Interspeech 2022
18-22 September 2022, Incheon, Korea

Enhancing Embeddings for Speech Classification in Noisy Conditions

Mohamed Nabih Ali [1], *Daniele Falavigna* [2], *Alessio Brutti* [2]

[1] Information Engineering and Computer Science School, University of Trento, 38121 Trento, Italy
[2] Fondazione Bruno Kessler, 38121 Trento, Italy

mohamed.nabih@unitn.it, falavi@fbk.eu, brutti@fbk.eu

Abstract

Robustness against noise is critical for several speech applications in real-world environments. In general, to improve the robustness, a speech enhancement front-end is integrated as a preprocessing stage, often jointly trained with the network backend to reduce the impact of distortions and artifacts on the performance. Recently, the use of speech representation computed using pre-trained models on large amounts of data, as *Wav2Vec*, has proved to be effective in a variety of speech processing and classification tasks. However, the performance of these models, although very robust, deteriorates in presence of environmental noise. In this paper, we investigate how enhancement can be applied in neural speech classification architectures employing pre-trained speech embeddings. We investigate two approaches: one applies time-domain enhancement prior to extracting the embeddings; the other employs a convolutional neural network to map the noisy embeddings to the corresponding clean ones. Exhaustive experiments on the Fluent Speech Commands and Google Speech Commands corpora, contaminated with noises from the Microsoft Scalable Noisy Speech Dataset, sheds light and provide insights about the most promising enhancement training approaches.

Index Terms: speech embeddings, speech enhancement, joint training

1. Introduction

Despite the large improvements achieved in recent years in the fields of automatic speech recognition (ASR) and understanding, mainly thanks to the usage of effective models pre-trained on large sets of data [1, 2], the problem of understanding speech in the presence of diverse acoustic noises is still a big challenge. As a matter of fact, when the speech signals are contaminated with either environmental noise, reverberation, or in presence of many competing speakers, causing the so called cocktail party effect [3], the resulting intelligibility can be unacceptable (see [4] for some performance figures on this issue). As a consequence, also the corresponding word accuracy and semantic accuracy will not be enough to face with many of the application areas of interests.

To mitigate the effects of these adverse conditions, several speech enhancement (SE) approaches [5, 6], as well as corresponding evaluation metrics [7, 8], have been proposed in the past that try to increase the intelligibility of noisy speech signals. Recently, thanks to the extraordinary improvements in deep neural networks (DNNs), SE has been approached as a deep learning task [9], in particular paving the way towards processing the input waveform in the time domain [10, 11]. In a previous work by us [12] we employed Wave-U-Net [11],

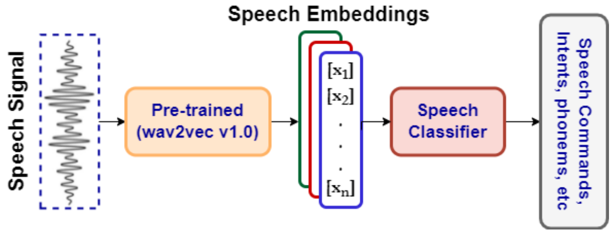

Figure 1: *Speech classification based on pre-trained speech embeddings*

a time domain SE model, as a front-end for an intent classification task. In the paper we demonstrated the effectiveness of joint training a pipeline formed by a SE module (the Wave-U-Net) followed by a time-domain intent classification module (also based on a DNN model): improvements have been reached both in terms of SE quality metrics and accuracy of the detected intents. Recently the use of pre-trained speech embeddings has become extremely popular since the model design and training does not involve the backbone, and is hence much simpler as in [13]. In addition, it makes also effective the implementation of unsupervised training of audio data [14]. Therefore, in this paper we extend our previous contribution investigating speech enhancement strategies for speech classification models employing pre-trained speech embeddings (as *Wav2Vec* [2]) as shown in figure 1. In particular we propose an approach for directly enhancing embeddings using convolutional neural networks (CNN) architectures inside a joint training framework. More in details, we compare two different SE architectures: *a)* speech embeddings are extracted (using *Wav2Vec*) on enhanced waveforms on top of a SE network; *b)* speech embeddings extracted from noisy speech are enhanced. The two architectures will be explained in the details in section 3 and in section 5. We point out that, to the best of our knowledge, this is the first attempt to apply an enhancement network to speech embeddings derived from noisy speech. Note that working directly on the embeddings, whose dimensionality is typically relatively reduced, allows the design of rather small and easy to train networks. To summarize, the novelty of this work lies in: *a)* the direct enhancement of *Wav2Vec* embeddings extracted from noisy speech, and *b)* the joint training of the (embedding based) front-end and the classifier back-end.

2. Related Work

In supervised SE, neural networks are trained using pairs of clean and noisy speech signals without any statistical assumptions [15]. Many of these approaches employ the short-time Fourier transform (STFT), either estimating the clean signal magnitude [16], or the Time-Frequency (T-F) masks used to separate the T-F bins associated to the target speech signal and noise [17, 18, 19]. The main drawback of all of these methods is that they reconstruct the magnitude of the spectrum, while the phase remains noisy. In addition, masks often introduce ar-

This work has been supported by the AGEVOLA project (SIME code 2019.0227), funded by the Fondazione CARITRO.

10.21437/Interspeech.2022-10707

tifacts which may affect the signal quality as well as the performance of following processing components. As an alternative, time-domain methods, which estimate enhanced signals directly in the time domain, have emerged recently [20, 21, 22, 11]. Their major advantage is that they can mitigate the phase estimation problem, which improves the speech quality and intelligibility [23].

In several applications involving noisy speech, SE is applied as a "front-end" module followed by a "back-end" one, which addresses the actual task, e.g. classifying speech. Often the front-end and the back-end are trained separately, however recent works have proven that joint training both modules provides better performance, both in terms of SE and classification or regression accuracy. Below we will summarize some of the (to our knowledge) most relevant recent works on joint training, the majority of which focuses on keyword spotting (KWS)

In [24], the authors addressed the problem of supervised noise reduction for KWS by employing a microphone-array SE front-end in the STFT domain. The front-end supervised by the KWS led to an approximately 32% improvement over their baseline. In an analogous way, the work in [25], proposes to jointly train a pre-trained front-end speech SE model with a CNN-based KWS classifier. In [26, 27], the wake-up word detection task in noisy conditions is addressed by optimizing the whole system using a linear combination of a reconstruction-based loss (either at log-mel spectrogram level or waveform level) with the cross-entropy loss for wake-up word detection. Exhaustive experiments with different classifiers show that jointly training improves the overall performance. Slightly different is the work in [28] which addresses KWS in multi-speaker environments. In particular, the authors propose a joint training approach incorporating a multi-look enhancement front-end that combines spectral, inter-channel phase difference, and directions associated features with a back-end KWS. All the reported works do not involve speech embeddings in the training procedure. This is instead done in [29] that proposes a weighted loss function that combines three losses, including the mean squared error between the embeddings extracted from the noisy speech and the enhanced output.

Whit respect to the current state of the art, our work investigates joint training with speech embeddings in the time domain and it presents an attempt to apply enhancement directly to the embeddings.

3. Proposed enhancement schemes and joint training losses

As mentioned in the Sec. 1 we investigate on two different neural architectures, shown in Figure 2, to classify noisy speech. Both architectures consist of a stack formed by a SE front-end and a back-end that classifies speech according to the task at hands (i.e. intent or keywords classification). The difference between the two architectures lies in where embeddings are computed: in Figure 2a) the *Wav2Vec* module is applied on top of the enhancing network, we refer to this approach as **Wave-Enh**; in Figure 2b) *Wav2Vec* is applied before (i.e. on bottom of) the speech enhancement module, we denote this solution as **Embeds-Enh**. Note that in the latter case the enhancement module operates directly on the embedded representations.

Both architectures of Figure 2 are trained by optimizing a joint loss function that combines a speech enhancement loss ($\mathcal{L}_{SE}$) with a classification loss ($\mathcal{L}_{cl}$), that is:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{SE} + (1-\alpha)\mathcal{L}_{cl} \qquad (1)$$

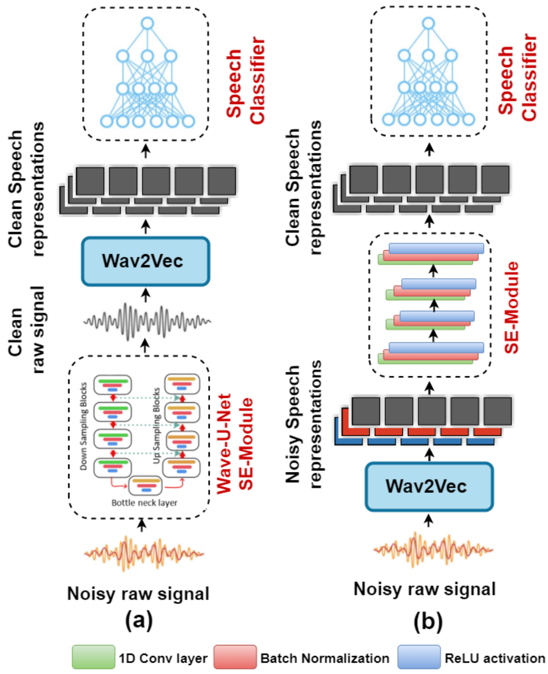

Figure 2: *The two proposed enhancement strategies: (**a**) Wave-Enh, speech waveforms are enhanced prior to embedding extraction; (**b**) Embeds-Enh, embedding are enhanced instead.*

In this paper we adopt the mean squared error (MSE) as the SE loss and the cross-entropy for the speech classification loss. The coefficient $\alpha \in (0, 1)$ is a hyper-parameter that adjusts the weight of each component in the joint loss. In this research, we experiment on a grid of values for α, i.e. (0.1, 0.5, 0.9). More details about other hyper-parameters are given in Section 6.

Note that, from equation 1, the gradient of the classification loss (i.e. $\mathcal{L}_{cl}$) is back-propagated across the layers of the SE module, so that the update of the related parameters, namely Θ_{SE}, can be expressed as follows:

$$\Theta_{SE} \leftarrow \Theta_{SE} - \lambda_1[\alpha\nabla\mathcal{L}_{SE} + (1-\alpha)\nabla\mathcal{L}_{cl}] \qquad (2)$$

where $\nabla\mathcal{L}_{SE}$ and $\nabla\mathcal{L}_{cl}$ represent the gradients of SE and classifier loss respectively, and λ_1 is the learning rate used to train the SE front-end. In this way the SE module provides enhanced signals that maximize the classifier performance. Unlike the front-end, the back-end network parameters depend only on the classification loss and they are updated as follows:

$$\Theta_{cl} \leftarrow \Theta_{cl} - [(1-\alpha)\lambda_2\nabla\mathcal{L}_{cl}], \qquad (3)$$

where Θ_{cl} denotes the classifier parameters, and λ_2 is the classifier learning rate.

4. Datasets and tasks

We have evaluated the proposed enhancement architectures on: an intent classification task, using the Fluent Speech Commands (FSC) dataset [30] and a keyword spotting task, using the Google Speech Commands dataset v1 (GSC) [31].

FSC includes 30,043 English utterances, recorded at 16 kHz. Overall, 97 native and non native speakers (52 males, and 49 females) are recorded while simulating an interaction with voice-enabled appliances or smart devices with no overlapping

Table 1: *Statistics (number of utterances and duration) of the FSC and GSC datasets.*

Data	FSC	Hours	GSC	Hours
Train Set	23132	14.7	45931	12.7
Validation SEt	3119	1.9	6799	1.8
Evaluation Set	3793	2.4	6836	1.9

speech in each signal. The dataset provides 248 different utterances that are mapped in 31 different intents.

GSC has 65,000 clips of one-second-long duration recorded as 16 kHz sampling frequency. The audio files are organized into folders based on the word they contain, resulting into 30 different keywords. For both datasets we adopt the official splits as reported in [30, 12] and in Table 1.

4.1. Noisy Datasets

To emulate realistic application scenarios, where environmental noise may affect the classification accuracy, we contaminated both FSC, and GSC datasets by superimposing 6 different types of noise out of the 25 noises available in the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [32], i.e. air conditioner, airport announcements, traffic, neighbor speaking, shutting door, and restaurant. Both datasets are contaminated using the "maracas" library available in [1]. A noise signal is randomly selected from those available and added to each clean signal with an SNR randomly selected from 3 possible values: -5dB, 0dB, 5dB.

5. System description

The computation of the speech embeddings depicted in Figure 1 is carried out with the pre-trained *Wav2Vec* model [1]. In particular we used the model trained on the "LibriSpeech" data set [33][2]. has been demonstrated effective for tackling ASR tasks, even when few supervised data are available[3] for fine tuning. The enhancement and classification networks are described in the next subsections. It is worth noting that, recently, other models pre-trained on very large clean and noisy speech datasets have been made available [34, 35] and will be investigated in future work.

5.1. Enhancement

Wave-U-Net [11], a U-Net architecture [36] specifically designed for speech processing, has been demonstrated to be very effective for time-domain SE, achieving promising results in comparison with other methods. U-Net is a fully convolutional neural network, with a contracting path (encoder network), a bottleneck layer and an expansive path (decoder network). The encoder part follows a typical architecture of a CNN, with multiple of 1D convolutions blocks without padding followed by a ReLU activation function and a max-pooling operation for downsampling. While the model downsample the space in factor of 2, it doubles the number of feature channels in the network [36]. The decoder network consists of an upsampling operation of the feature map (usually a transposed convolution or bi-linear interpolation) that sequentially halves the number of feature channels. Finally, skip-connections are applied to avoid information loss between encoder and decoder blocks.

Wave-U-Net follows the same architecture of U-Net, designed to process fixed-length input signals with sample length 16834 [4]. Therefore, signals longer than that are first split in segments and then concatenated at the output of Wave-U-Net. The model has 12 layers for the encoder part uses $kernelsize = 15$, $stride = 1$, and $padding = 7$, while the decoder has the same number of layers but with $kernelsize = 5$, $stride = 1$, and $padding = 2$. Each layer in both encoder and decoder is followed by a 1D batch normalization layer, and leaky ReLU activation function with negative slope $= 0.1$.

Consider, as input to Wave-U-Net, a noisy signals $y[n] \in [-1, 1]^{L\times C}$, where C is the number of speech channels and L is the number of audio samples ($C = 1$ in our case). In case of monaural speech enhancement, the network is trained to map $y[n]$ to its enhanced counterpart $\hat{x}[n]$ using the clean signal $x[n]$ as the training targets, by minimizing the MSE between $x[n]$ and $\hat{x}[n]$, i.e.:

$$\mathcal{L}_{SE} = \sum_n \|x[n] - \hat{x}[n]\|^2. \quad (4)$$

It is worth noting that, although Wave-U-Net is a well established architecture for time-domain waveform enhancement, there is no evidence of its efficacy for enhancing the speech embeddings. Therefore we experimented with 4 additional architectures namely: the same U-Net described above, U-Net-2, CNN-2, and CNN-4 described below.

U-Net-2 follows the same architecture of the U-net model, but without skip connections. In addition, the encoder has 4 1D-convolutional layers where the output feature channels are successively doubled from layer to layer. The decoder has the same architecture of the encoder, but the output feature channels are halved sequentially.

We investigated the performance of two CNN models with different sizes, called respectively CNN-2, and CNN-4, which has been shown to achieve competitive performance as SE front-end [37]. **CNN-2** consists of a stack of 2 1D-convolutional layers with dimensions ($inchannel$ = 512, $outchannel$ = 256), and ($inchannel$ = 256, $outchannel$ = 512) respectively $kernelsize = 3$, $stride = 1$, and $padding = 1$. **CNN-4** is an extension of the previous **CNN-2** formed by a cascade of 4 1D-convolutional with the same parameters.

In all of these architectures batch normalization and either leaky ReLU or ReLU activation functions are applied after each convolution operation. Models are trained using the Adam optimizer with learning rate $\lambda_1 = 10^{-4}$.

5.2. Classifier

For both tasks, i.e. intent classification and keywords spotting, we propose to use a multi-class architecture to directly map each utterance of either FSC or GSC dataset into the corresponding 31 possible intents or 30 keywords, respectively. This architecture is based on Conv-TAS Net, firstly introduced for time audio separation [38]. The model processes the enhanced speech embeddings from the SE front-end, as depicted in Figure 1. Conv-TAS Net consists of a 1-D convolutional layer that maps the input features into bottleneck features. This layer is followed by 2 convolutional blocks, each of them includes 5 residual blocks. Each residual block consists of 1-D dilated convolutional layer followed by normalization layer and Parametric ReLU activation function (pReLU). The dilation factor is increased in every successive residual block. The back-end classifier inputs are the speech embeddings of dimensions 512. Then, 64 channels

[1] https://github.com/jfsantos/maracas

[2] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

[3] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

[4] https://github.com/haoxiangsnr/Wave-U-Net-for-Speech-Enhancement

for bottleneck features and 128 channels for depth-wise separable convolutional layers are used, respectively. The classifier is trained by minimizing the cross entropy loss between the target and the predicted labels (intents or keywords, respectively), i.e:

$$\mathcal{L}_{cl} = -\frac{1}{T}\sum_{t}\log(p_t) \tag{5}$$

where T is the number of training samples and p_t is the probability of the t^{th} target sample.

The learning rate adopted in both tasks is $\lambda_2 = 10^{-3}$ and the optimization is carried out with the ADAM optimizer with decay rates: $\beta 1 = 0.9$ and $\beta 2 = 0.999$.

6. Experimental analysis

We evaluate our proposed enhancement strategies for speech embeddings considering both dis-joint and joint training. In the dis-joint approach the SE component is first trained individually and the classifier is successively trained on the enhanced data. For the joint training we explore three possible values of the hyperparamter α: 0.1, 0.5, and 0.9. All experiments make use of 100 training epochs. For the Wave-Enh strategy the batch size is 4, and 10 for FSC, and GSC respectively. While of Embeds-Enh the batch size is set to 10, and 20 for FSC and GSC respectively.

Table 2: *Classification accuracy using dis-joint training on FSC for different embedding enhancing models (no. of model parameters are also reported). The performance using Wave-Enh is reported as reference.*

	Wave-Enh	Embeds-Enh			
	Wave-U-Net	U-Net	U-Net-2	CNN-2	CNN-4
Acc	93.40%	79.88%	81.07%	92.51%	92.96%
Pars	10 M	1 M	38 M	788 K	986 K

6.1. Results

First of all we compare in Table 2 the performance achieved on the FSC dataset, applying disjoint training using the different embedding enhancement architectures described in section 5.1. Best performance is reached with the CNN-4 topology, therefore it will be used in the following experiments.

Table 3: *Accuracy for the two speech classification tasks using different enhancement strategies. The enhancement based on embeddings uses the CNN-4 model.*

Data		Acc - FSC	Acc - GSC
Clean		98.94%	96.22%
Noisy		89.63%	88.00%
Wave-Enh	Dis-joint training	93.40%	89.42%
	$\alpha = 0.1$	90.95%	89.78%
	$\alpha = 0.5$	**94.88%**	**90.13%**
	$\alpha = 0.9$	93.30%	89.52%
Embeds-Enh	Dis-joint training	92.96%	89.20%
	$\alpha = 0.1$	92.32%	89.08%
	$\alpha = 0.5$	92.56%	**90.28%**
	$\alpha = 0.9$	**93.30%**	89.81%

Table 3 compares the classification accuracy on the two tasks considering joint and disjoint training of both waveforms (Wave-Enh) and embeddings (Embeds-Enh) enhancement. The row "clean" reports the upper-bound performance on clean data.

Table 4: *Speech enhancement evaluation metrics on FSC dataset using Wave-Enh strategy.*

Metric	PESQ	STOI
Noisy	1.28	0.84
Dis-Joint	2.26	0.890
Joint training ($\alpha = 0.5$)	2.34	0.894

Note that the performance of our classifier is in-line with the state of the art in both tasks. The row "noisy" reports the (lower-bound) performance obtained on the noisy data when no enhancement is applied. The other rows of the table give the performance on the noisy data applying both dis-joint and joint training. First of all, it is worth noting the significant improvement achieved with speech enhancement, especially when joint training strategies are used (the relative improvement w.r.t. dis-joint training is always above 10%). In addition, the **Embeds-Enh** approach shows a competitive performance with respect to its time domain counterpart. Note that the **Embeds-Enh** strategy significantly reduces the computational complexity, being the total number of parameters of the CNN-4 model much lower ($\approx \frac{1}{10}$) than that of Wave-U-Net. This is related to the smaller dimensionality of the speech embeddings with respect to the audio waveforms. Going more into details, the Wave-Enh architecture obtains the highest classification accuracy when $\mathcal{L}_{SE}$ and $\mathcal{L}_{cl}$ equally contributes to the joint loss (i.e. $\alpha = 0.5$). A similar trend is observed for **Embeds-Enh** strategy only in the GSC dataset, while for the FSC dataset the optimal value is: $\alpha = 0.9$, meaning that for this task the SE component is predominant in loss optimization. In general, we can claim that apply enhancement directly on the embeddings is preferable with respect to enhancing the speech waveforms, in terms of performance as well as computational complexity.

Finally, in Table 4 is reported the performance, in terms of PESQ and STOI defined in [39] obtained on the FSC dataset. Interestingly, the joint approach not only improve the final classification accuracy, but also improves the SE metrics w.r.t the dis-joint approach.Our code will be available in [5].

7. Conclusion

In this paper we proposed two joint training approaches namely **Wave-Enh**, and **Embeds-Enh** to robust intent/keywords classification in noisy environment. The difference between the two approaches is where the speech embeddings are computed. The jointly compositional scheme consists of a neural speech enhancement front-end based on Wave-U-Net, and CNN model for **Wave-Enh**, and **Embeds-Enh** respectively combined an with intent/keywords classifier.

All experiments are conducted on FSC, and GSC datasets contaminated with a set of noises obtained from MS-SNSD. Exhaustive experiments prove the efficiency of the joint training approach in terms of final classification accuracy. In addition, the **Embed-Enh** strategy shows a competitive performance not only in terms classification accuracy, but also in terms of computational resources. However, CNN-based embedding enhancement model is rather simple and we believe that there is room for substantial improvement. Our future work will therefore investigate alternative effective neural architectures. Another research direction will explore other pre-trained speech embeddings (as Wav2Vec 2.0, HuBERT and WavLM), and integrate the Connectionist Temporal Classification, in similar way to what we proposed in [40], as an additional loss.

[5]https://github.com/mnabihali/Joint-training-embeddings

8. References

[1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[2] 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.

[3] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.

[4] P. Hoang, J. M. De Haan, Z.-H. Tan, and J. Jensen, "Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2022.

[5] N. Saleem and M. I. Khattak, "A review of supervised learning algorithms for single channel speech enhancement," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 1051–1075, 2019.

[6] N. Saleem and M. I. Khattak, "Unsupervised single-channel speech enhancement based on phase aware time-frequency mask estimation," in *Applied Speech Processing*. Elsevier, 2021, pp. 75–99.

[7] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[8] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *IEEE international conference on acoustics, speech and signal processing*, 2018, pp. 6697–6701.

[9] W. Tai, T. Lan, Q. Wang, and Q. Liu, "Idanet: An information distillation and aggregation network for speech enhancement," *IEEE Signal Processing Letters*, vol. 28, pp. 1998–2002, 2021.

[10] A. Pandey and D. Wang, "Dense cnn with self-attention for time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1270–1279, 2021.

[11] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[12] M. N. Ali, D. Falavigna, and A. Brutti, "Time-domain joint training strategies of speech enhancement and intent classification neural models," *Sensors*, vol. 22, no. 1, p. 374, 2022.

[13] D. Seo, H.-S. Oh, and Y. Jung, "Wav2kws: Transfer learning from speech representations for keyword spotting," *IEEE Access*, vol. 9, pp. 80 682–80 691, 2021.

[14] A. Baevski, W.-N. Hsu, and A. Conneau, "Unsupervised speech recognition," in *NeurIPS*, 2021.

[15] M. Z. Alom *et al.*, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, no. 3, p. 292, 2019.

[16] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, 2018, pp. 3229–3233.

[17] X. Li, J. Li, and Y. Yan, "Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions." in *Interspeech*, 2017, pp. 1203–1207.

[18] Y. Zhao *et al.*, "Dnn-based enhancement of noisy and reverberant speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 6525–6529.

[19] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[20] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 696–700.

[21] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[22] D. Rethage *et al.*, "A wavenet for speech denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5069–5073.

[23] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[24] Y. Huang, T. Hughes, T. Z. Shabestary, and T. Applebaum, "Supervised noise reduction for multichannel keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5474–5478.

[25] Y. Gu, Z. Du, H. Zhang, and X. Zhang, "A monaural speech enhancement method for robust small-footprint keyword spotting," *arXiv preprint arXiv:1906.08415*, 2019.

[26] D. Bonet, G. Cámbara, F. López, P. Gómez, C. Segura, and J. Luque, "Speech enhancement for wake-up-word detection in voice assistants," *arXiv preprint arXiv:2101.12732*, 2021.

[27] G. Cámbara, F. López, D. Bonet, P. Gómez, C. Segura, M. Farrús, and J. Luque, "Tase: Task-aware speech enhancement for wake-up word detection in voice assistants," *Applied Sciences*, vol. 12, no. 4, p. 1974, 2022.

[28] M. Yu, X. Ji, B. Wu, D. Su, and D. Yu, "End-to-end multi-look keyword spotting," *arXiv preprint arXiv:2005.10386*, 2020.

[29] V. A. Trinh and S. Braun, "Unsupervised speech enhancement with speech recognition embedding and disentanglement losses," *arXiv preprint arXiv:2111.08678*, 2021.

[30] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[31] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[32] C. K. Reddy *et al.*, "A scalable noisy speech dataset and online subjective test framework," *arXiv preprint arXiv:1909.08050*, 2019.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2015, pp. 5206–5210.

[34] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 7097–7101.

[35] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[37] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "An experimental analysis of deep learning architectures for supervised speech enhancement," *Electronics*, vol. 10, no. 1, p. 17, 2020.

[38] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[39] M. N. Ali, V. J. Schmalz, A. Brutti, and D. Falavigna, "A speech enhancement front-end for intent classification in noisy environments," in *European Signal Processing Conference*. IEEE, 2021, pp. 471–475.

[40] E. Mekkonen, A. Brutti, and D. Falavigna, "End-to-end low resource keyword spotting through character recognition and beam-search re-scoring," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.