A Speech Enhancement Front-End for Intent Classification in Noisy Environments

Mohamed Nabih Ali
University of Trento
Trento, Italy
mohamed.nabih@unitn.it

Veronica Juliana Schmalz
Free University of Bolzano
Bolzano, Italy
vschmalz@unibz.it

Alessio Brutti
Fondazione Bruno Kessler
Trento, Italy
brutti@fbk.eu

Daniele Falavigna
Fondazione Bruno Kessler
Trento, Italy
falavi@fbk.eu

2021 29th European Signal Processing Conference (EUSIPCO) | 978-9-0827-9706-0/21/$31.00 ©2021 | DOI: 10.23919/EUSIPCO54536.2021.9616322

***Abstract*—Recently, several neural time-domain speech denoising and speech separation approaches have been investigated in literature, considerably progressing the state-of-the-art in the field. Among these methods, Wave-U-Net is particularly appealing because it allows an integrated modelling of the phase information and can handle large temporal contexts. In this paper, we present an evolution of the original Wave-U-Net architecture, that features a deeper model with exponentially increasing dilation rate from layer to layer in the downsampling blocks.**

Experiments on a contaminated version of Librispeech show that the proposed architecture outperforms the original one in terms of intelligibility metrics. In addition, we evaluate the performance of the proposed enhancement scheme on a simple intent classification task based on a noisy version of the Fluent Speech Commands dataset. Results show that, also in this case, the proposed method outperforms the baseline and substantially improves the classification accuracy in noisy conditions.

***Index Terms*—Intent classification, Speech Enhancement, Deep Learning**

I. INTRODUCTION

The term Spoken Language Understanding (SLU) is conventionally related to the task of identifying an intent, or, more generally, extracting meaning from a spoken utterance [1]. This is important for many applications such as voice user interfaces (e.g for implementing spoken virtual assistants, chatbots, etc) and smart home applications, in which the speaker's intent has to be converted into an action. According to scientific literature, an intent is represented as a set of conceptual slots [2]. In smart home applications (which is the domain addressed in this paper), an utterance like ”switch on two lamps in the living room” might correspond to an intent represented with the following filled slots: action: ”switch”, type:”lamp”, count:”two”, place:”living room”.

Recently, end-to-end (E2E) SLU approaches got a lot of attention. In these approaches, a single model is trained to map the speech signals directly into the speaker intents without any intermediate phases [3]–[6], namely conversion from audio to text with automatic speech recognition (ASR). Basically, E2E models optimize the intent recognition accuracy directly, reducing both model size and error propagation.

However, as in ASR applications, noise and other distortions significantly aggravate the quality and intelligibility of the speech signals, producing a negative effect on the intent classification performance [7]. One way to alleviate the impact of noise in the speech signals is to train, or adapt, the model on matching noisy data [8] or to resort to data augmentation strategies [9]. However, training models in all possible conditions is not always feasible. An alternative approach is to use a speech enhancement front-end, that improves the speech quality and intelligibility by reducing the effects of the interfering signals. Although today speech enhancement components are employed in many applications (e.g. mobile communication systems, headphones, voice over IP, ASR, etc), this research area still presents several open issues.

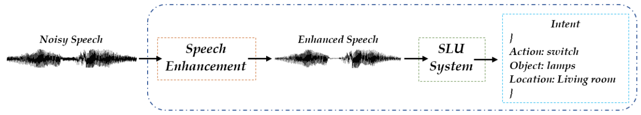

Fig. 1. Proposed system for Speech Enhancement for Intent classification

In this paper we experiment with a pipeline that integrates a time domain neural speech enhancement component with an E2E intent classification model, as depicted in Fig. 1. More in details, we introduce an improved version of the Wave-U-Net: a deep learning speech enhancement front-end [10] which is an extension of the model introduced for audio source separation in [11]. In addition, we exploit a convolutional deep neural network with residual layers (see section IV) as intent classification model which achieves state-of-the-art performance. Finally, we experiment with a multi-task learning framework which improves the robustness of the intent classification against out-of-vocabulary sentences by disjointly predicting the elements of intents.

II. SPEECH ENHANCEMENT: RELATED WORKS

In literature, many techniques based on statistical assumptions have been proposed for reducing the noise in audio signals [12]. Classical algorithms, such as spectral subtraction [13], [14], analyze the relation between speech signals and noise based on statistical assumptions. Typically, these techniques operate on the short-time Fourier transform (STFT), and use a frequency bin-wise gain function, derived from specific model assumptions for speech or noise signal distributions [14], [15]. These bin-wise functions often rely on an a-priori estimation of the signal-to-noise ratio (SNR) [16], [17], or of the noise power [18], [19]. The underlying assumption in most of these approaches is that noise is more stationary than speech in a given temporal segment. Therefore, these techniques are effective when applied to highly noisy environments or in case of stationary noise [20], but their performance degrades in presence of non-stationary noise [21].

Unlike classical approaches, in supervised deep learning a Deep Neural Network (DNN) is trained using pairs of clean and noisy speech signals without any statistical assumptions [22], [23]. Many of these techniques operate using the STFT, either estimating the clean signal magnitude [24], [25], the ideal ratio mask (IRM) [26], [27] or complex ratio masks [28]. The main drawback of these approaches is that they reconstruct the magnitude, while phase remains noisy. In addition, masks often introduce artifacts which may affect the signal quality as well as the performance of following processing components.

As an alternative to approaches operating in the frequency domain, several techniques working in the time domain have emerged recently [10], [29]–[31]. The major advantage of time domain approaches with respect to the frequency domain ones, is that they can mitigate the phase estimation problem, which improves the speech quality and intelligibility [32]. One example is SEGAN [30], which was the first attempt to employ generative adversarial networks (GAN) for speech enhancement. In [33], inspired by the performance of fully convolutional networks, the authors propose U-Net: a model that maps the noisy signal to its corresponding clean signal based on raw waveform directly. This model was later improved towards Wave-U-Net, that was firstly proposed in [10] for audio source separation, achieving promising results in comparison with other classical (e.g. Wiener filtering and spectral subtraction) and deep learning based methods (e.g. GANs).

III. Proposed Approach

A. Wave-U-Net

Wave-U-Net operates directly on time domain speech signals. Typically, this model consists of a series of fully convolutional downsampling blocks, followed by a bottleneck 1-D convolutional layer and, finally, by a series of fully convolutional upsampling blocks with skip connections from the downsampling to the upsampling blocks. In the downsampling blocks, a number of higher level features are computed on time scales, then are concatenated with the local, high resolution features computed from the same level upsampling block. This concatenation results into multi-scale features for predictions.

Suppose that a mixture of noisy signals $y[n] \in [-1,1]^{L\times C}$ is input to Wave-U-Net, where C is the number of speech channels and L is the number of audio samples. The network is trained to separate these mixture signals into K source waveforms, related either to clean speech, i.e. $x^1,\ldots,x^K$ with $x^k \in [-1,1]^{L\times C}$, or noise. In case of monaural speech enhancement, i.e. when $K = 1$, although it is possible to retrieve the noise, we are only interested in the clean speech signal $x^1[n]$. Fig.2 shows the entire architecture of a Wave-U-Net model.

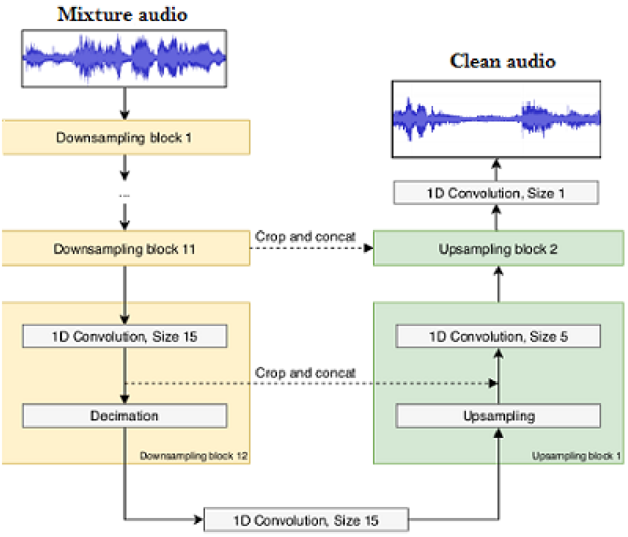

Fig. 2. Architecture of the Wave-U-Net model [34]

B. Dilated Encoder Wave-U-Net

The proposed modified model has the same architecture of the basic Wave-U-Net Recently, dilated convolution showed promising results with time series, demonstrates that,using dilated convolution captures the long-term information, without increasing network complexity [35], [36]. In particular we increase the model adding a fourth downsampling and a fourth upsampling blocks. Each downsampling block consists of three 1D convolutional layer with kernel size = 15, stride = 1, padding = 7, 14, 28, while the padding in the original Wave-U-Net is constant and equal to 7. As the original Wave-U-Net is not dilated i.e. dilation rate = 1, in this model we increase the dilation rate exponentially from layer to layer i.e. (1, 2, 4). Each convolutional layer is then followed by 1D-Batch normalization layer followed by leaky ReLU activation function with negative slope equals to 0.1. The bottleneck layer is a 1D convolutional layer with kernel size = 15, stride = 1 and padding = 7. The network right side consists of four upsampling blocks with the same number of layers as in the downsampling blocks, but without dilation. Finally, on the top of the network the output layer, which is a 1D convolutional layer with kernel size = 1 and stride = 1 followed by Tanh activation function, produces the enhanced audio samples.

IV. Back-End: Intent Classification on FSC

We evaluate the impact of the proposed speech enhancement scheme on an intent classification task. Intent classification (IC) is a functionality of SLU, which enables the interpretation of the information transmitted by speech signals. The goal of an IC task is the selection of the intent encoded in a spoken utterance from a closed set of categories [1], [5]. This task is closely related to dialogue-based natural language understanding, where the speaker does not need to use a predefined set of commands in order to successfully convey the desired intention.

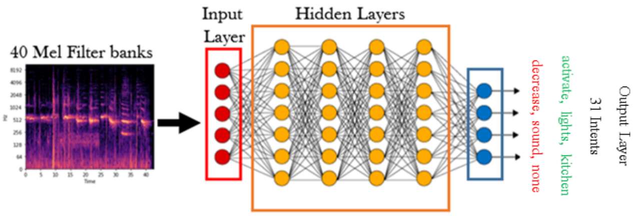

Fig. 3. Intent classification on FSC

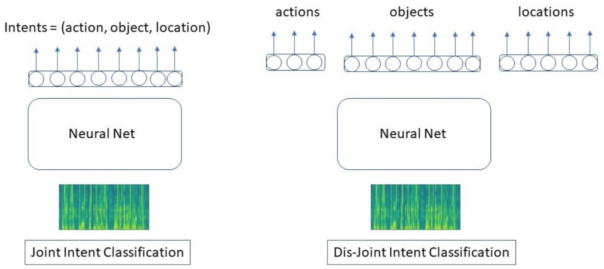
Intents = (action, object, location)
actions
objects
locations
Neural Net
Neural Net
Joint Intent Classification
Dis-Joint Intent Classification

Fig. 4. Joint and disjoint intent classification strategies on FSC

Recently, E2E approaches predicting the intent directly from the audio signal have been proposed [37], [38]. In general, these approaches are rather articulated. In [37], an E2E model is obtained by combining the word posteriors of a pre-trained acoustic modeling with an NLU model. A similar strategy is used in [39], where a pre-trained language model is employed to improve the performance. Nevertheless, solutions that attempt to directly classify the intent using a single, eventually pre-trained, classifier have been appearing lately. Examples are: [3]; [40], where a transformer model is used; [41], which implements a recurrent architecture trained with a reptile strategy and [42].

A. Fluent Speech Command

The Fluent Speech Commands (FSC) dataset [37] contains 30,043 utterances in English from 97 native and non native speakers interacting with smart-home devices or communicating with virtual assistants (e.g. “turn on the heat”, “switch on lights”, etc). All the signals are recorded as 16 kHz single-channel audio files. Overall, 248 different utterances are available in the dataset. Utterances are mapped into 31 different intents consisting of three items: action, object and location. For example, ”switch on the light” is labelled as: {`action: "activate", object: "lights", location: "none"`}. In total, 6 different actions, 14 objects and 4 locations are included in the dataset. The combination of these three slots represents the intent of an utterance. On average, for each intent 8 different utterances are present in the dataset. We use the official partitions: 23,132 utterances for training, 3,118 for validation and 3,793 for test.

The state of the art on clean speech for this dataset is around 99% intent classification accuracy [3], [37], [40], [41].

B. Intent Classification Model

Given the structure of the FSC data, we follow the same approach as in [3], [40], [41] and consider an E2E multi-class intent classification task which maps each utterance directly into one of the 31 possible intents. Fig. 3 shows a graphical illustration of the IC system.

Our model is based on the architecture of the encoder of Conv-TasNet, originally introduced for speech separation [43]. The network inputs are 40-Mel filter banks computed on a 20ms window, with 10ms step. The signals are limited to 4 seconds. The model consists of a normalization layer and 1D convolutional layer, that maps the 40 Mel features into 64 channels for bottleneck features, followed by 2 blocks of 5 residual blocks with 1-D dilated convolutions and skip connections. Each residual block has the same structure, with increasing dilation factor. We use 128 channels for the depth-wise separable convolutional layers.

We consider two classification strategies: joint classification, where the model outputs are the 31 logits associated to the intents; disjoint classification, where the three items of an intent are handled independently [44]. The latter is a multi-task learning approach, implemented by splitting the final classification layer in 3 as shown in Fig. 4. During inference, the three predicted objects are combined to form the predicted intent. On one hand, this is a more difficult task as it allows the prediction of non-existing intents when joining the three parts. On the other hand, it is expected to be more robust in case of out-of-vocabulary utterances (e.g. ways to express an intent not available in the training set) or in case unseen intents are present in the test set.

V. Experimental Results

The speech enhancement component is trained on a contaminated version of Librispeech-100 [45], which includes 100 hours of English read speech recorded 16 kHz. We randomly selected 10 speakers, resulting in approximately 10 hours of clean speech signals. The noisy signals are obtained by adding noise from the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [46], which provides noise sounds from 25 categories, e.g Air Conditioner, Bubble, and Cafe noise. The Librispeech dataset is contaminated randomly selecting for each file one out of five different SNRs: 5dB, 7.5dB, 10dB, 12.5dB and 15dB [47]. The dataset was divided into 6 hours for training, 2 hours for training and 2 hours for testing. The FSC dataset is also contaminated in the same way, by adding noise from MS-SNSD and using the same SNR range plus -5dB and 0dB.

Both the stock Wave-U-Net model [48] and the modified model are trained based on the noisy version of Librispeech-100 dataset. The network inputs are the mixture speech signals while the training targets are the clean speech signals. Due to the signals length variation, they are chunked taking 16384 continuous samples randomly selected from the noisy and clean signals. Both models are trained using Adam optimizer with learning rate $=10^{-4}$, decay rates $\beta 1 = 0.9$ and $\beta 2 = 0.999$. The batch size is 10 and the Leaky ReLU activation function

TABLE I
PESQ, STOI, AND SNR OF THE WAVE-U-NET MODELS USING LIBRISPEECH AND FSC DATASETS

Data sets		PESQ	STOI	SNR
Librispeech	Unproc.	1.30	0.70	11.52
	SE-GAN [30]	1.85	0.72	11.71
	Wave-U-Net [48]	2.13	0.76	13.42
	Dilated Wave-U-Net	2.37	0.78	13.95
FSC	Unproc.	1.79	0.62	8.68
	SE-GAN [30]	2.15	0.64	11.35
	Wave-U-Net [48]	2.68	0.67	11.06
	Dilated Wave-U-Net	3.09	0.73	11.30

is used with negative slope α = 0.1. Moreover, both models are trained using the mean square error (MSE) loss function. In inference, signals are split in 16384 sample long chunks, which are then re-concatenated once enhanced.

A. Speech Enhancement Results

The performance of the enhancement process is evaluated using a set of traditional intelligibility metrics: PESQ, STOI, and SNR. PESQ uses the wide-band version recommended in ITU-T, and its range lies between (-0.5 to 4.5) [49]. STOI is based on a correlation coefficient between time-aligned clean and enhanced signals, and its range is from 0 to 1 [50]. SNR measures the level of the desired signal with respect to the level of background noise and its unit is in dB [51].

Table I reports the results on the contaminated Librispeech-100 and FSC datasets using both the original Wave-U-Net model as implemented in [48] and our proposed deeper dilated Wave-U-Net. We also consider a SEGAN implementation as baseline [30]. For Librispeech, scores are computed on the 2-hours official testing partition. Conversely, for FSC we used the whole dataset (as the models are trained on the librispeech-100 training set). In both datasets the modified Wave-U-Net model outperforms the original Wave-U-Net model as well as SEGAN in all three metrics. Although in terms of SNR the improvement is very small and probably negligible form a statistical point of view, PESQ and STOI are clearly improved, in particular for the FSC dataset, indicating that the deeper and dilated network not only removes noise but also preserves the spectro-temporal properties of the signals.

B. Intent Classification

Finally, we evaluate the performance of the proposed enhancement strategy in terms of intent classification accuracy, which is measured as the actual match between the predicted intent slots and the ground-truth ones [37]. Table II reports the classification accuracy when applying the model trained on clean data to clean, noisy and enhanced signals. First of all, note that our best performance (98.8%) accuracy is in line with the current state-of-the-art. Therefore we can claim that our back-end model is sufficiently solid. Although the baseline Wave-U-Net improves the signal quality in terms of intelligibility metrics (as in Table I), it brings only a marginal improvement to the performance of the intent classifier with respect to the noisy data. Conversely, the proposed dilated models provides a clear improvement lifting the classification accuracy from 61.1% to 77.7%. Considering the two training strategies, the ”joint” approach is in general better, as expected, but the gap with the ”disjoint” approach is not that wide.

TABLE II
INTENT CLASSIFICATION ACCURACY ON FSC, USING CLEAN, NOISY AND ENHANCED SIGNALS. MODELS ARE TRAINED ON CLEAN DATA

	Full Data		50% out of voc.	
Evaluation Data	Disjoint	Joint	Disjoint	Joint
Clean	98.3%	98.8%	88.1%	84.8%
Noisy	63.2%	61.1%	42.3%	41.6%
Wave-U-Net [48]	61.6%	64.2%	50.4%	47.7%
Dilated Wave-U-Net	75.3%	77.7%	65.1%	62.5%

To evaluate the generalization capabilities of the proposed model, we consider an experimental set up where 50% of the utterances for each intent are removed from the training set. Therefore, for each intent an average of 4 utterances out of 8 in the test set haven’t been seen in training. This 50% is randomly selected and results are averaged on the two halves. Results are reported in the right-end part of Table II. In this case, obviously we observe a performance reduction with respect to using the full dataset. Speech enhancement is providing similar improvements as for the full data case. Note that the disjoint classification strategy provides a small but consistent improvement with respect to the joint approach. This confirms our hypothesis on the fact that predicting disjointly the three components of the intent helps in case of unseen utterances.

VI. Conclusion

In this paper, we propose a speech enhancement front-end module based on a dilated version of Wave-U-Net. We used the enhancement module as a pre-processing stage for intent classification in noisy conditions. In particular, we showed that our speech enhancement not only improves the signal quality in terms of intelligibility metrics but it also improves the intent classification accuracy on a contaminated version of the Fluent Speech Command corpus.

A natural extension of the proposed approach is joint training of speech enhancement and intent classification models. In particular, the key idea is to concatenate the speech enhancement model with the intent classification one and jointly adjust their parameters. This way, the intent classification model can guide the enhancement front-end to provide more suitable and more discriminative enhanced signals.

In the future, we plan to test the proposed intent classification model on different datasets and languages (e.g. ATIS [2], Almawave-SLU [52], SLURP [53]). We also intend to experiment on other SLU related tasks, such as dialog act recognition and slot filling.

References

[1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech.* John Wiley & Sons, 2011.

[2] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, “The ATIS spoken language systems pilot corpus,” in *Workshop on Speech and Natural Language*, 1990.

[3] Y. Qian *et al.*, "Exploring asr-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2017, pp. 569–576.

[4] D. Serdyuk *et al.*, "Towards end-to-end spoken language understanding," in *IEEE ICASSP*, 2018, pp. 5754–5758.

[5] Y.-P. Chen, R. Price, and S. Bangalore, "Spoken language understanding without speech recognition," in *IEEE ICASSP*, 2018, pp. 6189–6193.

[6] V. Renkens *et al.*, "Capsule networks for low resource spoken language understanding," *arXiv preprint arXiv:1805.02922*, 2018.

[7] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[8] S. Yin *et al.*, "Noisy training for deep neural networks in speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–14, 2015.

[9] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.

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

[11] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[12] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[15] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996, pp. 629–632.

[16] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori snr estimation approach based on selective cepstro-temporal smoothing," in *IEEE ICASSP*. IEEE, 2008, pp. 4897–4900.

[17] S. Elshamy *et al.*, "Instantaneous a priori snr estimation by cepstral excitation manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, 2017.

[18] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[19] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.

[20] C. Kwan *et al.*, "Enhanced speech in noisy multiple speaker environment," in *IEEE International Joint Conference on Neural Networks*, 2008, pp. 1640–1643.

[21] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2010.

[22] 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.

[23] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[24] Y. Xu *et al.*, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

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

[26] 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.

[27] Y. Zhao *et al.*, "Dnn-based enhancement of noisy and reverberant speech," in *IEEE ICASSP*. IEEE, 2016, pp. 6525–6529.

[28] 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.

[29] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE ICASSP*, 2018, pp. 696–700.

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

[31] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *IEEE ICASSP*, 2018, pp. 5069–5073.

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

[33] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[34] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *IEEE ICASSP*, 2019, pp. 181–185.

[35] O. Yazdanbakhsh and S. Dick, "Multivariate time series classification using dilated convolutional neural network," *arXiv preprint arXiv:1905.01697*, 2019.

[36] A. Bosca, A. Guérin, L. Perotin, and S. Kitić, "Dilated u-net based approach for multichannel speech enhancement from first-order ambisonics recordings," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 216–220.

[37] L. Lugosch *et al.*, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[38] N. Tomashenko, A. Caubrière, and Y. Estève, "Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech," in *Interspeech*. ISCA, 2019, pp. 824–828.

[39] W. I. Cho *et al.*, "Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation," in *Proc. Interspeech*, 2020, pp. 896–900.

[40] M. Radfar, A. Mouchtaris, and S. Kunzmann, "End-to-End Neural Transformer Based Spoken Language Understanding," in *Proc. Interspeech*, 2020, pp. 866–870.

[41] Y. Tian and P. J. Gorinski, "Improving End-to-End Speech-to-Intent Classification with Reptile," in *Proc. Interspeech*, 2020, pp. 891–895.

[42] R. Price, "End-to-end spoken language understanding without matched language speech model pretraining data," in *IEEE ICASSP*, 2020, pp. 7979–7983.

[43] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 8, p. 1256–1266, Aug. 2019.

[44] M. Firdaus *et al.*, "A deep multi-task model for dialogue act classification, intent detection and slot filling," *Cognitive Computing*, 2020.

[45] V. Panayotov *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE ICASSP*, 2015, pp. 5206–5210.

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

[47] M. N. Ali, A. Brutti, and D. Falavigna, "Speech enhancement using dilated wave-u-net: an experimental analysis," in *27th Conference of Open Innovations Association (FRUCT)*. IEEE, 2020, pp. 3–9.

[48] X. Haos, "Wave-U-Net-for-Speech-Enhancement," 2020. [Online]. Available: https://github.com/haoxiangsnr/Wave-U-Net-for-Speech-Enhancement

[49] R. I.-T. P. ITU, "862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. itu-telecommunication standardization sector, 2007."

[50] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE ICASSP*. IEEE, 2010, pp. 4214–4217.

[51] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice-Hall, 1988.

[52] V. Bellomaria *et al.*, "Almawave-SLU: a new dataset for SLU in italian," *arXiv preprint arXiv:1907.07526*, 2019.

[53] E. Bastianelli *et al.*, "SLURP: A spoken language understanding resource package," *arXiv preprint arXiv:2011.13205*, 2020.