A Brief Overview of Unsupervised Neural Speech Representation Learning

Lasse Borgholt, [1,3] **Jakob D. Havtorn,** [2,3] **Joakim Edin,** [3] **Lars Maaløe,** [2,3] **Christian Igel** [1, 4]

[1]Department of Computer Science, University of Copenhagen, Denmark
[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[3]Corti, Copenhagen, Denmark
[4]Pioneer Centre for AI, Copenhagen, Denmark
borgholt@di.ku.dk

Abstract

Unsupervised representation learning for speech processing has matured greatly in the last few years. Work in computer vision and natural language processing has paved the way, but speech data offers unique challenges. As a result, methods from other domains rarely translate directly. We review the development of unsupervised representation learning for speech over the last decade. We identify two primary model categories: self-supervised methods and probabilistic latent variable models. We describe the models and develop a comprehensive taxonomy. Finally, we discuss and compare models from the two categories.

1 Introduction

Representation learning has shaped modern computer vision (Simonyan and Zisserman 2015) and natural language processing (Devlin et al. 2019), and more recently speech processing has been subject to the same development (Baevski et al. 2020). Representation learning has been defined as *"learning representations of the data that make it easier to extract useful information when building classifiers or other predictors"* (Bengio, Courville, and Vincent 2013). Unsupervised representation learning is concerned with learning useful representations without the use of human annotations. Usually, a model is first pre-trained on a task where plenty of data is available. The model is then fine-tuned, or used to extract input representations for a smaller model, targeting a task with limited training data. In computer vision, both supervised (Simonyan and Zisserman 2015; Szegedy et al. 2015; He et al. 2016) and unsupervised (Pathak et al. 2016; Doersch, Gupta, and Efros 2015) representation learning have gained attention with supervised representation learning driven by the availability of large annotated datasets (Deng et al. 2009). For text and speech, pre-training is usually unsupervised as labeled data is difficult to obtain. Although work on text has paved the way, and the two fields share many characteristics, learning representations from speech is a problem faced with a unique set of challenges.

In this paper, we survey work on unsupervised representation learning for speech processing from within the last decade. From a methodological perspective, we identify two primary model categories, namely models based on self-supervised learning and probabilistic latent variable models. We provide a methodological review of the design choices related to each of the model categories and develop a model taxonomy that highlights the different directions of work. Finally, we compare and discuss models from the two categories and their respective evaluation procedures.

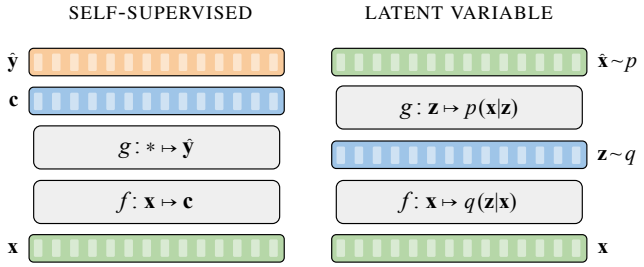

Figure 1: A schematic overview of the two groups of models covered in this survey. *Left*: A model trained with self-supervised learning. We take these models to consist of two functions $f(\cdot)$ and $g(\cdot)$ (sec. 2). After pre-training, $f(\cdot)$ is fine-tuned or used for extracting features $\mathbf{c}$. $g(\cdot)$ is an auxiliary function used to accommodate the self-supervised pre-training task. *Right*: A probabilistic latent variable model. In contrast to the self-supervised model, the functions $f(\cdot)$ and $g(\cdot)$ learn the parameters of distributions q and p. The latent variable $\mathbf{z}$ is commonly used for representation learning.

2 Unsupervised representation learning

In the following, we group previous work into *self-supervised models* and *probabilistic latent variable models*, and take a *model* to comprise a neural architecture and a corresponding learning algorithm. A schematic overview is found in figure 1. These categories are neither exhaustive nor mutually exclusive, but allow us to focus on the characteristics that have shaped different branches of research.

With emphasis on recent successes in the field, we cover literature from the last 10 years. While a complete description of all relevant models is not within the scope of this work, we sketch important technicalities when they are particularly descriptive of certain models. We first define our high-level notation and conventions to ease discussion.

Notation We use the subscript $i{:}j$ with $i{\leq}j$ to denote a vector sequence $\mathbf{a}_{i:j}$ containing elements $\mathbf{a}_i$ through $\mathbf{a}_j$. We denote model input as $\mathbf{x}_{1:T}$ which, in practice, might be either a spectrogram or the raw speech signal, but we do not distinguish between the two in notation as it is not essential to understand the models. Also, models commonly downsample the temporal dimension, but again, this is not crucial to understand the models, so we maintain a notation based on a single temporal dimension $t \in \{1, \ldots, T\}$.

When discussing self-supervised models, we use $\mathbf{c}_{1:T}$ to denote a contextualized representation. For stochastic latent variable models, we use $\mathbf{z}_{1:T}$ as is customary to the field. While some models are frozen and produce representations used as input for downstream tasks (**FRZ**, table 1), others are designed to be fine-tuned (**FTN**, table 1). In either case, we use $f(\cdot)$ to denote the model that is used for the downstream task. We use $g(\cdot)$ to denote any auxiliary model components (e.g., for a reconstruction task we might have $g : \mathbf{c}_t \mapsto \hat{\mathbf{x}}_t$). When a model can be naturally subdivided into multiple components, we simply use $f_*(\cdot)$ where $*$ may be any convenient descriptor. Finally, we often use a subscript when defining a loss, $\mathcal{L}_i$, to imply that the total loss is computed as a sum over i.

2.1 Self-supervised models

Self-supervised learning is a subset of unsupervised learning (Tsai et al. 2020). Where other unsupervised methods can be seen as a means to an end in itself (e.g., clustering or data generation), self-supervised learning takes the form of a pretext task that only adds value when associated with a downstream task. This makes self-supervised learning tie naturally with semi-supervised learning, but it may also be part of a fully unsupervised setup (Baevski et al. 2021). Self-supervised learning is often characterized by automatically deriving the target from the input or other unlabeled examples (Ouali, Hudelot, and Tami 2020).

Predictive models Similar to classic autoregressive language models (Mikolov et al. 2010), contextualized speech representations can be learned by predicting future values of a simple representation (van den Oord, Li, and Vinyals 2018; Chung et al. 2019; Schneider et al. 2019; Chung and Glass 2020a; Jiang et al. 2021) (**PRD**, table 1). Modeling spectrograms directly, autoregressive predictive coding (APC, Chung et al. 2019) is perhaps the simplest example in this category. The forward pass and loss are computed as

$$\mathbf{c}_t = f(\mathbf{x}_{1:t}) \tag{1}$$

$$\hat{\mathbf{x}}_{t+k} = g(\mathbf{c}_t) \tag{2}$$

$$\mathcal{L}_t = \|\hat{\mathbf{x}}_{t+k} - \mathbf{x}_{t+k}\|_1 \ . \tag{3}$$

Here, $f(\cdot)$ and $g(\cdot)$ are parameterized by neural networks such that each $\mathbf{c}_t$ is only conditioned on previous inputs $\mathbf{x}_{1:t}$ and $\hat{\mathbf{x}}_{t+k}$ is computed step-wise. Chung et al. (2019) use a stack of unidirectional LSTMs for $f(\cdot)$ and a linear regression layer for $g(\cdot)$. Tasks that seek to predict or reconstruct the input are very common. In the literature, these are often jointly referred to as reconstruction tasks (**REC**, table 1) (Liu, Li, and Lee 2021; Wang et al. 2021), although this is somewhat misleading in the case of prediction.

Contrary to generative models, such as WaveNet (van den Oord et al. 2016), the APC model is not restricted to next-step prediction. Instead, it predicts $k > 0$ steps ahead in order to ensure that the model does not learn a trivial solution by exploiting the smoothness of the signal. Depending on the downstream task, we are often interested in learning so-called slow features that will typically span multiple input frames (Wiskott and Sejnowski 2002). Even the smallest linguistic units of speech, phonemes, tend to span 0.1 seconds on average (Garofolo 1993), whereas spectrogram frames $\mathbf{x}_t$ are typically computed at 0.01 second intervals. However, sometimes local smoothness is explicitly used to define the task (Badino et al. 2014; Jati and Georgiou 2017, 2019).

Contrastive models Speech contains localized noise (e.g., phase shifts) that does not inform slow feature learning. Thus, directly modeling speech might not be the best way to learn contextualized representations. Contrastive predictive coding (CPC, van den Oord, Li, and Vinyals 2018) targets a local variable $\mathbf{v}_{1:T}$, learned from the model input $\mathbf{x}_{1:T}$, instead of the input itself. The forward pass is

$$\mathbf{v}_t = f_v(\mathbf{x}_{t-r:t+r}) \tag{4}$$

$$\mathbf{c}_t = f_c(\mathbf{v}_{1:t}) \tag{5}$$

$$\hat{\mathbf{v}}_{t,k} = g_k(\mathbf{c}_t) \ , \tag{6}$$

where $f_v(\cdot)$ is a convolutional neural network, such that each $\mathbf{v}_t$ only encodes information from a limited receptive field $2r + 1$. Again, $f_c(\cdot)$ should be limited to condition each $\mathbf{c}_t$ on previous time-steps $\mathbf{v}_{1:t}$ and $g_k(\cdot)$ is a step-wise transformation. The loss is based on noise constrastive estimation (Gutmann and Hyvärinen 2010) and is given by

$$\mathcal{L}_{t,k} = -\log\left(\frac{\exp(\hat{\mathbf{v}}_{t,k}^{\mathrm{T}}\mathbf{v}_{t+k})}{\sum_{n\sim\mathcal{D}}\exp(\hat{\mathbf{v}}_{t,k}^{\mathrm{T}}\mathbf{v}_n)}\right) \ . \tag{7}$$

Here, $\mathcal{D}$ is a set of indices including the target index $t{+}k$ and negative samples drawn from a proposal distribution, which is typically taken to be a uniform distribution over the set $\{1, \ldots, T\}$. Note that the loss is also indexed by k to show that CPC targets multiple offsets. The APC model is easily extended in a similar way (Chung and Glass 2020b).

Crucially, we cannot simply predict $\mathbf{v}_{t+k}$ from $\mathbf{c}_t$ with an ℓ_1 loss. This would cause $f_v(\cdot)$ to collapse to a trivial solution, such as setting all $\mathbf{v}_t$ equal. With a contrastive loss on the other hand, setting all $\mathbf{v}_t$ equal would cause $\mathcal{L}_{k,t}$ to be constant at a value no better than a random baseline.

A model closely related to the original CPC model is wav2vec (Schneider et al. 2019). It uses a different parameterization of the functions $f_v(\cdot)$ and $f_c(\cdot)$, and modifies the loss to consider a binary prediction task, such that we have

$$\mathcal{L}_{t,k} = -\log(\sigma(\hat{\mathbf{v}}_{t,k}^{\mathrm{T}}\mathbf{v}_{t+k})) - \sum_{n\sim\mathcal{D}}\log(\sigma(-\hat{\mathbf{v}}_{t,k}^{\mathrm{T}}\mathbf{v}_n)) \ . \tag{8}$$

This model was among the first to show that learned representations can be used to improve end-to-end speech recognition. As we will see, the wav2vec framework has evolved to shape state-of-the-art representation learning for speech.

Masking-based models One downside of predictive tasks is that models are primarily unidirectional. Some work has extended APC and CPC inspired models with separate encoders operating in opposite directions (Ling et al. 2020; Kawakami et al. 2020; Borgholt et al. 2021b), but these models are still restricted to process left and right context separately. Inspired by the masked language model task used for text-based representation learning (Devlin et al. 2019), several papers have used masking to overcome this challenge (**MSK**, table 1). Masking refers to replacing parts of the input with zeros or a learned masking vector. For zero-masking (Jiang et al. 2019; Liu et al. 2020; Wang, Tang, and Livescu 2020; Chi et al. 2021; Ling and Liu 2020), we have

$$\mathbf{c}_t = f(\mathbf{x}_{1:T} \circ \mathbf{m}_{1:T}) \tag{9}$$

$$\hat{\mathbf{x}}_t = g(\mathbf{c}_t) \tag{10}$$

$$\mathcal{L}_t = \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_1 \ , \tag{11}$$

where the $\circ$ operator denotes the Hadamard product, $f(\cdot)$ is typically a transformer encoder or a bidirectional recurrent neural network, $g(\cdot)$ is a step-wise transformation, and $\mathbf{m}_{1:T}$ is a mask such that $m_{t,i} \in \{0, 1\}$. Alternatively, $\mathbf{m}_{1:T}$ is used to select which $\mathbf{x}_t$ are replaced by a learned masking vector. The entries of $\mathbf{m}_{1:T}$ are determined by some stochastic policy. One frequent inspiration is SpecAugment (Park et al. 2019), which was originally proposed for supervised speech recognition and applies frequency and time masking to spectrogram representations. While temporal masking is most common, frequency masking has also been adopted for representation learning (Wang, Tang, and Livescu 2020). A simple, yet popular, masking strategy is to draw a proportion of input indices $t_i \sim \{1, \dots, T - M\}$ without replacement, and then mask $\{t_i, \dots, t_i + M\}$ (Baevski et al. 2020; Hsu et al. 2021; Ling and Liu 2020).

Combining masking with a contrastive loss, wav2vec 2.0 was the first work to show that a competitive speech recognition model can be learned by fine-tuning a pre-trained model with as little as 10 minutes of labeled data. For this model

$$\mathbf{v}_t = f_v(\mathbf{x}_{t-r:t+r}) \tag{12}$$

$$\mathbf{c}_t = f_c(\mathbf{v}_{1:T} \circ \mathbf{m}_{1:T}) \tag{13}$$

$$\mathbf{q}_t = g_q(\mathbf{v}_t) \ . \tag{14}$$

Here, $f_v(\cdot)$ is a convolutional neural network, $f_c(\cdot)$ is a transformer encoder (Vaswani et al. 2017) and $g_q(\cdot)$ is a quantization module used to learn targets from the localized variable $\mathbf{v}_{1:T}$. Computing quantized targets this way requires an extra loss term, which we will present when we discuss quantization in general below. The contrastive loss for wav2vec 2.0 is similar to that of the CPC model,

$$\mathcal{L}_t = -\log\left(\frac{\exp(S_{\mathrm{c}}(\mathbf{c}_t, \mathbf{q}_t))}{\sum_{n\sim\mathcal{D}} \exp(S_{\mathrm{c}}(\mathbf{c}_t, \mathbf{q}_n))}\right) \ , \tag{15}$$

where $S_{\mathrm{c}}(\cdot)$ is the cosine similarity and the negative samples in $\mathcal{D}$ are sampled from other masked time-steps.

In general, masking is less data efficient than prediction, as only the masked portion of the input is non-trivial to reconstruct. For this reason, the loss might be computed as

$$\mathcal{L}_t = \|(\hat{\mathbf{x}}_t - \mathbf{x}_t) \circ (\mathbf{1} - \mathbf{m}_t)\|_1 \ . \tag{16}$$

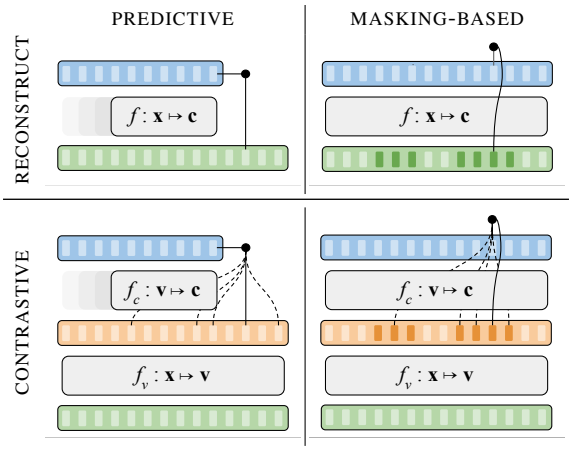

Figure 2: Schematic of self-supervised methods. Each subfigure illustrates the loss computation for a single time-step. The temporal subscript has been left out for simplicity.

Non-autoregressive predictive coding (NPC, Liu, Chung, and Glass 2021) tries to resolve this by using a convolutional neural network where the kernel is masked instead of the input. This allows for complete data utilization, but limits the amount of context encoded in the learned representation. Figure 2 summarizes the models discussed so far.

Quantization Several models enforce a discrete latent space by quantizing the vector representation (**QTZ**, table 1). The two most popular approaches are the Gumbel-softmax (Jang, Gu, and Poole 2017; Maddison, Mnih, and Teh 2017) and the quantization used in the VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2018).

Gumbel-softmax approach: Say we want to quantize a vector $\mathbf{v}$ such that it takes one of K possible values. We first map $\mathbf{v}$ to $\mathbf{l} \in \mathbb{R}^K$ and then map $\mathbf{l}$ to a probability vector $\mathbf{p} \in \mathbb{R}^K$ via the Gumbel softmax given by

$$p_i = \frac{\exp(l_i + n_i)/\tau}{\sum_k^K \exp(l_k + n_k)/\tau} \tag{17}$$

for $i = 1, \dots, K$. Here τ is a temperature parameter and $\mathbf{n} \in \mathbb{R}^K$ is a random vector with $n_i = -\log(-\log(u_i))$ for $u_i \sim \mathcal{U}(0, 1)$. As $\tau \to 0$, $\mathbf{p}$ approaches a one-hot vector. The Gumbel noise $\mathbf{n}$ is a practical way to sample from the untempered categorical distribution (i.e., $\tau = 1$). $\mathbf{p}$ is mapped to a one-hot vector using a function $\varphi(\cdot)$, such that $\varphi(\mathbf{p})_i = 1$ if $i = \arg\max_j p_j$ and 0 otherwise. As this function is non-differentiable, we must rely on the straight-through gradient estimator (Bengio, Léonard, and Courville 2013) which assumes that the Jacobian $\partial\,\varphi/\partial\,\mathbf{p}$ equals the identity matrix. The one-hot vector can then be used for a codebook lookup to obtain the final quantized vector (e.g., $\mathbf{q}_t$ in eq. 14).

The wav2vec 2.0 quantization module (eq. 14) uses the Gumbel softmax. To ensure utilization of codebook vectors, a diversity loss is added to the task specific loss (eq. 15)

$$\mathcal{L} = -H(\widetilde{\mathbf{p}})/K \ , \tag{18}$$

where $H(\cdot)$ is the entropy and $\widetilde{\mathbf{p}}$ is the untempered version of $\mathbf{p}$ without Gumbel noise.

Table 1: Selected models classified according to the binary attributes identified throughout the text. The models are sorted according to first publication date on arXiv which might differ from the citation year. **MSK**: masking, **PRD**: prediction, **CON**: contrastive, **REC**: reconstruction, **QTZ**: quantization, **GEN**: generative, **FRZ**: frozen, **FTN**: fine-tuned, **LOC**: local, **GLO**: global.

		Model and task design						Resolution			Usage	
Model	**Pub. date**	**MSK**	**PRD**	**CON**	**REC**	**QTZ**	**GEN**	**LOC**	**GLB**	**VAR**	**FRZ**	**FTN**
Self-supervised models												
Audio Word2vec (Chung et al. 2016)	2016 Mar.	✓	✗	✗	✓	✗	✗	✗	✓	✗	✓	✗
Speech2Vec (Chung and Glass 2018)	2018 Mar.	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗
Unspeech (Milde and Biemann 2018)	2018 Apr.	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓	✗
CPC (van den Oord et al. 2018)	2018 Jul.	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
APC (Chung et al. 2019)	2019 Oct.	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
wav2vec (Schneider et al. 2019)	2019 Apr.	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗
Mockingjay (Liu et al. 2020)	2019 Oct.	✓	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓
wav2vec 2.0 (Baevski et al. 2020)	2020 Jun.	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✓
NPC (Liu, Chung, and Glass 2021)	2020 Nov.	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
DeCoAR 2.0 (Ling and Liu 2020)	2020 Dec.	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓	✗
SCPC (Bhati et al. 2021a)	2021 Jun.	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓	✗
HuBERT (Hsu et al. 2021)	2021 Jun.	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✓
Probabilistic latent variable models												
VRNN (Chung et al. 2015)	2015 Jun.	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
SRNN (Fraccaro et al. 2016)	2016 May	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
HMM-VAE (Ebbers et al. 2017)	2017 Mar.	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
ConvVAE (Hsu, Zhang, and Glass 2017a)	2017 Apr.	✗	✗	✗	✓	✗	✓	✗	✓	✗	✓	✗
FHVAE (Hsu, Zhang, and Glass 2017b)	2017 Sep.	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
VQ-VAE (van den Oord et al. 2018)	2017 Nov.	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗
BHMM-VAE (Glarner et al. 2018)	2018 Mar.	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
STCN (Aksan and Hilliges 2019)	2019 Feb.	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗
FDMM (Khurana et al. 2019)	2019 Oct.	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✗
ConvDMM (Khurana et al. 2020)	2020 Jun.	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗

VQ-VAE approach: Instead of directly parameterizing a probability distribution, as in the Gumbel softmax, a vector $\mathbf{v}$ can be quantized by replacing it with the closest codebook vector $\mathbf{e}_k$. Specifically, given a learned codebook $\mathbf{e} \in \mathbb{R}^{K \times D}$, where K is the codebook size and D is the dimensionality of each codebook vector $\mathbf{e}_k$, the quantized representation $\mathbf{q}$ of $\mathbf{v}$ is obtained as,

$$\mathbf{q} = \mathbf{e}_k \ , \quad \text{where} \quad k = \arg\min_j \|\mathbf{v} - \mathbf{e}_j\|_2 \quad . \tag{19}$$

As $\arg\min$ is non-differentiable, the straight-through estimator is used as for the Gumbel-softmax. Codebook learning is facilitated by a two-term auxiliary loss similar to classical vector quantization dictionary learning (Burton, Shore, and Buck 1983; Soong, Rosenberg, and Juang 1985). Gradients for the codebook vectors are given solely by a vector quantization term. A so-called commitment term ensures that non-quantized vectors do not grow unboundedly.

$$\mathcal{L} = \underbrace{\|\text{sg}\left[\mathbf{v}\right] - \mathbf{e}\|_2^2}_{\text{vq}} + \underbrace{\beta \|\mathbf{v} - \text{sg}\left[\mathbf{e}\right]\|_2^2}_{\text{commitment}} \ , \tag{20}$$

where $\text{sg}[\mathbf{x}] = \mathbf{x}$ is the stop-gradient operator with the property $\frac{d}{dx_i}\text{sg}[\mathbf{x}] \equiv 0$ for all i and β is a hyperparameter. Although vector quantization was introduced by the VQ-VAE which is, in some ways, a latent variable model, it has been applied to self-supervised methods (van Niekerk, Nortje, and Kamper 2020; Baevski, Schneider, and Auli 2019).

Motivation: Similar to how quantization approaches differ between works, so do the motivations provided for employing them. The vq-wav2vec (Baevski, Schneider, and Auli 2019; Baevski, Auli, and Mohamed 2019) learn quantized representations in order to apply natural language processing models, like BERT (Devlin et al. 2019), afterwards. Other works use quantization for speech segmentation (Kamper and van Niekerk 2021; Chorowski et al. 2019a) or as a bottleneck in order to "*limit model capacity*" (Chung, Tang, and Glass 2020; Ling and Liu 2020). Finally, Chung, Tang, and Glass (2020) explore quantization between different layers in the APC model, but find that continuous representations consistently perform better than their quantized counterparts on a downstream phoneme classification task

Given our previous discussion of how it might not be beneficial to model localized noise, quantization in wav2vec 2.0 seems well motivated, as it enforces the target representation $\mathbf{q}_{1:T}$ to discard such noise. Taking this idea further, the HuBERT model (Hsu et al. 2021) uses offline quantization to learn categorical targets. Initially, spectrogram features are used to learn frame-wise labels with k-means clustering. A model similar to wav2vec 2.0, but without online quantization, is then trained to infer labels for masked time-steps. Since quantization is offline, this model does not need to rely on a contrastive loss, but can infer the target class directly. The offline quantization also ensures more stable training, as targets do not change abruptly.

Global representations The models covered so far learn representations that maintain a temporal resolution proportional to the input resolution. We say that they learn local representations (**LOC**, table 1). Now, we cover models that learn global representations (**GLB**, table 1).

Early work on global speech representation learning takes inspiration from the autoencoder framework (Kramer 1991). Chung et al. (2016) propose a simple sequence-to-sequence autoencoder for learning acoustic word embeddings:

$$\mathbf{c} = f(\mathbf{x}_{1:T}) \tag{21}$$

$$\hat{\mathbf{x}}_{1:T} = g(\mathbf{c}) \ , \tag{22}$$

where $f(\cdot)$ and $g(\cdot)$ are a recurrent neural networks, such that $\mathbf{c}$ is taken to be the hidden state at the last time-step T of $f(\cdot)$ and used as initial hidden state of $g(\cdot)$. The authors also propose a denoising autoencoder with masked inputs $f(\mathbf{x}_{1:T} \circ \mathbf{m}_{1:T})$. Similar RNN-based autoencoders have also been explored (Kamper 2019; Holzenberger et al. 2018).

Prior to this work, Kamper et al. (2015) and Renshaw et al. (2015) introduced the *correspondence autoencoder*. This method uses dynamic time warping to align input-target segment pairs extracted with unsupervised term discovery. In more recent work, the need for alignment has been alleviated by adopting the sequence-to-sequence framework (Kamper 2019; Jacobs, Matusevych, and Kamper 2021).

Inspired by the work on semantic word embeddings for text (Mikolov et al. 2013), the sequence-to-sequence framework has also been used to implement speech-based versions of the skipgram and continuous bag-of-words models (Chung and Glass 2017, 2018). Given a segment corresponding to a single word $\mathbf{x}_{(n)} = \mathbf{x}_{t_n:t_{n+1}}$, the skipgram model is trained to predict neighboring words $\mathbf{x}_{(n+k)}$ where $k \neq 0$. That is, instead of a single decoder, as in eq. 22, the skipgram model employs multiple decoders

$$\hat{\mathbf{x}}_{(n+k)} = g_k(\mathbf{c}) \ . \tag{23}$$

Conversely, the continuous bag-of-words model is trained to predict the target word from the neighboring words, so here multiple encoders sum over several offsets $\mathcal{K}$ to obtain $\mathbf{c}$:

$$\mathbf{c} = \sum_{k \in \mathcal{K}} f_k(\mathbf{x}_{(n+k)}) \tag{24}$$

The sequence-to-sequence models described above rely on speech segments corresponding to words. The segments are obtained by supervised forced alignment, but similar models have been explored without this requirement (Jati and Georgiou 2017; Tagliasacchi et al. 2020).

Contrastive learning has also been explored for global speech representation learning (Milde and Biemann 2018; Jati and Georgiou 2019; Jansen et al. 2018). And prior to the widespread adoption of neural networks, Levin et al. (2013) explore principal component analysis and Laplacian eigenmaps for learning fixed-sized acoustic embeddings.

Other work Some models learn local representations with a variable temporal resolution that is not proportional to the input resolution (**VAR**, table 1). In practice, this is often achieved implicitly, by learning segment boundaries or by taking repeated quantized values to belong to the same segment (Kamper and van Niekerk 2021; Chorowski et al. 2019a; Michel et al. 2017; Kreuk, Keshet, and Adi 2020; Wang, Chung, and Lee 2017; Dieleman et al. 2021). An exception is the recently proposed *segmental contrastive predictive coding* (SCPC, Bhati et al. 2021a,b). With this approach, the model explicitly learns segment boundaries, which are used to downsample the representations during training. The same segmentation strategy has subsequently been applied in other models (Cuervo et al. 2021).

Most of the work presented so far fits neatly into the taxonomy presented in table 1. One exception is the problem-agnostic speech encoder (PASE, Pascual et al. 2019; Ravanelli et al. 2020) that combines multiple pre-training tasks. Furthermore, many of the presented models have been successfully applied to other use cases. For instance, wav2vec 2.0 and related models have been applied to learn cross-lingual and multi-lingual representations (Riviere et al. 2020; Conneau et al. 2020; Khurana, Laurent, and Glass 2021) and proven well-suited for concurrently learning with labeled data (Talnikar et al. 2021; Wang et al. 2021).

2.2 Probabilistic latent variable models

Another prominent class of models are probabilistic latent variable models (LVMs). Before surveying their application to speech, we briefly review LVMs and their usual specification when applied for representation learning in general. We disregard any specific temporal notation without loss of generality. We then introduce the variational autoencoder framework (VAE, Kingma and Welling 2014). We focus on different dependency structures between data and learned representations, in contrast to the more practical view on self-supervised models taken above.

LVMs and inference Fundamental to LVMs is the assumption that the data is produced by a generative process that involves unobserved stochastic latent variables $\mathbf{z}$. An LVM aims to model this generative process to enable generation of new data $\mathbf{x}$ (**GEN**, table 1) and inference of the latent variable associated with a given observed variable $\mathbf{x}$. For representation learning, the inference of latent variables is of primary interest. An LVM is defined by the observation model $p(\mathbf{x}|\mathbf{z})$, which defines the relationship between the observed and latent variables, and the prior $p(\mathbf{z})$, which defines the relationship among the latent variables (Bartholomew, Knott, and Moustaki 2011). An LVM models the generative process via the joint observation and prior model $p(\mathbf{x}, \mathbf{z})$ often referred to as the generative model. The likelihood of an LVM given an example $\mathbf{x}$ can be written as

$$\log p(\mathbf{x}) = \log \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z} \ . \tag{25}$$

The latent variable can be inferred with e.g. Markov Chain Monte Carlo (MCMC) methods (Mohamed et al. 2020) or variational inference (Jordan et al. 1999).

For representation learning, LVMs are commonly defined using the VAE framework (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014)) which is also the focus of our exposition. In the VAE framework, the observation model $p(\mathbf{x}|\mathbf{z})$ is parameterized using a deep neural

network. This choice allows modeling complex and high-dimensional data but also makes the integral in eq. 25 analytically intractable. MCMC methods can be used to estimate it and the true model posterior $p(\mathbf{z}|\mathbf{x})$, but these methods are usually computationally expensive in this setting (Mohamed et al. 2020). To counter this and make gradient-based maximum likelihood training feasible, the VAE instead employs variational inference (Jordan et al. 1999). It approximates the intractable true model posterior by introducing a variational posterior distribution $q(\mathbf{z}|\mathbf{x})$, also parameterized by a deep neural network. From eq. 25, via Jensen's inequality, this gives rise to a variational lower bound on the likelihood, also known as the evidence lower bound (ELBO).

$$\log p(\mathbf{x}) \geq \int q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} \, \mathrm{d}\mathbf{z} \equiv \mathcal{L}_{\text{ELBO}} \quad . \quad (26)$$

The bound can be efficiently evaluated and optimized with Monte Carlo (MC) estimation by sampling from $q(\mathbf{z}|\mathbf{x})$. Low-variance gradient estimates are usually obtained via reparameterization of $q(\mathbf{z}|\mathbf{x})$ (Kingma and Welling 2014) although alternatives exist (e.g., inverse CDF sampling) (Mohamed et al. 2020). The ELBO can also be written as

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z})\right] - D_{\text{KL}} \left(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right) \quad , \quad (27)$$

where $\mathbb{E}\left[\log p(\mathbf{x}|\mathbf{z})\right]$ can be seen as a reconstruction loss and $D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right)$ is the Kullback-Leibler (KL) divergence between the variational posterior distribution and the prior.

In brief, LVMs of the VAE-type consist of a approximate posterior, $q(\mathbf{z}|\mathbf{x})$, an observation model, $p(\mathbf{x}|\mathbf{z})$, and a prior, $p(\mathbf{z})$. With reference to probabilistic coding theory, the approximate posterior is often referred to as the encoder and the observation model as the decoder (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014). From a theoretical perspective, the encoder exists solely as the result of choosing to use variational inference to train the decoder rather than e.g. MCMC. As such, it is also referred to as the inference model. However, from a representation learning perspective, the encoder is essential as it can be used to efficiently obtain the representation $\mathbf{z}$ commonly used for downstream tasks. It is still possible to evaluate and sample the true posterior distribution $p(\mathbf{z}|\mathbf{x})$ by applying MCMC methods such as Hamiltonian Monte Carlo on the decoder, but for computational reasons this is rarely done in practice.

We next review VAEs applied to speech. We consider the choices of observation, prior and inference models. We provide a model taxonomy for selected LVMs in table 3.

Observation models A common choice for the observation model $p(\mathbf{x}|\mathbf{z})$ is to include an autoregressive dependency on the observed variable (**ARX**, table 2) that is, $p(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \cdot)$ where $\cdot$ represents some dependency on the latent variable (Chung et al. 2015; Fraccaro et al. 2016; van den Oord, Vinyals, and Kavukcuoglu 2018; van den Oord, Li, and Vinyals 2018). This allows the latent representation to focus on correlations that cannot easily be predicted from the observed variable at previous time-steps (van den Oord, Li, and Vinyals 2018). In practice, the dependency on $\mathbf{x}_{1:t-1}$ is often assumed to be Markovian and hence only on $\mathbf{x}_{t-1}$. Another common choice is to depend on a local window $\mathbf{x}_{t-r:t-1}$ where $r > 1$ is an integer denoting some receptive field. We will take a dependency on $\mathbf{x}_{1:t-1}$ to mean any one of these choices unless otherwise specified.

Table 2: A comprehensive overview of observation, prior and inference models for VAE type latent variable models with a single latent variable. The observation, prior and inference models may all belong to one or more of the categories listed under them as detailed in section 2.2. The types listed here serve as primitives from which more complex structures can be constructed including models with hierarchies of multiple latent variables.

TYPE		**FORM**
	OBSERVATION MODEL	
ARX	Autoregressive on $\mathbf{x}_t$	$p(\mathbf{x}_t\|\mathbf{x}_{1:t-1})$
LOC	Local latent variable	$p(\mathbf{x}_t\|\mathbf{z}_{1:t})$
GLB	Global latent variable	$p(\mathbf{x}_t\|\mathbf{z})$
	PRIOR	
ARX	Autoregressive on $\mathbf{x}_t$	$p(\mathbf{z}_t\|\mathbf{x}_{1:t-1})$
ARZ	Autoregressive on $\mathbf{z}_t$	$p(\mathbf{z}_t\|\mathbf{z}_{1:t-1})$
IND	Locally independent $\mathbf{z}_t$	$p(\mathbf{z}_t)$
GLB	Global latent variable	$p(\mathbf{z})$
	INFERENCE MODEL	
ARZ	Autoregressive on $\mathbf{z}_t$	$q(\mathbf{z}_t\|\mathbf{z}_{1:t-1})$
FLT	Filtering	$q(\mathbf{z}_t\|\mathbf{x}_{1:t})$
LSM	Local smoothing	$q(\mathbf{z}_t\|\mathbf{x}_{t-r:t+r})$
GSM	Global smoothing	$q(\mathbf{z}_t\|\mathbf{x}_{1:T})$
GLB	Global latent variable	$q(\mathbf{z}\|\mathbf{x}_{1:T})$

While the autoregressive dependency might be important for learning a powerful generative model, it might not benefit the learned latent representations. Specifically encouraging the latent representation to discard correlations across the temporal dimension might degrade the quality of the latent representation. Furthermore, since such a decoder can perform quite well by simply solving an autoregressive prediction problem, similar to WaveNet (van den Oord et al. 2016), it can make the model prone to suffer from posterior collapse. This problem arises when the approximate and true posterior distributions collapse into the prior which renders the representations non-informative (Bowman et al. 2016; Sønderby et al. 2016). Notably, posterior collapse is a local minimum of the ELBO since the KL-divergence becomes zero. Some works alleviate this problem with tricks like KL-annealing and free bits (Bowman et al. 2016; Sønderby et al. 2016; Kingma et al. 2016). The VQ-VAE uses a quantized latent space that is not susceptible to posterior collapse *per se* (van den Oord, Vinyals, and Kavukcuoglu 2018). How to equip LVMs with powerful decoders while avoiding posterior collapse is an open problem.

Some LVMs do not use autoregressive observation models (Ebbers et al. 2017; Glarner et al. 2018; Hsu, Zhang, and Glass 2017b,a; Khurana et al. 2019, 2020). These more closely follow the assumption of *local independence* which

Table 3: Selected latent variable models classified according the attributes defined throughout section 2.2. See table 2 for the probability distributions that correspond to each of the attribute short-hands. **HIE** indicates a hierarchical representation.

		Observation			Prior				Inference					
Model	**Pub. Date**	**ARX**	**LOC**	**GLB**	**ARX**	**ARZ**	**IND**	**GLB**	**ARZ**	**FLT**	**LSM**	**GSM**	**GLB**	**HIE**
VRNN (Chung et al. 2015)	2015 Jun.	✓	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
SRNN (Fraccaro et al. 2016)	2016 May	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
HMM-VAE (Ebbers et al. 2017)	2017 Mar.	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓
ConvVAE (Hsu, Zhang, and Glass 2017a)	2017 Apr.	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
FHVAE (Hsu, Zhang, and Glass 2017b)	2017 Sep.	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✓	✗	✓
VQ-VAE (van den Oord et al. 2018)	2017 Nov.	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
BHMM-VAE (Glarner et al. 2018)	2018 Mar.	✗	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗
STCN (Aksan and Hilliges 2019)	2019 Feb.	✗	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓
FDMM (Khurana et al. 2019)	2019 Oct.	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓
ConvDMM (Khurana et al. 2020)	2020 Jun.	✗	✓	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗

states that observed variables are conditionally independent given the local (**LOC**, table 2) and/or global (**GLB**, table 2) latent variables (Bartholomew, Knott, and Moustaki 2011). However, this forces the latent variable to encode details about the observed variable to achieve a good reconstruction. This is opposite to contrastive self-supervised learning which allows models to discard details in $\mathbf{x}_{1:T}$ that do not inform the training objective (Baevski et al. 2020).

Priors Priors can be said to belong to one or more of four broad categories. See table 2. Priors that are autoregressive on the observed variable (**ARX**, table 2) take the form $p(\mathbf{z}_t|\mathbf{x}_{1:t-1})$. This generally results in a slow-down of the generative process which may be of concern if the use-case is data generation. Priors that are autoregressive on the latent variable (**ARZ**, table 2) take the form $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ and enable stochastic temporal transitions similar to hidden Markov models but with potentially nonlinear transition functions (Chung et al. 2015; Fraccaro et al. 2016; Khurana et al. 2019, 2020). Locally independent priors (**IND**, table 2) are rarely applied to sequential latent variables since they make the prior latent dynamics independent of the value of previous latent variables. Models that do impose such priors on sequential latents are quite limited in their generative power, unless they learn the prior dynamics post-hoc as done in the VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2018). Global latent variables (**GLB**, table 2) are fundamentally limited in the amount of information they can encode. Hence, models usually use them in combination with another local latent variable, or to encode fixed length input segments (Khurana et al. 2019; Hsu, Zhang, and Glass 2017a,b).

Inference models LVMs based on the VAE perform so-called amortized variational inference. Here, a single inference network is used to infer the latent variables of any $\mathbf{x}$. For this reason, all inference models covered here are conditioned on the observed sequence in some way. Generally, the inference model can be seen as solving either a filtering or smoothing problem. In filtering (**FLT**, table 2), the latent variables are assumed to depend only on past and current values of the observed variable, $q(\mathbf{z}_t|\mathbf{x}_{1:t})$ (Chung et al. 2015; Khurana et al. 2020). In global smoothing (**GSM**, table 2), this causal dependency is replaced with a dependency on all observed values, $q(\mathbf{z}_t|\mathbf{x}_{1:T})$ (Fraccaro et al. 2016; Hsu, Zhang, and Glass 2017a). Smoothing can also be done locally (**LSM**, table 2), where the latent variables then depend on $\mathbf{x}_{t-r:t+r}$ for some integer $r > 0$ (van den Oord, Vinyals, and Kavukcuoglu 2018). Compared to self-supervised models that often use transformer encoders it can be hypothesized that global smoothing offers a stronger case than local smoothing and filtering for representation learning.

The inference model may also be used to infer a global latent variable (**GLB**, table 2) that might encode global information about $\mathbf{x}$. While it must be included in the prior model it might not be in the observation model, if the model also has a local latent variable. Finally, latent variables are often made to depend autoregressively on past inferred values, e.g. $q(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:t})$ (**ARZ**, table 2) (Chung et al. 2015; Fraccaro et al. 2016).

Multiscale and hierarchical models Some work has explored using a hierarchy of latent variables (**HIE**, table 3). This allows encoding the inductive bias that speech contains information at different temporal scales by letting the latent variables operate at different temporal scales (Hsu, Zhang, and Glass 2017b). Khurana et al. (2019) propose using a temporal latent variable along with a global latent variable. Recent work has focused on learning a deeper latent hierarchy with five latent variables (Aksan and Hilliges 2019).

Other work Before the introduction of the VAE, models such as deep belief networks (DBN, Hinton, Osindero, and Teh 2006) built from stacks of restricted Boltzmann machines (Smolensky 1986; Fischer and Igel 2014) were popular. Lee et al. (2009) show the feasibility of using a two-layered DBN for discovering acoustic units of speech, while Deng et al. (2010) show that a DBN can learn a binary coding of spectrograms that has higher signal-to-noise ratio than classical vector-quantization techniques for speech coding. DBNs are however notoriously tricky to optimize requiring the use of expensive MCMC sampling techniques for inference or resort to biased gradient estimates (Hinton 2012; Fischer and Igel 2011). Non-neural LVMs for speech representation learning have also been explored (Lee and Glass 2012; Ondel, Burget, and Cernocký 2016; Heck, Sakti, and Nakamura 2017; Jansen, Thomas, and Hermansky 2013).

3 Discussion

From global to local In table 1, we see that work on global representations within self-supervised learning precedes work on local representations. However, we find that the core ideas underlying the recent successes in learning local representation models have also been used for global representation learning; masking (Chung et al. 2016), context prediction (Chung and Glass 2018), and contrastive training (Milde and Biemann 2018) have been applied in both settings. Furthermore, where work on global representation learning has taken inspiration from Word2vec (Mikolov et al. 2013), the techniques used for learning local representations are inspired by contextualized word embeddings (Devlin et al. 2019). Thus, the gap between these two model classes is largely a product of the developments in related fields and the general increase in computational resources.

Representations beyond inference Predictive tasks are commonly used for self-supervised models, but they are not directly compatible with LVM training. However, an LVM prior with an autoregressive parameterization, $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ or $p(\mathbf{z}_t|\mathbf{x}_{1:t-1})$, can be seen as predictive in the sense that it tries to match the approximate posterior. Hence, the prior might be considered for feature extraction. Jones and Moore (2020) examine the importance of the prior in the VQ-VAE and show that the ability of this model to estimate densities $p(\mathbf{x}_{1:T})$ lies solely in its prior. Other work has also explored representations beyond the latent variable such as hidden units of the observation model (Khurana et al. 2020; Chorowski et al. 2019b).

Masking and missing data Masking may also improve representations learned with VAEs. Masking in VAEs has already been explored in the literature in the context of missing data imputation. Here, $\mathbf{x}$ is only partially observed, and often represented as a segmentation into observed and missing parts and a mask $\mathbf{m}$ indicating where the data is missing. The model is then trained to infer the latent variable from the observed data. Reconstruction also deals only with the observed data. Previous work has largely focused on the ability of these models to yield high-quality imputations within the tabular and image data domains, without probing for the effects on the learned latent representation (Mattei and Frellsen 2019; Ipsen, Mattei, and Frellsen 2021). The idea of using VAEs to impute missing data was already examined in the seminal paper by Rezende, Mohamed, and Wierstra (2014). Here the model was trained with fully observed data and used to impute data in an iterative sampling approach post hoc leaving the learned representations unchanged.

Evaluating representations Although this review has a primarily methodological focus, we should briefly touch upon evaluation procedures. Training metrics for self-supervised tasks and the likelihood of LVMs offer little guidance as to the quality of the learned representations (Huszár 2017). Thus, a common approach is to evaluate the representations in terms of their usefulness for downstream tasks. Such tasks may be chosen to target specific attributes of the representation (e.g. semantic or speaker information).

The SUPERB benchmark (Yang et al. 2021) gathers multiple tasks grouped into categories such as *recognition*, *detection*, *semantics*, *speaker*, *paralinguistics* and *generation*. The recently proposed SLUE benchmark focuses on spoken language understanding (Shon et al. 2021). The long-standing zero resource speech challenge (ZeroSpeech) offers a new set of tasks for each edition (Versteegh et al. 2015; Dunbar et al. 2017, 2019, 2020, 2021) usually featuring a minimal-pair ABX task (Schatz et al. 2013, 2014).

Tasks that evaluate representations in terms of speaker-related information include speaker verification (Hsu, Zhang, and Glass 2017b; Khurana et al. 2019; Milde and Biemann 2018), speaker identification (van den Oord, Li, and Vinyals 2018; Jati and Georgiou 2019; Chung et al. 2019; Liu, Chung, and Glass 2021), dialect classification (Khurana et al. 2019), emotion recognition (Pascual et al. 2019; Yang et al. 2021) and gender classification (Lee et al. 2009). The semantic content of representations are evaluated using tasks such as intent classification (Morais et al. 2021; Yang et al. 2021), slot filling (Lai et al. 2021; Yang et al. 2021), sentiment analysis (Liu et al. 2020), question answering (Chung, Zhu, and Zeng 2020), named entity recognition (Shon et al. 2021; Borgholt et al. 2021a; Pasad et al. 2021) and speech translation (Bansal et al. 2017; Chung and Glass 2020a). Cardiac arrest detection for emergency calls has also been used to evaluate speech representations (Borgholt et al. 2021a). For local representations, phoneme classification is very common (Lee et al. 2009; Hsu, Zhang, and Glass 2017a; Chorowski et al. 2019b; Chung et al. 2019; Liu, Li, and Lee 2021). However, automatic speech recognition has become the *de facto* standard benchmark task (Ling and Liu 2020; Chung and Glass 2020a; Hsu et al. 2021).

Moving forward Most of the seminal work has focused on improving speech recognition (Schneider et al. 2019; Baevski et al. 2020). This focus has gained traction over the last couple of years, as computational resources have become more accessible and end-to-end models (Graves et al. 2006; Chan et al. 2016) have been established as the dominant approach to speech recognition (Gulati et al. 2020). It is important to stress that self-supervised models, such as wav2vec 2.0 (Baevski et al. 2020), represent a breakthrough, and recent successful approaches build upon this method. That is, deep self-attention models combined with masking (Hsu et al. 2021; Wang et al. 2021; Chen et al. 2021). This development mirrors years of rapid progress in masked language modeling within natural language processing (Devlin et al. 2019; Clark et al. 2020) and we expect this to continue for unsupervised neural speech representation learning.

4 Conclusion

We reviewed unsupervised representation learning for speech, focusing on two primary categories: self-supervised methods and probabilistic latent variable models. Inspired by the development of self-supervised learning and the dependency structures of latent variable models, we derived a comprehensive model taxonomy. Finally, we compare and discuss models from the two categories and their respective evaluation procedures.

References

Aksan, E.; and Hilliges, O. 2019. STCN: Stochastic Temporal Convolutional Networks. *International Conference on Learning Representations (ICLR)*.

Badino, L.; Canevari, C.; Fadiga, L.; and Metta, G. 2014. An auto-encoder based approach to unsupervised learning of subword units. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Baevski, A.; Auli, M.; and Mohamed, A. 2019. Effectiveness of self-supervised pre-training for speech recognition. *arXiv:1911.03912*.

Baevski, A.; Hsu, W.-N.; Conneau, A.; and Auli, M. 2021. Unsupervised Speech Recognition. *Neural Information Processing System (NeurIPS)*.

Baevski, A.; Schneider, S.; and Auli, M. 2019. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. *International Conference on Learning Representations (ICLR)*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Neural Information Processing System (NeurIPS)*.

Bansal, S.; Kamper, H.; Lopez, A.; and Goldwater, S. 2017. Towards speech-to-text translation without speech recognition. *European Chapter of the Association for Computational Linguistics (EACL)*.

Bartholomew, D. J.; Knott, M.; and Moustaki, I. 2011. *Latent variable models and factor analysis: a unified approach*. Wiley Series in Probability and Statistics. Wiley, 3rd ed edition.

Bengio, Y.; Courville, A. C.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 35(8): 1798–1828.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv:1308.3432*.

Bhati, S.; Villalba, J.; Żelasko, P.; Moro-Velazquez, L.; and Dehak, N. 2021a. Segmental Contrastive Predictive Coding for Unsupervised Word Segmentation. *arXiv:2106.02170*.

Bhati, S.; Villalba, J.; Żelasko, P.; Moro-Velazquez, L.; and Dehak, N. 2021b. Unsupervised Speech Segmentation and Variable Rate Representation Learning using Segmental Contrastive Predictive Coding. *arXiv:2110.02345*.

Borgholt, L.; Havtorn, J. D.; Abdou, M.; Edin, J.; Maaløe, L.; Søgaard, A.; and Igel, C. 2021a. Do We Still Need Automatic Speech Recognition for Spoken Language Understanding? *arXiv:2111.14842*.

Borgholt, L.; Tax, T. M.; Havtorn, J. D.; Maaløe, L.; and Igel, C. 2021b. On Scaling Contrastive Representations for Low-Resource Speech Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Józefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. *SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Burton, D. K.; Shore, J. E.; and Buck, J. T. 1983. A generalization of isolated word recognition using vector quantization. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2021. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *arXiv:2110.13900*.

Chi, P.-H.; Chung, P.-H.; Wu, T.-H.; Hsieh, C.-C.; Chen, Y.-H.; Li, S.-W.; and Lee, H.-y. 2021. Audio albert: A lite bert for self-supervised learning of audio representation. *IEEE Spoken Language Technology Workshop (SLT)*.

Chorowski, J.; Chen, N.; Marxer, R.; Dolfing, H.; Łańcucki, A.; Sanchez, G.; Alumäe, T.; and Laurent, A. 2019a. Unsupervised neural segmentation and clustering for unit discovery in sequential data. *NeurIPS workshop: Perception as generative reasoning*.

Chorowski, J.; Weiss, R. J.; Bengio, S.; and van den Oord, A. 2019b. Unsupervised speech representation learning using WaveNet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Chung, J.; Kastner, K.; Dinh, L.; Goel, K.; Courville, A. C.; and Bengio, Y. 2015. A Recurrent Latent Variable Model for Sequential Data. *Neural Information Processing System (NeurIPS)*.

Chung, Y.-A.; and Glass, J. 2017. Learning word embeddings from speech. *NeurIPS ML4Audio Workshop*.

Chung, Y.-A.; and Glass, J. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *INTERSPEECH*.

Chung, Y.-A.; and Glass, J. 2020a. Generative pre-training for speech with autoregressive predictive coding. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chung, Y.-A.; and Glass, J. 2020b. Improved speech representations with multi-target autoregressive predictive coding. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chung, Y.-A.; Hsu, W.-N.; Tang, H.; and Glass, J. R. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. *INTERSPEECH*.

Chung, Y.-A.; Tang, H.; and Glass, J. 2020. Vector-Quantized Autoregressive Predictive Coding. *INTERSPEECH*.

Chung, Y.-A.; Wu, C.-C.; Shen, C.-H.; Lee, H.-Y.; and Lee, L.-S. 2016. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *INTERSPEECH*.

Chung, Y.-A.; Zhu, C.; and Zeng, M. 2020. SPLAT: Speech-Language Joint Pre-Training for Spoken Language Understanding. *arXiv:2010.02295*.

Clark, K.; Luong, M.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *International Conference on Learning Representations (ICLR)*.

Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; and Auli, M. 2020. Unsupervised cross-lingual representation learning for speech recognition. *INTERSPEECH*.

Cuervo, S.; Grabias, M.; Chorowski, J.; Ciesielski, G.; Łańcucki, A.; Rychlikowski, P.; and Marxer, R. 2021. Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words. *arXiv:2110.15909*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, L.; Seltzer, M. L.; Yu, D.; Acero, A.; Mohamed, A.; and Hinton, G. E. 2010. Binary coding of speech spectrograms using a deep auto-encoder. *INTERSPEECH*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Dieleman, S.; Nash, C.; Engel, J.; and Simonyan, K. 2021. Variable-rate discrete representation learning. *arXiv:2103.06089*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dunbar, E.; Algayres, R.; Karadayi, J.; Bernard, M.; Benjumea, J.; Cao, X.-N.; Miskic, L.; Dugrain, C.; Ondel, L.; Black, A.; et al. 2019. The Zero Resource Speech Challenge 2019: TTS without T. *INTERSPEECH*.

Dunbar, E.; Bernard, M.; Hamilakis, N.; Nguyen, T.; de Seyssel, M.; Rozé, P.; Rivière, M.; Kharitonov, E.; and Dupoux, E. 2021. The Zero Resource Speech Challenge 2021: Spoken language modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dunbar, E.; Cao, X. N.; Benjumea, J.; Karadayi, J.; Bernard, M.; Besacier, L.; Anguera, X.; and Dupoux, E. 2017. The Zero Resource Speech Challenge 2017. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Dunbar, E.; Karadayi, J.; Bernard, M.; Cao, X.-N.; Algayres, R.; Ondel, L.; Besacier, L.; Sakti, S.; and Dupoux, E. 2020. The Zero Resource Speech Challenge 2020: Discovering discrete subword and word units. *INTERSPEECH*.

Ebbers, J.; Heymann, J.; Drude, L.; Glarner, T.; Haeb-Umbach, R.; and Raj, B. 2017. Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery. *INTERSPEECH*.

Fischer, A.; and Igel, C. 2011. Bounding the Bias of Contrastive Divergence Learning. *Neural Computation*, 23: 664–673.

Fischer, A.; and Igel, C. 2014. Training Restricted Boltzmann Machines: An Introduction. *Pattern Recognition*, 47: 25–39.

Fraccaro, M.; Sønderby, S. K.; Paquet, U.; and Winther, O. 2016. Sequential Neural Models with Stochastic Layers. *Neural Information Processing System (NeurIPS)*.

Garofolo, J. S. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download.

Glarner, T.; Hanebrink, P.; Ebbers, J.; and Haeb-Umbach, R. 2018. Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery. *INTERSPEECH*.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *International Conference on Machine Learning (ICML)*.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. *INTERSPEECH*.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heck, M.; Sakti, S.; and Nakamura, S. 2017. Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to ZeroSpeech 2017. *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017*.

Hinton, G. E. 2012. A Practical Guide to Training Restricted Boltzmann Machines. In Montavon, G.; Orr, G. B.; and Müller, K.-R., eds., *Neural Networks: Tricks of the Trade: Second Edition*, 599–619. Springer.

Hinton, G. E.; Osindero, S.; and Teh, Y. W. 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7): 1527–1554.

Holzenberger, N.; Du, M.; Karadayi, J.; Riad, R.; and Dupoux, E. 2018. Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments. *INTERSPEECH*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv:2106.07447*.

Hsu, W.-N.; Zhang, Y.; and Glass, J. 2017a. Learning Latent Representations for Speech Generation and Transformation. *INTERSPEECH*.

Hsu, W.-N.; Zhang, Y.; and Glass, J. 2017b. Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data. *Neural Information Processing System (NeurIPS)*.

Huszár, F. 2017. Is Maximum Likelihood Useful for Representation Learning? *inFERENCe*.

Ipsen, N. B.; Mattei, P.-A.; and Frellsen, J. 2021. not-MIWAE: Deep generative modelling with missing not at random data. *International Conference on Learning Representations (ICLR)*.

Jacobs, C.; Matusevych, Y.; and Kamper, H. 2021. Acoustic word embeddings for zero-resource languages using self-supervised contrastive learning and multilingual adaptation. *IEEE Spoken Language Technology Workshop (SLT)*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. *International Conference on Learning Representations (ICLR)*.

Jansen, A.; Plakal, M.; Pandya, R.; Ellis, D. P.; Hershey, S.; Liu, J.; Moore, R. C.; and Saurous, R. A. 2018. Unsupervised learning of semantic audio representations. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jansen, A.; Thomas, S.; and Hermansky, H. 2013. Weak top-down constraints for unsupervised acoustic model training. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jati, A.; and Georgiou, P. 2019. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27.

Jati, A.; and Georgiou, P. G. 2017. Speaker2Vec: Unsupervised Learning and Adaptation of a Speaker Manifold Using Deep Neural Networks with an Evaluation on Speaker Segmentation. *INTERSPEECH*.

Jiang, D.; Lei, X.; Li, W.; Luo, N.; Hu, Y.; Zou, W.; and Li, X. 2019. Improving transformer-based speech recognition using unsupervised pre-training. *arXiv:1910.09932*.

Jiang, D.; Li, W.; Zhang, R.; Cao, M.; Luo, N.; Han, Y.; Zou, W.; Han, K.; and Li, X. 2021. A Further Study of Unsupervised Pretraining for Transformer Based Speech Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jones, H. T.; and Moore, J. 2020. Is the Discrete VAE's Power Stuck in its Prior? *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2): 183–233.

Kamper, H. 2019. Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kamper, H.; Elsner, M.; Jansen, A.; and Goldwater, S. 2015. Unsupervised neural network based feature extraction using weak top-down constraints. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kamper, H.; and van Niekerk, B. 2021. Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks. *INTERSPEECH*.

Kawakami, K.; Wang, L.; Dyer, C.; Blunsom, P.; and van den Oord, A. 2020. Learning Robust and Multilingual Speech Representations. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Khurana, S.; Joty, S. R.; Ali, A.; and Glass, J. 2019. A Factorial Deep Markov Model for Unsupervised Disentangled Representation Learning from Speech. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Khurana, S.; Laurent, A.; and Glass, J. 2021. Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0. *arXiv:2110.03560*.

Khurana, S.; Laurent, A.; Hsu, W.-N.; Chorowski, J.; Lancucki, A.; Marxer, R.; and Glass, J. 2020. A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning. *INTERSPEECH*.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved Variational Inference with Inverse Autoregressive Flow. *Neural Information Processing System (NeurIPS)*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *International Conference on Learning Representations (ICLR)*.

Kramer, M. A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2): 233–243.

Kreuk, F.; Keshet, J.; and Adi, Y. 2020. Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation. *INTERSPEECH*.

Lai, C.-I.; Chuang, Y.-S.; Lee, H.-Y.; Li, S.-W.; and Glass, J. 2021. Semi-supervised spoken language understanding via self-supervised speech and language model pretraining. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lee, C.-y.; and Glass, J. 2012. A Nonparametric Bayesian Approach to Acoustic Model Discovery. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lee, H.; Largman, Y.; Pham, P.; and Ng, A. Y. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. *Neural Information Processing System (NeurIPS)*.

Levin, K.; Henry, K.; Jansen, A.; and Livescu, K. 2013. Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Ling, S.; and Liu, Y. 2020. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv:2012.06659*.

Ling, S.; Liu, Y.; Salazar, J.; and Kirchhoff, K. 2020. Deep contextualized acoustic representations for semi-supervised speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Liu, A. H.; Chung, Y.-A.; and Glass, J. 2021. Non-autoregressive predictive coding for learning speech representations from local dependencies. *INTERSPEECH*.

Liu, A. T.; Li, S.-W.; and Lee, H.-y. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2351–2366.

Liu, A. T.; Yang, S.-w.; Chi, P.-H.; Hsu, P.-c.; and Lee, H.-y. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2017. The concrete distribution: A continuous relaxation of discrete random variables. *International Conference on Learning Representations (ICLR)*.

Mattei, P.-A.; and Frellsen, J. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. *International Conference on Machine Learning (ICML)*.

Michel, P.; Rasanen, O.; Thiollière, R.; and Dupoux, E. 2017. Blind Phoneme Segmentation With Temporal Prediction Errors. *ACL Student Research Workshop*.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. *INTERSPEECH*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Neural Information Processing System (NeurIPS)*.

Milde, B.; and Biemann, C. 2018. Unspeech: Unsupervised speech context embeddings. *INTERSPEECH*.

Mohamed, S.; Rosca, M.; Figurnov, M.; and Mnih, A. 2020. Monte Carlo Gradient Estimation in Machine Learning. *Journal of Machine Learning Research*, 21(132): 1–62.

Morais, E.; Kuo, H.-K. J.; Thomas, S.; Tüske, Z.; and Kingsbury, B. 2021. End-to-end spoken language understanding using transformer networks and self-supervised pretrained features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ondel, L.; Burget, L.; and Cernocký, J. 2016. Variational Inference for Acoustic Unit Discovery. *SLTU-2016, 5th Workshop on Spoken Language Technologies for Underresourced languages*.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. An overview of deep semi-supervised learning. *arXiv:2006.05278*.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *INTERSPEECH*.

Pasad, A.; Wu, F.; Shon, S.; Livescu, K.; and Han, K. J. 2021. On the Use of External Data for Spoken Named Entity Recognition. *arXiv:2112.07648*.

Pascual, S.; Ravanelli, M.; Serrà, J.; Bonafonte, A.; and Bengio, Y. 2019. Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. *INTERSPEECH*.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ravanelli, M.; Zhong, J.; Pascual, S.; Swietojanski, P.; Monteiro, J.; Trmal, J.; and Bengio, Y. 2020. Multi-task self-supervised learning for Robust Speech Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Renshaw, D.; Kamper, H.; Jansen, A.; and Goldwater, S. 2015. A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge. *INTERSPEECH*.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *International Conference on Machine Learning (ICML)*.

Riviere, M.; Joulin, A.; Mazaré, P.-E.; and Dupoux, E. 2020. Unsupervised pretraining transfers well across languages. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Schatz, T.; Peddinti, V.; Bach, F.; Jansen, A.; Hermansky, H.; and Dupoux, E. 2013. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. *INTERSPEECH*.

Schatz, T.; Peddinti, V.; Cao, X.-N.; Bach, F.; Hermansky, H.; and Dupoux, E. 2014. Evaluating speech features with the Minimal-Pair ABX task (II): Resistance to noise. *INTERSPEECH*.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. *INTERSPEECH*.

Shon, S.; Pasad, A.; Wu, F.; Brusco, P.; Artzi, Y.; Livescu, K.; and Han, K. J. 2021. SLUE: New Benchmark Tasks for Spoken Language Understanding Evaluation on Natural Speech. *arXiv:2111.10367*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

Smolensky, P. 1986. Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. In D.E. Rumelhart, J. M., ed., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, chapter 6, 194–281. MIT Press.

Soong, F.; Rosenberg, A.; and Juang, L. R. B. 1985. A vector Quantization approach to Speaker Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sønderby, C. K.; Raiko, T.; Maaløe, L.; Sønderby, S. K.; and Winther, O. 2016. Ladder Variational Autoencoders. *Neural Information Processing System (NeurIPS)*.

Tagliasacchi, M.; Gfeller, B.; de Chaumont Quitry, F.; and Roblek, D. 2020. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27: 600–604.

Talnikar, C.; Likhomanenko, T.; Collobert, R.; and Synnaeve, G. 2021. Joint masked cpc and ctc training for asr. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tsai, Y.-H. H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2020. Self-supervised Learning from a Multi-view Perspective. *International Conference on Learning Representations (ICLR)*.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. *Proceedings of the 9th ISCA Speech Synthesis Workshop*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2018. Neural Discrete Representation Learning. *Neural Information Processing System (NeurIPS)*.

van Niekerk, B.; Nortje, L.; and Kamper, H. 2020. Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge. *INTERSPEECH*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Neural Information Processing System (NeurIPS)*.

Versteegh, M.; Thiolliere, R.; Schatz, T.; Cao, X. N.; Anguera, X.; Jansen, A.; and Dupoux, E. 2015. The Zero Resource Speech Challenge 2015. *INTERSPEECH*.

Wang, C.; Wu, Y.; Qian, Y.; Kumatani, K.; Liu, S.; Wei, F.; Zeng, M.; and Huang, X. 2021. Unispeech: Unified speech representation learning with labeled and unlabeled data. *International Conference on Machine Learning (ICML)*.

Wang, W.; Tang, Q.; and Livescu, K. 2020. Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wang, Y.-H.; Chung, C.-T.; and Lee, H.-y. 2017. Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries. *INTERSPEECH*.

Wiskott, L.; and Sejnowski, T. J. 2002. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4): 715–770.

Yang, S.-w.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I. J.; Lakhotia, K.; Lin, Y. Y.; Liu, A. T.; Shi, J.; Chang, X.; Lin, G.-T.; et al. 2021. SUPERB: Speech processing Universal PERformance Benchmark. *INTERSPEECH*.