Representation Learning for Automatic Speech Recognition: A Review of Speech-to-Text Methods

Maja Mitreska[1,2], Blagica Penkova[1,2], Kostadin Mishev[1,2] and Monika Simjanoska[1,2]

[1]iReason, LLC, Skopje, N. Macedonia
[2]Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, N. Macedonia
{maja.mitreska.1, blagica.penkova}@students.finki.ukim.mk,
{kostadin.mishev, monika.simjanoska}@finki.ukim.mk
,
{maja.mitreska, blagica.penkova, kostadin.mishev, monika.simjanoska}@ireason.mk

***Abstract*—Representation learning has emerged as a promising approach to overcoming the limitations of discriminative representations from the raw speech signal. In this review, we cover a range of speech-to-text methods that employ representation learning, including deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models. The advantages and limitations of each approach are described, as well as recent advances in pretraining techniques such as contrastive predictive coding (CPC) and masked language modelling (MLM). The reviewed papers are divided according to their novelty, their approaches and their type of representation learning models**

***Index Terms*—Speech-to-text, representation learning**

I. Introduction

Speech recognition also known as Automatic Speech Recognition (ASR) is the process of converting speech into text by a computer program. The field of speech recognition is quite explored as it exists for a long time. The thrill of building such systems was motivated by the desire to make communication with the machines as much human-like as possible as speech is the primary tool of communication between humans. As the field expanded, more and more challenges appeared. However, as technology advanced, various approaches could be applied to resolve the problems. At the beginning in 1952, the research started with the first speech recognition system which recognizes the digits of a single speaker. The system was created in the Bell Laboratories by the researchers Davis KH., Biddulph R. and Balashek S. [1]

In the period between 1950 and 1960, the focus was on the spectral resonances during the vowel region of each utterance extracted from the output signals of an analogue filter bank and logic circuits. In the '70s and '80s, the general problems were revolving around the dynamic programming paradigm, statistical modelling, hidden Markov models, n-grams, etc. The '90s were years when most innovations centred around pattern recognition and the application of these systems in telephone networks. With the beginning of the new millennium, the research headed into the exploration of continuous and robust speech recognition [2], [3].

With the progress of neural networks and especially deep neural networks and their application for speech recognition [4], a huge number of novel neural architectures were introduced to solve the challenges.

Representation learning is a class of machine learning approaches that means that an algorithm is able to learn the data representations from raw input and make it simpler to extract meaningful information when developing classifiers [5]. The representations that have been learned are interpretable, include latent features, and may be utilized for transfer learning. Often, deep neural networks are considered respectable representation learning models as they encode information into different subspaces. Moreover, representation learning can be divided into supervised and unsupervised learning and also self-supervised learning.

In this paper, we provide an overview of the recent advances in speech-to-text methods based on representation learning. Furthermore, we will review the different approaches to representation learning and discuss their strengths and weaknesses, and highlight some of the most promising applications of these methods, including their potential for developing more accurate and robust speech recognition systems for a variety of real-world applications.

II. Methods

This section presents a brief review of the recent achievements related to developing speech-to-text methods based on representation learning. Each method introduces a new approach that resolves or improves some of the existing challenges in the field.

As we mentioned, the speech-to-text methods depending on the training and representation learning approach can be divided into supervised, unsupervised, semi-supervised, self-supervised, etc.

In the next few paragraphs, multiple models which are proven to be influential in the field are chronologically described in each section. All of the models, except the ones in the section II-A, are published in the previous few years and are considered one of the most significant in this field. The models in II-A are as important as the others, but as the field progressed the research became more focused on exploring the unsupervised ways of training and learning due to the challenges and the lack of labelled data in the supervised approach.

A. Supervised Approach

The supervised approach gives promising results since the representations of the features are learnt from a labelled dataset. However, due to the limited labelled data and the challenges of acquiring transcriptions the supervised approach deals with a lack of available resources. Nevertheless, there are few papers that are exploring this technique.

In [6] a CNN model is used to learn the relationship between the raw speech signal and the phones. That way the CNN layers are modelling the phone-specific spectral envelope information of 2-4 ms speech. Moreover, the same authors are exploring other features learned from a raw speech by the CNN-based approach that could generalize across different databases in [7]. The Wall Street Journal (WSJ) corpus and the TIMIT corpus are used in the training phase. Two experiments were performed, the word recognition experiment is done over the WSJ corpus using features learned from the TIMIT corpus, and for the phoneme recognition task, the model is trained on the TIMIT corpus using features learned from WSJ. Another paper that focuses its research on phoneme classification using the TIMIT corpus is [8]. It can be noted that most of the research in the mentioned period concentrates on the representation learning of features using convolutional networks trained on labelled datasets for phonemes or word recognition [9]. In [10] a time convolutional layer is added in a CLDNN model for reducing the temporal variations and a frequency convolutional layer for reducing the frequency variations. Here, the representation of the features is in waveforms and the features are learned from an artificially corrupted dataset consisting of YouTube recordings.

As we have seen from the presented papers, most of them are using the same datasets since the resources for supervised training are limited. There are only a few datasets that contain labels for each input sample.

1) Unsupervised Approach: Encouraged by the scarce resources for supervised training, more and more researchers are discovering and exploring the benefits of the unsupervised approach. Generally, unsupervised training tries to find patterns in unlabelled data, so representation learning is quite suitable for this approach.

TABLE I: Training approaches in different papers

Learning Type	Papers
Supervised	[6], [7], [8], [10], [9]
Unsupervised	[11], [12], [13], [14], [15]
Self-Supervised	[16], [17], [18], [19], [20]
Semi-Supervised	[21], [22], [23], [24], [25], [26]

One of the most prominent methods in the unsupervised approach is the distinguished wav2vec model [11]. Wav2vec is an unsupervised method which is trained on large unlabeled audio corpora and the output representations are used to improve the training of an acoustic model. The model represents a convolutional neural network which learns the general representations of the raw audio inputs that are further fed as input to a speech recognition system. The representation learning is done via two networks. The first one is an encoder network that consists of a five-layer CNN and the second network is the context network consisting of nine layers which combines multiple latent representations of the encoder to acquire a single contextualized representation. After the representation learning, the features are evaluated on acoustic models for classification. The approach is evaluated with the Word Error Rate (WER) metric and it outperforms the state-of-the-art model at that time.

This method, wav2vec is an inspiration for other authors to base their research on improving the already proven architecture. On that note, vq-wav2vec [12] examines the method of learning discrete representations by a new quantization module of audio segments leveraging the original wav2vec prediction task and obtains results that have phoneme error rate (PER) by 3 less than the original work.

In [13] an unsupervised speech representation learning method is proposed that extracts features of more the 8000 hours of diverse data. Unlike most of the papers that evaluate the representations based on the performance of the speech recognition system, here the representations are evaluated in terms of their robustness to domain shifts. The model for representation learning is an acoustic feature extractor trained with a bidirectional variant of contrastive predictive coding (CPC) on unlabeled data. Another paper that uses CPC is [14]. Here, autoregressive models and negative sampling are utilized to predict features in latent space. The paper [15] introduced a new English audio dataset for speech recognition and proposed a baseline unsupervised system trained on CPC features that exceeds the results from the MFCC baseline model for comparison.

2) Self-Supervised Approach: The self-supervised approach can be considered as an approach that combines the benefits of the supervised and the unsupervised approaches. Following the criteria of using unlabeled data, the approach strives to learn the input using other parts from it. The models trained in a self-supervised way auto-generate labels and assign those labels to themselves, thus transforming the problem into a supervised one.

The authors of [11] are upgrading and improving their original wav2vec approach with the proposal of the wav2vec 2.0 model [16]. This time the model incorporates a self-supervised learning technique to learn the speech representations. The wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task specified over a quantization of the jointly learnt latent representations. The difference with the previous approach is that, here, spans of the encoded representations after the first network (CNN) are masked similarly to the techniques used in masked language modelling in BERT. The second network is a Transformer network whose output represents the contextualized representations. The contextualized representation is obtained when the Transformer network is trained on a contrastive task. The output is discretized via product quantization and the Gumbel softmax layer to represent the targets of the self-supervised objective. Furthermore, the whole pre-trained model on unlabeled audio data can be fine-tuned on downstream speech recognition tasks using labelled data with the Connectionist Temporal Classification (CTC) loss.

Another paper that follows the footprints of wav2vec 2.0 but introduces a different quantization process is the wav2vec-c model [17] which uses an additional consistency network for the quantization process. More precisely, it learns to reconstruct the input features to the wav2vec 2.0 network from the quantized representations in a manner similar to a VQ-VAE model. This approach solves the codebook utilization problems that are appearing in the wav2vec 2.0 approach.

Moreover, another wav2vec variant, wav2vec-U [18] leverages the adversarial training of the speech representations mapped to phonemes using unlabeled audio data. Because self-supervised representation learning is used the good representations of the unlabeled audio and text data are the key to the success of the method.

As we presented, most of the recent work done in this field bases its methods over the wav2vec [11] and explores different ways of improving it [12], [16]–[18], [21], [27]. Moreover, [27] present an interesting approach of combining the acoustic model wav2vec 2.0 and a language model BERT to fuse and utilize the contextual information in both speech and text. A specifically designed representation aggregation module is integrated to acquire the acoustic and linguistic representation to solve representation differences and an additional embedding attention module to forward the acoustic information into the language layer of BERT to solve the embedding inconsistencies between the two types of input.

On the other side, HuBERT [19] (Hidden-Unit Bert) is another architecture that changes the course of the field. This model is a self-supervised approach for speech recognition learning which incorporates an additional offline clustering phase so it can provide aligned target labels for a BERT prediction loss. The prediction loss is only applied over the masked regions to build a combined model of the acoustic and language continuous inputs. It is of great importance for the model to first learn the continuous latent representations from the unmasked inputs and then to learn the long relations between the representations. The authors are proposing an acoustic unit discovery system to provide frame-level targets because simple discrete latent variable models can infer hidden units. Even more, they are exploring the idea of learning with cluster ensembles and adding iterative refinement of the cluster assignments to improve the representation learning model. Moreover, the representation learning is enhanced via masked prediction where a masked prediction model predicts the distribution over the target instances. Similarly to HuBERT, DistilHuBERT [20] introduces a distilled version of the HuBERT model in order to reduce the required large memory and high pre-training costs. DistilHuBERT learns the distil hidden representations of the HuBERT model directly. To obtain these representations, the authors based their idea on the original HuBERT, but they proposed that the predicting of the hidden representations in the 4th, 8th and 12th hidden layer of the teacher is executed with separate prediction heads. This multi-task learning paradigm is one of the preliminary studies for the compression of large speech recognition systems.

3) Semi-supervised Approach: The semi-supervised approach is a technique that includes small amounts of labelled data during training or fine-tuning and a lot of unlabelled data.

The paper [21] explores a more general approach where the domain of the unlabeled data used for pre-training is different from the labelled data used for fine-tuning. This technique brings many benefits as it improves the performances of the models and enables training models on various unlabeled data. The purpose of this research is to analyze how the different types of in-domain and out-of-domain data for both unlabeled and labelled datasets either in pre-training or fine-tuning affect the accuracy of the models. The conclusion is that adding unlabeled in-domain data highly improves the performance gap between models trained on in-domain labelled data and supervised out-of-domain models. Moreover, the representations obtained this way enable the development of more robust and general models. Moreover, a different paper [22] presents a model built with noisy student training with SpecAugment using Conformer models pre-trained on wav2vec 2.0. The proposed model employs the semi-supervised learning approach to obtain state-of-the-art results on unlabeled speech audio data. Namely, the approach combines two training techniques, iterative self-training, and pre-training. The pre-trained models are used to initialize models for iterative self-training using the pipeline of noisy student training with adaptive SpecAugment. The model architecture is based on the Conformer (a convolution-augmented Transformer for speech recognition is a model that combines the benefits of both worlds, the CNN and the Transformers) meaning that the Conformer encoder and LSTM decoder are used as an ASR network. The network is pre-trained using the wav2.vec 2.0 training objective on the log-mel spectrograms instead of the raw waveforms as in the original approach.

One more study concentrates on the efficiency of the Conformer named Squeezeformer [23]. This study uses the same training scheme but incorporates a couple of different modules. Firstly, it integrates a Temporal U-Net structure and a modification in the feed-forward module where the original Macaron structure is replaced with simpler modules of multi-head attention or convolutional layers. Similarly to [28] it incorporates a depth-wise down-sampling layer to sub-sample the input and to reduce the temporal redundancy between adjacent features.

SpeechStew [24] is an end-to-end speech recognition model that leverages multi-domain and transfer learning. This is a model that is trained on all publicly available datasets without utilizing any domain-specific labels. The model is implemented using the Conformer RNN-T architecture without

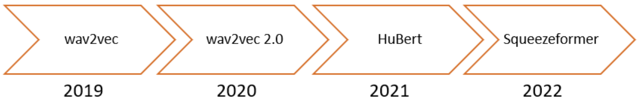

Fig. 1: Timeline of the most influential models in the past few years in the field of representation learning for ASR

incorporating any additional language models and trained on seven datasets leveraging all available resources.

Another unique technique in semi-supervised ASR uses representation learning to exploit a huge quantity of unlabeled data by reconstructing a temporal slice of filter-bank features from previous and future context frames. The technique is called DeCoAR [25] which comes from deep contextualized acoustic representations that are then fed to a CTC classifier (BLSTM layers with CTC loss). Furthermore in [26], the authors are enhancing DeCoAR's current approach with vector quantization. The presented novelties are the shift of the LSTM architecture to Transformer encoders, the addition of a vector quantization layer between the encoder and the reconstruction modules, and a new objective function that focuses on training and obtaining representations by combining the reconstructive loss with the vector quantization diversity loss. The conversion from LSTM to Transformer is obvious since the Transformer architecture is a proven methodology for obtaining deep contextualized representations.

The abundance of data is driving so many researchers to semi-supervised and self-supervised algorithms in speech recognition. In such systems, there is far more unlabeled data than labelled data. Although labelled data is more convenient to use, unlabeled data allows us to discover and extract representations that maintain far more knowledge and information than labelled data. All papers mentioned in this section are presented in Table I where each paper is assigned to its appropriate group depending on the training approach.

In Figure 1 we present the most influential models for representation learning. We can observe how the field made its progress starting with the original wav2vec model and its upgrade to wav2vec 2.0. Without a doubt, these models transformed the existing research and opened new horizons for experimenting with different architectures and models for obtaining feature representations suitable for speech recognition models. The previous couple of years produced important models that are considered state-of-the-art architectures.

III. Conclusion

In this paper, we presented a survey of the most recent advancements in the field of automatic speech recognition. Many of them are indeed improvements of the wav2vec model where various approaches and methods are explored. Considering that all of these papers are published in the previous two years, it can be concluded that the field of automatic speech recognition is still unexplored and more and more researchers are trying to enhance the existing systems in a self-supervised manner. In times when there is a lot of unlabeled data and the competition is fierce, self-supervised techniques are necessary. Moreover, the development of assistive technologies and their applications in the everyday life, makes the ASR systems an even more exciting topic regarding their conversion to cross-lingual models that can learn and serve multiple languages.

References

[1] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[2] S. Furui, "50 years of progress in speech and speaker recognition research," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 1, no. 2, pp. 64–74, 2005.

[3] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," 2010. DOI: 10.48550/ARXIV.1001.2267.

[4] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE access*, vol. 7, pp. 19 143–19 165, 2019.

[5] Y. Bengio, A. C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR*, vol. abs/1206.5538, 2012. arXiv: 1206.5538.

[6] D. Palaz, R. Collobert, *et al.*, "Analysis of cnn-based speech recognition system using raw speech as input," Idiap, Tech. Rep., 2015.

[7] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 4295–4299.

[8] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[9] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve, and E. Dupoux, "Learning filterbanks from raw speech for phone recognition," in *2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP)*, IEEE, 2018, pp. 5509–5513.

[10] T. Sainath, R. J. Weiss, K. Wilson, A. W. Senior, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," 2015.

[11] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[12] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[13] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord, "Learning robust and multilingual speech representations," *arXiv preprint arXiv:2001.11128*, 2020.

[14] A. van den Oord, Y. Li, and O. Vinyals, *Representation learning with contrastive predictive coding*, 2019. arXiv: 1807.03748 [cs.LG].

[15] J. Kahn, M. Riviere, W. Zheng, *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7669–7673.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[17] S. Sadhu, D. He, C.-W. Huang, *et al.*, "Wav2vec-c: A self-supervised model for speech representation learning," *arXiv preprint arXiv:2103.08393*, 2021.

[18] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 826–27 839, 2021.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. DOI: 10.1109/TASLP.2021.3122291.

[20] H. Chang, S. Yang, and H. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit BERT," *CoRR*, vol. abs/2110.01900, 2021. arXiv: 2110.01900.

[21] W. Hsu, A. Sriram, A. Baevski, *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *CoRR*, vol. abs/2104.01027, 2021. arXiv: 2104.01027.

[22] Y. Zhang, J. Qin, D. S. Park, *et al.*, *Pushing the limits of semi-supervised learning for automatic speech recognition*, 2020. DOI: 10.48550/ARXIV.2010.10504.

[23] S. Kim, A. Gholami, A. Shaw, *et al.*, *Squeezeformer: An efficient transformer for automatic speech recognition*, 2022. DOI: 10.48550/ARXIV.2206.00888.

[24] W. Chan, D. S. Park, C. A. Lee, Y. Zhang, Q. V. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *CoRR*, vol. abs/2104.02133, 2021. arXiv: 2104.02133.

[25] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6429–6433.

[26] S. Ling and Y. Liu, *Decoar 2.0: Deep contextualized acoustic representations with vector quantization*, 2020. DOI: 10.48550/ARXIV.2012.06659.

[27] G. Zheng, Y. Xiao, K. Gong, P. Zhou, X. Liang, and L. Lin, "Wav-bert: Cooperative acoustic and linguistic representation learning for low-resource speech recognition," *CoRR*, vol. abs/2109.09161, 2021. arXiv: 2109.09161.

[28] M. Burchi and V. Vielzeuf, "Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition," 2021. DOI: 10.48550/ARXIV.2109.01163.