E2E-SINCNET: TOWARD FULLY END-TO-END SPEECH RECOGNITION

Titouan Parcollet, Mohamed Morchid, Georges Linares

E2E-SINCNET: TOWARD FULLY END-TO-END SPEECH RECOGNITION

Titouan Parcollet[⋆‡†] *Mohamed Morchid*[⋆] *Georges Linarès*[⋆]

[⋆] Avignon Université, France
[‡] University of Oxford, UK
[†]Orkis, France

ABSTRACT

Modern end-to-end (E2E) Automatic Speech Recognition (ASR) systems rely on Deep Neural Networks (DNN) that are mostly trained on handcrafted and pre-computed acoustic features such as Mel-filter-banks or Mel-frequency cepstral coefficients. Nonetheless, and despite worse performances, E2E ASR models processing raw waveforms are an active research field due to the lossless nature of the input signal. In this paper, we propose the E2E-SincNet, a novel fully E2E ASR model that goes from the raw waveform to the text transcripts by merging two recent and powerful paradigms: SincNet and the joint CTC-attention training scheme. The conducted experiments on two different speech recognition tasks show that our approach outperforms previously investigated E2E systems relying either on the raw waveform or pre-computed acoustic features, with a reported top-of-the-line Word Error Rate (WER) of 4.7% on the Wall Street Journal (WSJ) dataset.

Index Terms— End-to-end speech recognition, SincNet.

1. INTRODUCTION

ASR systems are either hybrid DNN-HMM or end-to-end (E2E). The former set of ASR models provides state-of-the-art performances on numerous speech-related real-world tasks [1, 2] but involves multiple sub-blocks trained separately, and often requires separate and a strong human expertise. E2E systems, on the other hand, propose to directly transcribe a sequence of acoustic input features [3, 4] with a single model usually composed of different Neural Networks (NN) trained jointly in an *end-to-end* manner. In particular, a major challenge is to automatically generate an alignment from the raw signal that often contains several thousands of data point per second, to the text, only consisting of a single character or concept in the same time scale.

Recently, E2E approaches started to outperform traditional DNN-HMM baselines on common speech recognition tasks with the introduction of more efficient sequence training objectives [5, 6], more powerful architectures [7, 8] and attention mechanisms [9, 10, 11, 12]. Despite being named "E2E", the latter models still require pre-processed acoustic features such as Mel-filter-banks, alleviating a pure E2E pipeline based on the raw audio signal.

Processing raw waveforms in the specific context of ASR is an active challenge [13, 14, 15, 7]. Most of these works rely on modified Convolutional Neural Networks (CNNs) to operate over the signal. As an example, in [7], the authors propose to combine a log non-linearity with a CNN architecture that exactly matches an output dimension equivalent to standard Mel-filter-banks features, forcing the input layer to learn the latter signal transformation. Nevertheless, and as demonstrated in [16], CNNs are not efficient at learning common acoustic features due to the lack of constraint on the numerous trainable parameters. Consequently, the authors proposed SincNet, a specific convolutional layer that integrates common acoustic filters, such as band-pass filters, to replace the convolutional kernel weights drastically reducing the number of parameters. Furthermore, it is demonstrated that the learned filters have a much better frequency response than those learned with traditional CNNs, resulting in better performances in a speaker recognition task. Then, SincNet has been combined with a straightforward fully-connected DNN in the context of a DNN-HMM ASR system also outperforming CNNs trained with both pre-computed acoustic features and raw waveforms [17].

Unfortunately, there is no available model combining both the efficacy of SincNet to operate over raw signals, and the latest training scheme for E2E systems. Therefore, we propose to bridge this gap by investigating and releasing[1] a fully E2E model, named E2E-SincNet, combining SincNet with the joint CTC-attention training scheme [5] and resulting in a customizable, efficient and interpretable E2E ASR system. Contributions of the paper are summarized as:

1. Enhance the original SincNet to fit bi-directional recurrent neural networks (RNN).
2. Merge the later model with the joint CTC-attention method [5] to create E2E-SincNet[1] based on the well-known ESPnet toolkit [18] (Section 2).
3. Evaluate the model alongside with other baseline models on the WSJ and TIMIT speech recognition tasks (Section 3).

The conducted experiments show that E2E-SincNet obtains superior and state-of-the-art (SOTA) performances to both traditional E2E models operating on raw waveform with CNNs, and SOTA E2E architectures relying on pre-computed acoustic features.

2. END-TO-END SPEECH RECOGNITION

This section introduces the necessary building blocks to conceive a fully E2E automatic speech recognition system. First, latent acoustic features are extracted from the raw waveform signal with a specific kernelized CNN, also known as SincNet [16] (Section 2.1). The latter model is then merged with a joint CTC-attention [5] training procedure (Section 2.2.1), based on an encoder-decoder architecture [9] (Section 2.2.2).

2.1. Processing raw waveforms with SincNet

Traditional parametric CNNs operate over the raw waveform by performing multiple time-domain convolutions between the input signal and a certain finite impulse response [19] as:

[1]Code is available at: https://github.com/TParcollet/E2E-SincNet

$$y[n] = x[n] \times f[n] = \sum_{L}^{l=0} x[l].f[n-l], \quad (1)$$

with $x[n]$ a part of the speech signal, $f[n]$ a filter of length L, and y[n] the output finally filtered. In this case, all the elements of f are learnable parameters. SincNet proposes to replace f with a predefined function g that only depends on much fewer parameters to describe its behavior. In [16], the authors implemented g as a filter-bank composed of rectangular bandpass filters. Such function can be written in the time domain as:

$$g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n), \quad (2)$$

with f_1 and f_2 the two learnable parameters that describe the low and high cutoff frequencies of the bandpass filters, and $sinc(x) = \frac{sin(x)}{x}$. Such parameters are randomly initialized in the interval $[0, \frac{f_s}{2}]$, with f_s equal to the input signal frequency sampling. It is also important to notice that g is smoothed based on the Hamming window [20].

Other definitions of the filter g have been proposed including triangular, Gammatone, and Gaussian filters [21] demonstrating superior performances over Eq. 2 due to better filter responses to the signal. As a matter of fact, SincNet allows an important flexibility to efficiently enhance traditional acoustic-based CNNs with prior and well-investigated knowledge. Finally, SincNet filters are facilitating the interpretability of the model by being easily extracted and applied over any signal for further investigations of the transformations [16, 22, 21].

Unfortunately, SincNet has only been investigated with a mere fully-connected DNN based on a hybdrid DNN-HMM setup [17, 21]. We propose to connect SincNet to a recurrent encoder-decoder structure trained in a complete E2E manner following the joint CTC-attention procedure [5].

2.2. Joint CTC-attention models

2.2.1. Connectionist Temporal Classification

In E2E ASR systems, the task of sequence-to-sequence mapping from an input acoustic signal $X = [x_1, ..., x_n]$ to a sequence of symbols $T = [t_1, ..., t_m]$ is complex due to: 1) X and T could be in arbitrary length; 2) The alignment between X and T is unknown in most cases; 3) T is usually shorter than X in terms of symbols.

To alleviate these problems, connectionist temporal classification (CTC) has been proposed [23]. First, a softmax is applied at each timestep, or frame, providing a probability of emitting each symbol X at that timestep. This probability results in a symbol sequences representation $P(O|X)$, with $O = [o_1, ..., o_n]$ in the latent space O. A blank symbol $'-'$ is introduced as an extra label to allow the classifier to deal with the unknown alignment. Then, O is transformed to the final output sequence with a many-to-one function $z(.)$ defined as follows:

$$\left.\begin{array}{l} z(o_1, o_2, -, o_3, -) \\ z(o_1, o_2, o_3, o_3, -) \\ z(o_1, -, o_2, o_3, o_3) \end{array}\right\} = (o_1, o_2, o_3). \quad (3)$$

Consequently, the output sequence is a summation over the probability of all possible alignments between X and T after applying the function $z(O)$. Accordingly to [23] the parameters of the models are learned based on the cross entropy loss function:

$$\sum_{X,T \in train} -\log(P(O|X)). \quad (4)$$

During the inference, a best path decoding algorithm is performed. Therefore, the latent sequence with the highest probability is obtained by performing the $argmax$ of the $softmax$ output at each timestep. The final sequence is obtained by applying the function $z(.)$ to the latent sequence.

2.2.2. Attention-based encoder-decoder

Conversely to CTC, encoder-decoder models [9] do not suffer from a forced many-to-one mapping. Indeed, the input signal $X = [x_1, ..., x_n]$ is entirely consumed by a first encoder neural network (*e.g.* a recurrent neural network), before being fed to a second one that is free to emit any number of outputs $T = [t_1, ..., t_m]$ starting from the information contained in the last latent space of the encoder. Major bottlenecks are therefore related to the ability of the encoder to map correctly an entire sequence to an arbitrary latent space, and to the decoder that is not aware of the sequential order of the input signal. To alleviate these issues, attention-based encoder-decoder have been proposed [9].

From a high-level perspective, an attention-based decoder is able to look over the complete set of the hidden states generated by the encoder, making it feasible to "*choose*" the relevant information in the time-domain [9]. More precisely, an attention-based encoder-decoder consists of two RNNs. The encoder part remains mostly unaltered and maps an input sequence of arbitrary length n, $X = [x_1, ..., x_n]$ to n hidden vectors h, $(h_1, ..., h_n)$. Then, the attention-decoder generates m output distributions $O = [o_1, ..., o_m]$ corresponding to m timesteps, by attending to both the n encoded h and the previously generated token o_{t-1}. Two special tokens, denoting the *start-of-sentence* and *end-of-sentence* are added to integrate boundaries. The loss function is nearly identical to CTC, except that a condition on the previous ground-truth token (o_{t-1}^{truth}) is added [5]:

$$\mathcal{L}_{enc,dec} = \sum_{t=1}^{m} -\log(P(o_t|x_t, o_{t-1}^{truth})). \quad (5)$$

Following [5], our proposed E2E model relies on a location-based attention method [24]. The attention weight vector at time step t named a_t identifies the focus location in the entire encoded hidden sequence h at time step t with a context vector c_t:

$$c_t = \sum_i a_{t,i} h_i, \quad a_{t,i} = \frac{exp(\gamma e_{t,i})}{\sum_{k=1}^{m} exp(\gamma e_{t,k})}, \quad (6)$$

γ is the sharpening factor [25] and $e_{t,k}$ relates the importance or energy of the k annotation vector for predicting the i output token. In the case of a location-based attention, $e_{t,k}$ is computed as:

$$e_{t,k} = w^T tanh(W s_{t-1} + v h_k + U(F * a_{t-1}) + b, \quad (7)$$

with w, W,V,F,Y and b are different trainable parameters, s_{t-1} is the hidden state of the decoder at the previous time step, and $*$ is the convolution product. It is important to note that the implementation of e varies accordingly to the type of attention mechanism employed [9]. Then, the decoder generates an output token o_t from an input vector based on both the context vector c_t and the previous state

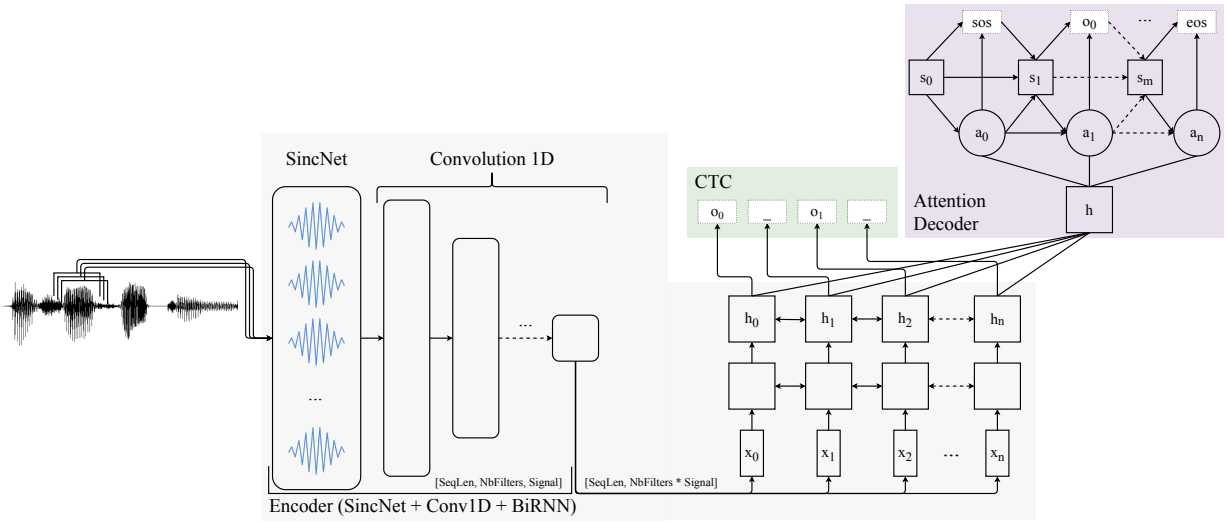

Fig. 1. Illustration of the proposed E2E-SincNet. The batch dimension is omitted for readability. The raw signal is first encoded with a SincNet layer, followed by multiple 1D convolutions and a bidirectional RNN. Then a CTC and an attention-based decoders emit a sequence of text symbols and are trained jointly.

s_{t-1} alongside with updating the current state s_t following RNNs equations with s_{t-1}, c_t and o_t. Unfortunately, ASR systems solely relying on an attention mechanism are highly perturbed by noisy data that generate wrong alignments [5]. Furthermore, it has been shown that it is difficult to train such models from scratch on wide input sequences [25, 9].

2.2.3. Joint CTC-attention

To overcome the limitations of both CTC training and attention-based encoder-decoder models and to benefit from their strengths, [5] introduced the joint CTC-attention paradigm. The key idea of the latter method relies on the introduction of the CTC loss as an auxiliary task to the attention-based encoder-decoder training. More precisely, both losses are combined and controlled with a fixed hyperparameter λ ($0 \leq \lambda \leq 1$) as:

$$\mathcal{L}_{joint} = (1-\lambda)\mathcal{L}_{enc,dec} + \lambda\mathcal{L}_{CTC}. \quad (8)$$

2.3. E2E-SincNet

SincNet has only been combined with mere feed-forward NNs [16, 17], while the joint CTC-attention approach has only been applied to pre-computed acoustic features such as MFCCs and Mel-filter-banks [5, 18]. We propose to combine SincNet to the latter training procedure in a efficient, interpretable and fully E2E ASR approach (Figure 1).

The E2E architecture is composed of three components: 1) An encoder that operates over the raw audio signal with a first SincNet layer followed by N one-dimensional convolutional layers. The latent features are then consumed by a traditional bidirectional RNN. 2) A simple CTC decoder that produces a token for each time step encoded. 3) An attention-based decoder that looks out over the entire encoded hidden sequence to output the right symbol. The model is trained following the joint CTC-attention loss function (Eq. 8).

3. EXPERIMENTS

In this Section, E2E-SincNet is compared to other state-of-the-art end-to-end ASR systems with two different speech recognition tasks. First, datasets alongside with pre-computed and raw acoustic features are detailed (Section 3.1). Then, baselines and proposed models architectures are described (Section 3.2). Finally, we report and discuss the results in Section 3.3.

3.1. Speech recognition datasets and acoustic features

E2E-SincNet is evaluated in two different tasks of phoneme recognition with the TIMIT dataset, and word recognition with the Wall Street Journal corpus.

3.1.1. The TIMIT phoneme recognition task

The TIMIT [26] dataset is composed of a standard 462-speaker training dataset, a 50-speakers development dataset and a core test dataset of 192 sentences for a total of 5 hours of clean speech. During the experiments, the SA records of the training set are removed and the development set is used for early stopping. The accuracy is reported in terms of Phoneme Error Rate (PER). TIMIT is considered as a challenging task for E2E systems due to its very limited amount of available training data (less than 5 hours).

3.1.2. The Wall Street Journal speech recognition task

Only the full "train-si284" dataset is considered as a training set (81 hours), due to the fact that the models have already been evaluated on the smaller TIMIT dataset. The usual "test-eval92" is used at testing time, while "test-dev93" is considered as a validation dataset. The accuracy is reported in terms of Word Error Rate (WER).

3.1.3. Acoustic features

In the original SincNet proposal [16], chunks of raw signal are created every 400ms with a 10ms overlapping. Instead, we propose to split the waveform of each speech sentence into blocks of 25ms. Indeed, [16] introduce a SincNet followed by a DNN that requires both right and left contexts to be trained properly. Our approach relies on a combination of SincNet with a RNN allowing the latter context to be captured within the recurrent connections, making it feasible to drastically reduce both the input dimension and the VRAM consumption at training time (*i.e.* by a factor of 5). Then, other E2E systems usually process either pre-computed acoustic features. Therefore, 23 and 80 Mel-filter-banks are extracted for the TIMIT and WSJ datasets respectively, based on windows of size 25ms with a 10ms overlapping.

3.2. Models architectures

Two different E2E ASR models operating on the raw waveform and relying on the encoder-decoder approach with the joint CTC-attention training scheme are introduced (see Figure 1).

E2E-SincNet. The encoder is made of a specific SincNet layer and 3 one-dimensional convolutional layers with $256 - 128 - 128$ filters, followed by a bidirectional LSTM composed with 4 or 6 layers of size 512 for the TIMIT and WSJ tasks respectively. A one-dimensional maxpooling of length 3 is applied after the convolutional and SincNet layers to reduce the signal dimension. In [16], the authors introduced a SincNet layer composed of 128 filters of size 251. We propose to increase the number of filters to 512 and to decrease their size to 129 to enhance the local resolution of the filters, better fitting to the task of speech recognition. Finally, the decoder relies on a simple attention layer of size 512 combined with the CTC loss (Section 2.2.3).

E2E-CNN. This architecture is proposed to highlight the impact of the SincNet layer in E2E-SincNet. More precisely, E2E-CNN is identical to E2E-SincNet but with a traditional convolutional layer with 512 filters to replace the SincNet one.

Models are trained based on the Adadelta optimizer with vanilla hyperparameters [27] for 20 and 15 epochs during the TIMIT and WSJ tasks respectively. The joint CTC-attention loss control hyperparameter λ (Eq. 8) is set to 0.5 for the TIMIT experiments and decreased to 0.2 with WSJ. No dropout is applied and the results observed on the test dataset are reported with respect to the best performances obtained on the validation dataset.

3.3. Results and discussions

Table 1 reports the results obtained by our approaches compared to a more traditional E2E model operating on Mel-filter-banks on the TIMIT dataset. First, it is worth underlining that the E2E-SincNet obtains the best performances with a PER of 19.3% on the test dataset, compared to 20.5% for the baseline and 21.1% for the non-SincNet alternative representing a relative improvement of 1.2% and 1.3% respectively. Unfortunately, TIMIT is a very challenging task for E2E systems due to the small amount of available training data (less than 5 hours), resulting in worse performances in comparison to hybrid DNN-HMM ASR systems [2]. Therefore, it is of crucial interest to scale the E2E-SincNet model to a larger dataset to validate its suitability to real-world tasks.

Table 2 reports the performances obtained by various SOTA E2E models on the WSJ dataset by integrating a 3-gram recurrent language model (RNNLM) [18]. "Jasper"[8] uses a transformerXL language model. First, the proposed E2E-SincNet obtains a top-of-line WER of 4.5% on the "test_eval92" dataset, outperforming all the baselines. Indeed, a previous best score of 5.9% was reported in [12], highlighting a relative improvement of 1.2%.

Table 1. Results obtained with different E2E ASR systems on the TIMIT phoneme recognition tasks. "Fea." details the type of input features employed, and "Valid." denotes the validation dataset. Results are expressed in Phoneme Error Rate (*i.e.* lower is better).

Models	Fea.	Valid. %	Test %
E2E-CNN	RAW	18.9	21.1
ESPnet (VGG) [18]	FBANK	17.9	20.5
E2E-SincNet	**RAW**	**17.3**	**19.3**

Table 2. Results obtained with different E2E ASR systems on the WSJ dataset. "Fea." details the type of input features employed, "Valid." denotes the validation dataset, "-ASG" is the auto segmentation criterion (*i.e* a variation of CTC) and "-Att." is attention only. Results are expressed in Word Error Rate (*i.e.* lower is better).

Models	Fea.	Valid.	Test
BiGRU-Att. [9]	FBANK	-	9.3
Wav2Text [28]	FBANK	12.9	8.8
Jasper [8]	FBANK	9.3	6.9
E2E-CNN	RAW	9.8	6.5
ESPnet (VGG) [18]	FBANK	9.7	6.4
CNN-GLU-ASG [7]	RAW	8.3	6.1
SelfAttention-CTC [12]	FBANK	8.9	5.9
E2E-SincNet	**RAW**	**7.8**	**4.7**

The E2E-SincNet outperforms the E2E-CNN with a relative gain of 1.8% on both TIMIT and WSJ task. This demonstrates the efficacy of the SincNet layer to learn an expressive filtered signal enabling a better and lossless latent representation of the raw waveform. It is interesting to note that "transformer" models have recently obtained better performances on multiple ASR tasks [29]. Nonetheless, transformers are a specific architecture that differ significantly from the presented models and are therefore not considered in our benchmarks.

4. CONCLUSION

Summary. In this paper, we introduced E2E SincNet, a fully end-to-end automatic speech recognition system able to process the raw waveform based on an adaptation of the recent SincNet with the powerful joint CTC-attention training paradigm. The conducted experiments on two different speech recognition-related tasks have demonstrated the superiority of our approach over various other E2E systems based on both pre-computed acoustic features and the raw waveform, achieving one of the best result observed so far with an E2E ASR model on the Wall Street Journal dataset.

Future work. SincNet currently suffers from various issues. First, it is important to investigate other filters to efficiently operate over the raw signal. Then, an alternative to the maxpooling must be explored to alleviate the risk of aliasing in the filtered signal.

Acknowledgments. This work was supported by the AISSPER project through the French National Research Agency (ANR) under Contract AAPG 2019 ANR-19-CE23-0004-01 and by the Engineering and Physical Sciences Research Council (EPSRC) under Grant: MOA (EP/S001530/).

5. REFERENCES

[1] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[2] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *In Proc. of ICASSP*, 2019.

[3] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[4] Ying Zhang, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, Cesar Laurent Yoshua Bengio, and Aaron Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv preprint arXiv:1701.02720*, 2017.

[5] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[6] Yiming Wang, Tongfei Chen, Hainan Xu, Shuoyang Ding, Hang Lv, Yiwen Shao, Nanyun Peng, Lei Xie, Shinji Watanabe, and Sanjeev Khudanpur, "Espresso: A fast end-to-end neural speech recognition toolkit," *arXiv preprint arXiv:1909.08723*, 2019.

[7] Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, and Emmanuel Dupoux, "End-to-end speech recognition from the raw waveform," in *Interspeech 2018*, 2018.

[8] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv preprint arXiv:1904.03288*, 2019.

[9] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[10] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[11] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[12] Julian Salazar, Katrin Kirchhoff, and Zhiheng Huang, "Self-attention networks for connectionist temporal classification in speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7115–7119.

[13] Dimitri Palaz, Ronan Collobert, and Mathew Magimai Doss, "End-to-end phoneme sequence recognition using convolutional neural networks," *arXiv preprint arXiv:1312.2137*, 2013.

[14] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr," in *Fifteenth annual conference of the international speech communication association*, 2014.

[15] Yedid Hoshen, Ron J Weiss, and Kevin W Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4624–4628.

[16] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," *arXiv preprint arXiv:1808.00158*, 2018.

[17] Mirco Ravanelli and Yoshua Bengio, "Speech and speaker recognition from raw waveform with sincnet," *arXiv preprint arXiv:1812.05920*, 2018.

[18] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[19] Lawrence R Rabiner and Ronald W Schafer, *Theory and applications of digital speech processing*, vol. 64, Pearson Upper Saddle River, NJ, 2011.

[20] Sanjit Kumar Mitra and Yonghong Kuo, *Digital signal processing: a computer-based approach*, vol. 2, McGraw-Hill New York, 2006.

[21] Erfan Loweimi, Peter Bell, and Steve Renals, "On learning interpretable cnns with parametric modulated kernel-based filters," *Proc. Interspeech 2019*, pp. 3480–3484, 2019.

[22] Mirco Ravanelli and Yoshua Bengio, "Interpretable convolutional filters with sincnet," *arXiv preprint arXiv:1811.09725*, 2018.

[23] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[24] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[25] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[26] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[27] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[28] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Attention-based wav2text with feature transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 309–315.

[29] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *INTERSPEECH 2019*, 2019.