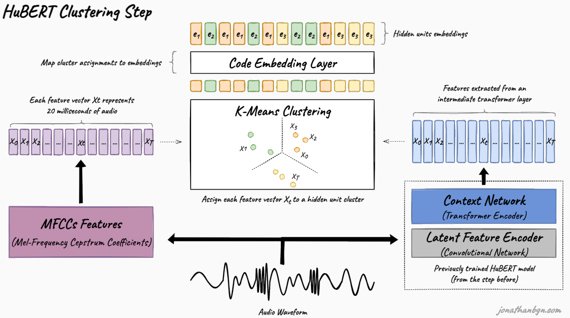
HuBERT Step 1: Clustering Speech Segments

- Audio divided into **25ms segments**
- Extract MFCC features from each segment



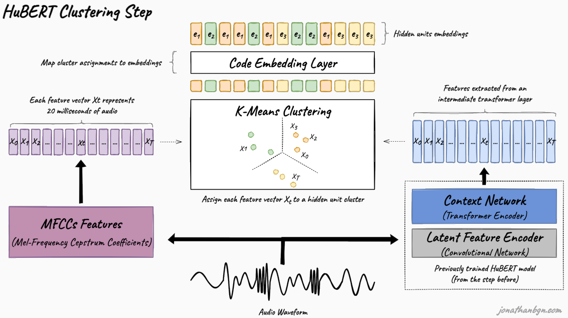
HuBERT Step 1: Clustering Speech Segments

- Audio divided into **25ms segments**
- Extract MFCC features from each segment
- **K-means clustering** groups similar segments



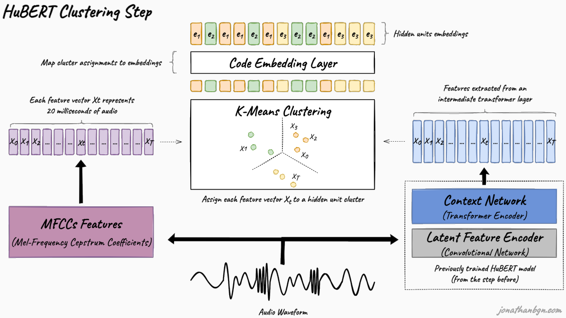
HuBERT Step 1: Clustering Speech Segments

- Audio divided into **25ms segments**
- Extract MFCC features from each segment
- **K-means clustering** groups similar segments
- Each segment assigned a **cluster ID**



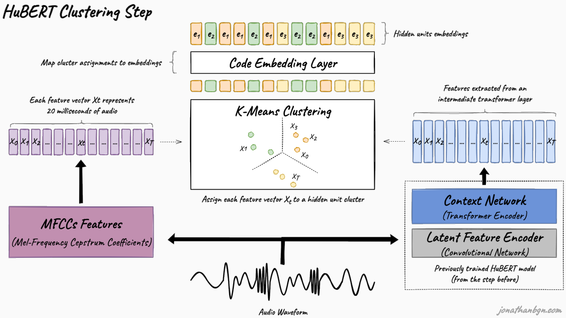
HuBERT Step 1: Clustering Speech Segments

- Audio divided into **25ms segments**
- Extract MFCC features from each segment
- **K-means clustering** groups similar segments
- Each segment assigned a **cluster ID**
- These IDs become "pseudo-labels" for training



HuBERT Step 1: Clustering Speech Segments

- Audio divided into **25ms segments**
- Extract MFCC features from each segment
- **K-means clustering** groups similar segments
- Each segment assigned a **cluster ID**
- These IDs become "pseudo-labels" for training

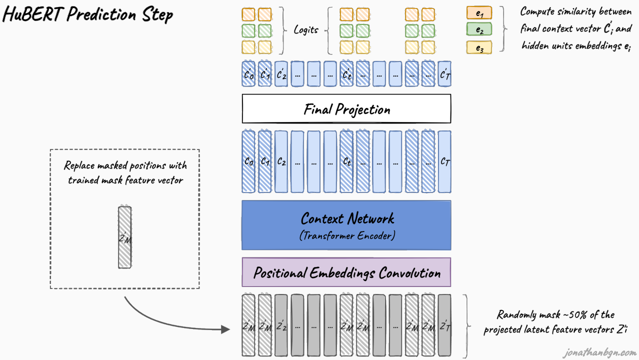


Why Clustering?

Creates discrete targets from continuous audio, enabling BERT-style training

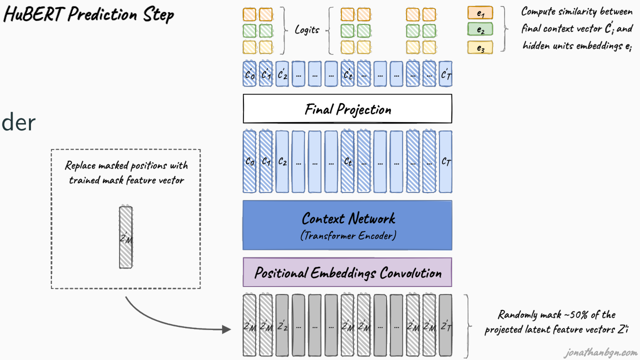
HuBERT Step 2: Masked Prediction Training

- Randomly **mask** ∼**50%** of audio segments



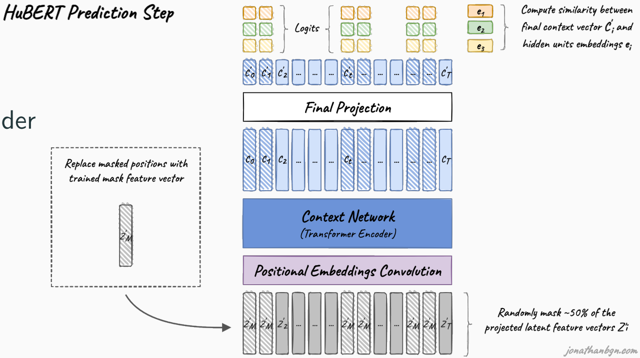
HuBERT Step 2: Masked Prediction Training

- Randomly **mask** ~**50%** of audio segments
- Transformer encoder processes the sequence



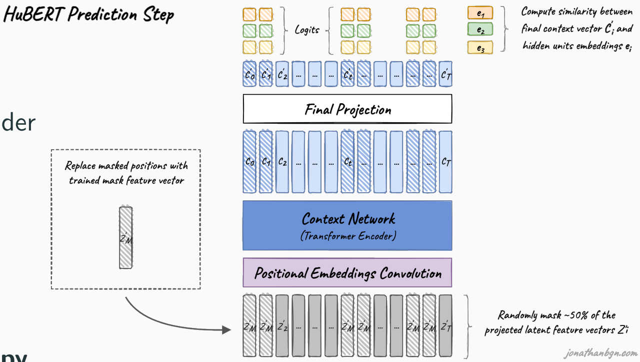
HuBERT Step 2: Masked Prediction Training

- Randomly **mask** ∼**50%** of audio segments
- Transformer encoder processes the sequence
- Model predicts **cluster IDs** of masked segments



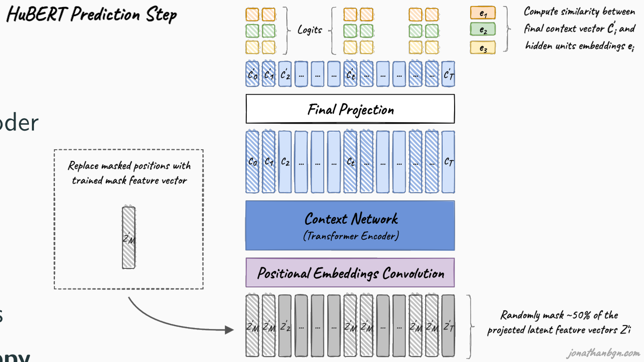
HuBERT Step 2: Masked Prediction Training

- Randomly **mask** ~**50%** of audio segments
- Transformer encoder processes the sequence
- Model predicts **cluster IDs** of masked segments
- Uses **cross-entropy loss**



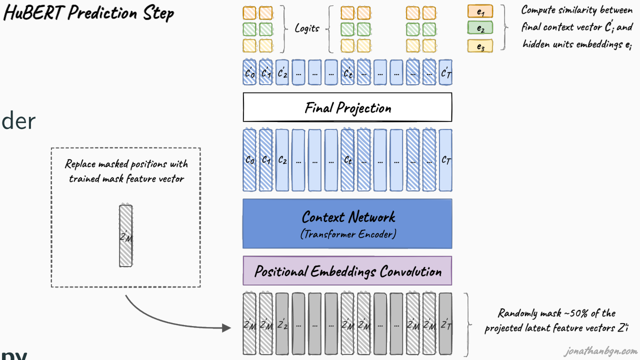
HuBERT Step 2: Masked Prediction Training

- Randomly **mask** ~**50%** of audio segments
- Transformer encoder processes the sequence
- Model predicts **cluster IDs** of masked segments
- Uses **cross-entropy loss**
- Model learns contextualized representations



HuBERT Step 2: Masked Prediction Training

- Randomly **mask** ~**50%** of audio segments
- Transformer encoder processes the sequence
- Model predicts **cluster IDs** of masked segments
- Uses **cross-entropy loss**
- Model learns contextualized representations



Key Difference from wav2vec 2.0

Simpler loss: predict discrete targets instead of contrastive learning

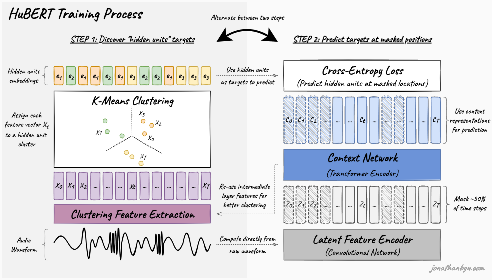
HuBERT: Iterative Refinement

Iteration 1

- Cluster using **MFCC features**
- Train model with these labels
- Extract features from **layer 6**

Iteration 2

- Re-cluster using **layer 6 features**
- Train new model with refined labels
- Extract features from **layer 9**



Why Iterate?

Each iteration produces **higher quality** features that capture more semantic information

Credit: Jonathan Bgn

HuBERT vs wav2vec 2.0: Key Differences

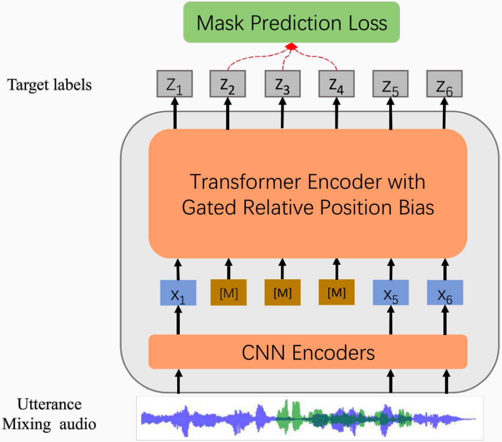
Aspect	wav2vec 2.0	HuBERT
Target	Quantized latents	Cluster IDs
Loss	Contrastive	Cross-entropy
Training	Single pass	Iterative refinement
Complexity	Higher	Simpler
Performance	Excellent	**Better on most tasks**

HuBERT Advantages

- Simpler training objective (no negative sampling)
- Better transfer to non-ASR tasks (speaker ID, emotion, etc.)
- More stable training
- Iterative refinement improves quality

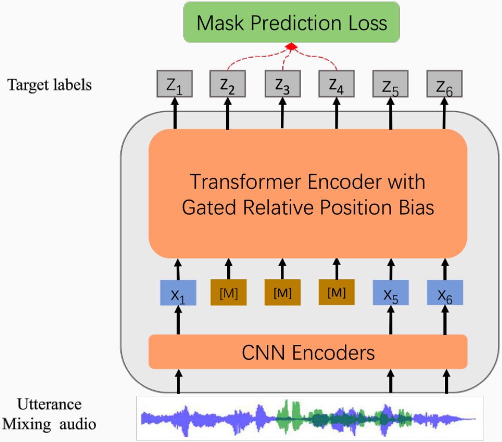
wavLM [5]

- Same as HuBERT but with Noise



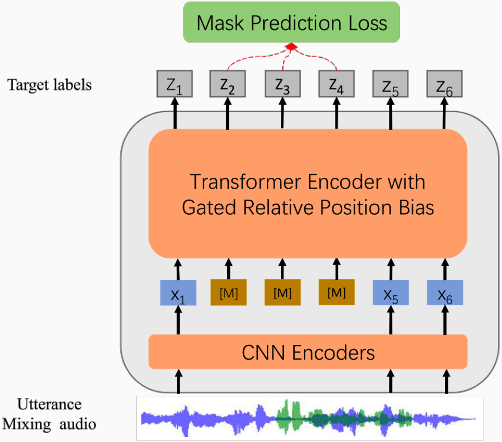
wavLM [5]

- Same as HuBERT but with Noise
- Model needs to find the representation of the original audio

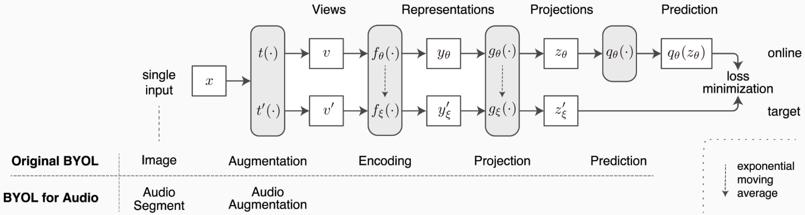


wavLM [5]

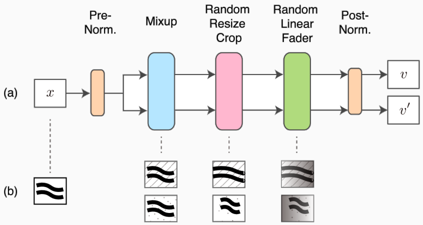
- Same as HuBERT but with Noise
- Model needs to find the representation of the original audio
- Allows extending pre-trained speech models to non-ASR tasks: models information needed for speaker identification, separation, or diarization



Boostrap Your Own Latent - Audio: BYOL-A [12]



- Same principle as BYOL [12] but apply the augmentation on the spectrograms
- Simple CNN
- Obtain one vector per sound
- Very good for general sounds:
    - Sound Event Recognition
    - Non Semantic Speech
    - Music tasks



Audio Masked Auto-Encoder: AudioMAE [16]

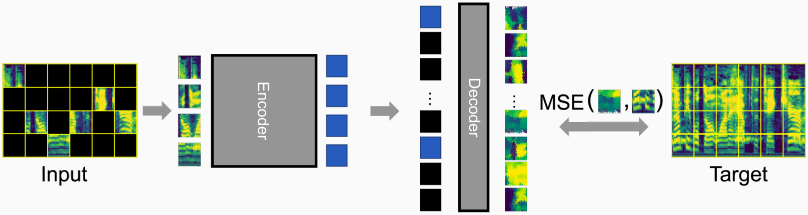
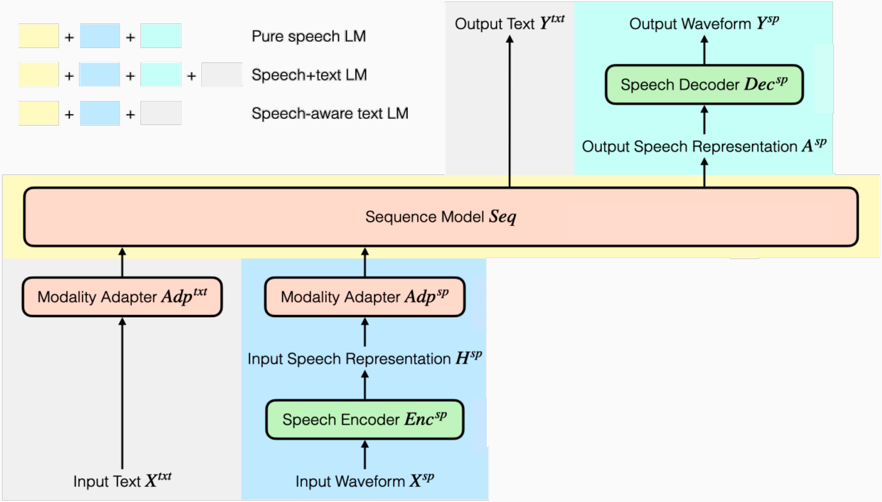


Figure 6: As simple as it sounds

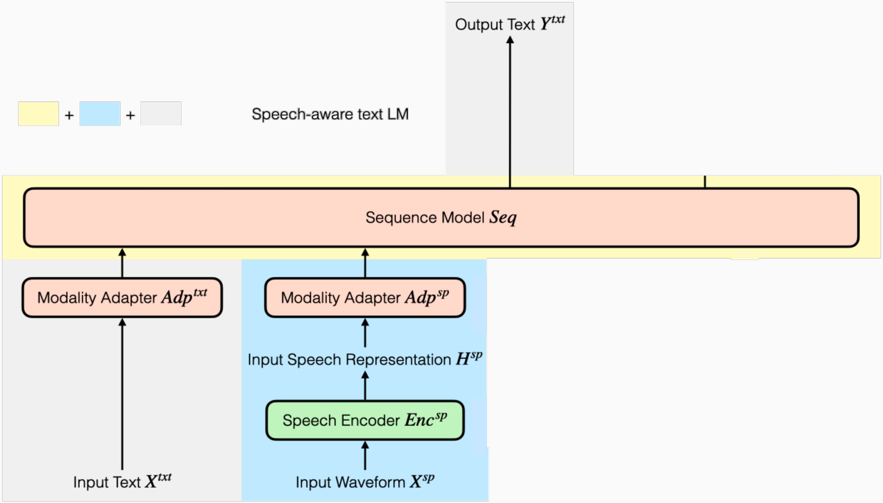
- Self-supervised learning
- Spectrogram is split into patches
- Mask 80% of the patches
- Restore the input, minimizing MSE on the masked portion
- Use a ViT as backbone [8]

Outline : Speech LLMs

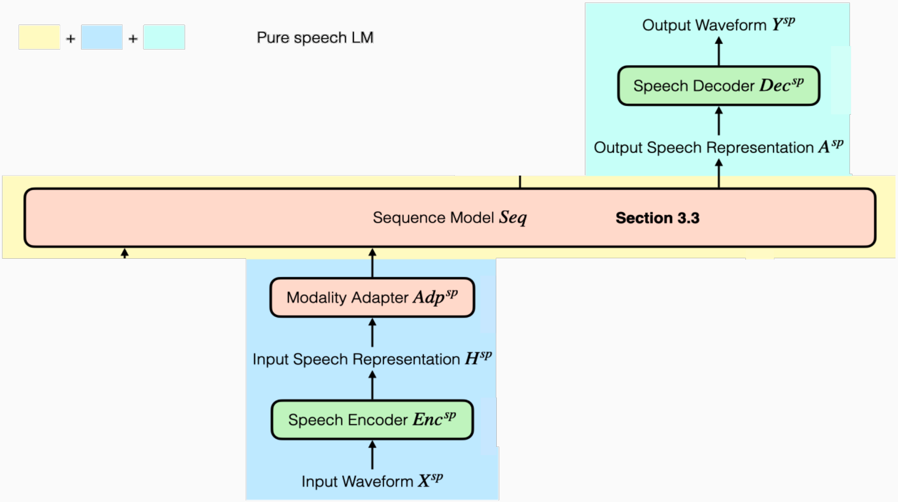
Speech LLMs [1]



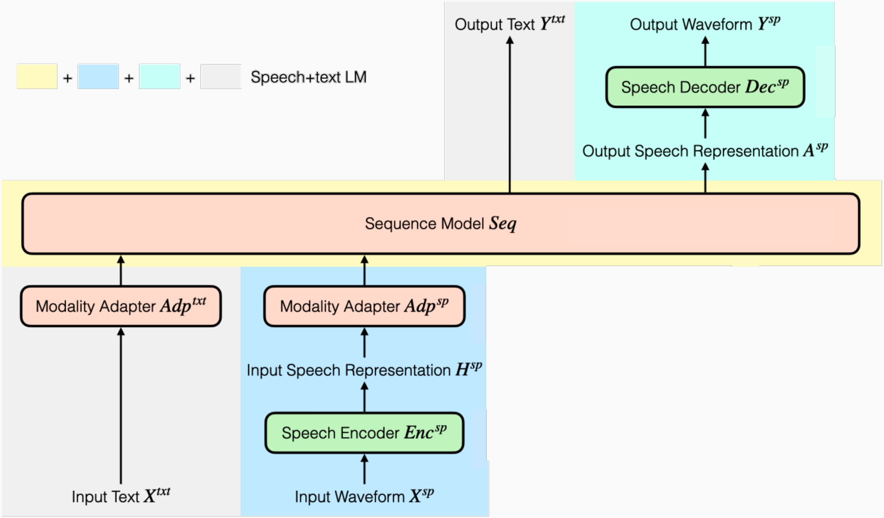
Speech LLMs [1]



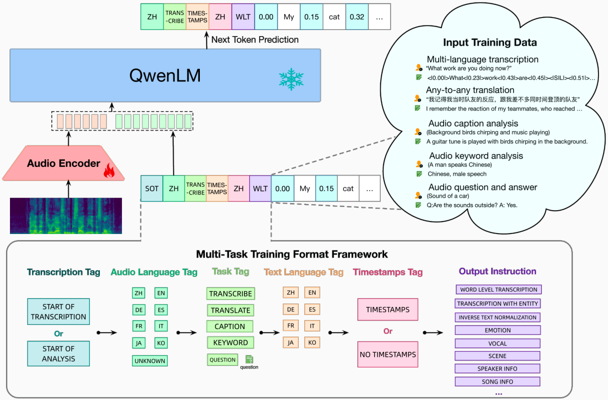
Speech LLMs [1]



Speech LLMs [1]



Qwen(2)-Audio [7, 6]



Types	Task	Description	Hours
Speech	ASR	Automatic speech recognition (multiple languages)	30k
	S2TT	Speech-to-text translation	3.7k
	OSR	Overlapped speech recognition	<1k
	Dialect ASR	Automatic dialect speech recognition	2k
	SRWT	English speech recognition with word-level timestamps	10k
		Mandarin speech recognition with word-level timestamps	11k
	DID	Dialect identification	2k
	LID	Spoken language identification	11.7k
	SGC	Speaker gender recognition (biologically)	4.8k
	ER	Emotion recognition	<1k
	SV	Speaker verification	1.2k
	SD	Speaker diarization	<1k
	SER	Speech entity recognition	<1k
	KS	Keyword spotting	<1k
	IC	Intent classification	<1k
	SF	Slot filling	<1k
	SAP	Speaker age prediction	4.8k
	VSC	Vocal sound classification	<1k
Sound	AAC	Automatic audio caption	8.4k
	SEC	Sound event classification	5.4k
	ASC	Acoustic scene classification	<1k
	SED	Sound event detection with timestamps	<1k
	AQA	Audio question answering	<1k
Music&Song	SID	Singer identification	<1k
	SMER	Singer and music emotion recognition	<1k
	MC	Music caption	25k
	MIC	Music instruments classification	<1k
	MNA	Music note analysis such as pitch, velocity	<1k
	MGR	Music genre recognition	9.5k
	MR	Music recognition	<1k
	MQA	Music question answering	<1k

- Use the embeddings from Whisperv2/3-large [20]
- LLM pre-trained weights from Qwen-7B [4]
- Freeze LLM and optimize audio encoder: `Qwen-Audio`, then freeze the audio encoder and train the LLM: `Qwen-Audio-Chat`

Audio Flamingo [17, 10]

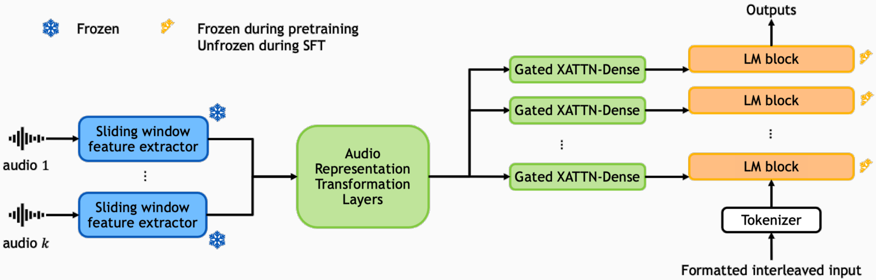


Figure 7: Interleaved audio and text as input and outputs free-form text.

- ClapCap as audio feature extractor (7s clips) and merge them with small transformer [9]
- **Pre-train**: learn the audio representation transformation layers and the gated xattn-dense layers → obtain a good set of initialization weights for these layers
- **Fine-Tune**: unfreeze the entire LM, and train all modules (except ClapCap)

SUPERB [24]

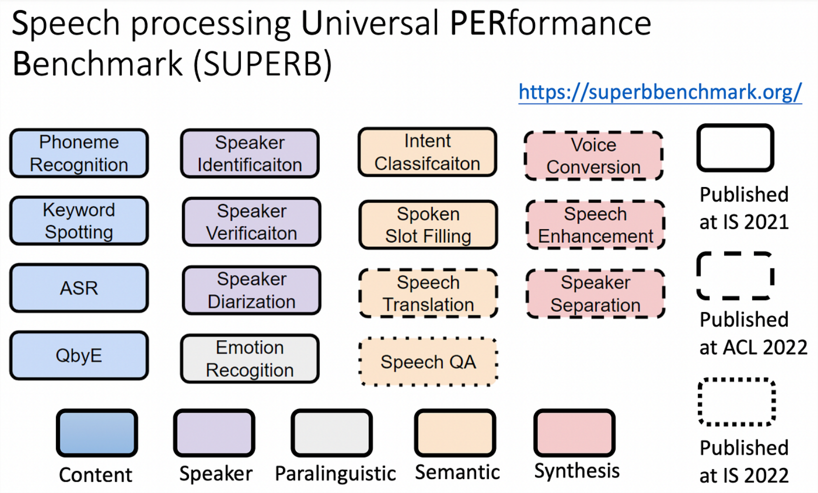
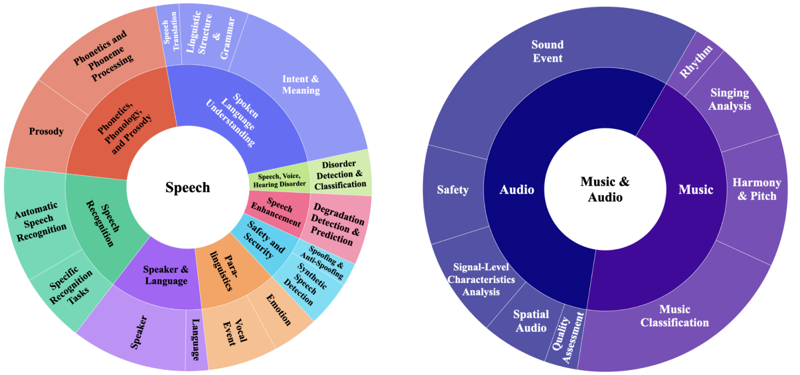


Figure 8: Models are assessed through a variety of downstream tasks. Such as NLP models on GLUE [22]

Dynamic-SUPERB [15, 14]



(a) Task Taxonomy for speech tasks.

(b) Task Taxonomy for audio and music tasks.

Figure 9: Dynamic-SUPERB is an evolving collection of 180 speech + audio "understanding" tasks (audio in, text out).

- Instruction + audio input → text output
- Evaluated with LLM-as-judge

Outline : Pre-training datasets

Speech Dataset

LibriSpeech

- **LibriSpeech:** 1,000 hours of speech from audio books.
- **VoxCeleb:** Speaker recognition dataset with diverse voices: 7k+ hours in-the-wild conditions.
- **AudioSet:** Over 2 million labeled audio clips from 600+ classes

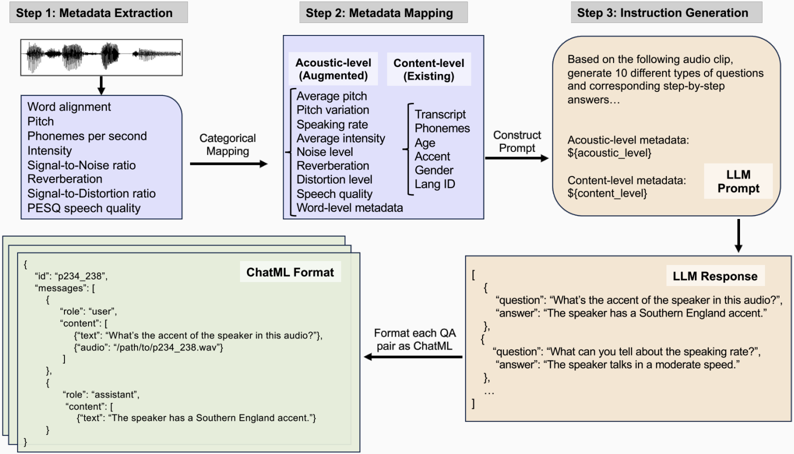
Speech and Audio Pre-training Datasets

Dataset	Size	Type	Key Features
LibriSpeech	1,000h	Clean speech	English audiobooks, high quality, widely used for ASR
LibriLight	60,000h	Unlabeled speech	Extended LibriSpeech for self-supervised learning
VoxPopuli	400,000h	Multilingual speech	23 languages from European Parliament recordings
VoxCeleb 1/2	2,000h	Speaker ID	7,000+ speakers, in-the-wild conditions, diverse accents
Common Voice	20,000h+	Crowdsourced	100+ languages, diverse speakers, community-driven
AudioSet	2M clips	General audio	632 classes, environmental sounds, music, speech
FSD50K	51,000 clips	Sound events	Freesound dataset, diverse everyday sounds
MusicCaps	5,500 clips	Music	Text-captioned music for music generation

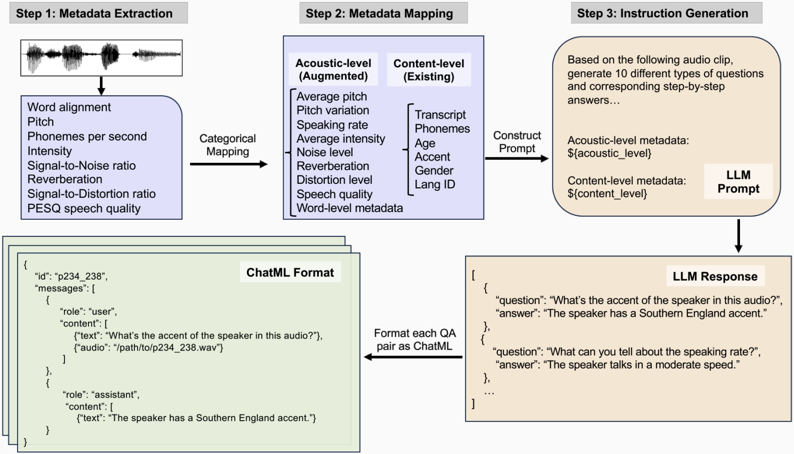
Dataset Specialization

Speech-only: LibriSpeech, LibriLight, VoxPopuli, Common Voice — **Speaker recognition:** VoxCeleb — **General audio:** AudioSet, FSD50K — **Music:** MusicCaps

Instruction Fine Tuning: SIFT-50m [19]



Instruction Fine Tuning: SIFT-50m [19]



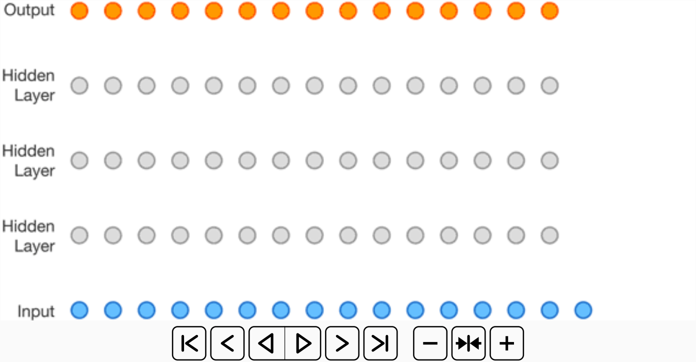
Model	Closed-Ended		Open-Ended		Dynamic-Superb Tasks					
	DS-1	EvalSIFT	AB-Chat	EvalSIFT	Audio	PL	Semt.	Degrd.	Content	Speaker
SALMONN-7B	34.7	21.9	6.4	6.0	31.7	30.5	47.5	30.0	45.2	31.9
Qwen2-Audio-Inst.	48.0	25.1	7.2	7.3	**53.5**	28.9	40.3	43.9	70.6	43.6
O-ASQA-LLM	45.9	22.9	6.6	4.7	28.5	30.0	38.6	45.9	72.3	40.7
SIFT-LLM (ours)	**57.4**	**46.1**	**7.3**	**7.8**	37.5	**42.8**	**51.3**	**63.6**	**75.6**	**47.7**

Outline : Voice synthesis

WaveNet [18]

Advantages of Dilated Stacked causal convolutions

- **Dilated convolutions** → exponentially growing receptive field
- **Parallelizable** over time → fast training
- **Causal** → no future leakage



First model to generate [decent voices]

Outline : Music and Audio Generation

Jukebox: Neural Music Generation

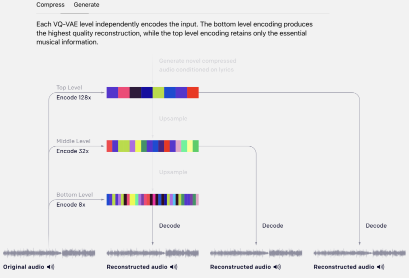
Overview, and examples

Generates music as **raw audio** with artist styles, genres, and singing.
First large-scale neural music generation.

Key Challenge: 4-min song at 44kHz = **10M+ timesteps** ⇒ Must learn long-range dependencies

Dataset: 1.2M songs, lyrics, metadata (Artist, genre, year, mood tags)

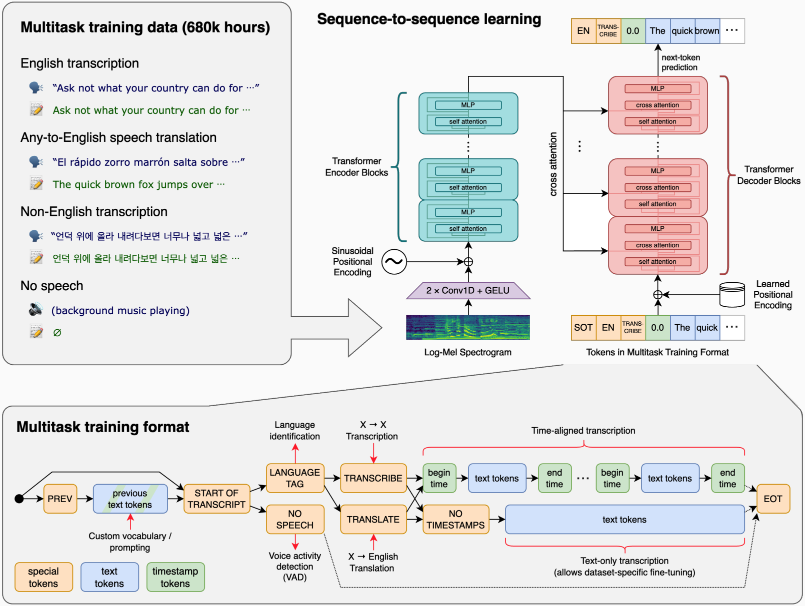
Conditioning: Artist, genre, and lyrics via encoder-decoder attention



VQ-VAE: 3 hierarchical levels compress audio by 8x, 32x, 128x

Outline : Audio Speech Recognition

Whisper[20]



Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ⋯"
Ask not what your country can do for ⋯
Any-to-English speech translation
"El rápido zorro marrón salta sobre ⋯"
The quick brown fox jumps over ⋯
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
⋯
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
NO TIMESTAMPS
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Questions?

S. Arora, K.-w. Chang, C.-m. Chien, Y. Peng, H. Wu, Y. Adi, E. Dupoux, H.-Y. Lee, K. Livescu, and S. Watanabe.
On The Landscape of Spoken Language Models: A Comprehensive Survey.
TMLR, pages 1–40, 2025.

A. Baevski, S. Schneider, and M. Auli.
Vq-Wav2Vec: Self-Supervised Learning of Discrete Speech Representations.
8th International Conference on Learning Representations, ICLR 2020, pages 1–12, 2020.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli.
wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.
arXiv, (Figure 1):1–19, 2020.

J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, and B. Xu.
Qwen Technical Report.
pages 1–59, 2023.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei.
WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing.
IEEE Journal on Selected Topics in Signal Processing, 16(6):1505–1518, 2022.

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou.
Qwen2-Audio Technical Report.
pages 1–16, 2024.

Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou.
Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models.
pages 1–18, 2023.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby.
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
In ICLR, pages 1–21, 2021.

B. Elizalde, S. Deshmukh, and H. Wang.
NATURAL LANGUAGE SUPERVISION FOR GENERAL-PURPOSE AUDIO REPRESENTATIONS.
In ICASSP, 2024.

References v

A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro.
Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models.
pages 1–61, 2025.

Y. Gong, Y. A. Chung, and J. Glass.
Ast: Audio spectrogram transformer.
In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 1, pages 56–60, 2021.

J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko.
Bootstrap your own latent a new approach to self-supervised learning.
In Advances in Neural Information Processing Systems, volume 2020-Decem, 2020.

W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed.
HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.
IEEE/ACM Transactions on Audio Speech and Language Processing, 29(Cv):3451–3460, 2021.

C.-y. Huang, W.-C. Chen, S.-w. Yang, A. T. Liu, C.-a. Li, Y.-X. Lin, W.-c. Tseng, A. Diwan, Y.-j. Shih, J. Shi, W. Chen, C.-k. Yang, W. Ren, X. Chen, C.-Y. Hsiao, P. Peng, S.-H. Wang, C.-Y. Kuan, K.-H. Lu, K.-W. Chang, F. Ritter-Gutierrez, K.-P. Huang, S. Arora, Y.-K. Lin, M. T. Chuang, E. Yeo, K. Chang, C.-m. Chien, K. Choi, J.-y. Wang, C.-h. Hsieh, Y.-C. Lin, C.-E. Yu, I.-H. Chiu, H. R. Guimarães, J. Han, T.-Q. Lin, T.-Y. Lin, H. Chang, T.-W. Chang, C. W. Chen, S.-J. Chen, Y.-h. Chen, H.-c. Cheng, K. Dhawan, J.-L. Fang, S.-x. Fang, K.-y. F. Chiang, C. A. Fu, H.-f. Hsiao, C. Y. Hsu, S.-S. Huang, L. C. Wei, H.-C. Lin, H.-H. Lin, H.-T. Lin, J.-r. Lin, T.-c. Liu, L.-c. Lu, T.-m. Pai, A. Pasad, S.-Y. S. Kuan, S. Shon, Y. Tang, Y.-S. Tsai, J.-C. Wei, T.-C. Wei, C. Wu, D.-R. Wu, C.-H. H. Yang, C.-C. Yang, J. Q. Yip, S.-X. Yuan, V. Noroozi, Z. Chen, H. Wu, K. Livescu, D. Harwath, S. Watanabe, and H.-y. Lee.

Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities of Spoken Language Models with 180 Tasks.
In ICLR, pages 1–67, 2025.

C. Y. Huang, K. H. Lu, S. H. Wang, C. Y. Hsiao, C. Y. Kuan, H. Wu, S. Arora, K. W. Chang, J. Shi, Y. Peng, R. Sharma, S. Watanabe, B. Ramakrishnan, S. Shehata, and H. Y. Lee.
DYNAMIC-SUPERB: TOWARDS A DYNAMIC, COLLABORATIVE, AND COMPREHENSIVE INSTRUCTION-TUNING BENCHMARK FOR SPEECH.
In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pages 12136–12140, 2024.

- P. Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer.
  Masked Autoencoders that Listen.
  Advances in Neural Information Processing Systems, 35(NeurIPS):1–13, 2022.
- Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro.
  Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities.
  Proceedings of Machine Learning Research, 235:25125–25148, 2024.
- A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu.
  WaveNet: A Generative Model for Raw Audio.
  pages 1–15, 2016.

References x

P. Pandey, R. V. Swaminathan, K. V. V. Girish, A. Sen, J. Xie, G. Strimel, and A. Schwarz.
SIFT-50M: A Large-Scale Multilingual Dataset for Speech Instruction Fine-Tuning.
In ACL, volume 1, pages 13921–13942, 2025.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever.
Robust Speech Recognition via Large-Scale Weak Supervision.
Proceedings of Machine Learning Research, 202:28492–28518, 2023.

A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu.
Neural discrete representation learning.
Advances in Neural Information Processing Systems, 2017-Decem(Nips):6307–6316, 2017.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman.
GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.
In EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop, pages 353–355, 2018.

Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov.
Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation.
ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2023-June, 2023.

S. W. Yang, P. H. Chi, Y. S. Chuang, C.-i. I. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. T. Lin, T. H. Huang, W. C. Tseng, K. T. Lee, D. R. Liu, Z. Huang, S. Dong, S. W. Li, S. Watanabe, A. Mohamed, and H. Y. Lee.
SUPERB: Speech processing Universal PERformance Benchmark.
In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 4, pages 3161–3165, 2021.