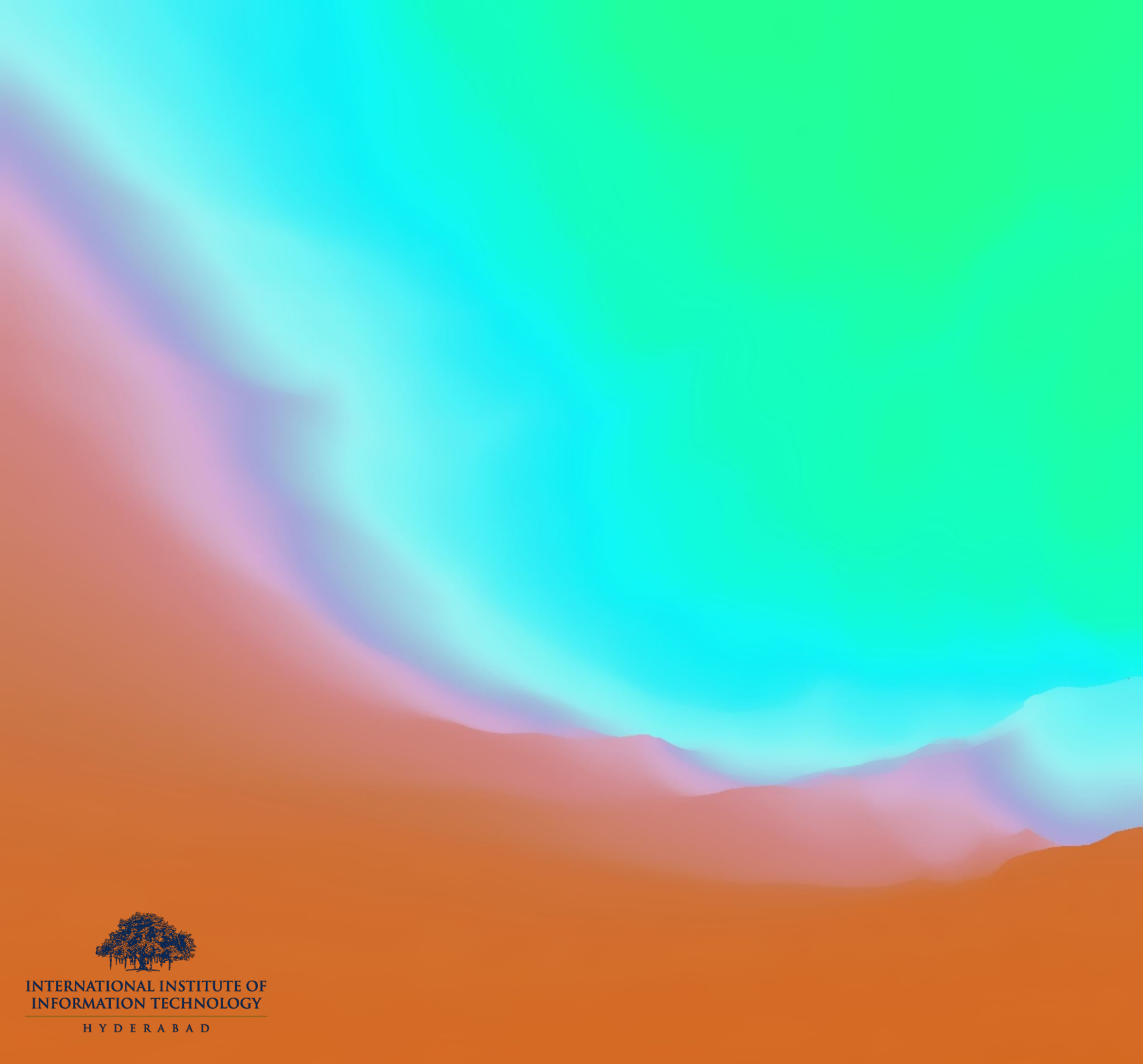


The Sound Dimension: Speech and Audio in Multimodal AI

Vineet Gandhi

Associate Professor

Center for Visual Information Technology

Why?
How?
What?
Where?

Speech and Audio in MAI

Excellent survey paper

Baltrusaitis et al. Multimodal Machine Learning: A Survey and Taxonomy. **TPAMI 2019**

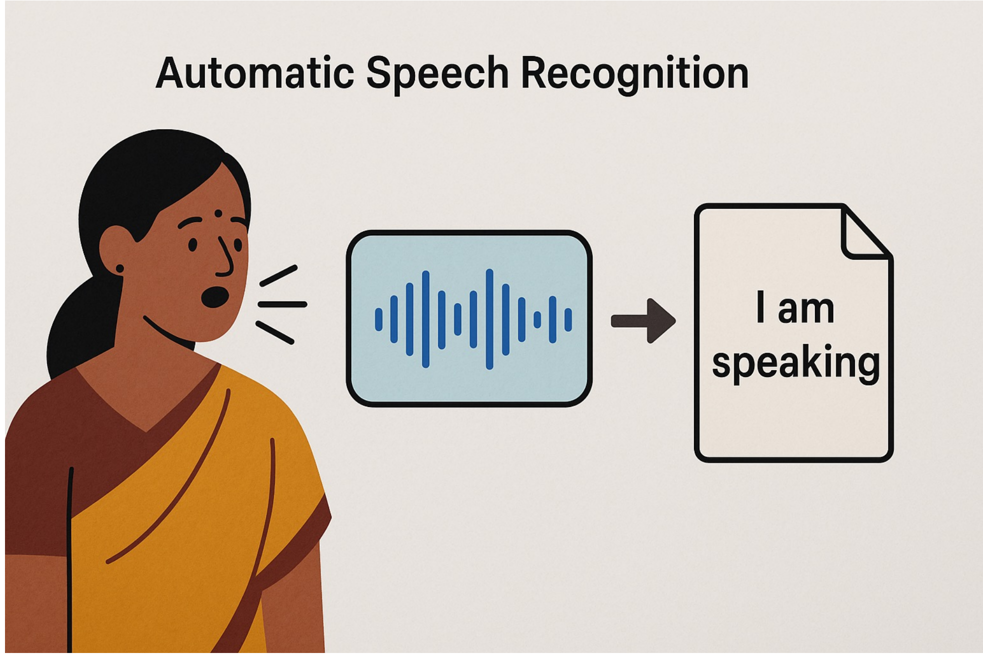

BBC FOUR HD

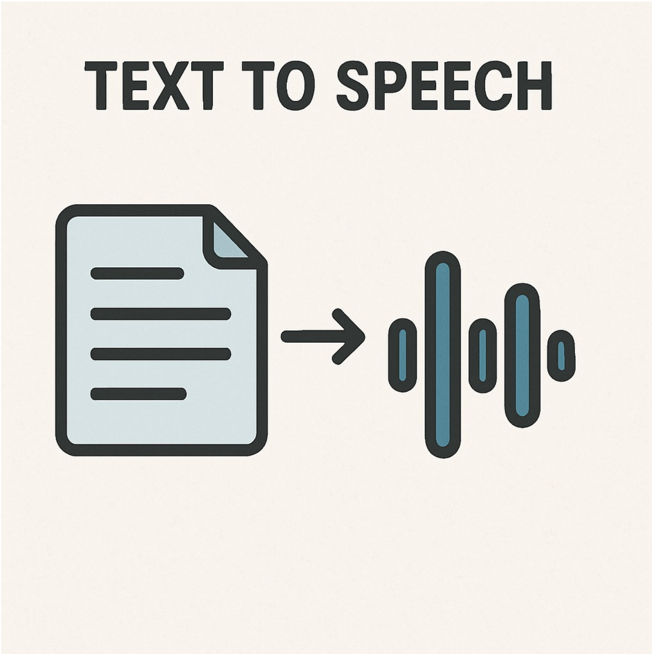

VOICE OR TEXT?

HUMAN LEARNING IS INHERENTLY MULTIMODAL

And more importantly, we will miss out a lot of fun applications/use-cases

Key Speech problems

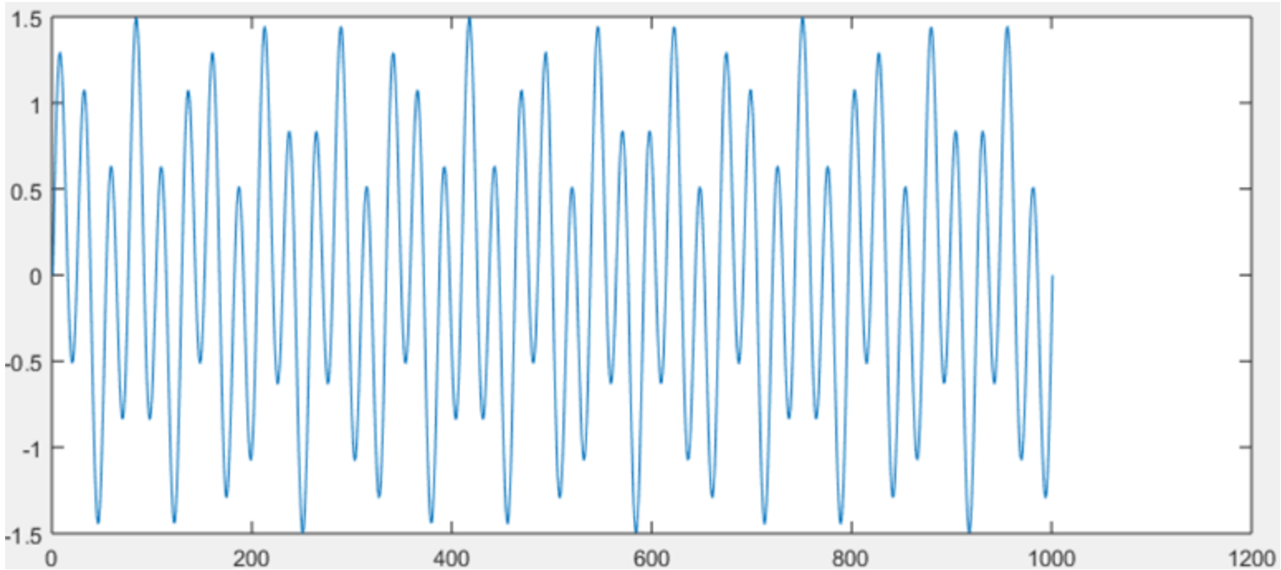


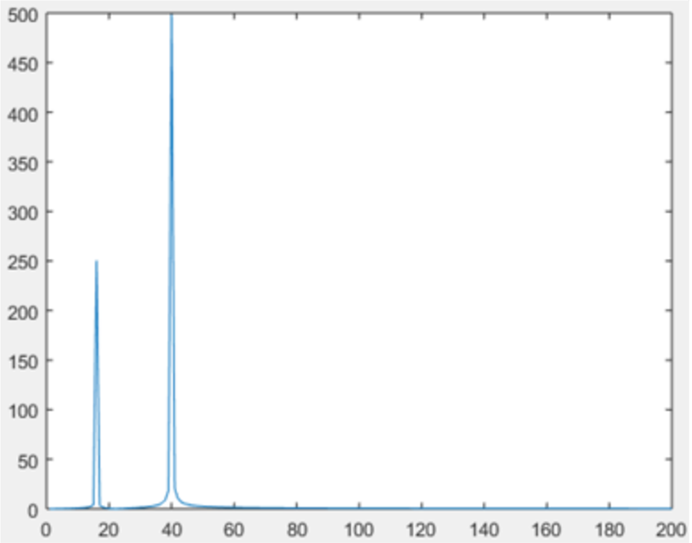


Multimodal processing has transformed Speaker Recognition

Some basics to help my argument

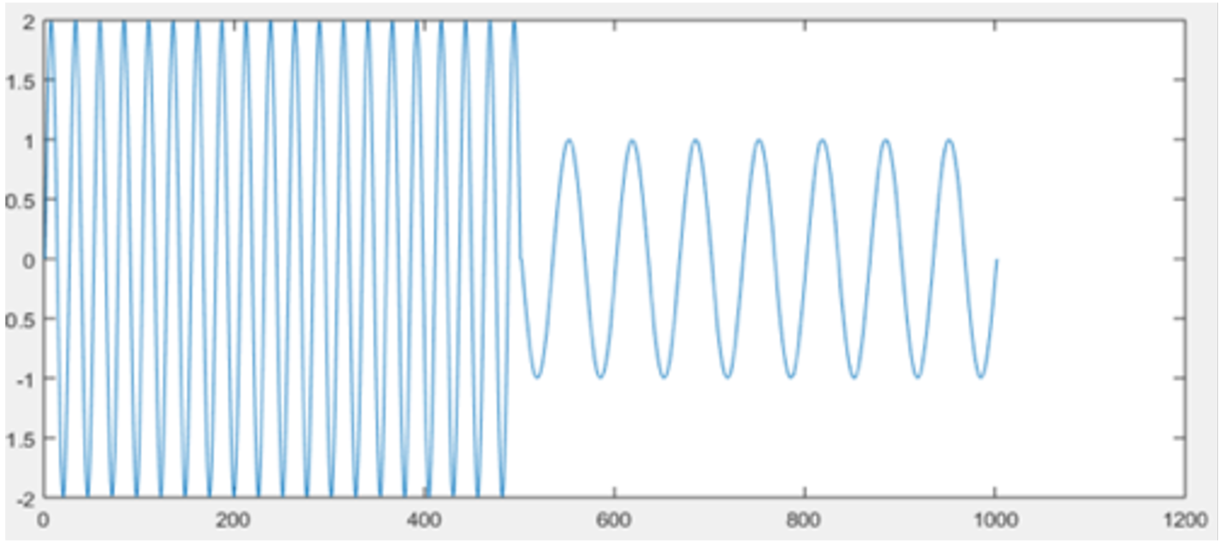
$$f(t) = \sin(2\pi \cdot 39t) + 0.5\sin(2\pi \cdot 15t)$$

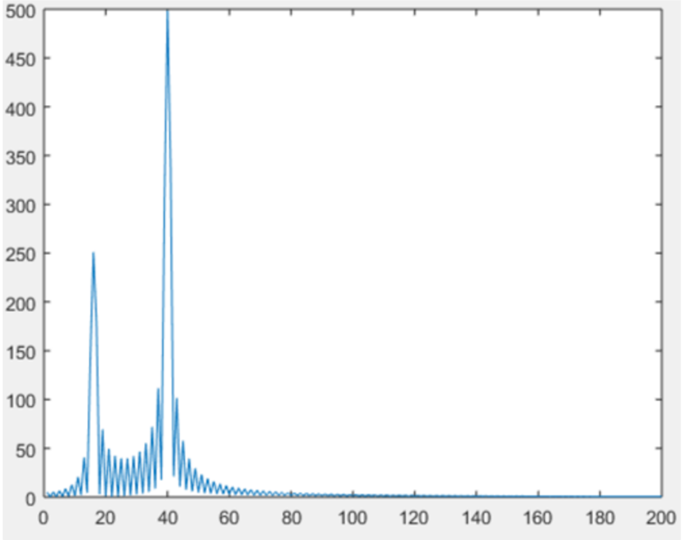




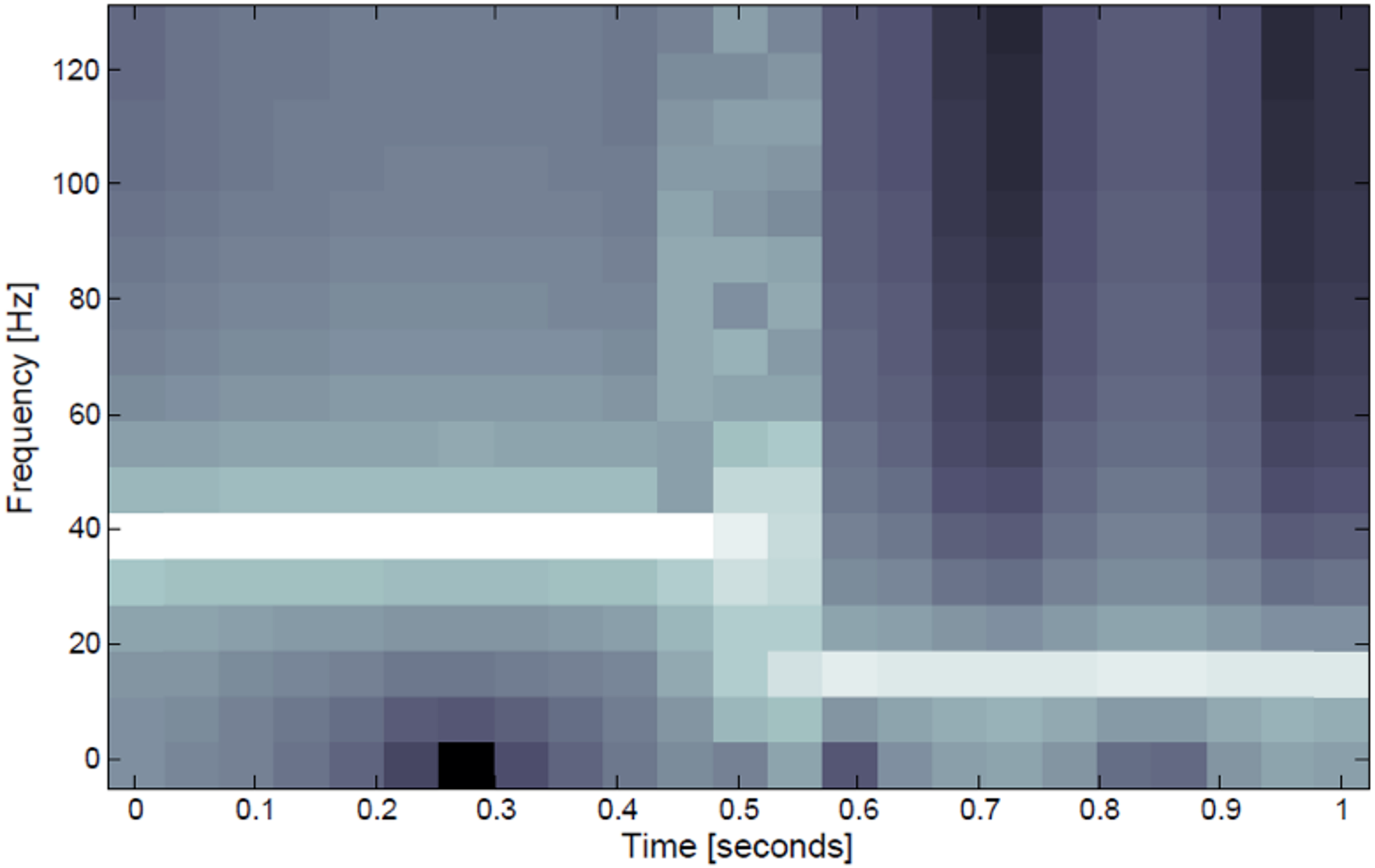
Another example sound signal

$$g(t) = \begin{cases} 2 * \sin(2\pi \cdot 39t)\,, 0 \le t \le 1/2 \\ \sin(2\pi \cdot 15t)\,, 1/2 < t \le 1 \end{cases}$$





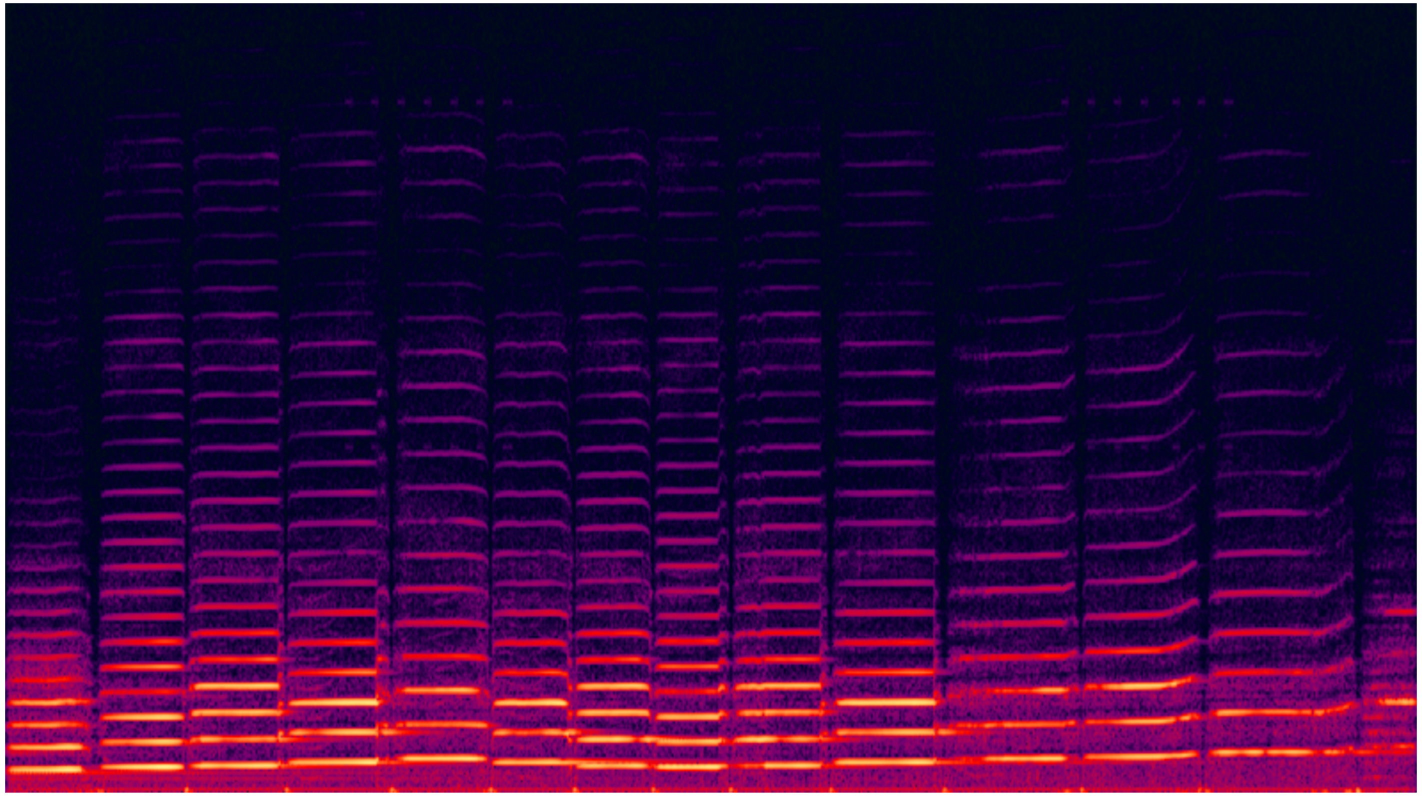
Spectrogram



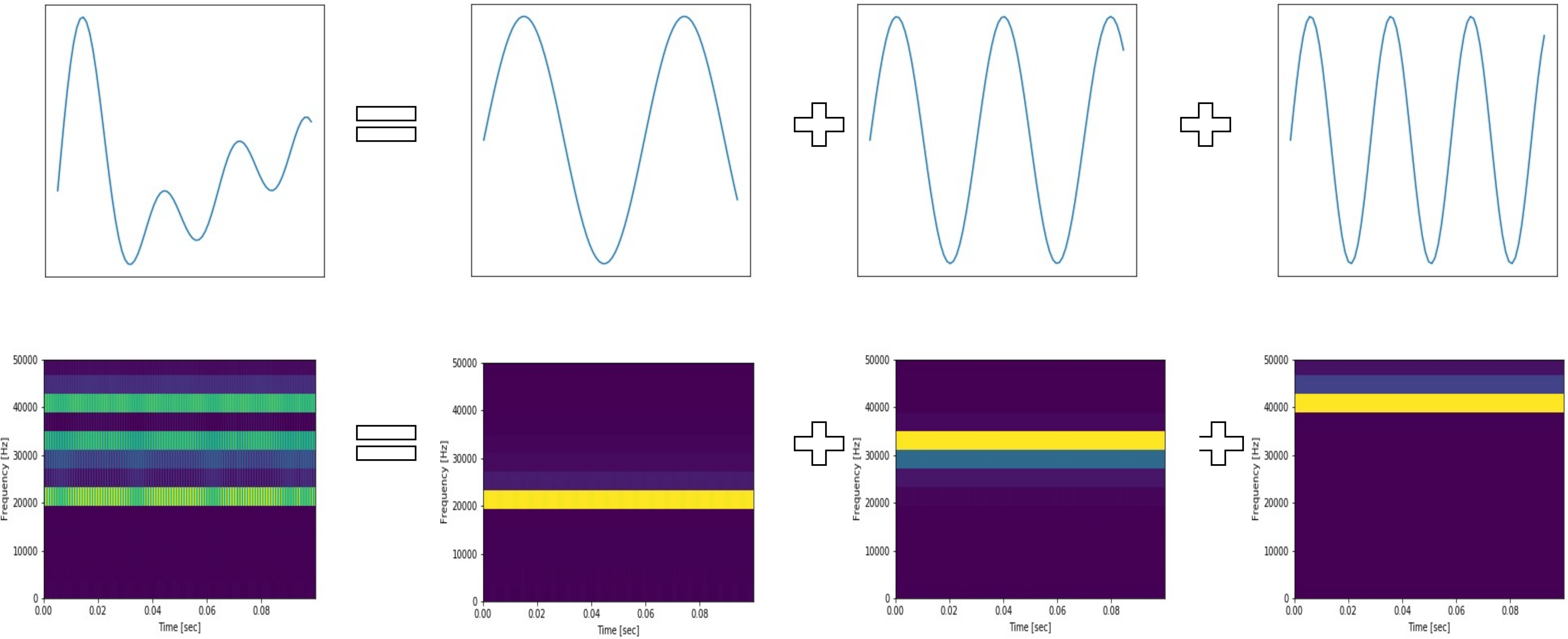
Spectrogram of a piecewise monochromatic signal.

Lighter color indicates greater DFT magnitude

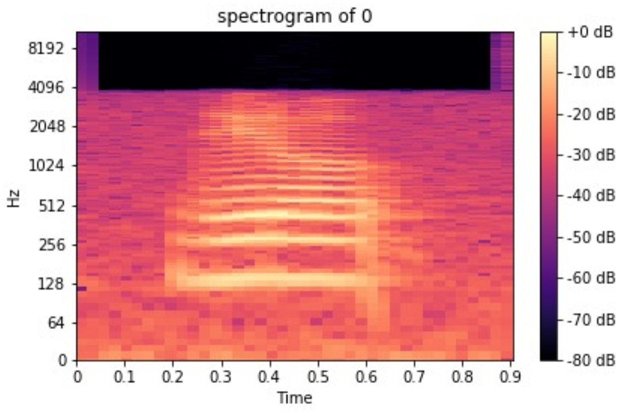
Spectrogram

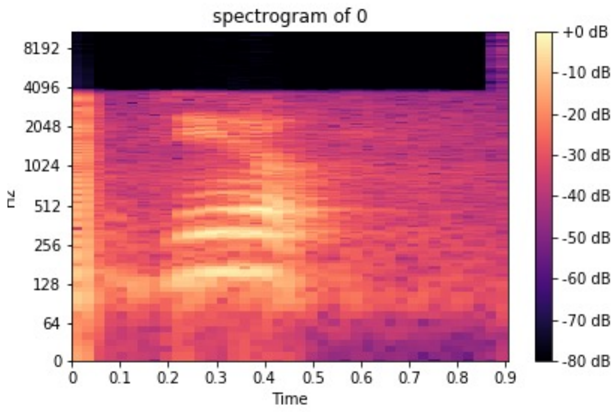


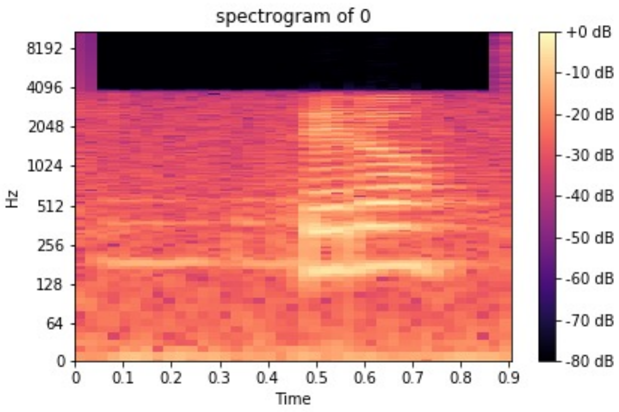
Wave as a combination of sine waves

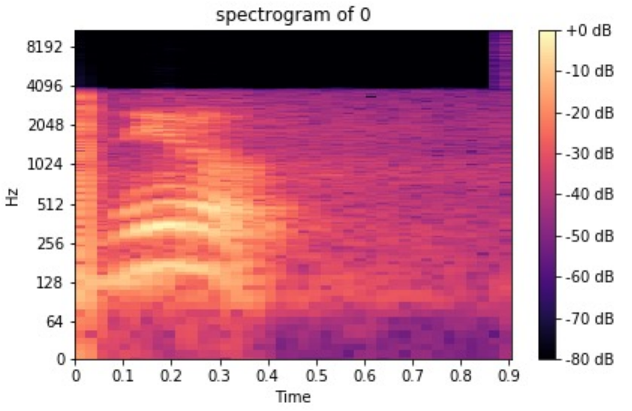


Utterance of word zero





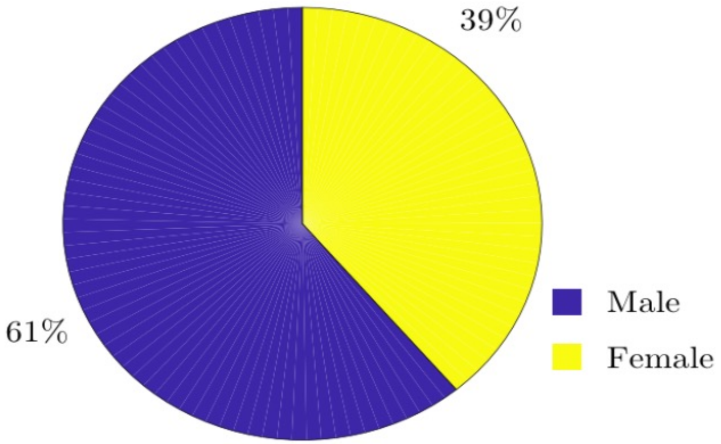


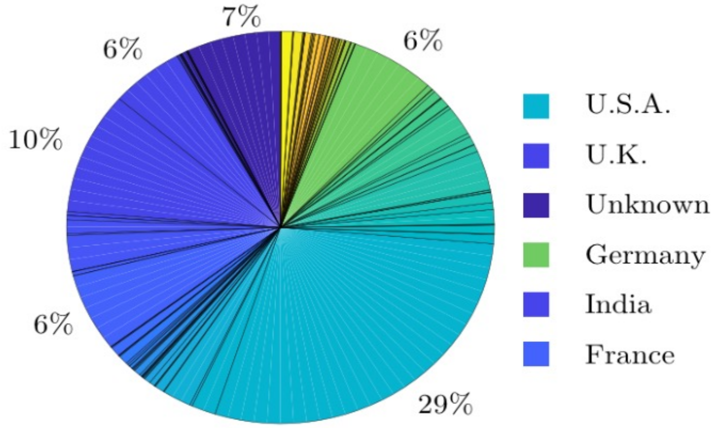


Transforming speaker verification/identification

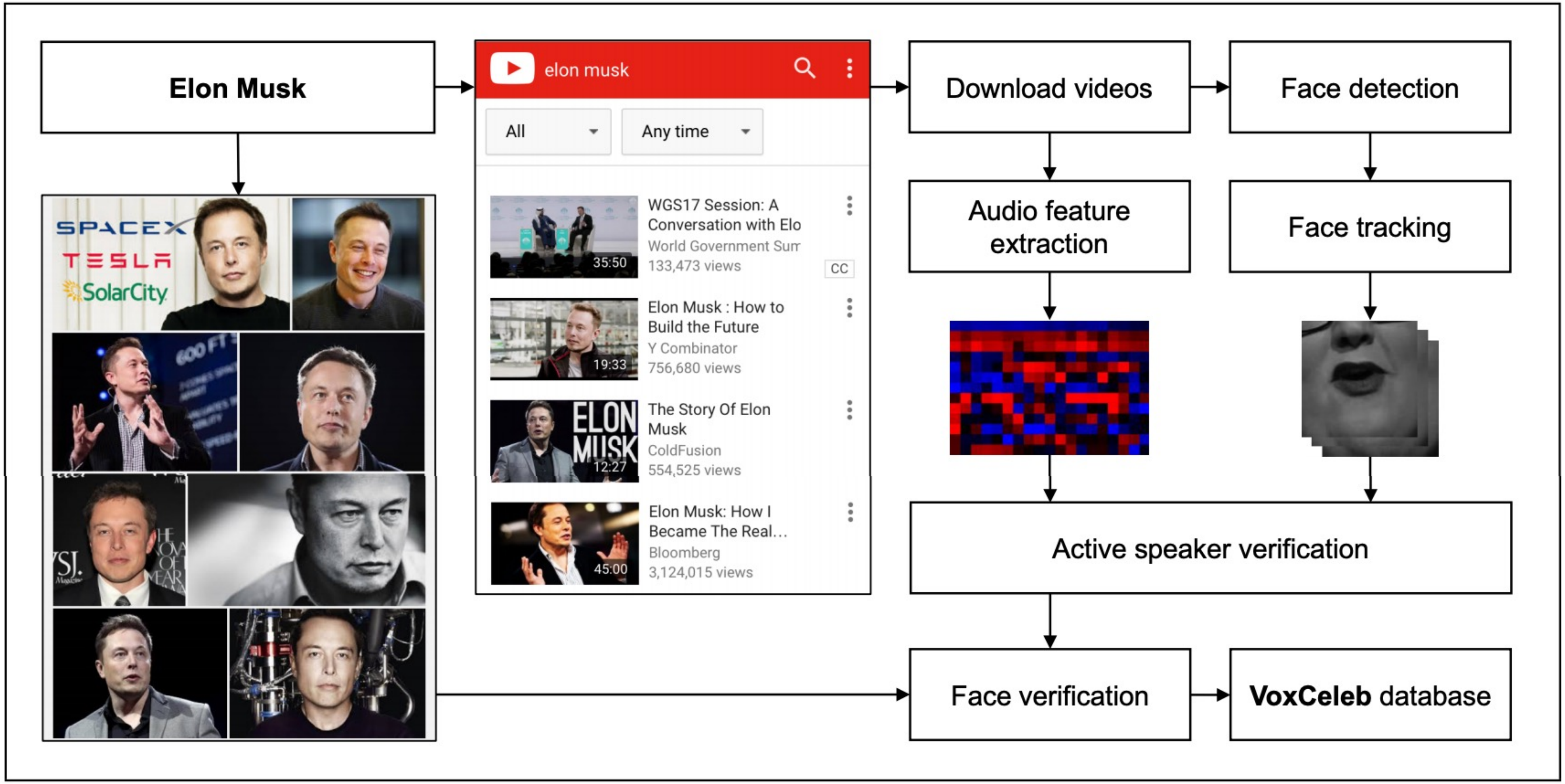
VoxCeleb2

VoxCeleb2 contains over a million utterances for 6,112 identities.

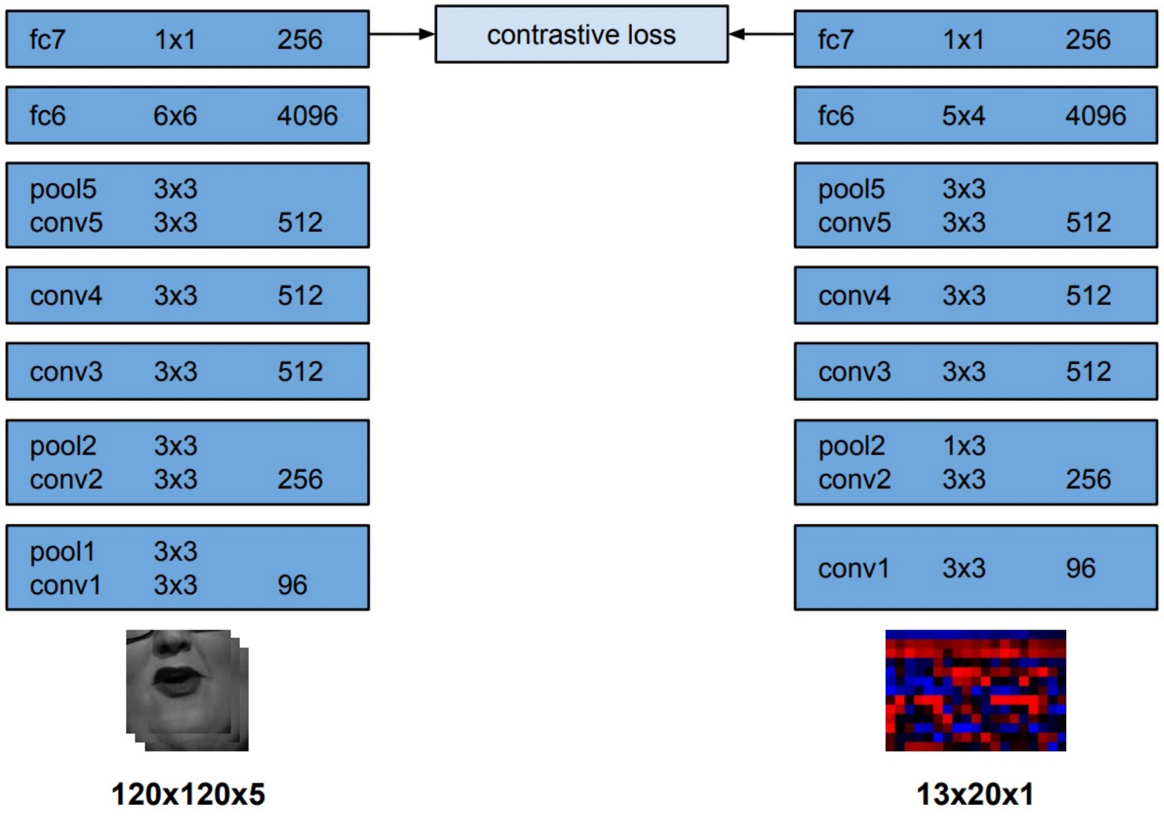




VoxCeleb: automated data collection



SyncNet



Solves three tasks

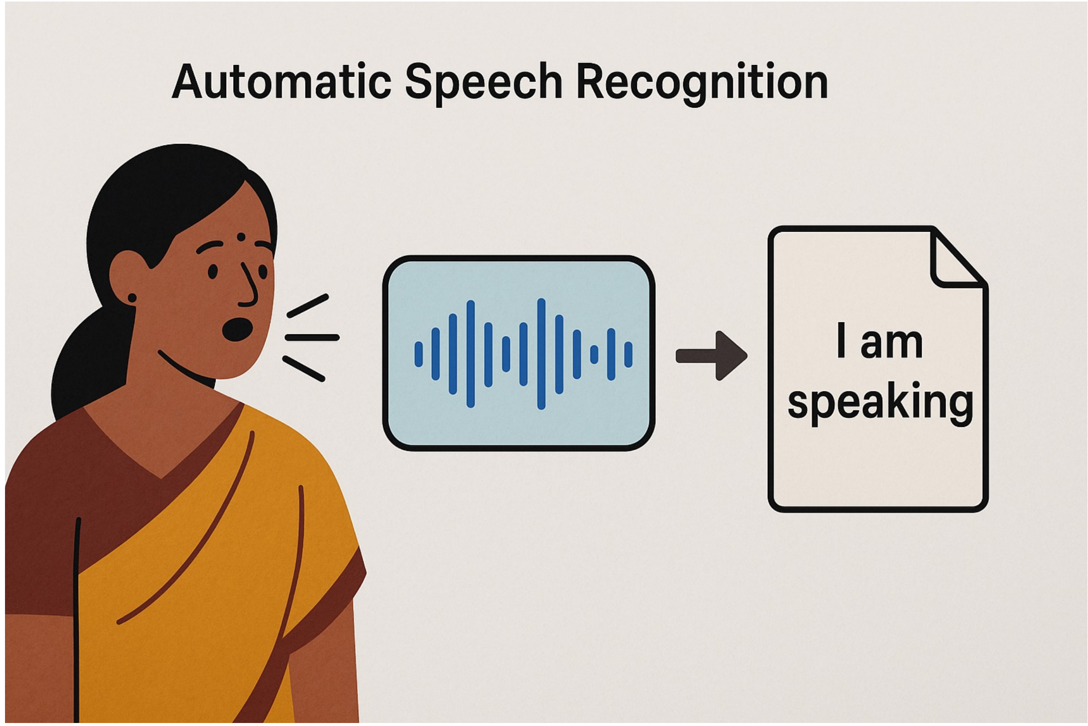
- Determining the lip-sync error in videos
- Detecting the speaker in a scene with multiple faces
- Lip reading

J. S. Chung and A. Zisserman, “Out of time: automated lip sync in the wild,” in Workshop on Multi-view Lip-reading, ACCV, 2016.

VoxCeleb performance

Accuracy	Top-1 (%)	Top-5 (%)
I-vectors + SVM	49.0	56.6
I-vectors + PLDA + SVM	60.8	75.6
CNN-fc-3s no var. norm.	63.5	80.3
CNN-fc-3s	72.4	87.4
CNN	**80.5**	**92.1**

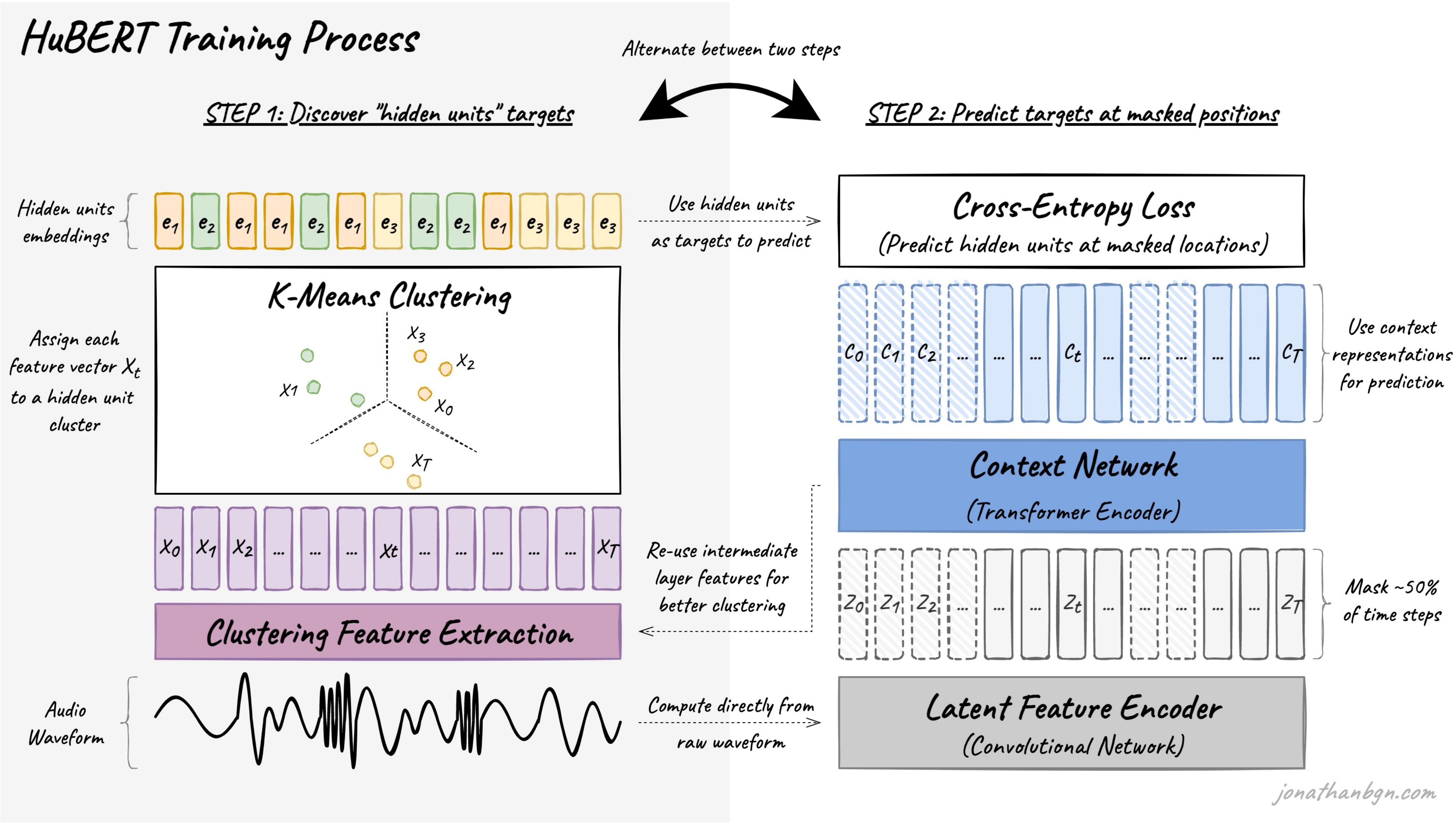
Transformation through transformers



Transformers and Self-Supervision

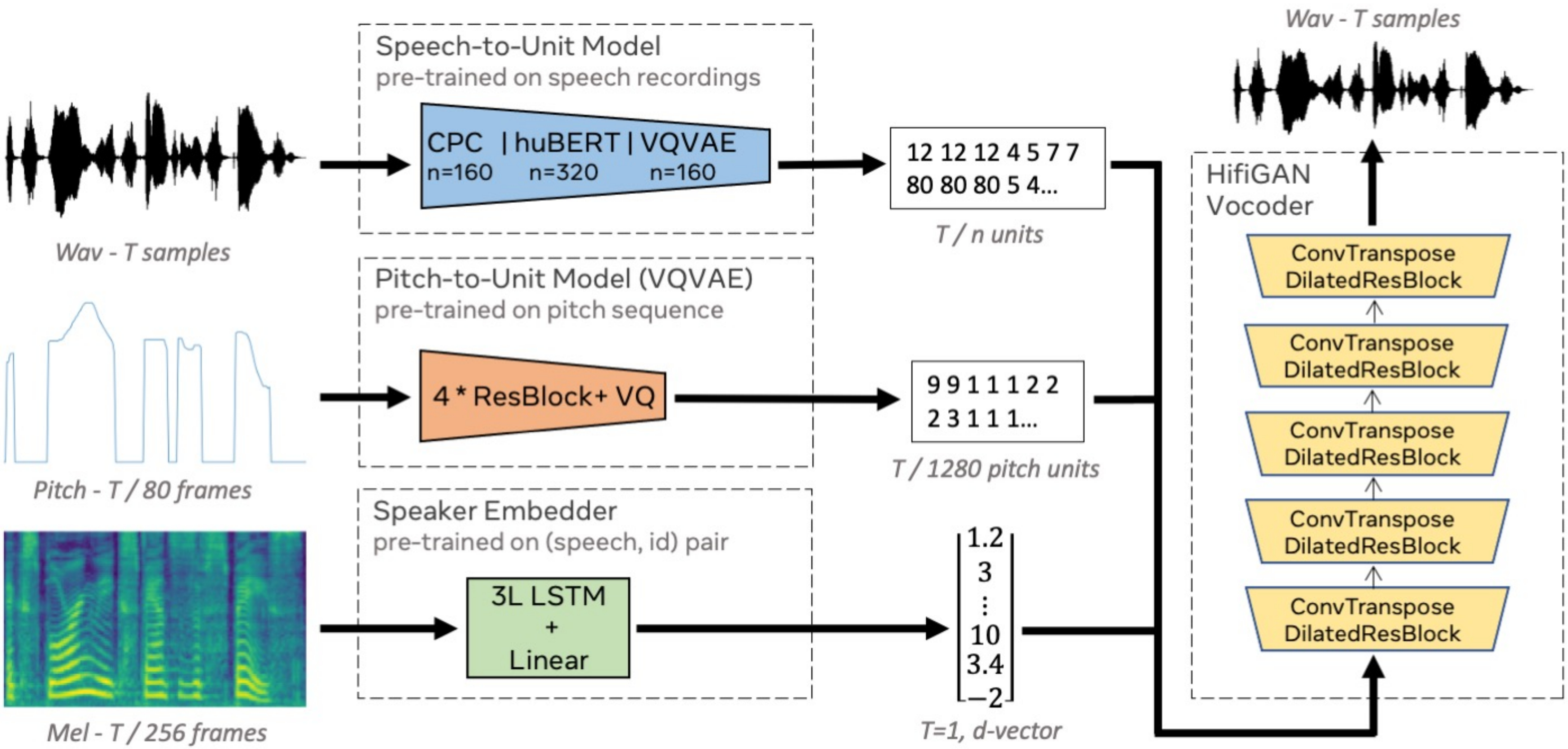
1. Shared pool of architectural insights
    - Models like **HuBERT, wav2vec 2.0, and Whisper** borrow architectural insights (e.g., Transformers, masked modeling) from NLP

2. Language Models Enhance ASR Decoding
    - Leads to **better handling of rare words, disfluencies, and long-range dependencies**



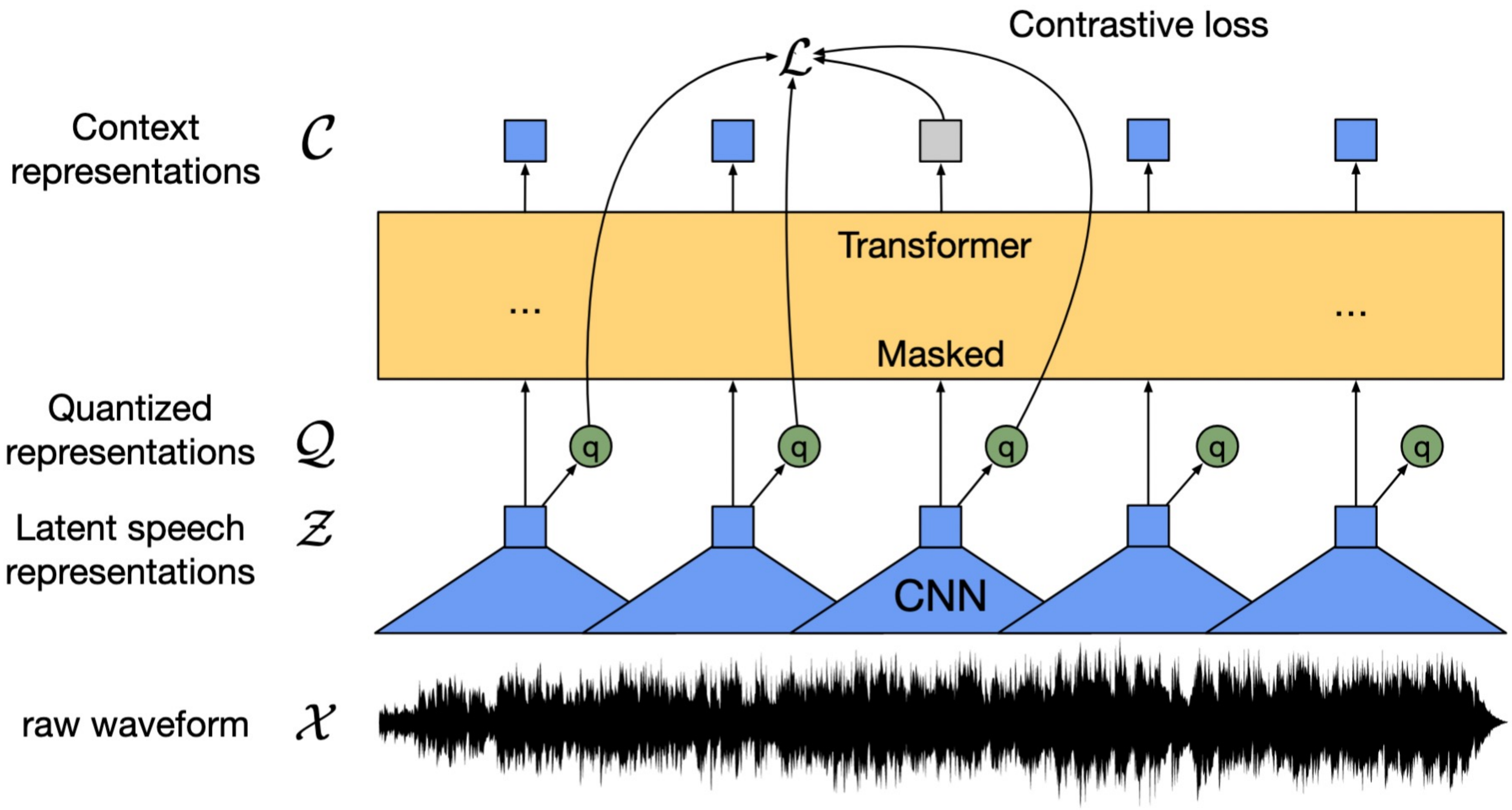
HuBERT Training Process
Alternate between two steps
STEP 1: Discover "hidden units" targets
STEP 2: Predict targets at masked positions
Hidden units embeddings
e1 e2 e1 e1 e2 e1 e3 e2 e2 e1 e3 e3 e3
Use hidden units as targets to predict
Cross-Entropy Loss
(Predict hidden units at masked locations)
K-Means Clustering
Assign each feature vector Xt to a hidden unit cluster
X3 X2 X1 X0 XT
C0 C1 C2 ... ... Ct ... ... CT
Use context representations for prediction
Context Network
(Transformer Encoder)
X0 X1 X2 ... ... Xt ... ... XT
Re-use intermediate layer features for better clustering
Z0 Z1 Z2 ... ... Zt ... ... ZT
Mask ~50% of time steps
Clustering Feature Extraction
Audio Waveform
Compute directly from raw waveform
Latent Feature Encoder
(Convolutional Network)
jonathanbgn.com

Hubert (speech resynthesis)



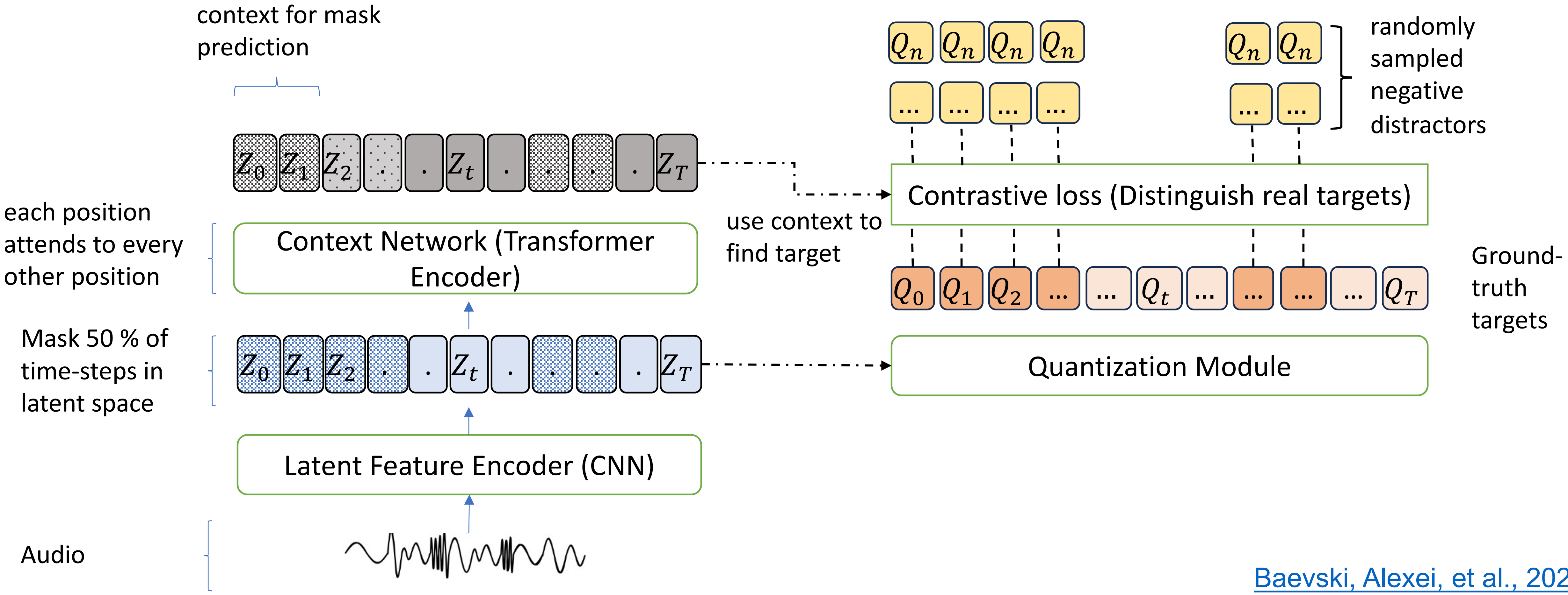
Polyak et al. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. Interspeech 2021

Wav2Vec 2.0



Contrastive loss
$\mathcal{L}$
Context representations $\mathcal{C}$
Transformer
...
...
Masked
Quantized representations $\mathcal{Q}$
q
q
q
q
q
Latent speech representations $\mathcal{Z}$
CNN
raw waveform $\mathcal{X}$

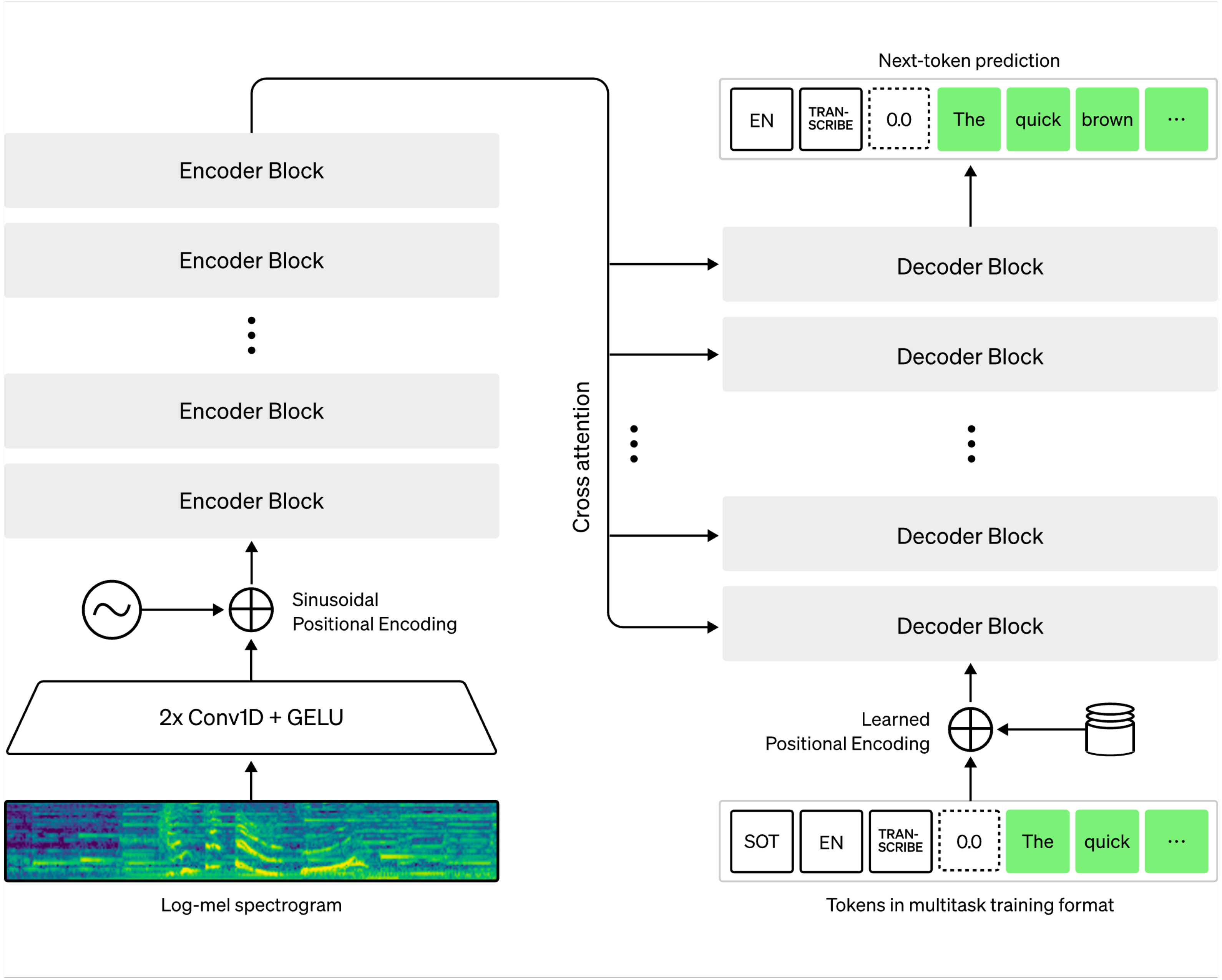
Wav2Vec 2.0



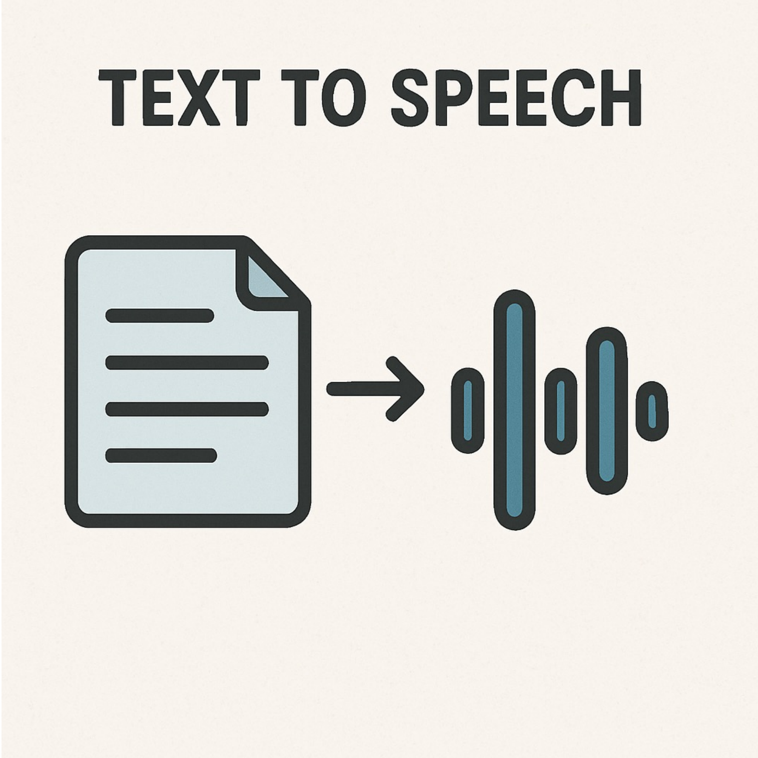
Baevski, Alexei, et al., 2020

Whisper

- Trained on 680K hours of multilingual, multitask web data
- Robust to accents, noise, and technical terms
- Supports multilingual transcription and translation to English
- Whisper doesn't outperform models on LibriSpeech but is significantly more robust in zero-shot settings, with 50% fewer errors across varied datasets.



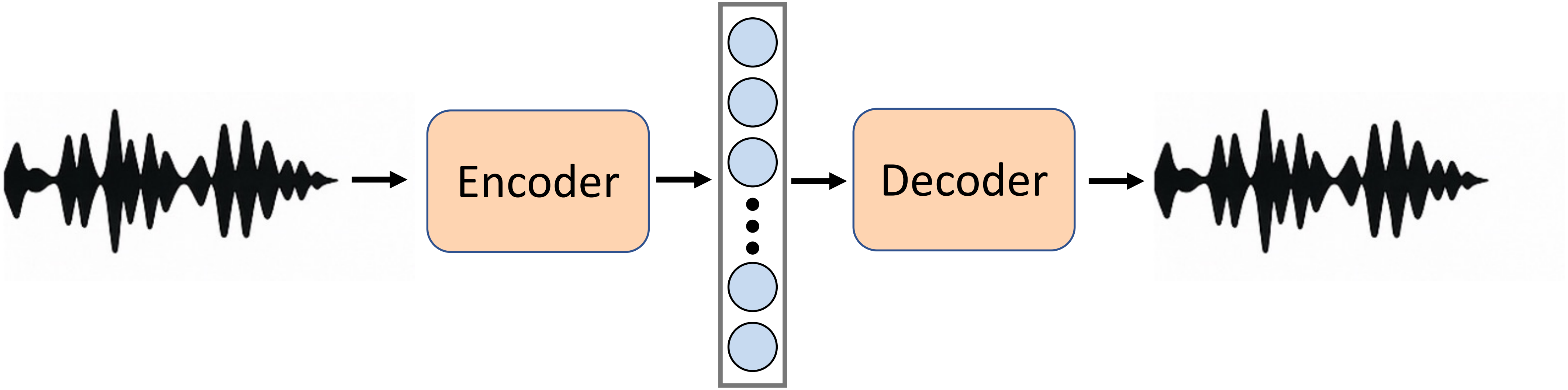
Thinking multimodal, opens-up possibilities

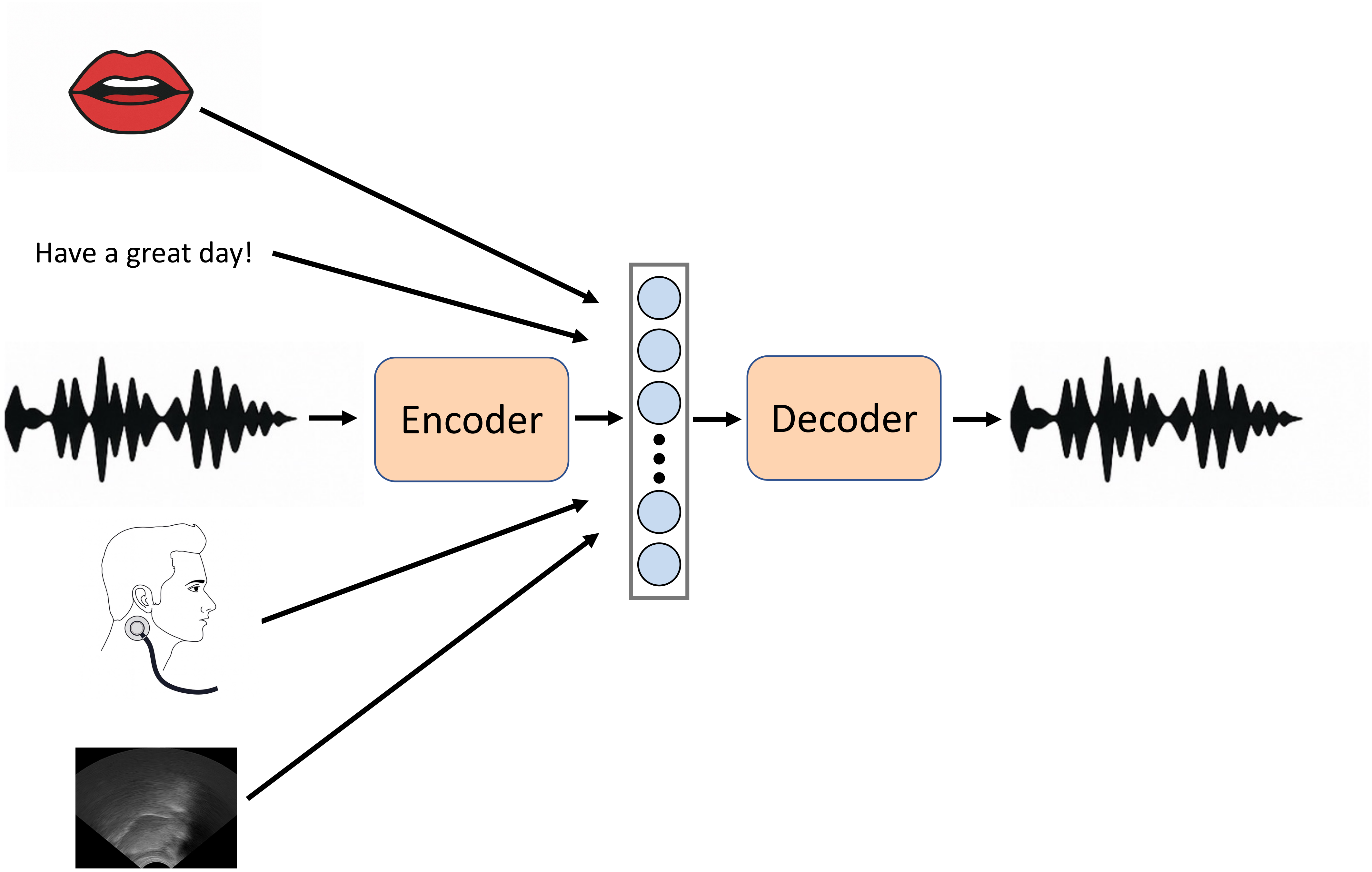


ParrotTTS

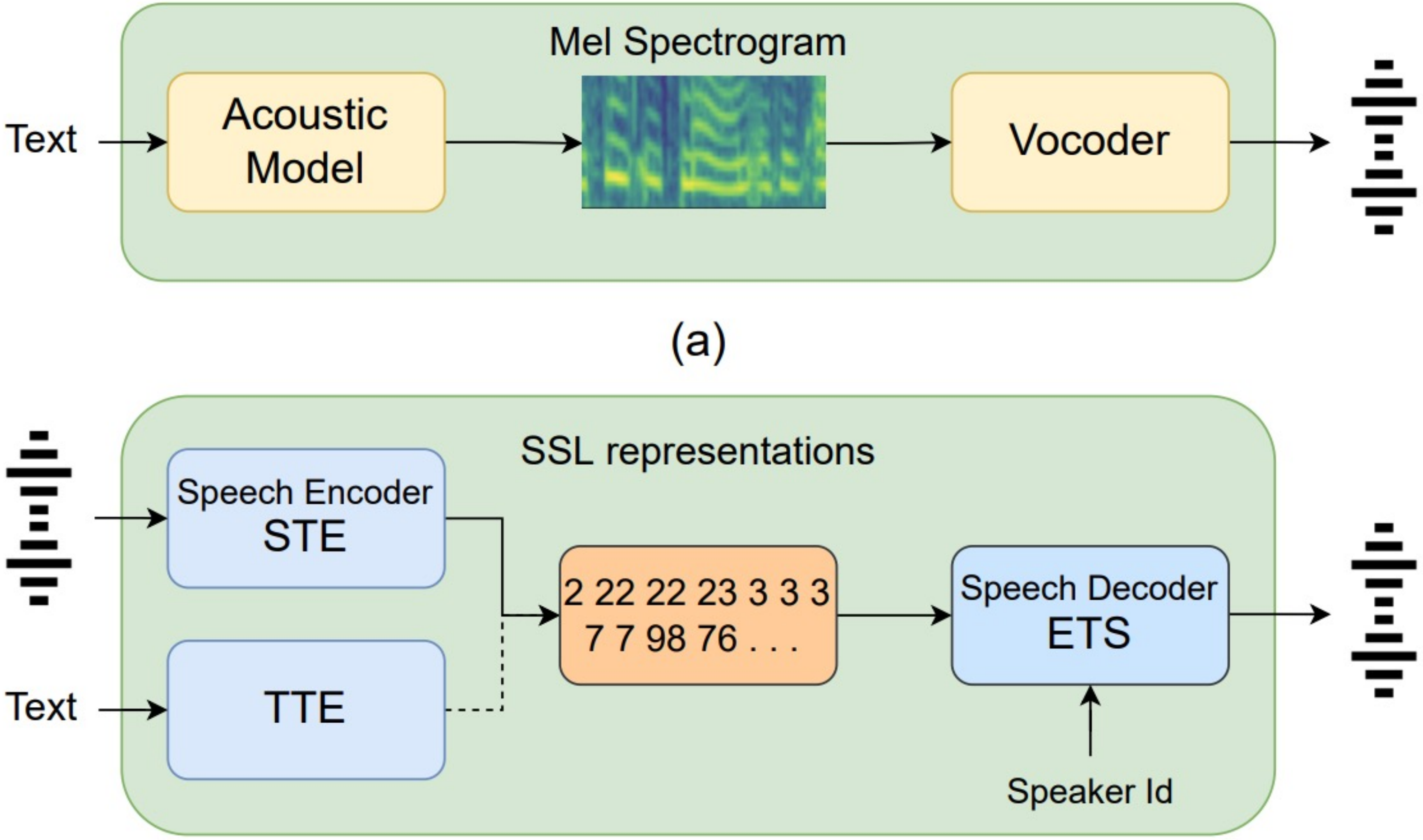
Vocal learning forms the first phase of infants starting to talk (Locke, 1996, 1994) by simply listening to sounds/speech. It is hypothesized (Kuhl and Meltzoff, 1996) that infants listening to ambient language store perceptually derived representations of the speech sounds they hear, which in turn serve as targets for the production of speech utterances. Interestingly, in this phase, the infant has no conception of text or linguistic rules, and speech is considered sufficient to influence speech production (Kuhl and Meltzoff, 1996) as can parrots (Locke, 1994).

Shah et al. ParrotTTS: Text-to-speech synthesis exploiting disentangled self-supervised representations. EACL Findings 2024


Encoder
Decoder


Have a great day!
Encoder
Decoder

ParrotTTS

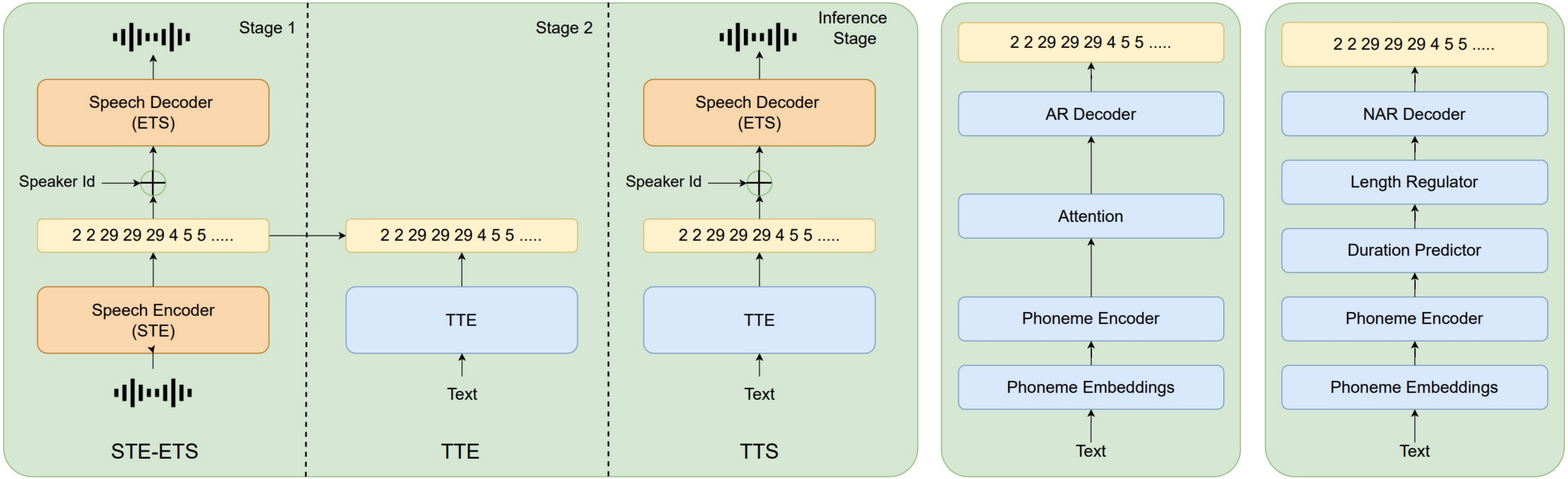


(a)

(b)

Shah et al. ParrotTTS: Text-to-speech synthesis exploiting disentangled self-supervised representations. EACL Findings 2024

ParrotTTS



(a) ParrotTTS (b) Autoregressive TTE (c) Non Autoregressive TTE

ParrotTTS

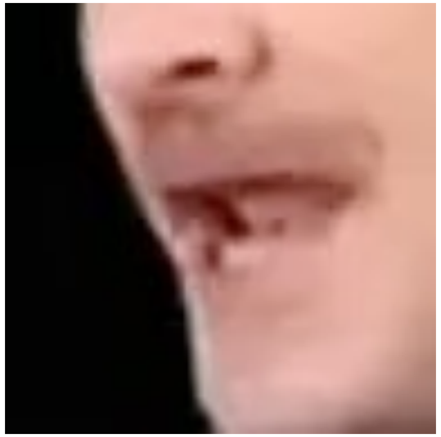
	Model	MOS ↑	WER ↓
Traditional TTS	SS-FastSpeech2	3.87	4.52
	SS-Tacotron2	3.90	4.59
	FastSpeech2-SupASR	3.78	4.72
	Tacotron2-UnsupASR	3.50	11.3
WavThruVec	SS-WavThruVec	3.57	6.27
VQ-VAE	SS-VQ-VAES	3.12	21.78
ParrotTTS	AR-TTE$_{\text{LJS}}$+SS-ETS	3.85	4.80
	NAR-TTE$_{\text{LJS}}$+SS-ETS	3.86	4.58
	NAR-TTE$_{\frac{1}{2}\text{LJS}}$+SS-ETS	3.81	6.14

Table 1: Subjective and objective comparison of TTS models in the single speaker setting.

Model	VCTK	MOS ↑	WER ↓	EER ↓
GT-Mel+Vocoder	Yes	4.12	2.25	2.12
MS-FastSpeech2	Yes	3.62	5.32	3.21
MS-FastSpeech2-SupASR	No	3.58	6.65	3.85
VC-FastSpeech2	No	3.41	7.44	8.18
WavThruVec-MS	No	3.17	6.79	5.08
NAR-TTE$_{\text{LJS}}$+MS-ETS	No	3.78	6.53	4.38

Table 2: Comparison of the multi-speaker TTS models on the VCTK dataset. Column 2 indicates if the corresponding method uses VCTK transcripts while training.

Lip Reading



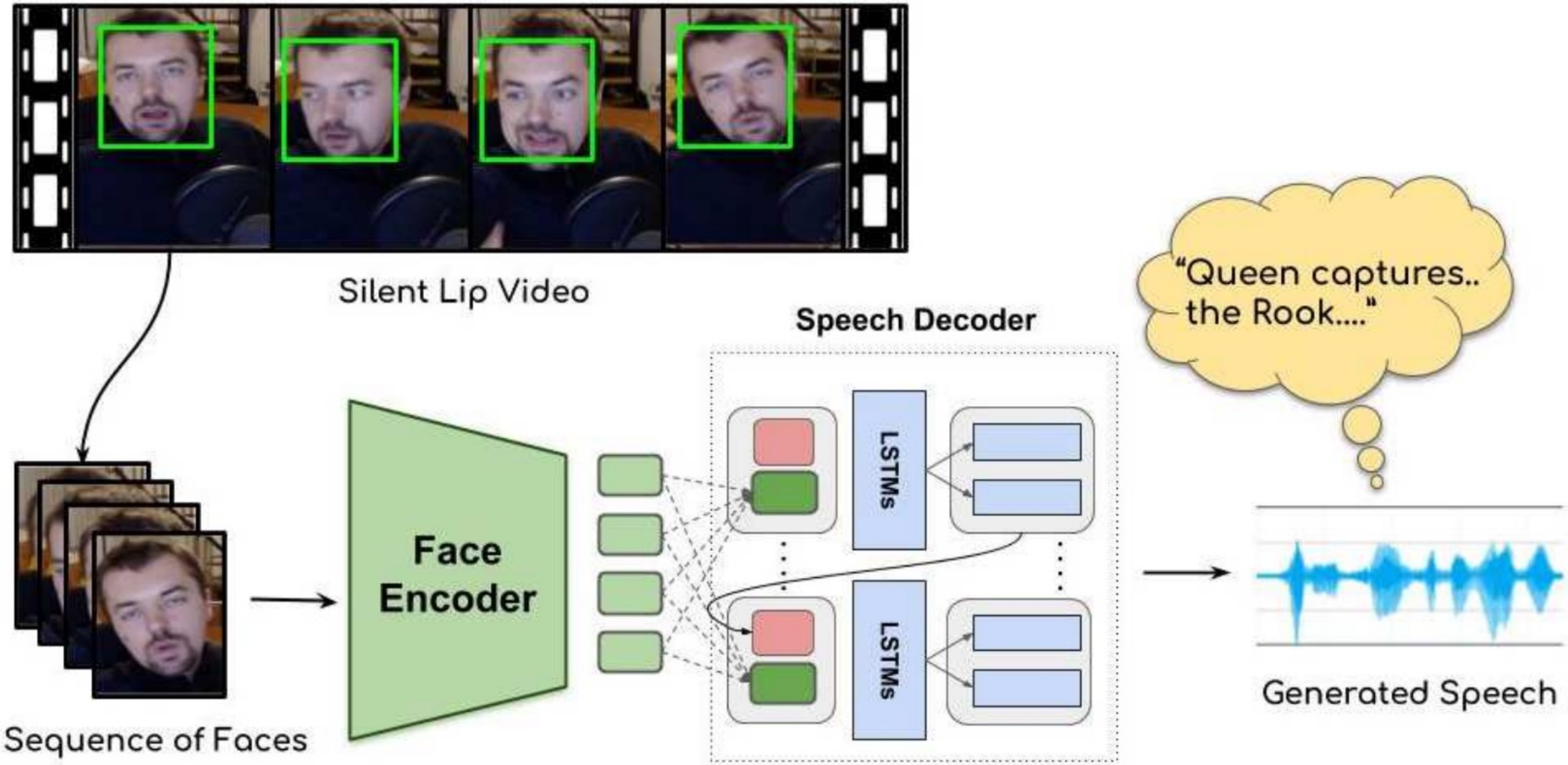
Sahipjohn et al. *RobustL2S: Speaker-Specific Lip-to-speech Synthesis exploiting Self-Supervised Representations*. **APSIPA 2023**

Lip2Wav

The speech you are hearing is completely generated from the lip movements

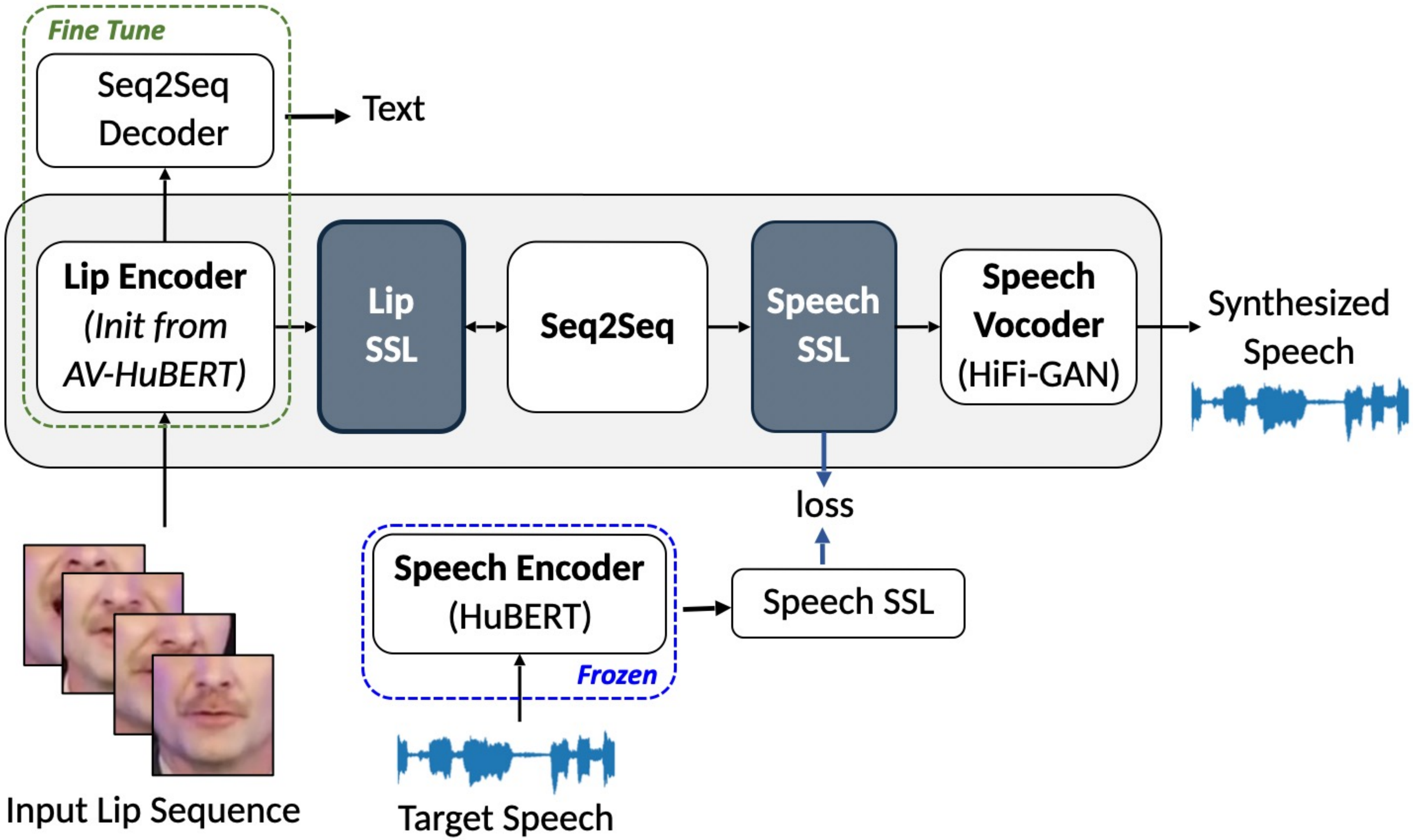
Rudrabha et al. *Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis*. **CVPR 2020**

Lip2Wav



Rudrabha et al. *Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis*. **CVPR 2020**

RobustL2s



Sahipjohn et al. *RobustL2S: Speaker-Specific Lip-to-speech Synthesis exploiting Self-Supervised Representations*. **APSIPA 2023**

RobustL2S

TABLE II
PERFORMANCE COMPARISON IN CONSTRAINED-SPEAKER SETTING ON GRID-4S DATASET

Method	**STOI ↑**	**ESTOI ↑**	**WER ↓**
Vid2speech [13]	0.491	0.335	44.92 %
Lip2AudSpec [15]	0.513	0.352	32.51 %
1D GAN-based [17]	0.564	0.361	26.64 %
Vocoder-based [40]	0.648	0.455	23.33 %
Ephrat *et al.* [14]	0.659	0.376	27.83 %
Lip2Wav [5]	0.731	0.535	14.08 %
VAE-based [16]	0.724	0.540	-
VCA-GAN [19]	0.724	**0.609**	12.25 %
kim *et al.* [28], [46]	0.738	0.579	-
RobustL2S	**0.754**	0.571	**11.21** %

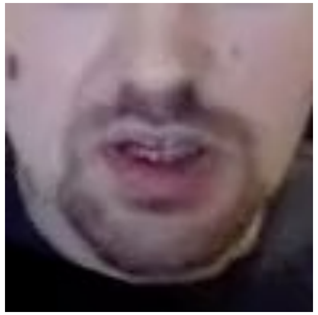
TABLE III
PERFORMANCE COMPARISON IN CONSTRAINED-SPEAKER SETTING ON TCD-TIMIT-3S DATASET

Method	**STOI ↑**	**ESTOI ↑**	**WER ↓**
Vid2speech [13]	0.451	0.298	75.52 %
Lip2AudSpec [15]	0.450	0.316	61.86 %
1D GAN-based [17]	0.511	0.321	49.13 %
Ephrat *et al.* [14]	0.487	0.310	53.52 %
Lip2Wav [5]	0.558	0.365	31.26 %
VCA-GAN [19]	0.584	0.401	-
RobustL2S	**0.596**	**0.452**	**29.03** %

TABLE IV
PERFORMANCE COMPARISON IN SPEAKER-DEPENDENT SETTING ON LIP2WAV DATASET

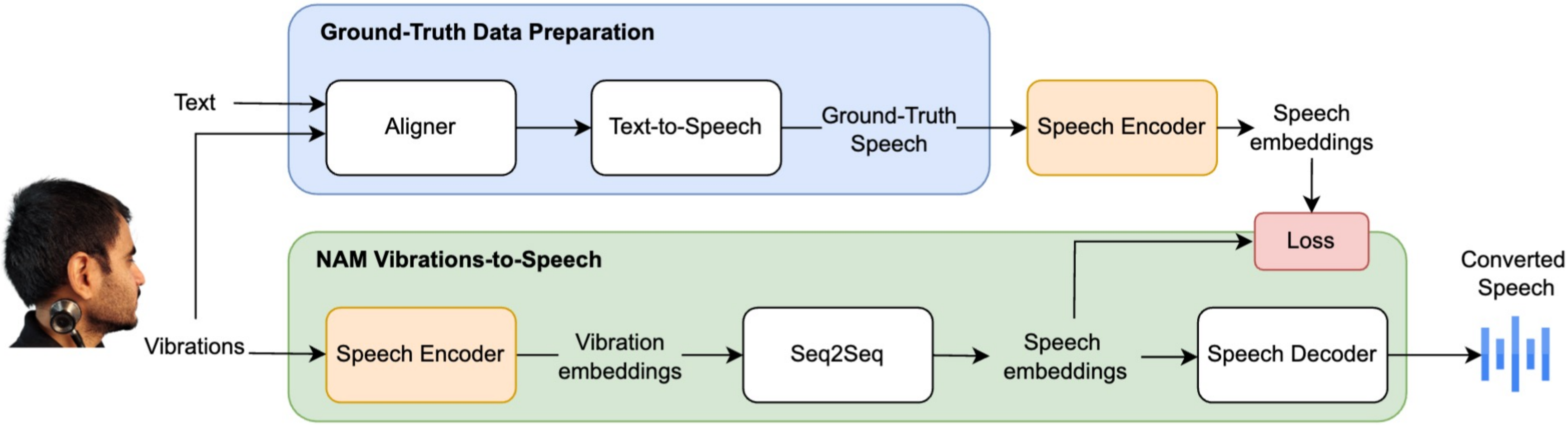
Speaker	**Method**	**STOI ↑**	**ESTOI ↑**
Chemistry Lectures (chem)	Ephrat *et al.* [5]	0.165	0.087
	GAN-based [47]	0.192	0.132
	Lip2Wav [5]	0.416	0.284
	Hong et al. [28]	0.566	**0.429**
	RobustL2S	**0.583**	0.397
Chess Analysis (chess)	Ephrat *et al.* [5]	0.184	0.098
	GAN-based [47]	0.195	0.104
	Lip2Wav [5]	0.418	0.290
	Hong et al. [28]	0.506	0.334
	RobustL2S	**0.517**	**0.340**
Deep Learning (dl)	Ephrat *et al.* [5]	0.112	0.043
	GAN-based [47]	0.144	0.070
	Lip2Wav [5]	0.282	0.183
	Hong et al. [28]	0.576	0.402
	RobustL2S	**0.627**	**0.419**
Hardware Security (hs)	Ephrat *et al.* [5]	0.192	0.064
	GAN-based [47]	0.251	0.110
	Lip2Wav [5]	0.446	0.311
	Hong et al. [28]	0.504	0.337
	RobustL2S	**0.511**	**0.337**
Ethical Hacking (eh)	Ephrat *et al.* [5]	0.143	0.064
	GAN-based [47]	0.171	0.089
	Lip2Wav [5]	0.369	0.220
	Hong et al. [28]	0.463	**0.304**
	RobustL2S	**0.493**	0.277

RobustL2S

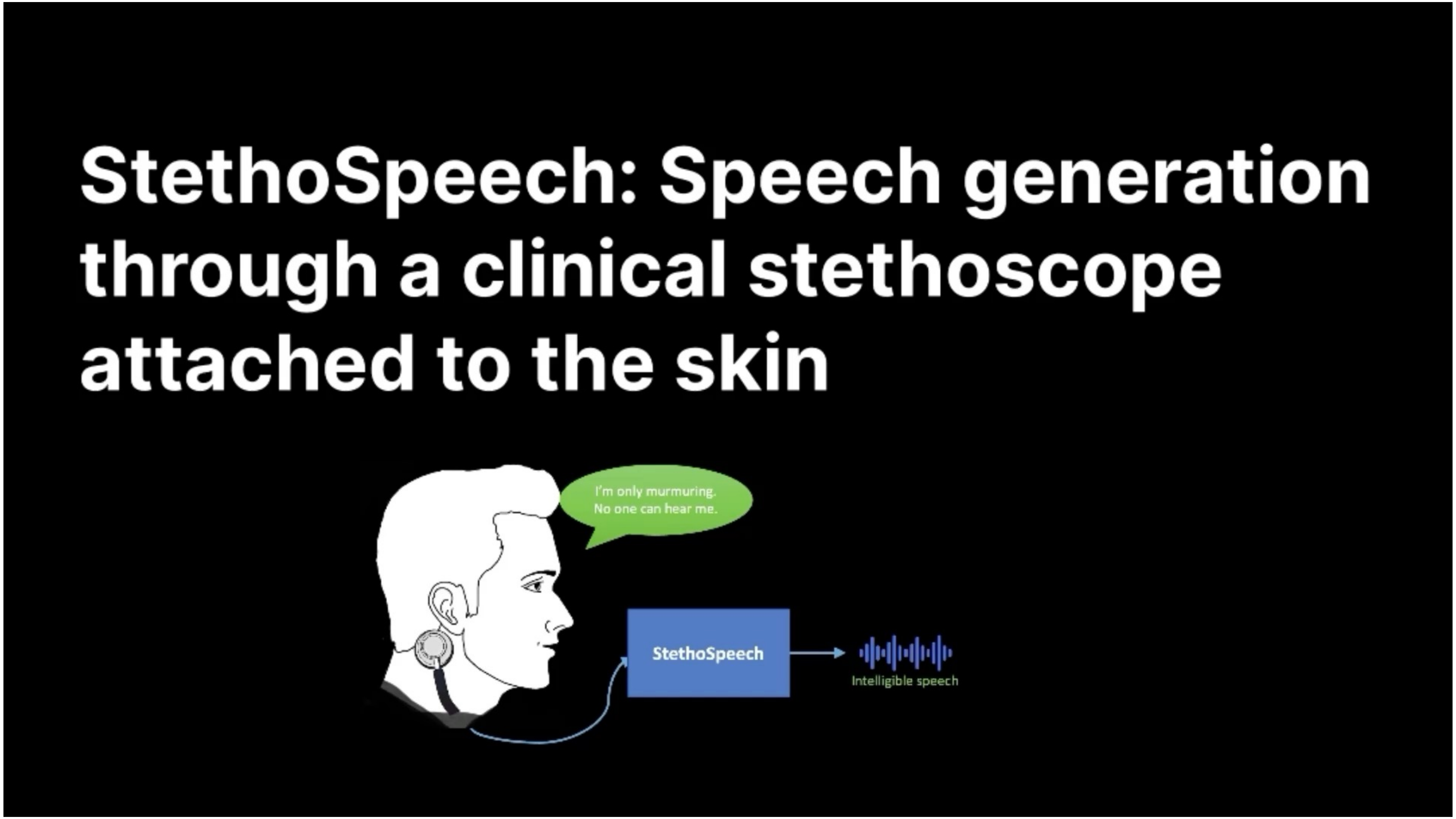


Sahipjohn et al. *RobustL2S: Speaker-Specific Lip-to-speech Synthesis exploiting Self-Supervised Representations*. **APSIPA 2023**

StethoSpeech



StethoSpeech



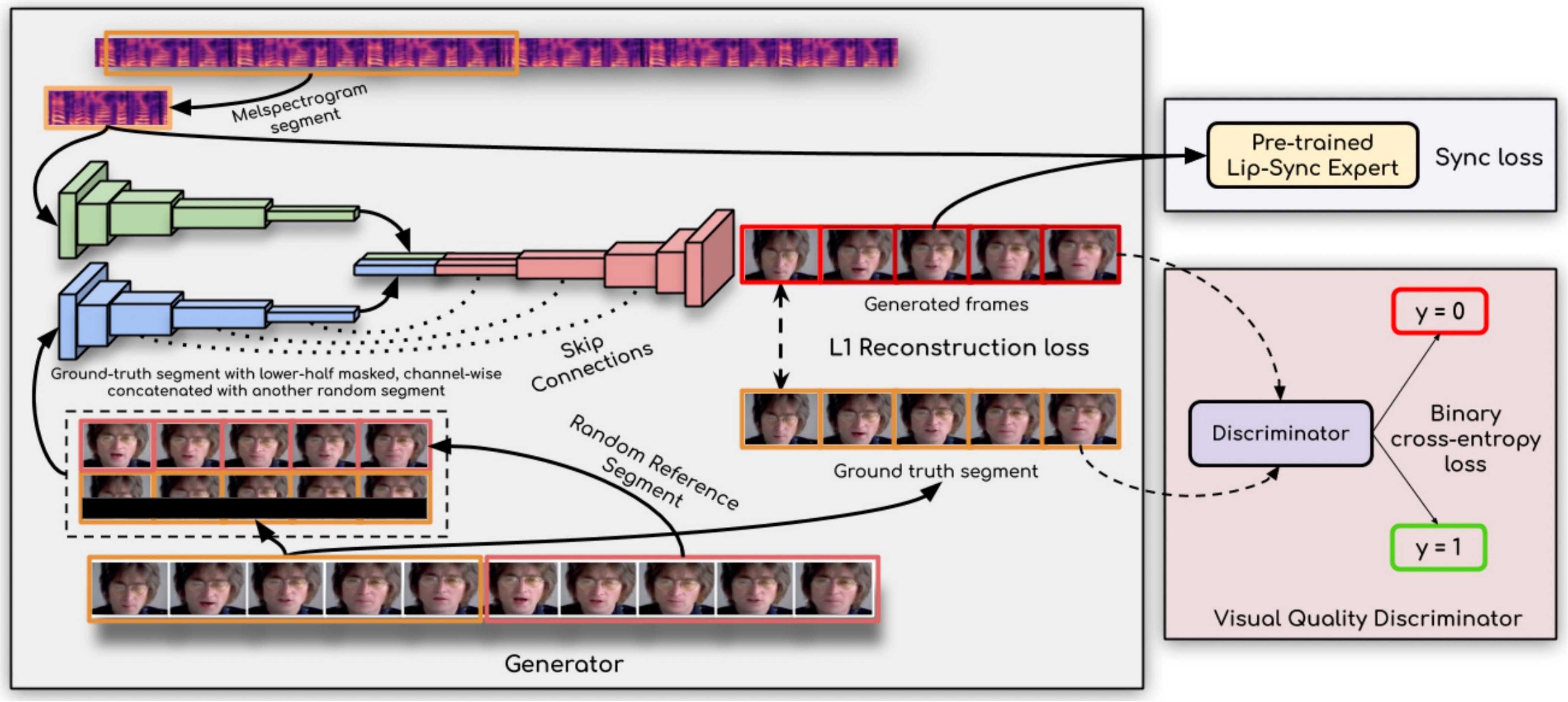
Tongue Ultrasound to speech

Prediction

Ground Truth

Many fun/useful ideas and possibilities

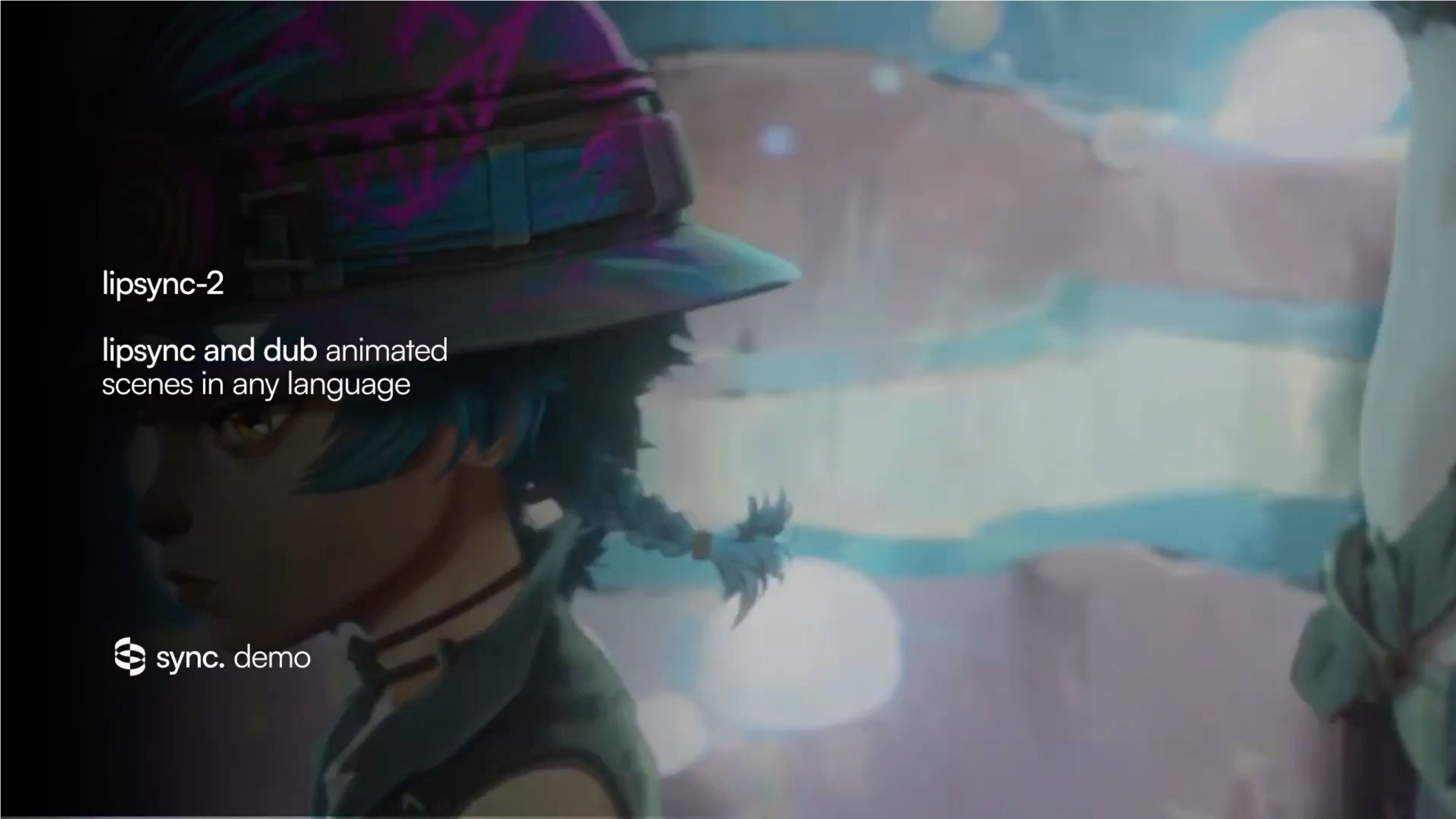
Wav2Lip



Prajwal et al. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. ACM Multimedia 2020

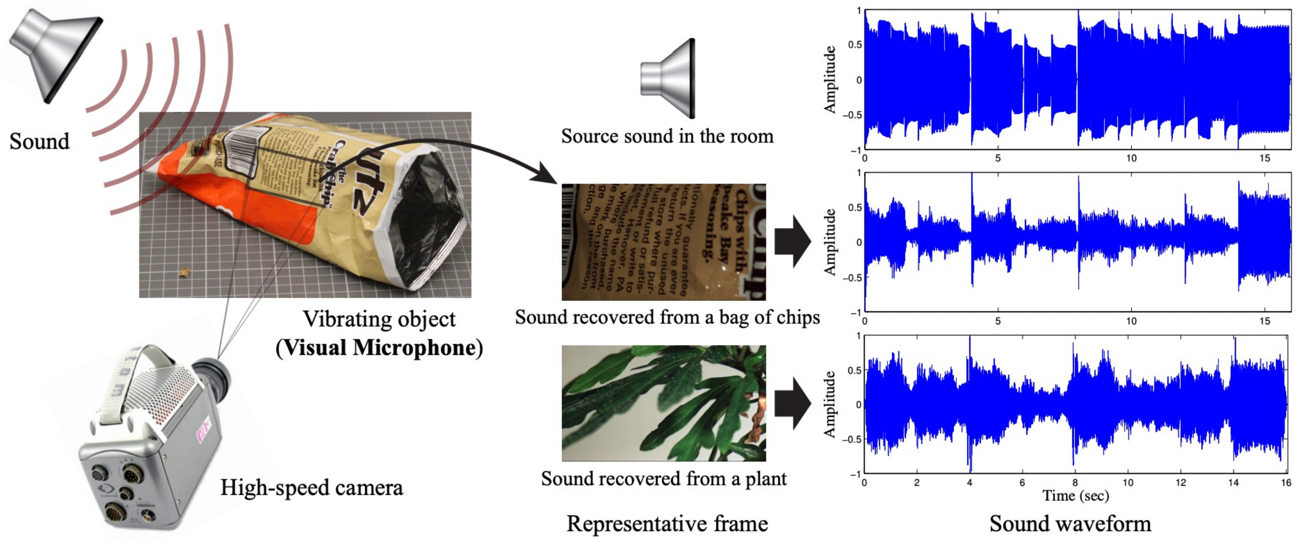
preserve **speaker style**
while lipsyncing

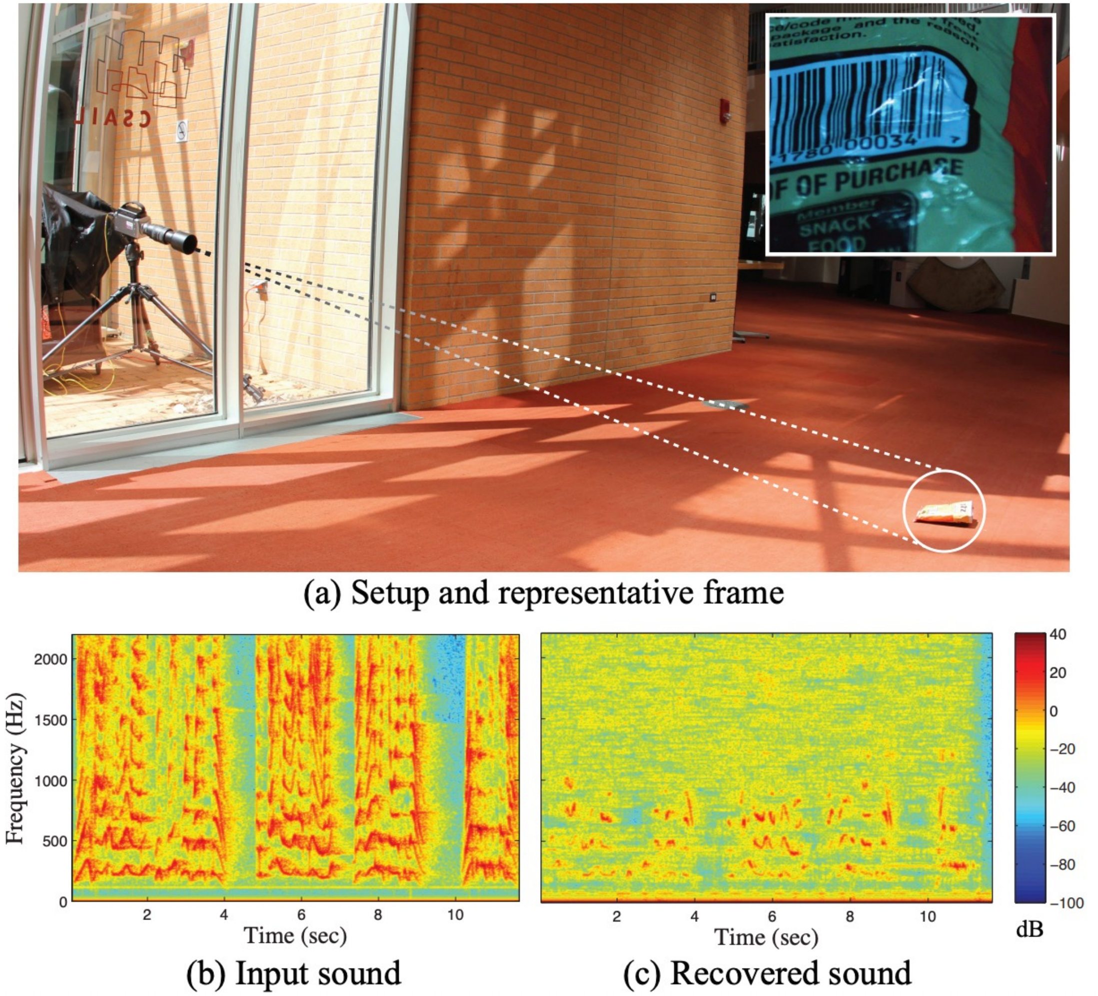
sync. demo


lipsync-2
lipsync and dub animated scenes in any language
sync. demo

Avatars: Synthesia and others

Visual microphone





Davis et al. The Visual Microphone: Passive Recovery of Sound from Video. ACM Transactions on Graphics. 2014

The Visual Microphone: Passive Recovery of Sound from Video

Abe Davis
Michael Rubinstein
Neal Wadhwa
Gautham J. Mysore
Fredo Durand
William T. Freeman

Automated Cinematography

Original (Wide Angle Static)

Edited

Moorthy et al. GAZED - Gaze-guided Cinematic Editing of Wide-Angle Monocular Video Recordings. CHI 2020
Girmaji et al. EditIQ: Automated Cinematic Editing of Static Wide-Angle Videos via Dialogue Interpretation and Saliency Cues. IUI 2025

AudioSet

There are 2,084,320 YouTube videos containing 527 labels

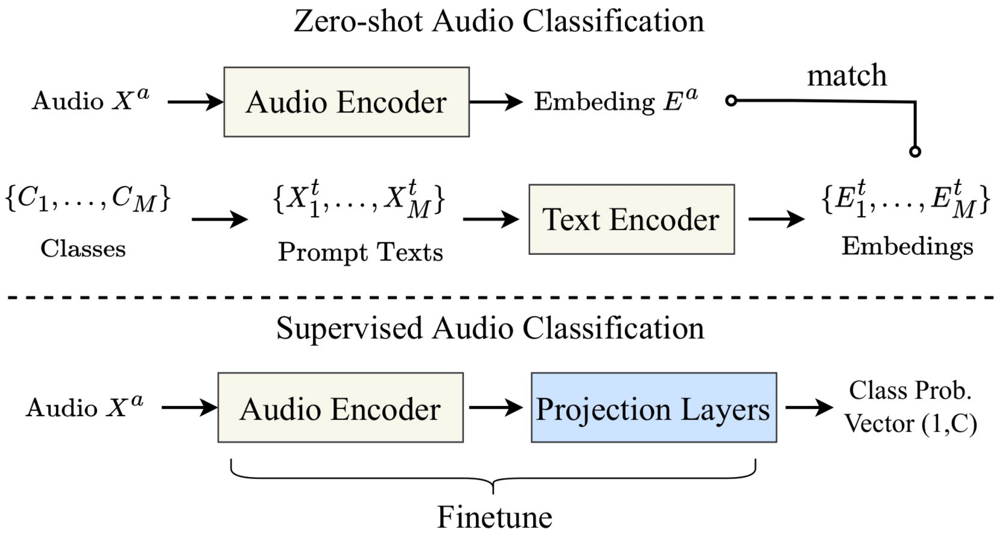
Type a sound to filter the dataset

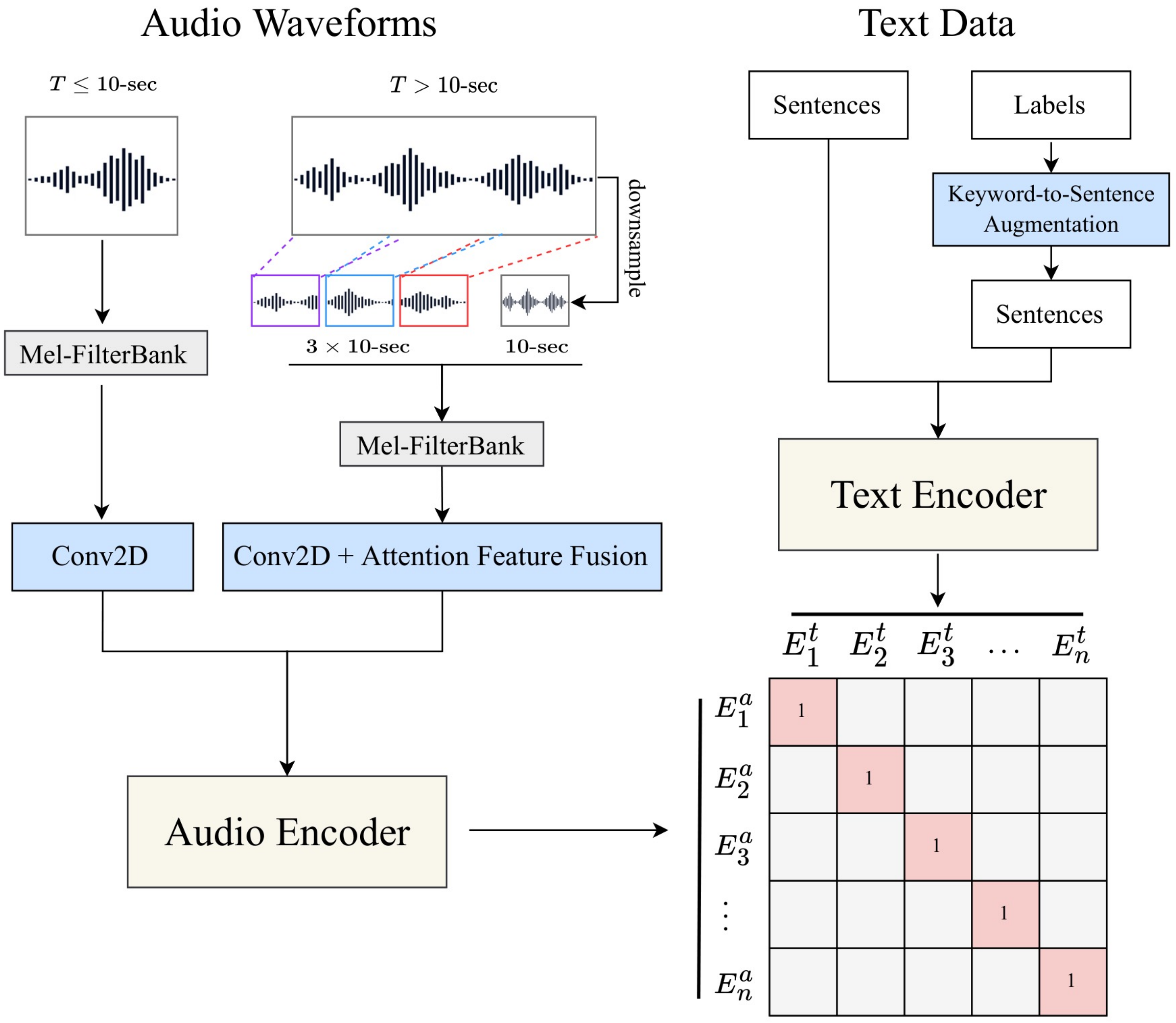
Show detailed breakdown?

Label	Quality estimate?	Number of videos
Music	100%	1,011,305
Speech	100%	1,010,480
Vehicle	100%	128,051
Musical instrument	100%	117,343
Plucked string instrument	100%	44,565
Singing	100%	42,493
Car	100%	41,554
Animal	100%	40,758
Outside, rural or natural	100%	35,731
Violin, fiddle	100%	28,125
Bird	100%	26,894
Drum	100%	20,246
Engine	100%	16,245
Narration, monologue	100%	15,590
Drum kit	100%	15,169
Acoustic guitar	100%	14,568
Dog	100%	13,705
Child speech, kid speaking	100%	11,816
Bass drum	100%	9,292

https://research.google.com/audioset/

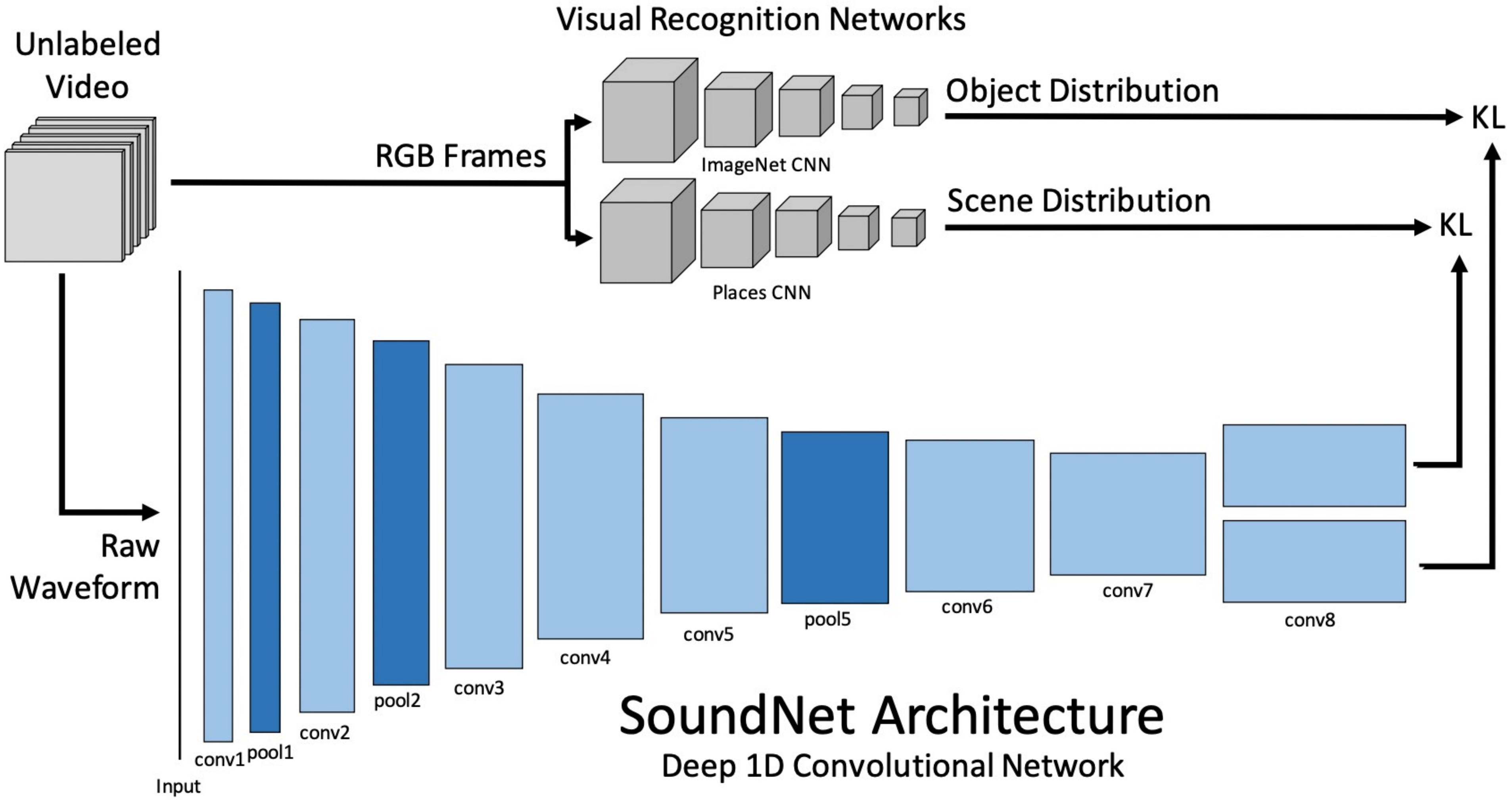
CLAP model





Yusng et al. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. ICASSP 2023

Multimodal distillation



Aytar et al. SoundNet: Learning Sound Representations from Unlabeled Video. Neurips 2016

Thank you!