Self Supervised Representation Learning from Speech

Laurent Besacier

Outline

Self supervised representation learning

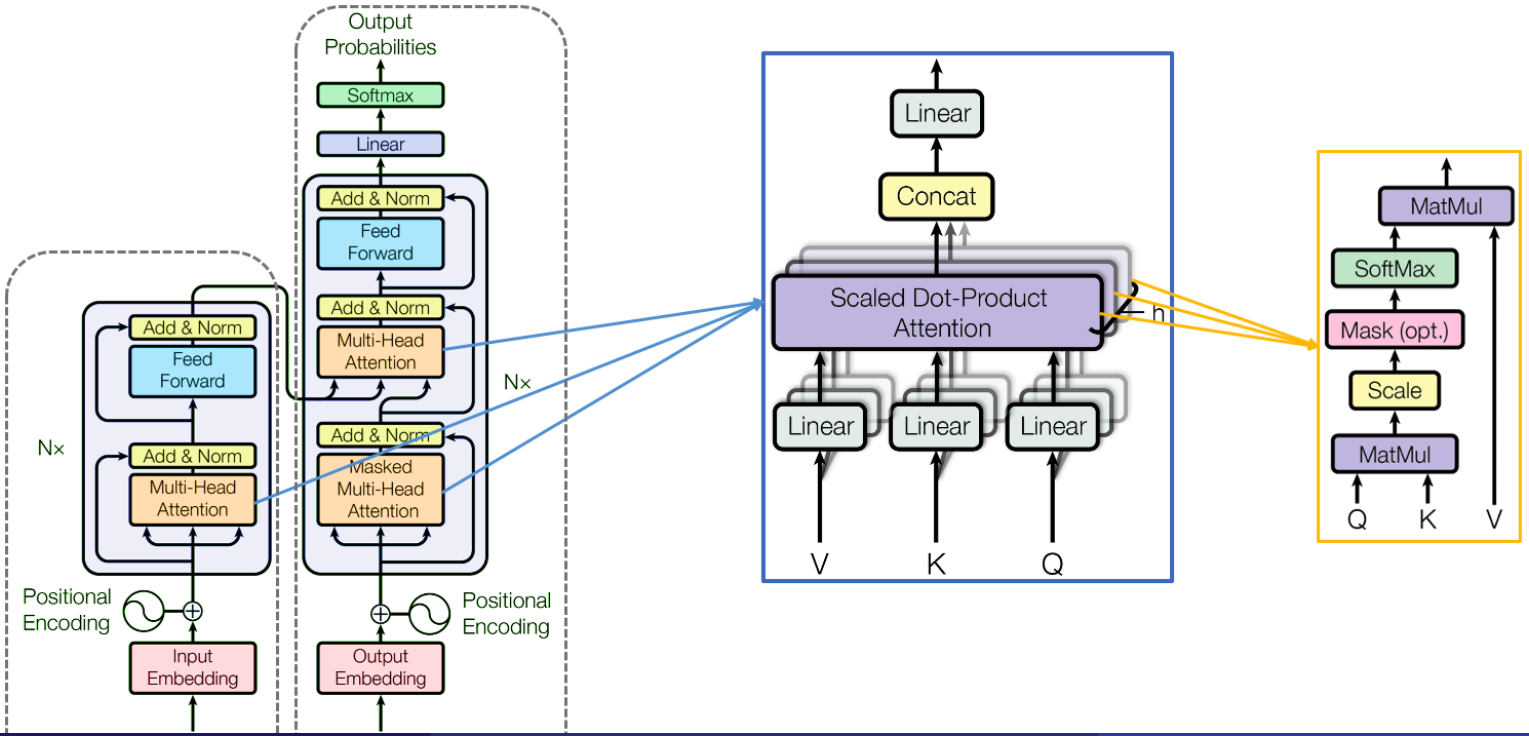
- Using huge unlabeled data for training ; targets are computed from the signal itself
  - *"learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset"* (from Chen et al. (2020))
- Introduced for vision: see for instance (Chen et al., 2020)
  - learn representations by contrasting positive pairs against negative pairs
- Introduced also in NLP: see for instance (Devlin et al., 2018)
  - more details in next slides

Previous works

- Stacked restricted Boltzman machines (RBM) (Hinton and Salakhutdinov, 2006)
  - hidden layer extracts relevant features from the observations that serve as input to next RBM that is stacked on top of it forming a deterministic feed-forward neural network
- Denoising autoencoders (AE) (Vincent et al., 2008)
  - networks which are tasked with reconstructing outputs from their (noisy) input versions
- Variational autoencoders (VAE) (Kingma and Welling, 2013)
  - VAE is like a traditional AE in which the encoder produces distributions over latent representations (rather than deterministic encodings) while the decoder is trained on samples from this distribution
  - both encoder and decoder are trained jointly
  - VQ-VAE (van den Oord et al., 2017) replaces continuous latent vectors with deterministically quantized versions

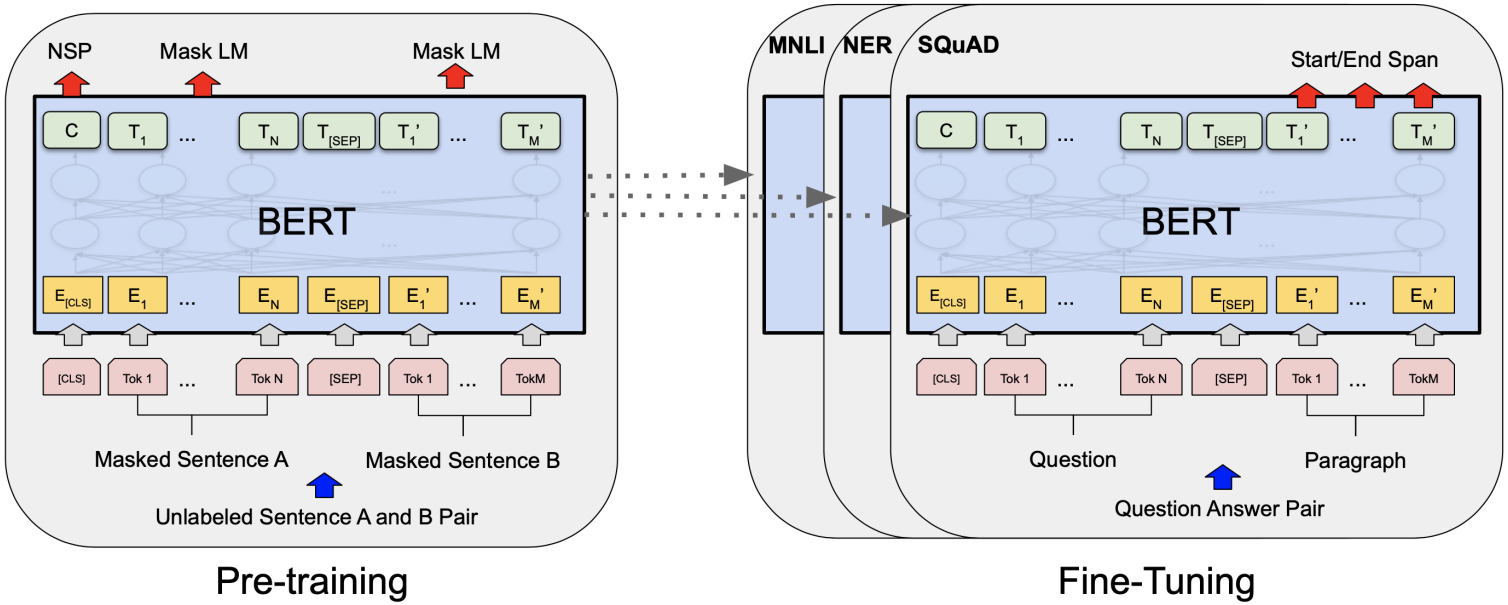
Pre-trained language models

- Leverage large amount of freely available unlabeled text to facilitate transfer learning in NLP
- Yield state-of-the-art results on a wide range of NLP tasks + save time and computational resources
- Example of BERT (Devlin et al., 2018) based on the Transformer model (Vaswani et al., 2017)



Pre-training: example of BERT

Following the BERT Devlin et al. (2018) paradigm



- **Masked Language Model**: learn to predict randomly masked input tokens.
- **Next Sentence Prediction**: learn to predict if B is the next sentence to A, given an input pair (A, B).

Self supervised representation learning from speech

- Autoregressive predictive coding (APC) (Chung et al., 2019; Chung and Glass, 2020)
  - Considers the sequential structure of speech and predicts information about a future frame
- Contrastive Predictive Coding (CPC) (Baevski et al., 2019; Schneider et al., 2019a; Kahn et al., 2019)
  - Easier learning objective which consists in distinguishing a true future audio frame from negatives
- Other approaches for feature representation learning using multiple self supervised tasks (Pascual et al., 2019; Ravanelli et al., 2020) or bidirectional encoders (Song et al., 2019; Liu et al., 2020; Wang et al., 2020)

Autoregressive predictive coding (APC)

- Predicting the spectrum of a future frame (rather than a wave sample)
- Largely inspired by language models (LMs) for text, which are typically a probability distribution over sequences of T tokens $(t_1, t_2, ..., t_T)$

$$P(sequence) = \prod_{k=1}^{T} P(t_k | t_1, t_2, ..., t_{k-1}) \quad (1)$$

$$P(sequence) = \prod_{k=1}^{T} P(t_k | h) \quad (2)$$

- Recurrent neural network LM:
  $h = rnn\_state(E(t_1), E(t_2), ..., E(t_{k-1}))$
- For speech, each token t_k corresponds to a frame rather than a word or character token

Autoregressive predictive coding (APC)

- Differences between speech and text tokens
  - Input speech representations (MFCCs for instance) are already in a vector form (no embedding layer)
  - No final set of target tokens (softmax layer replaced by a regression layer)
  - Learnable parameters in APC are the RNN parameters θ_{rnn} and the regression layer parameters θ_r
  - Encourage APC to infer more global structures rather than the local information in the signal
    - ask the model to predict a frame n steps ahead of the current one
  - Model is optimized by minimizing the L1 loss between input sequence $(t_1, t_2, ..., t_T)$ and the predicted sequence $(y_1, y_2, ..., y_T)$:

$$\sum_{i=1}^{T-n} |t_{i+n} - y_i| \tag{3}$$

Autoregressive predictive coding (APC)

- Chung et al. (2019) models APC with a multi-layer unidirectional LSTM with residual connections
- After training, RNN hidden states are taken as the learned representations
- A follow-up work (Chung and Glass, 2020) adds an auxiliary objective that serves as regularization to improve generalization

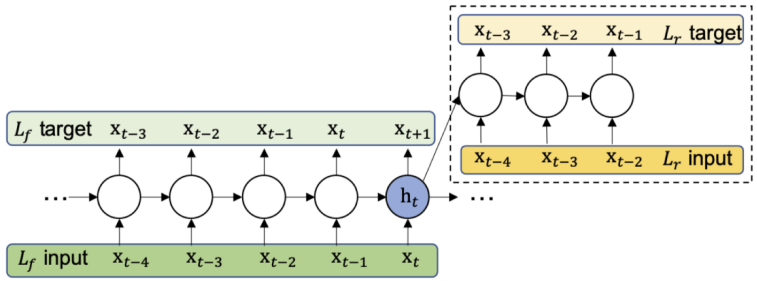


Figure 1: Overview of our method. L_f is the original APC objective that aims to predict x_{t+n} given a context $(x_1, x_2, \ldots, x_t)$ with an autoregressive RNN. Our method first samples an anchor position, assuming it is time step t. Next, we build an auxiliary loss L_r that computes L_f of a past sequence $(x_{t-s}, x_{t-s+1}, \ldots, x_{t-s+\ell-1})$ (see Section 3.1 for definitions of s and ℓ), using an auxiliary RNN (dotted line area). In this example, $(n, s, \ell) = (1, 4, 3)$. In practice, we can sample multiple anchor positions, and averaging over all of them gives us the final L_r.

Figure: Figure from (Chung and Glass, 2020)

Autoregressive predictive coding (APC)

- Speech transcription and translation experiments on Librispeech
- Architecture is a RNN model with attention

Feature	ASR (WER ↓)	ST (BLEU ↑)
log Mel	18.3	12.9
APC w/ L_f	15.2	13.8
APC w/ L_m	14.2	14.5

Table 2: Automatic speech recognition (ASR) and speech translation (ST) results using different types of features as input to a seq2seq with attention model. Word error rates (WER, ↓) and BLEU scores (↑) are reported for the two tasks, respectively.

Contrastive Predictive Coding (CPC)

- Use of a contrastive loss that distinguishes a true future audio sample from negatives
- Example of *wav2vec* (Schneider et al., 2019b) that relies on a fully convolutional architecture
- Applied the learned representations to improve a supervised ASR system

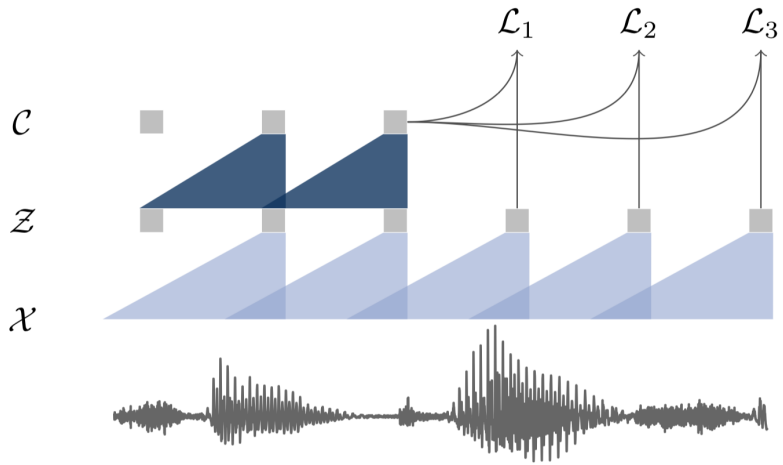
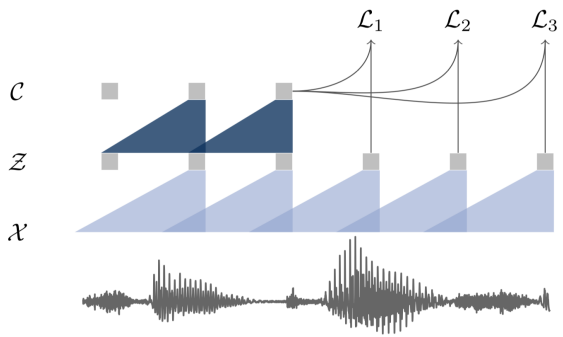


Figure: Figure from (Schneider et al., 2019b)

Contrastive Predictive Coding (CPC)

- Encoder network $Z = f(X)$; 5 (causal) convolution layers ; local feature representations z_i encode 30 ms of audio every 10ms
- Context network $C = g(Z)$; 9 (causal) convolution layers ; mix multiple z_i (receptive field of dimension v corresponding to 210ms) into a single contextualized representation c_i
- Model trained to distinguish a sample z_{i+k} that is k steps in the future from distractor samples $\tilde{z}$ drawn from a proposal distribution p_n by minimizing a contrastive loss for each step $k = 1, ..., K$
- Negatives examples sampled by uniformly choosing distractors from each audio sequence: is $p_n(z) = 1/T$ where T is the sequence length



Contrastive Predictive Coding (CPC)

Experiments

- After pre-training: input the representations c_i produced by the context network to the acoustic model instead of log-mel filterbank features
- ASR tasks: phoneme recognition on TIMIT and word recognition on WSJ
- Librispeech corpus (1000h) used for unsupervised pre-training

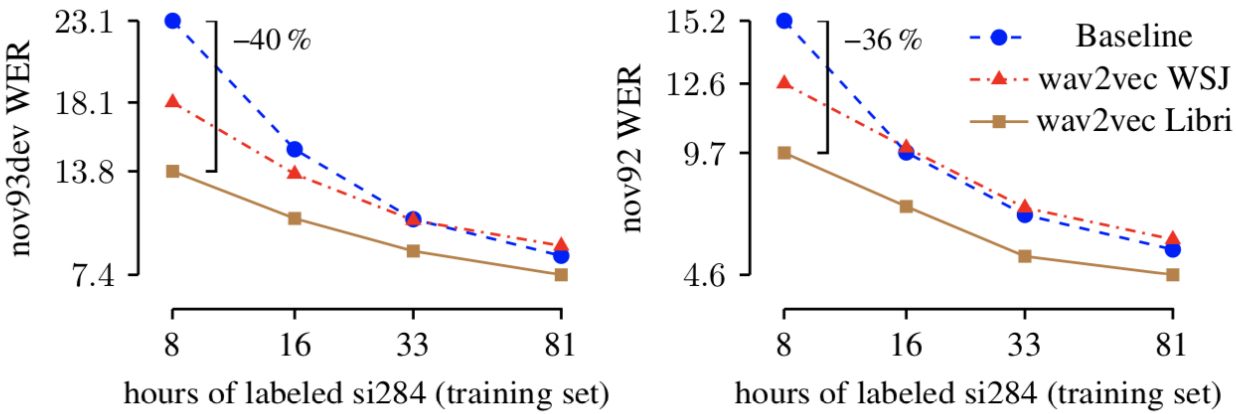


Figure 2: Pre-training substanstially improves WER in simulated low-resource setups on the audio data of WSJ compared to wav2letter++ with log-mel filterbanks features (Baseline). Pre-training on the audio data of the full 960 h Librispeech dataset (wav2vec Libri) performs better than pre-training on the 81 h WSJ dataset (wav2vec WSJ).

Contrastive Predictive Coding (CPC)

Experiments

- Results on TIMIT (phone recognition)
- Focus on acoustic modeling

	dev	test
CNN + TD-filterbanks (Zeghidour et al., 2018a)	15.6	18.0
Li-GRU + MFCC (Ravanelli et al., 2018)	–	16.7 ± 0.26
Li-GRU + FBANK (Ravanelli et al., 2018)	–	15.8 ± 0.10
Li-GRU + fMLLR (Ravanelli et al., 2018)	–	14.9 ± 0.27
Baseline	16.9 ± 0.15	17.6 ± 0.11
wav2vec (Librispeech 80h)	15.5 ± 0.03	17.6 ± 0.12
wav2vec (Librispeech 960h)	13.6 ± 0.20	15.6 ± 0.23
wav2vec (Librispeech + WSJ)	$\mathbf{12.9 \pm 0.18}$	$\mathbf{14.7 \pm 0.42}$

Table 2: Results for phoneme recognition on TIMIT in terms of PER. All our models use the CNN-8L-PReLU-do0.7 architecture (Zeghidour et al., 2018a).

Figure: Figure from (Schneider et al., 2019b)

Contrastive Predictive Coding (CPC)

Take home message

- Experiments on WSJ show that this approach not only improves resource-poor setups but also settings where all WSJ training data is used
- More data for pre-training improves performances
- Could also be used for other tasks (speech translation)
- Pre-trained model on English Librispeech made available[1]
- (I could not find details on pre-training time...)

[1] `https://github.com/pytorch/fairseq/tree/master/examples/wav2vec`

Representation learning with multiple self-supervised tasks

- Problem-agnostic speech encoder (PASE) (Pascual et al., 2019)
- PASE+: robust speech recognition in noisy and reverberant environments (Ravanelli et al., 2020)

Problem-agnostic speech encoder (PASE)

- Problem-agnostic speech encoder (PASE) (Pascual et al., 2019)
- Jointly tackle multiple self-supervised tasks using an ensemble of neural networks that cooperate to discover good speech representations
- Approach requires consensus across tasks, more likely to learn general, robust, and transferable features
- Authors find that such representations outperform more traditional hand-crafted features in different speech classification tasks such as speaker identification, emotion classification, and ASR

Problem-agnostic speech encoder (PASE)

- Encoder: SincNet (Ravanelli and Bengio, 2018) + Convblocks (receptive field 150ms)
- Workers: one for each task (see next slide)

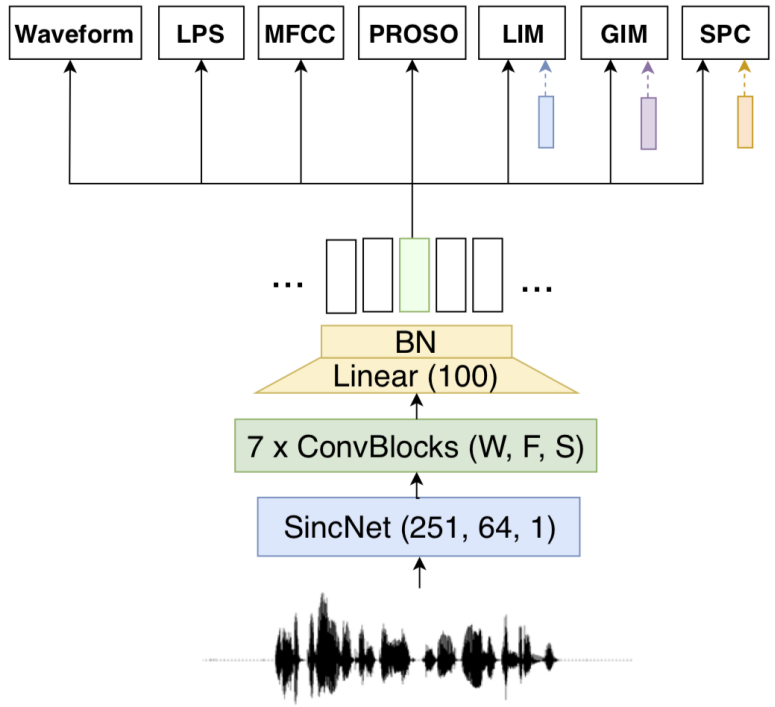


Figure 1: *The PASE architecture, with the considered workers.*

Figure: Figure from (Pascual et al., 2019)

Problem-agnostic speech encoder (PASE)

- Regression workers that solve 7 self-supervised tasks
- Trained to minimize the mean squared error (MSE) between the target features and the network predictions
  - **Waveform** learns to reconstruct waveforms
  - **LPS** reconstruct log power spectrum
  - **MFCC** reconstruct mel-frequency cepstral coefficients
  - **Prosody** predicts 4 basic prosodic features per frame
  - **LIM** (local info max) contrastive task where positive sample is drawn from the same utterance and a negative sample is drawn from another random utterance (that likely belongs to a different speaker)
  - **GIM** (global info max) similar to LIM using global representations (averaged over 1s) instead of local ones
  - **SPC** sequence predicting coding: similar to contrastive predictive coding (CPC) introduced earlier

Problem-agnostic speech encoder (PASE)

- Experiments on speaker identification, emotion recognition and ASR

Table 2: *Accuracy comparison on the considered classification tasks using MLPs and RNNs as classifiers.*

Model	Classification accuracy [%]					
	Speaker-ID *(VCTK)*		Emotion *(INTERFACE)*		ASR *(TIMIT)*	
	MLP	RNN	MLP	RNN	MLP	RNN
MFCC	96.9	72.3	90.8	91.1	81.1	84.8
FBANK	98.4	75.1	94.1	92.8	80.9	85.1
PASE-Supervised	97.0	80.5	93.8	92.8	82.1	84.7
PASE-Frozen	97.3	82.5	91.5	92.8	81.4	84.7
PASE-FineTuned	**99.3**	**97.2**	**97.7**	**97.0**	**82.9**	**85.3**

Table 3: *Word error rate (WER) obtained on the DIRHA corpus.*

	WER [%]
MFCC	35.8
FBANK	34.0
PASE-Supervised	33.5
PASE-Frozen	32.5
PASE-FineTuned	**29.8**

Figure: Table from (Pascual et al., 2019)

PASE+: Robust Speech Recognition in Noisy and Reverberant Environments

- PASE+: robust speech recognition in noisy and reverberant environments (Ravanelli et al., 2020)
- Speech distortion to "contaminate" the input signals with a variety of random disturbances

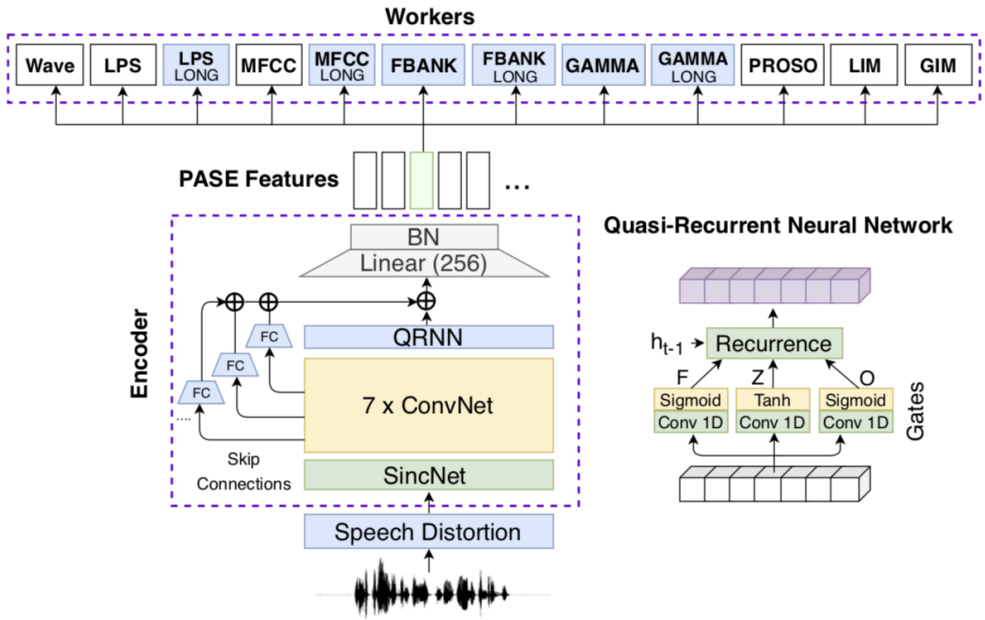


Figure: Figure from (Ravanelli et al., 2020)

PASE+: Robust Speech Recognition in Noisy and Reverberant Environments

- Different kinds of distorsions applied

Disturbance	*p*	*Description*
Reverberation	0.5	Convolution with a large set of impulse responses derived with the image method.
Additive Noise	0.4	Non-stationary noises from the FreeSound and the DIRHA datasets.
Frequency Mask	0.4	Convolution with band-stop filters that randomly drops one band of the spectrum.
Temporal Mask	0.2	Replacing a random number of consecutive samples with zeros.
Clipping	0.2	Adding a random degree of saturation to simulate clipping conditions.
Overlap Speech	0.1	Adding another speech signal in background that overlaps with the main one.

Table 1. List of the distortions used in the speech contamination module (each one activated independently with probability of p).

Figure: Table from (Ravanelli et al., 2020)

Unsupervised speech representation learning using WaveNet autoencoders (Chorowski et al., 2019)

- Learn latent representations that encode phonetic content only
- Rely on Wavnet decoder (used in TTS) to infer information that was rejected by the encoder
  - acts as an inductive bias to allow encoder focusing on high level semantic features
- Made up of 3 parts
  - an encoder that computes a stream of latent vectors from speech input
  - a bottleneck that encourages the encoder to discard portions of the latent representation which the decoder can recover
  - a decoder which reconstructs the speech waveform

Unsupervised speech representation learning using WaveNet autoencoders

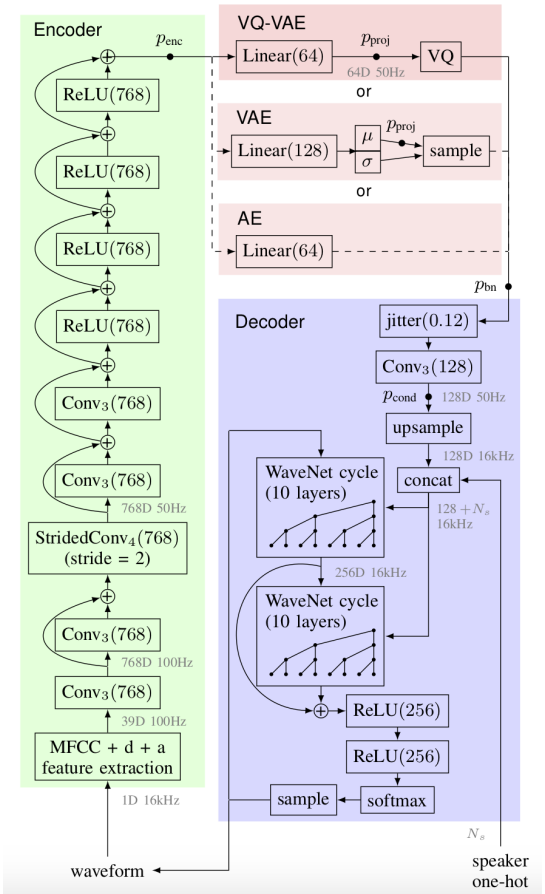


Figure: Figure from (Chorowski et al., 2019)

Unsupervised speech representation learning using WaveNet autoencoders

- Three bottlneck variants are proposed
  - a simple dimensionality reduction (AE)
  - a Gaussian VAE with different latent representation dimensionalities and different capacities
  - a VQ-VAE with different number of quantization prototypes
- bottlenecks optionally followed by the dropout inspired time-jitter regularization
  - during training, each latent vector can replace either one or both of its neighbors
  - promotes latent representation stability over time

Unsupervised speech representation learning using WaveNet autoencoders

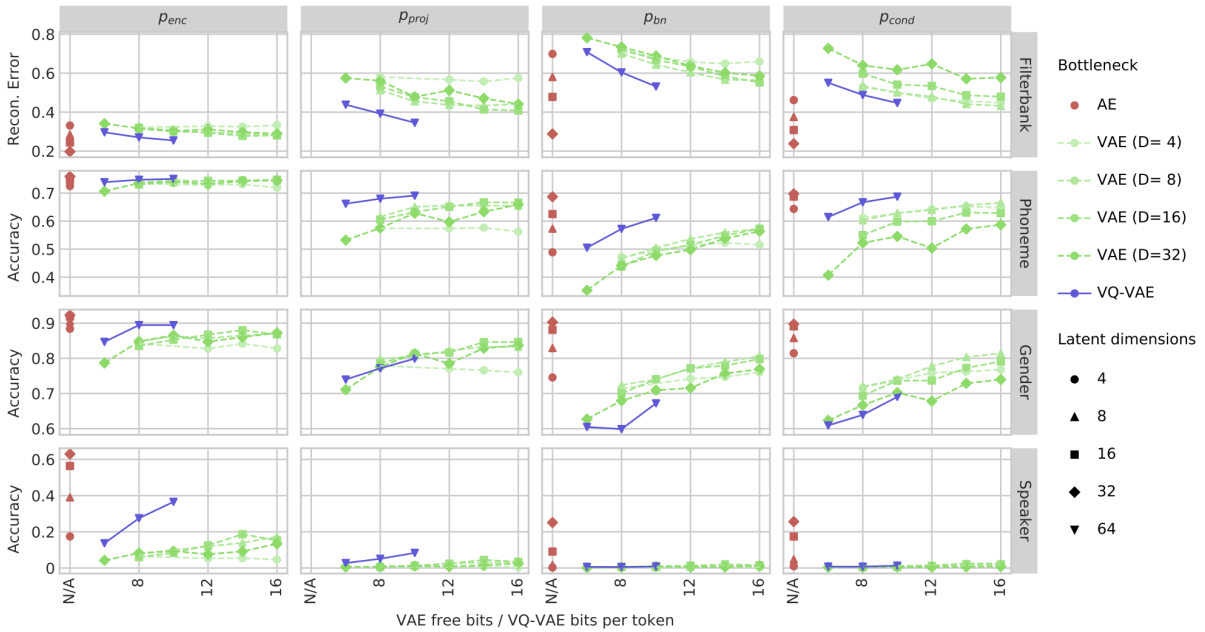


Fig. 2. Accuracy of predicting signal characteristics at various probe locations in the network. Among the three bottlenecks evaluated, VQ-VAE discards the most speaker-related information at the bottleneck, while preserving the most phonetic information. For all bottlenecks, the representation coming out of the encoder yields over 70% accurate framewise phoneme predictions. Both the simple AE and VQ-VAE preserve this information in the bottleneck (the accuracy drops to 50%-60% depending on the bottleneck's strength). However, the VQ-VAE discards almost all speaker information (speaker classification accuracy is close to 0% and gender prediction close to 50%). This causes the VQ-VAE representation to perform best on the acoustic unit discovery task – the representation captures the phonetic content while being invariant to speaker identity.

Figure: Table from (Chorowski et al., 2019)

Beyond left-to-right

- Speech-XLNet: Unsupervised Acoustic Model Pretraining For Self-Attention Networks (Song et al., 2019)
- Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders (Liu et al., 2020)
- Unsupervised pre-training of bidirectional speech encoders via masked reconstruction (Wang et al., 2020)

Speech-XLNet

- Speech-XLNet (Song et al., 2019)
- Learn speech representations with self-attention networks
- BERT-like autoencoding (AE) scheme to train a bi-directional speech representation model (not only left-to-right)
- Mask and reconstruct speech frames rather than word tokens (regression instead of classification task)
- Encourage network to learn global structures by shuffling speech frame orders (can be also seen as dynamic data augmentation)
- Training using a Mean Absolute Error (MAE) loss (that combines L1 and L2 losses) over several permutations of the input frames

Speech-XLNet

- Experiments on Hybrid and end-to-end ASR
- Hybrid ASR on TIMIT are reported below

Table 2: *PER comparison with previous pretrain methods. We approximate the number of parameters based on the description in the previous studies.*

Pretrain Method	Pretrain Data	Pretrain Params	Dev/Test PER(%)
VQ-Wav2vec ([8])	libri (960h)	34M	15.34 / 17.78
RBM-DBN ([21])	timit (8h)	$\approx$ 34.2M	15.90 / 16.80
Ours (Randomly Init)	-	19.9M	13.20 / 15.10
Wav2vec ([7])	libri+wsj (1041h)	34M	12.90 / 14.70
Ours (Pretrained)	libri+wsj+ted (1248h)	19.9M	11.70 / 12.80
VQ-Wav2vec+BERT ([8])	libri (960h)	$\approx$ 71.8M	9.64 / 11.64

Figure: Table from (Song et al., 2019)

Unsupervised speech representation learning with deep bidirectional transformer encoders

- Predict the current frame through jointly conditioning on both past and future contexts (Mockingjay (Liu et al., 2020))
- Masked acoustic modeling task (rand. mask 15% of input frames)[2]
- Use multi-layer transformer encoders and multi-head self-attention
- Add a prediction head (2 layers of feed-forward network with layer-norm) using last encoder layer as input

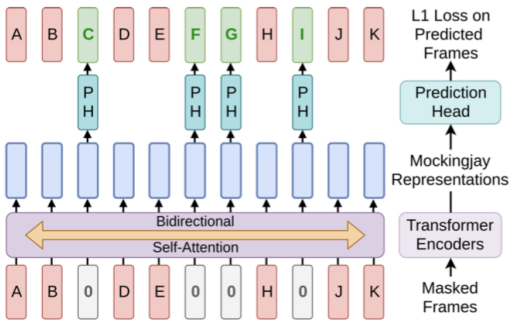


Figure: Table from (Liu et al., 2020)

[2]Use of additional consecutive masking where they mask consecutive frames C_{num} to zero. The model is required to infer on global rather than local structure.

Unsupervised speech representation learning with deep bidirectional transformer encoders

- Experiments on phoneme classification
- With different amount of annotated data for training

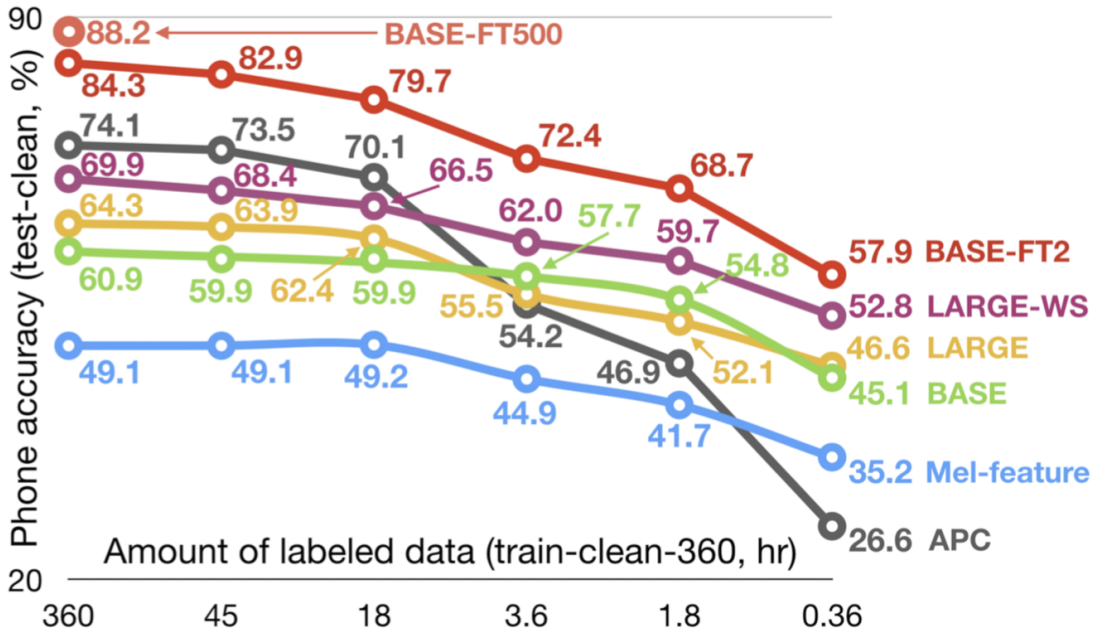


Figure: Figure from (Liu et al., 2020)

Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction

- Pre-training speech representations via a (BERT-style) masked reconstruction loss (Wang et al., 2020)
- Masking in both frequency and time to encourage model to exploit spatio-temporal info
- Elegant extension of data augmentation technique SpecAugment (Park et al., 2019)

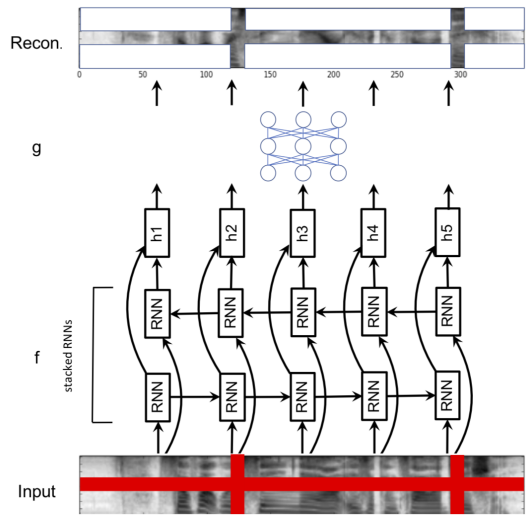


Figure: Figure from (Wang et al., 2020)

Unsupervised Pre-training of Bidirectional Speech Encoders via Masked Reconstruction

Table 3: Dev set %CERs of character-based systems pre-trained on LibriSpeech, and fine-tuned with different amounts of supervised data.

	Baseline	Pre-train *Libri.* w/o LIN	Pre-train *Libri.* w/ LIN
si84	15.23	14.02	**13.29**
+ SpecAug	12.98	12.26	**11.70**
si284	7.01	6.90	**6.48**
+ SpecAug	6.29	6.19	**5.61**

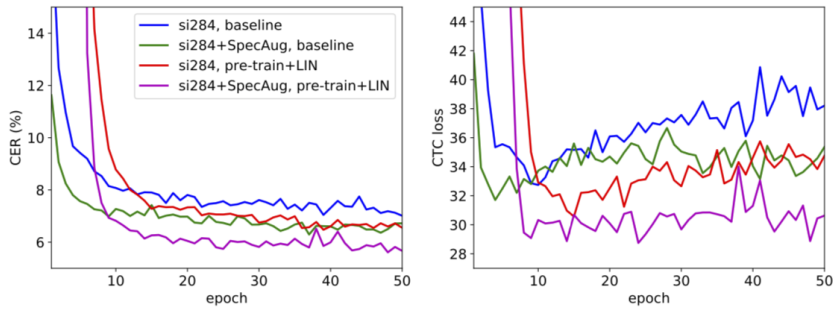


Fig. 2: Dev set learning curves (%CER and CTC loss) of different systems pre-trained on *LibriSpeech*. The first 5 epochs of fine-tuning update only the LIN and softmax layers.

Figure: From (Wang et al., 2020)

A Convolutional Deep Markov Model for Unsupervised Speech Representation Learning

- A generative model for speech representation learning (Khurana et al., 2020)
- Convolutional deep Markov model (ConvDMM)[3]
- A Gaussian state-space model with non-linear emission and transition functions modelled by deep neural networks
- Directly based on the Deep Markov Model proposed by (Krishnan et al., 2016)
- Outperform multiple self-supervised methods on ASR and is complementary with self-supervised methods like Wav2Vec and PASE

[3] Read `https://arxiv.org/pdf/2006.02547.pdf` for more details!

Application to Speech Translation in Low Resource Conditions (submitted to IS2020)

- Investigate the possibility to leverage unlabeled speech for end-to-end automatic speech translation (AST)
- We focus on scenarios where
  - (a) recordings in source language are not transcribed[4] (no ASR pre-training is possible),
  - (b) only a small-medium amount of training data (speech aligned to translations) is available,
  - (c) a larger amount of unlabeled speech can be used.
- Scenario typical of situations when one builds a system that translates from speech in a language with poorly standardized orthography or even from an unwritten language

[4] Transcription not available or language poorly written

Application to Speech Translation in Low Resource Conditions (submitted to IS2020)

- Speech translation experiments on How2 En-Pt corpus
- Fbank versus Wav2Vec features
- Impact of fine tuning Wav2Vec model on untranscribed speech
- Self supervised representations have strong impact in low resource conditions
- Validation on other language pairs (en-fr and en-de) confirm this trend

No.	Feature	10% (28h)	20% (56h)	30% (84h)	60% (169h)	100% (281h)
1	wav2vec	11.33	26.75	30.83	36.33	41.02
2	wav2vec + FT	12.52	27.30	32.11	37.78	42.32
3	wav2vec + norm	16.52	27.33	31.27	37.62	41.08
4	wav2vec + FT + norm	18.50	27.68	32.17	37.75	41.30
5	fbanks	1.03	18.61	27.32	37.23	41.63
6	fbanks + norm	2.11	24.58	30.21	37.56	42.51
7	Ensemble [5, 6]		25.28	31.90	40.39	44.35
8	Ensemble [4, 6]		29.87	34.67	41.22	45.02
9	**Ensemble [1,2,3,4,5,6]**		**31.88**	**36.80**	**42.62**	**46.16**

Table 2: *Detokenized case-sensitive BLEU scores measured on How2 val set of different models trained on different partitions of How2 corpus (EN-PT) with different speech features.* ***FT*** *means fine-tuned and* ***norm*** *stands for MVN normalization.*

Takeaways

- Self-supervised representation learning from speech using future or past-future context
- Allows to adopt the pre-training + fine tuning methodology widely used in (text) NLP
- Reduce dependence on labeled data for building speech systems (speech transcription, translation, paralinguistics)
- Particularly efficient in low resource scenarios (few annotated data)
- A rapidly evolving sub-domain (from APC/CPC to masked reconstruction models)
- Very recent approaches based on generative models (Khurana et al., 2020)
- **Not shared benchmark to evaluate all these approaches** - need something equivalent to GLUE/SuperGLUE in NLP !!

Questions?

Thank you

References I

Baevski, A., Auli, M., and Mohamed, A. (2019). Effectiveness of self-supervised pre-training for speech recognition.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations.

Chorowski, J., Weiss, R. J., Bengio, S., and van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. CoRR, abs/1901.08810.

Chung, Y., Hsu, W., Tang, H., and Glass, J. R. (2019). An unsupervised autoregressive model for speech representation learning. CoRR, abs/1904.03240.

Chung, Y.-A. and Glass, J. (2020). Improved speech representations with multi-target autoregressive predictive coding.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786):504 – 507.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., and Dupoux, E. (2019). Libri-light: A benchmark for asr with limited or no supervision.

References II

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.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes.

Krishnan, R. G., Shalit, U., and Sontag, D. (2016). Structured inference networks for nonlinear state space models.

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. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

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 2019.

Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., and Bengio, Y. (2019). Learning problem-agnostic speech representations from multiple self-supervised tasks. CoRR, abs/1904.03416.

Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet.

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.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019a). wav2vec: Unsupervised Pre-Training for Speech Recognition. In Proc. Interspeech 2019, pages 3465–3469.

References III

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019b). wav2vec: Unsupervised pre-training for speech recognition. CoRR, abs/1904.05862.

Song, X., Wang, G., Wu, Z., Huang, Y., Su, D., Yu, D., and Meng, H. (2019). Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. (2017). Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems 30, pages 6306–6315. Curran Associates, Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. CoRR, abs/1706.03762.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders.

Wang, W., Tang, Q., and Livescu, K. (2020). Unsupervised pre-training of bidirectional speech encoders via masked reconstruction.