Self-Supervised Learning for Speech Representations

Automatic Speech Processing, Master Cycle

Motivation

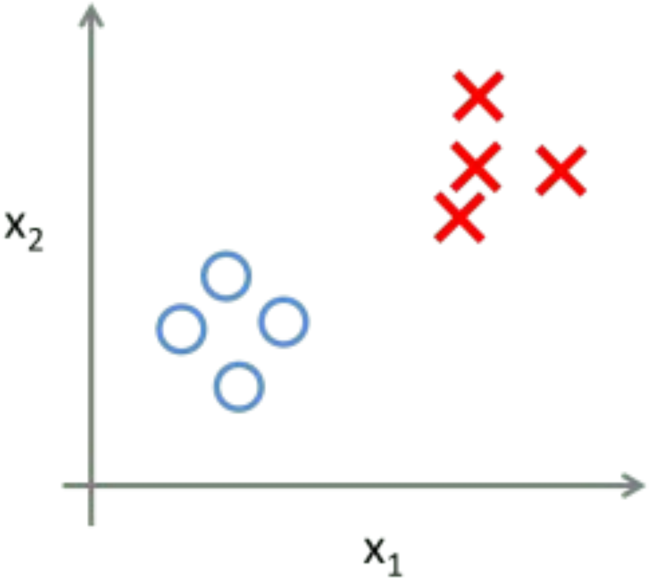
Speech datasets with transcriptions can be scarce / costly to create, depending on the task

Meanwhile, there is an abundance of unlabeled speech data

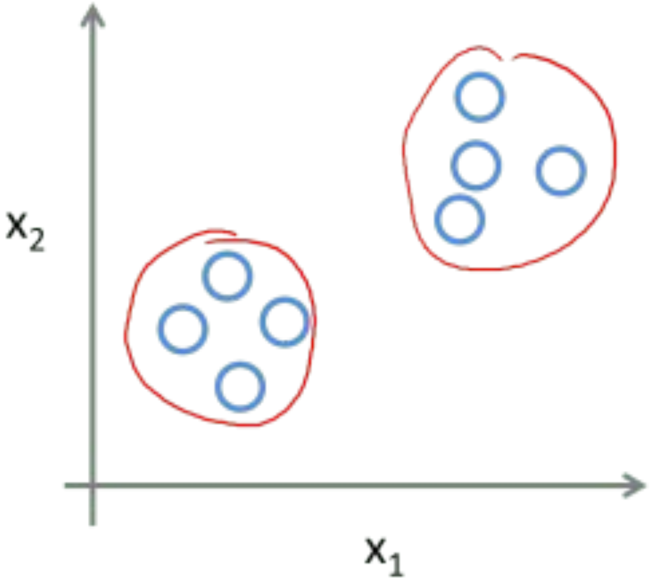
How can we benefit from unlabeled data?

→ Self-supervised Learning

Supervised vs. Unsupervised Learning



<u>Supervised</u>
Learn from labeled examples



<u>Unsupervised</u>
Find structure in unlabeled data

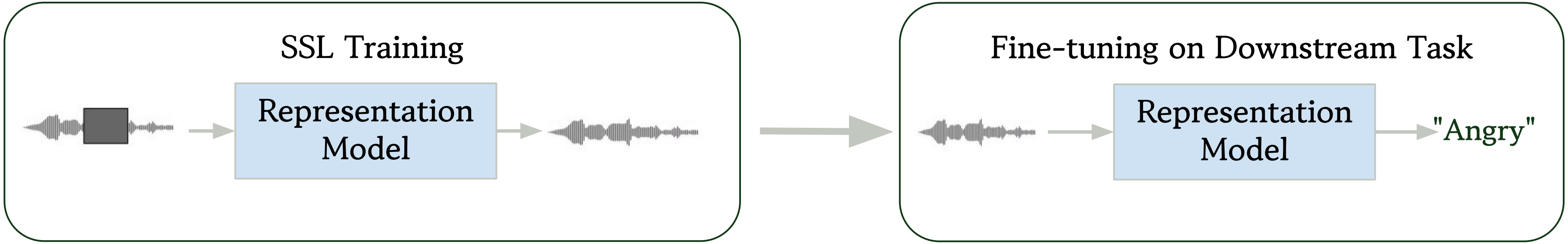
Figure from https://lakshaysuri.wordpress.com/2017/03/19/machine-learning-supervised-vs-unsupervised-learning/

Self-Supervised Learning (SSL)

Data is still unlabeled but we design a pretext task to generate pseudo-labels.

A supervision signal is created from the unlabeled data itself

Goal: Learn useful speech representations which can perform well for a wide-range of downstream tasks



Typical Speech SSL Model Architecture

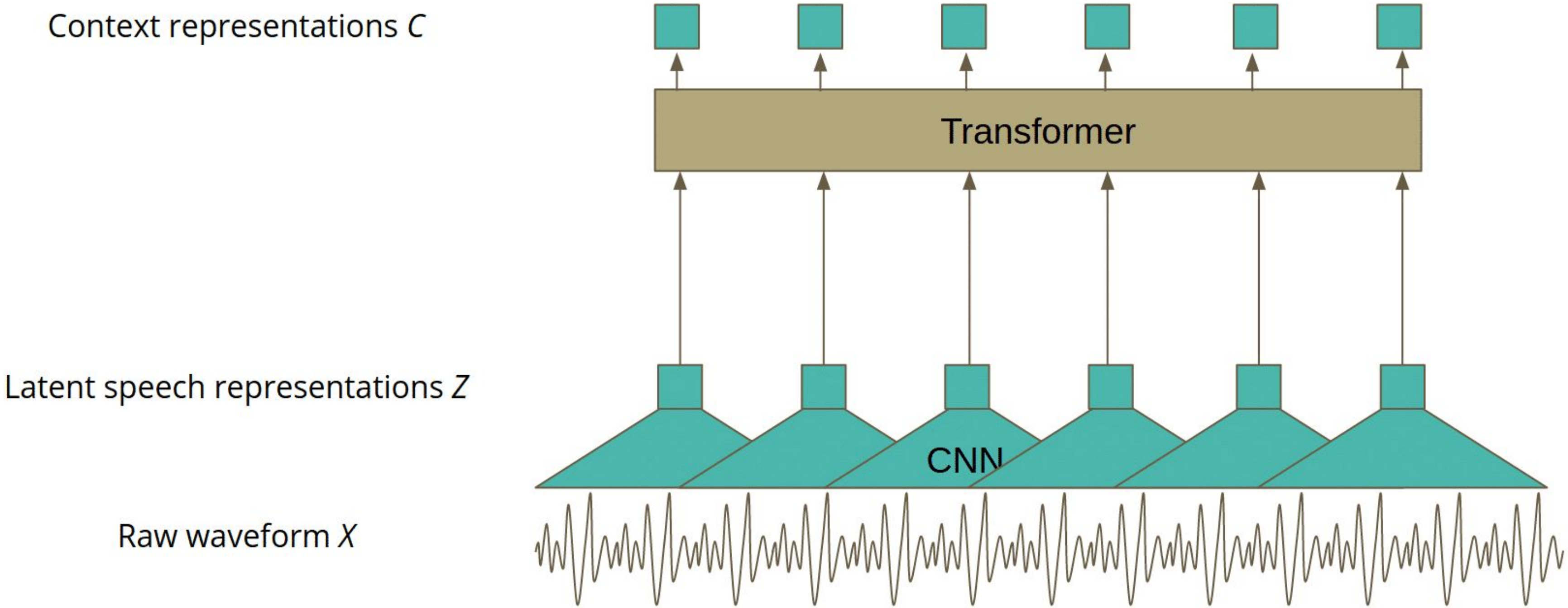
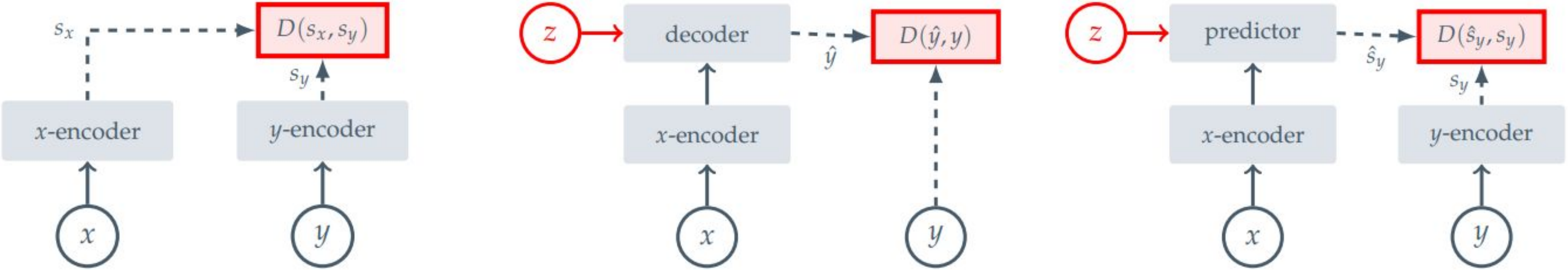


Figure from https://neurosys.com/blog/wav2vec-2-0-framework

SSL Training Architectures



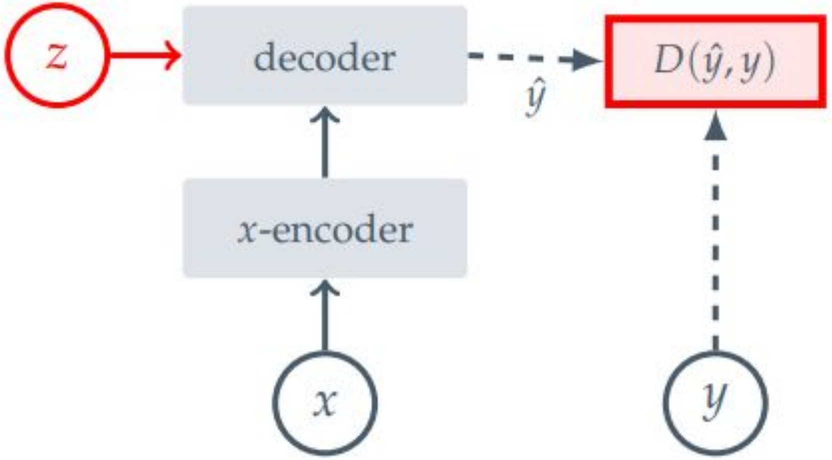
(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

Assran, Mahmoud, et al. "Self-supervised learning from images with a joint-embedding predictive architecture." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Generative Architecture

Learns to directly reconstruct a signal y from a compatible signal x

Decoder conditioned on additional variables z to facilitate reconstruction



(b) **Generative Architecture**

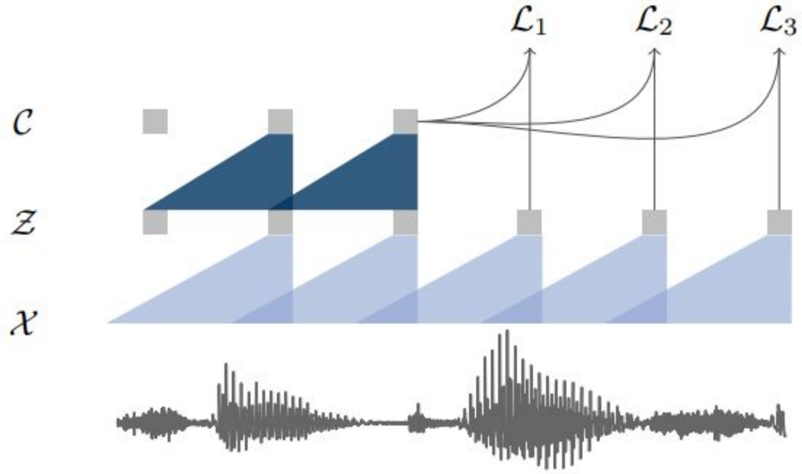
Generative Architecture

Wav2vec

Solves a next time-step prediction task

Predicts input features (raw audio encoded by the CNN encoder network)

<u>Objective</u>: Learns to distinguish a sample z_{i+k} that is k steps into the future from distractor samples drawn from a proposal distribution



Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition." arXiv preprint arXiv:1904.05862 (2019).

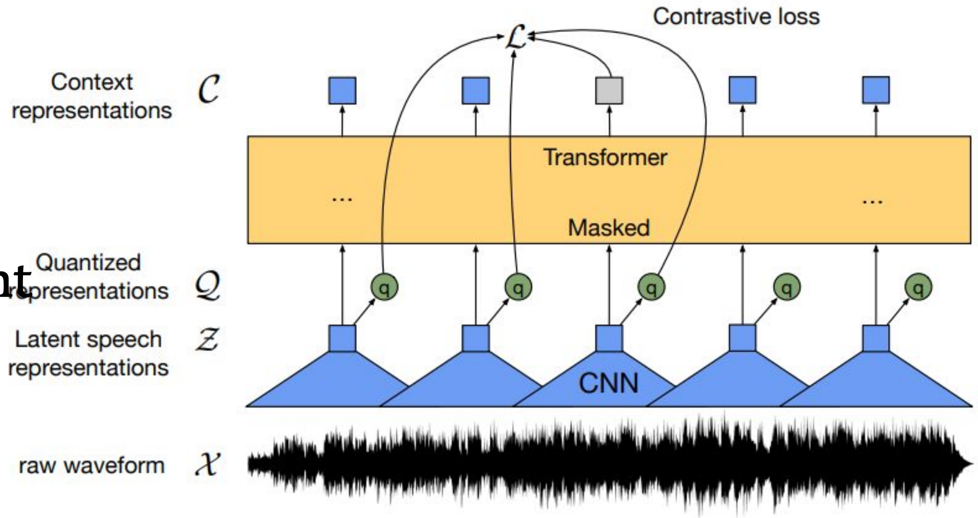
Generative Architecture

Wav2vec2.0

Solves a contrastive task defined over a quantization of the latent representations z

Masks the speech input in the latent space (CNN encoder output)

<u>Objective</u>: Learns to distinguish a true latent from distractors using quantized targets



Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

Generative Architecture

Bert (Text)

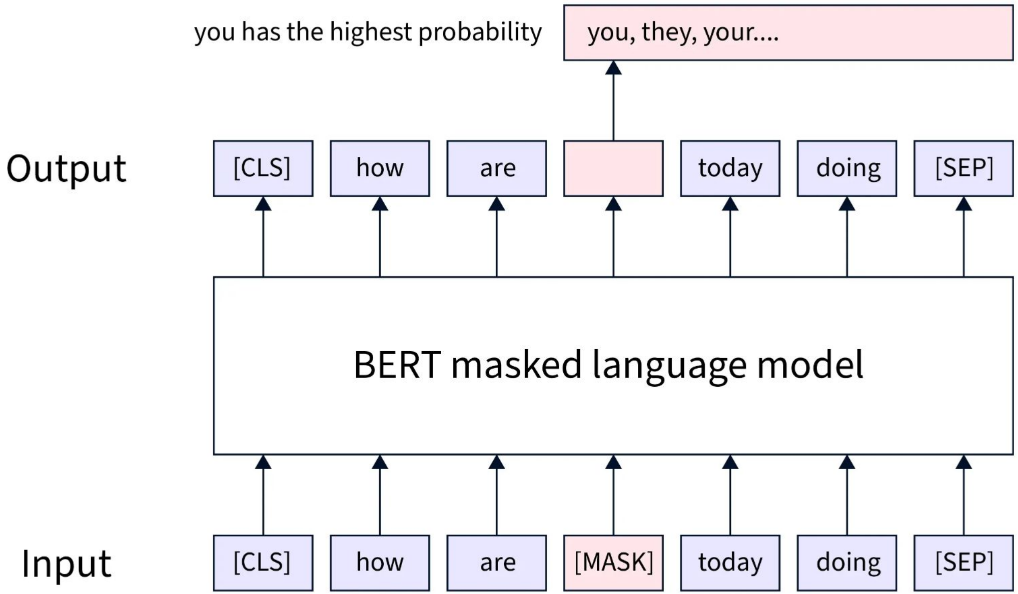


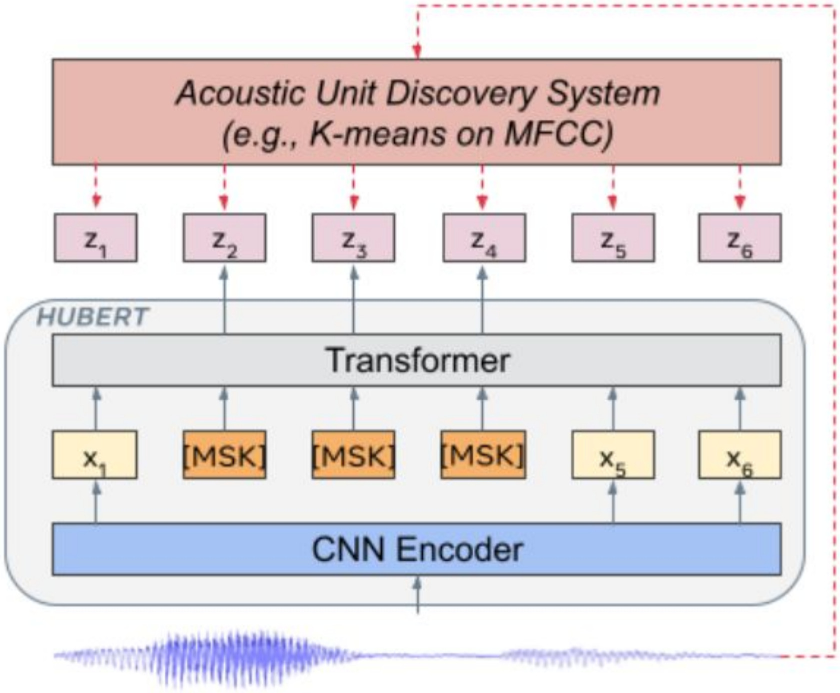
Figure from: https://www.scaler.com/topics/nlp/pre-training-bert/

Generative Architecture

HuBERT

Uses offline clustering to provide aligned target labels for a BERT-like prediction loss

Prediction loss is applied over the masked regions only



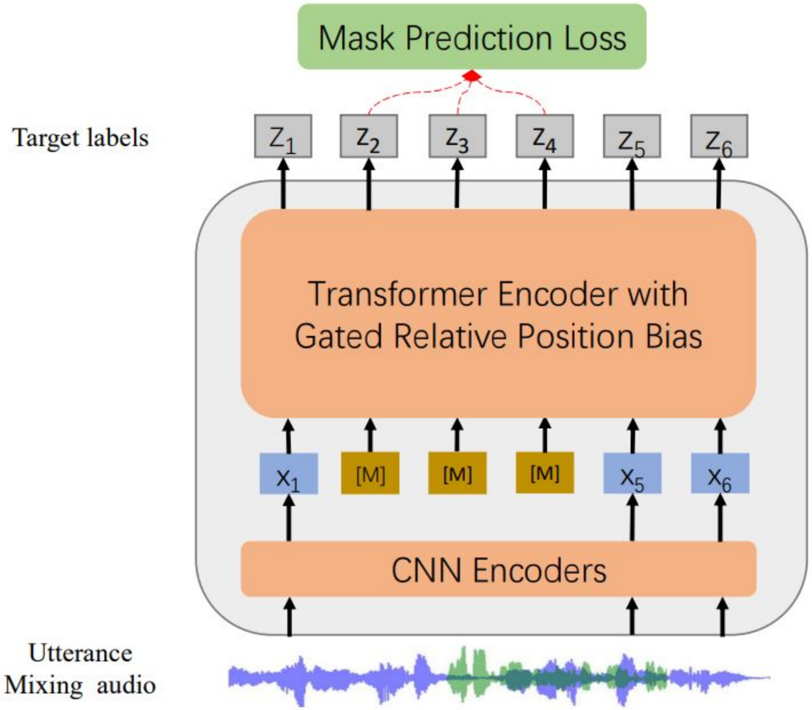
Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM transactions on audio, speech, and language processing 29 (2021): 3451-3460.

Generative Architecture

WavLM

Jointly learns masked speech prediction and denoising

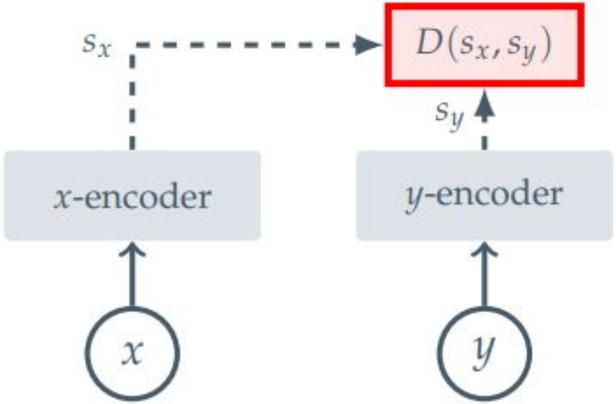
Diversifies data sources for better generalization (vs. other models which are trained on podcast data only)



Chen, Sanyuan, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1505-1518.

Joint-Embedding Architecture

Learns to output similar embeddings for compatible inputs x, y and dissimilar embeddings for incompatible inputs

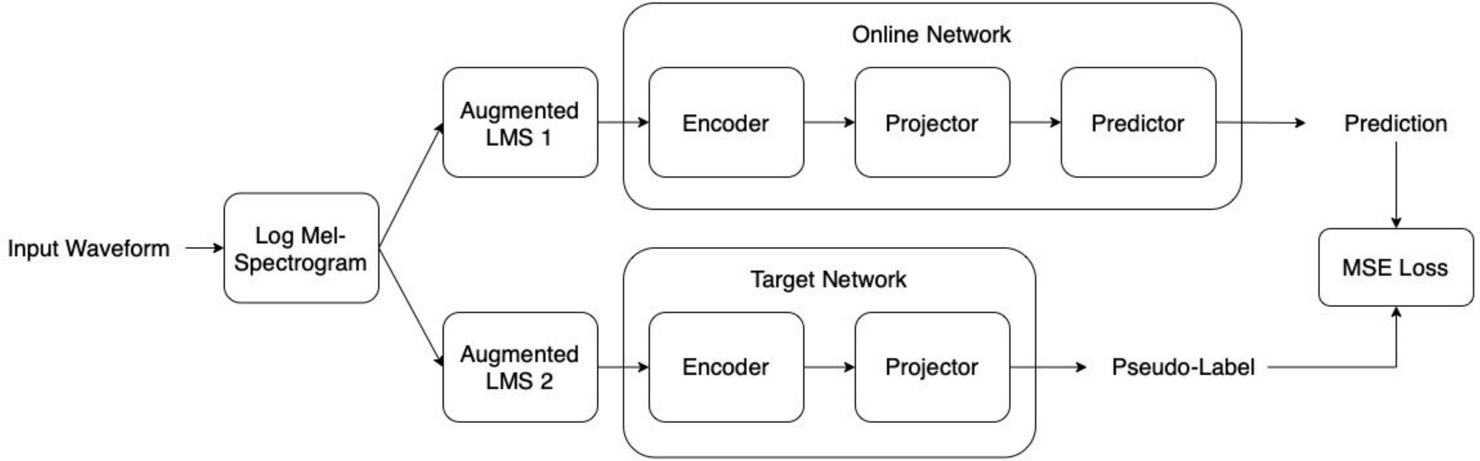


(a) **Joint-Embedding Architecture**

Joint-Embedding Architecture

BYOL

From an augmented view of an input waveform, we train the online network to predict the target network representation of the same image under a different augmented view

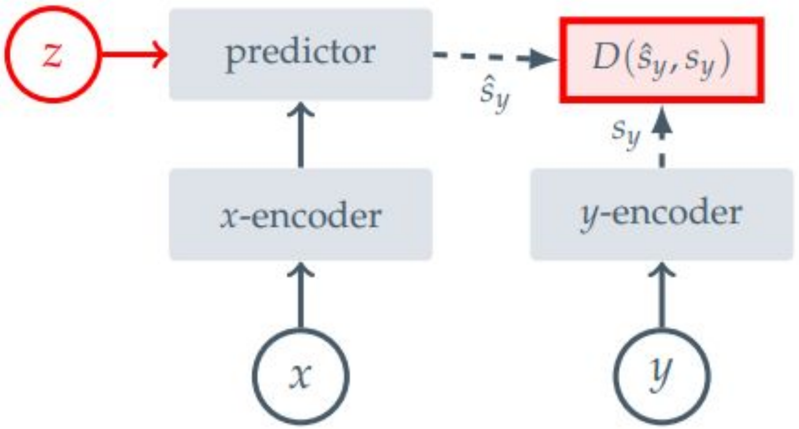


Elbanna, Gasser, et al. "Byol-s: Learning self-supervised speech representations by bootstrapping." HEAR: Holistic Evaluation of Audio Representations. PMLR, 2022.

Joint-Embedding Predictive Architecture

Learns to predict the embeddings of a signal y from a compatible signal x

Predictor conditioned on additional variables z to facilitate prediction



(c) **Joint-Embedding Predictive Architecture**

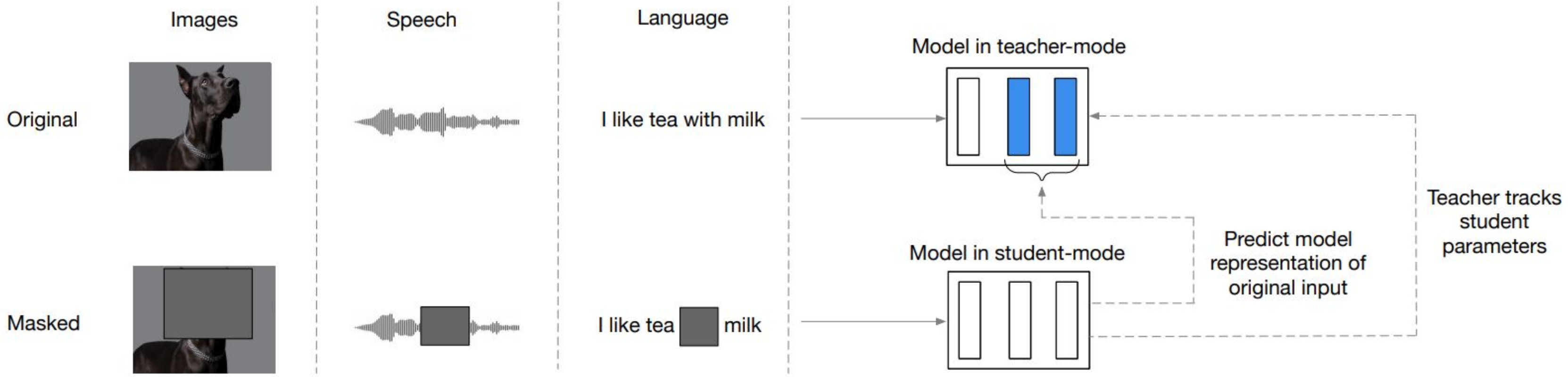
Joint-Embedding Predictive Architecture

Data2vec

Combines masked prediction with the learning of latent target representations

Teacher: builds representation of the full input data which serves as targets in the learning task

Student: encodes a masked version of the input sample with which we predict the full data representations



Baevski, Alexei, et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language." International conference on machine learning. PMLR, 2022.

Why does SSL work?

Can make use of very large amounts of data

Transformer architecture is well suited to learn contextual information

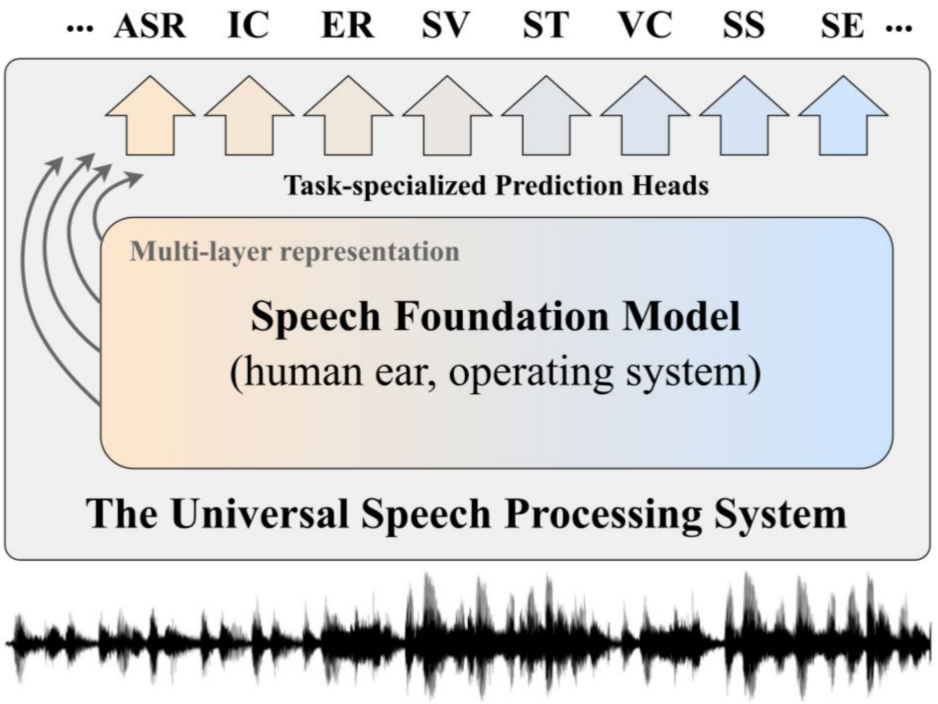
Pre-text tasks allow learning representations that encode general information that is useful for a wide-range of tasks

For example, to predict what's missing, the model must grasp the structure of the data
→ The model builds rich internal representations

SSL Model Evaluation

Models are compared to each other by evaluating the representations across a wide variety of tasks

Benchmarks have been set up to perform this evaluation



Yang, Shu-wen, et al. "A large-scale evaluation of speech foundation models." IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024): 2884-2899.

SUPERB Benchmark

Category	Tasks
Content	Phoneme Recognition (PR), Automatic Speech Recognition (ASR), Out-of-Domain ASR (OOD-ASR), Keyword Spotting (KS), Query-by-Example (QbE)
Speaker	Speaker Identification (SID), Speaker Verification (SV), Speaker Diarization (SD)
Prosody	Emotion Recognition (ER)
Semantics	Intent Classification (IC), Slot Filling (SF), Speech Translation (ST)
Generation	Voice Conversion (VC), Source Separation (SS), Speech Enhancement (SE)

Yang, Shu-wen, et al. "A large-scale evaluation of speech foundation models." IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024): 2884-2899.

HEAR Benchmark

Category	Tasks
Speech	Speech Command Classification; Emotion Recognition; Language ID; Speaker Count
Music	Pitch Classification; Music Genre Classification; Music/Speech Classification; Music Transcription; Percussion Classification; Instrument Classification
Environmental / Other Audio	Office Sound Detection; Environmental Sound Classification; General Audio Tagging; Beehive Condition Classification; Gunshot Triangulation; Imitated Sound Type Classification

Turian, Joseph, et al. "Hear: Holistic evaluation of audio representations." NeurIPS 2021 Competitions and Demonstrations Track. PMLR, 2022.

Leaderboards

https://superbbenchmark.github.io/#/leaderboard

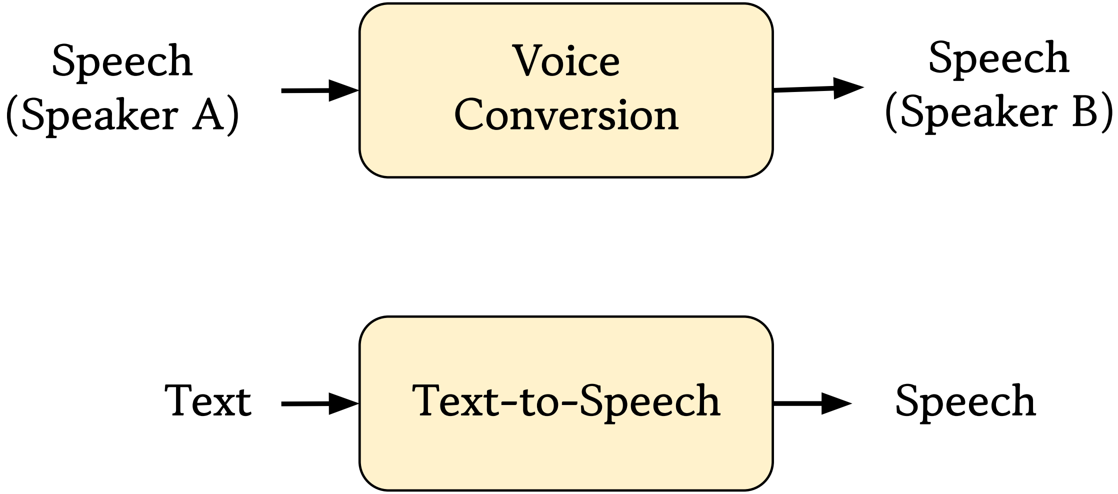
https://hearbenchmark.com/hear-leaderboard.html

Turian, Joseph, et al. "Hear: Holistic evaluation of audio representations." NeurIPS 2021 Competitions and Demonstrations Track. PMLR, 2022.

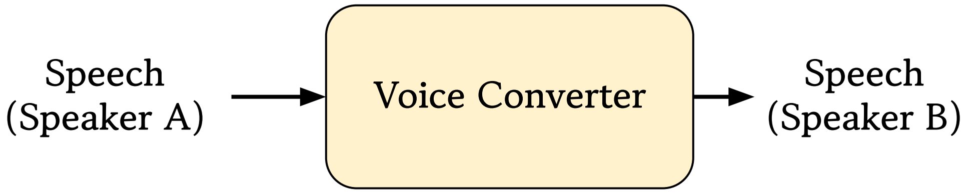
SSL for Synthesis

In a synthesis task, we generate new speech waveforms given inputs

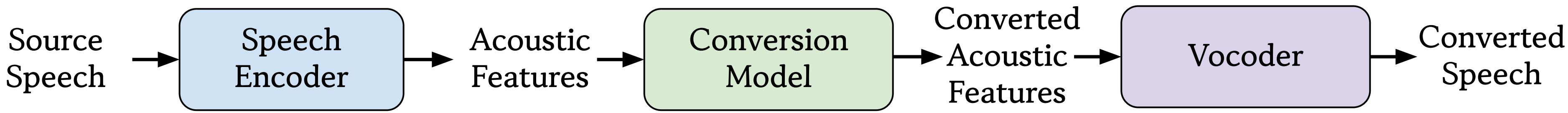
Examples:



Voice Conversion



Typical Pipeline:



Copy-Synthesis

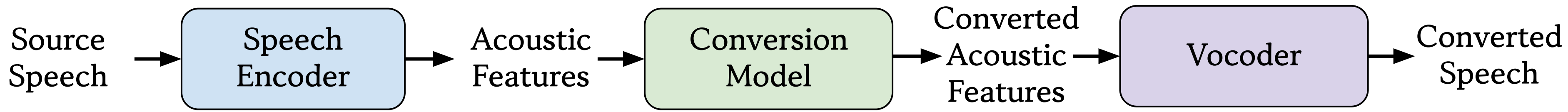
	Features	Vocoders
Signal Processing-based Examples	Mel-spectrogram, MFCCs	Griffin-Lim, WORLD
Neural Examples	SSL Representations	HiFiGAN, BigVGAN

Voice Conversion

How can we leverage the strengths of SSL representations for voice conversion?

SSL models encode speech into a sequence of frames, each corresponding to a window of 25ms of speech, with a 20ms hop between windows.

<u>Property:</u> frames are linearly close if they contain similar phonetic, linguistic information, even if they are spoken by very different voices

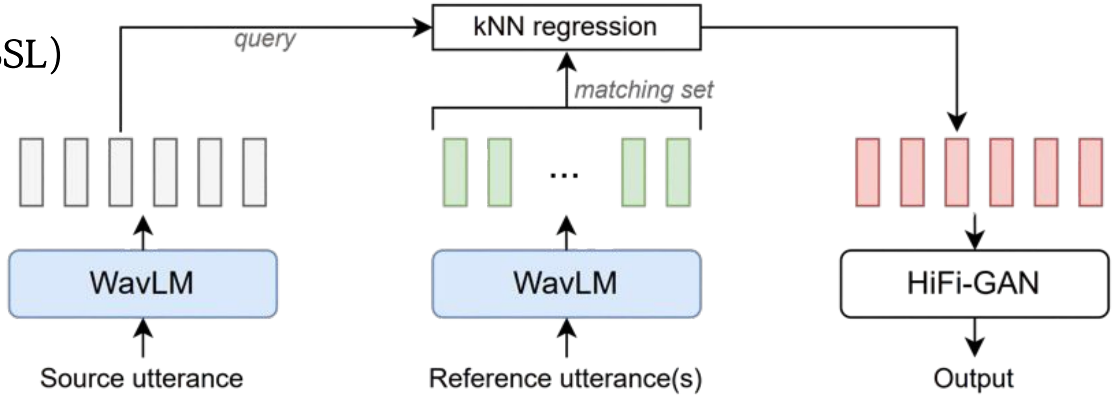


Voice Conversion: kNN-VC

Any-to-Any VC method

Leveraging self-supervised learning (SSL) features for zero-shot conversion

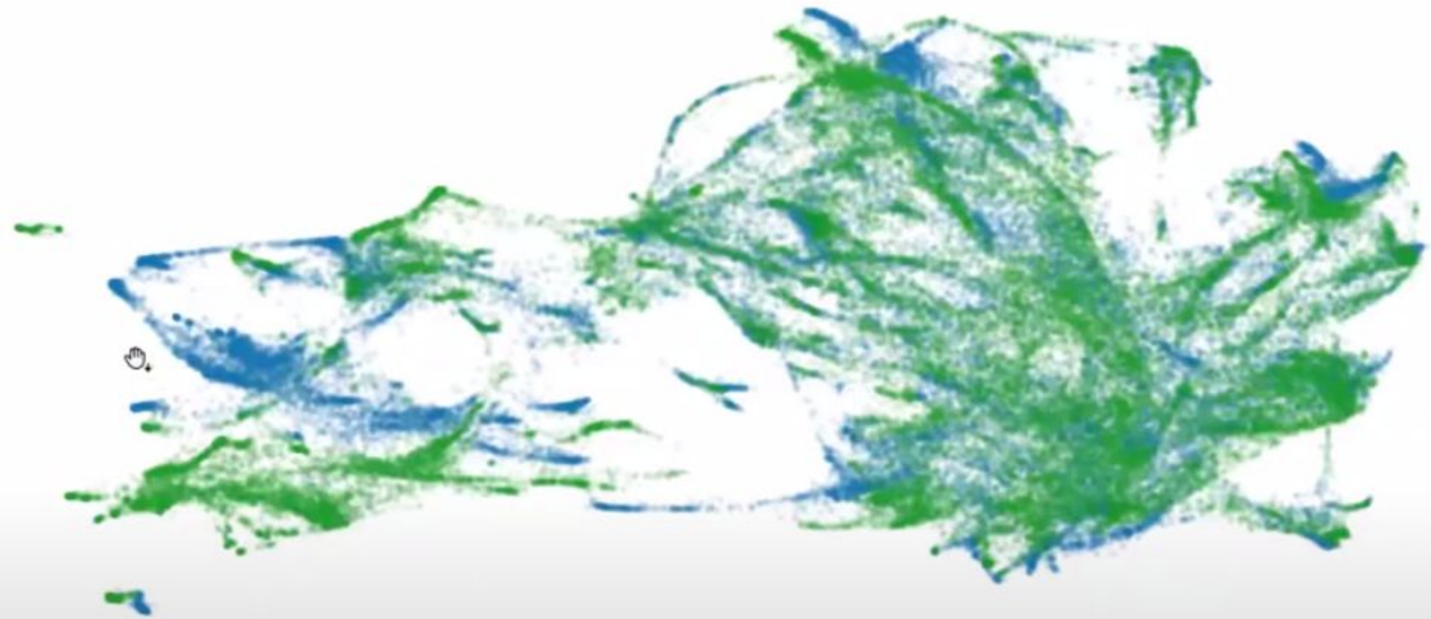
Requires no training



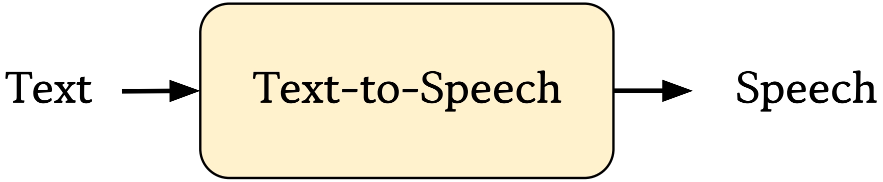
Source	*Target*	*Converted*

Baas, M., van Niekerk, B., Kamper, H. (2023) Voice Conversion With Just Nearest Neighbors. Proc. Interspeech 2023, 2053-2057

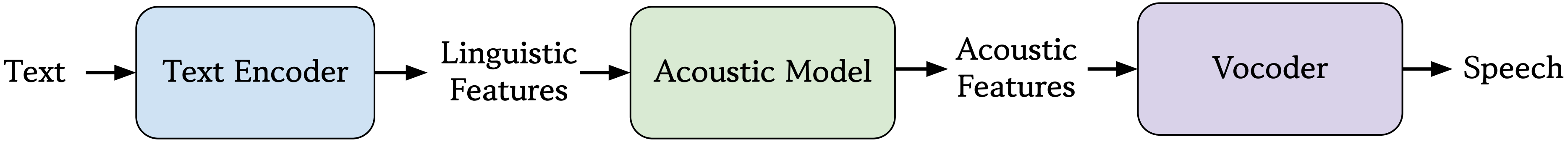
3D Projection of WavLM Features for Two Speakers



Text-to-Speech

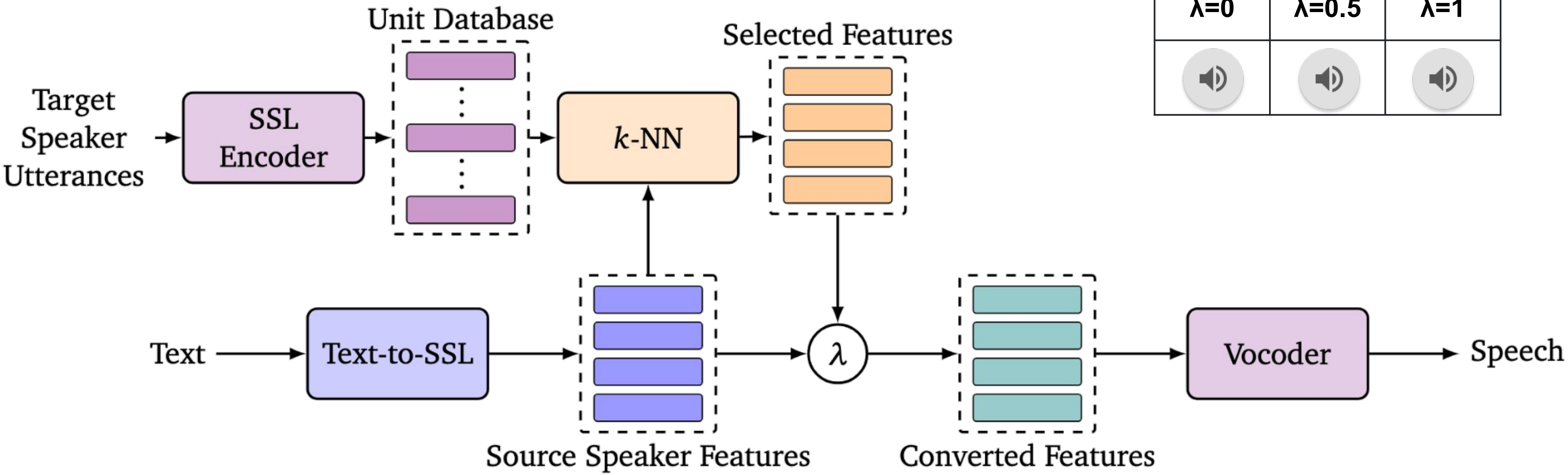


Typical Pipeline:



kNN-TTS

λ=0	λ=0.5	λ=1
🔊	🔊	🔊



$$y_{\text{converted}} = \lambda\, y_{\text{selected}} + (1-\lambda)\, y_{\text{source}}$$

Hajal, Karl El, et al. "kNN retrieval for simple and effective zero-shot multi-speaker text-to-speech." Proceedings of the Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL) (2025).

Summary

Self-supervised learning enables learning useful general representations from unlabeled data

Evaluations show that SSL representations enable performance improvements for wide range of tasks;

→ Especially whenever labeled data is scarce

As seen in Voice Conversion for example, good representations can simplify task-specific models