Neural Speech Recognition summary

I. Papastratis, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 1.0.1
Date: 6/6/2021

Artificial Intelligence &
Information Analysis Lab

Neural Speech Recognition

- Introduction
- Neural Speech Recognition Datasets
- Neural Speech Recognition Methods
- Deep Neural Networks (DNN)
  - Recurrent Neural Networks (RNN)
  - Convolutional Neural Networks (CNN)
  - Transformers

Automatic Speech Recognition

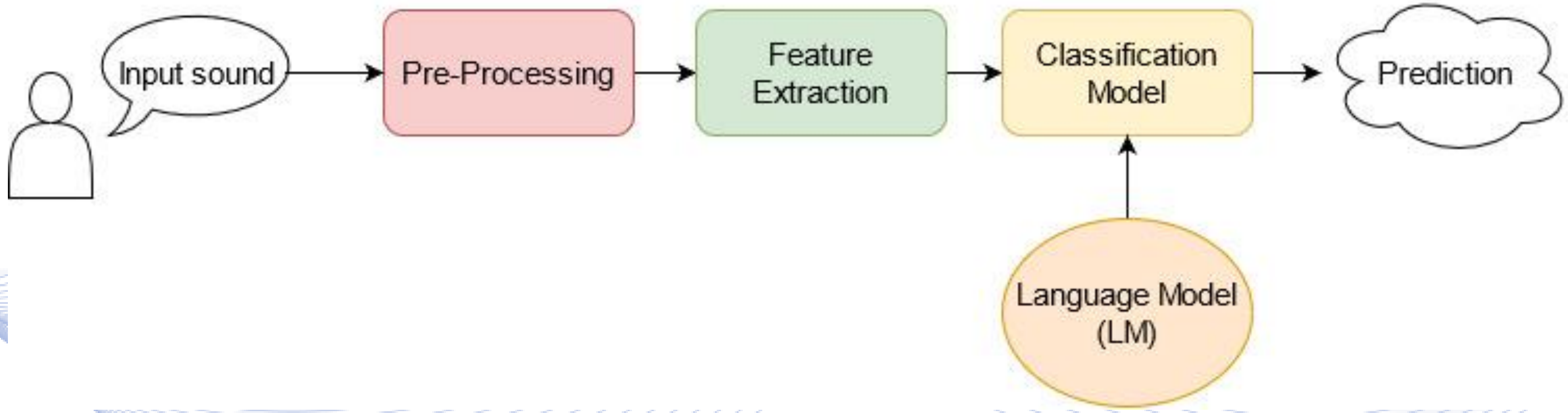
Automatic Speech Recognition

Applications

- Workplace: increase efficiency of simple tasks
    - Dictate the information you want to be incorporated into a document
    - Print documents on request
- Smart assistants:
    - Apple’s Siri, Amazon’s Alexa, Google Assistant, Microsoft’s Cortana
- Behavior /emotion recognition

Automatic Speech Recognition

- Given an input audio sequence $\mathbf{x}$ an Automatic Speech Recognition (ASR) system tries to predict the output sequence of the spoken language $\mathbf{y}$



Automatic Speech Recognition

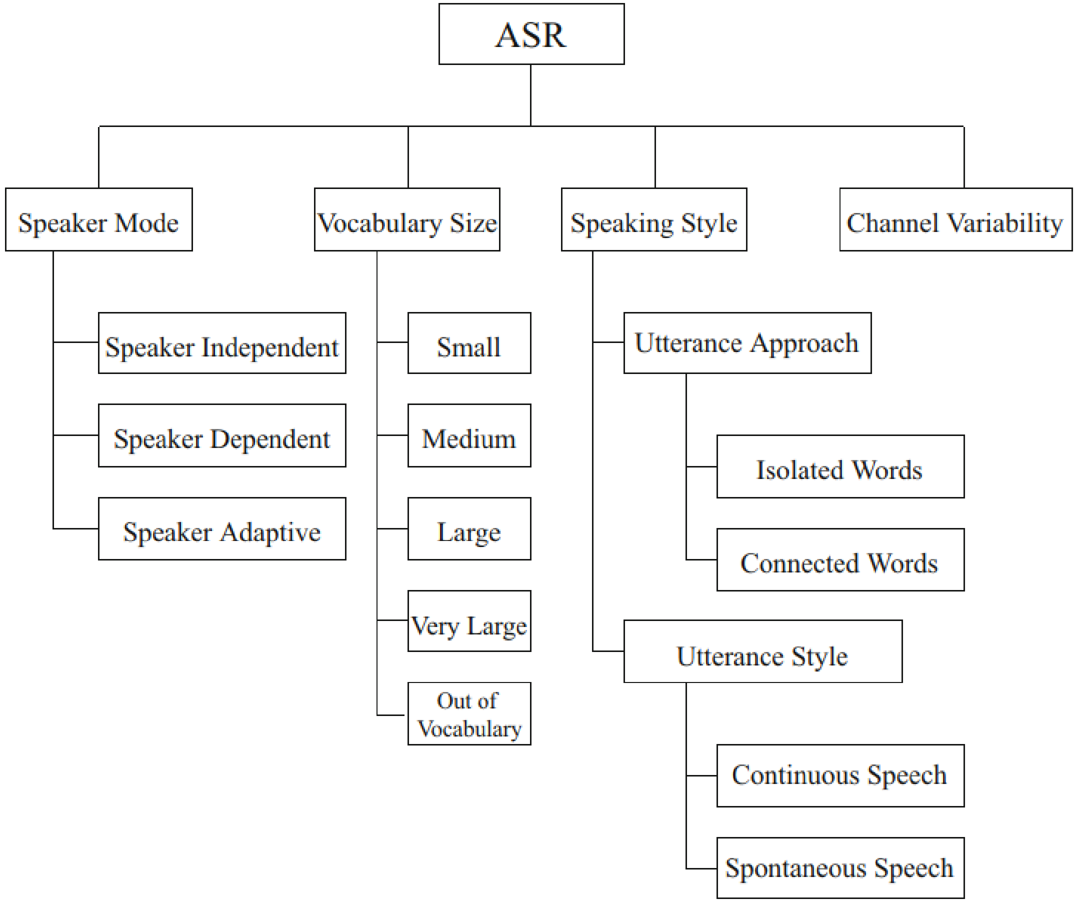
- ***Pre-Processing***: The pre-processing step aims to improve the audio signal by reducing the signal-to-noise ratio, reduce the noise and filter the signal
- ***Feature extraction*** : Features are usually the predefined number of coefficients or values that are obtained by applying various methods on the input speech signal. This step should be robust to different factors, such as noise and echo effect. Most commonly used feature extraction methods are Mel-frequency cepstral coefficients (MFCCs), and discrete wavelet transform (DWT)

Automatic Speech Recognition

- ***Classification Model:*** this model aims to predict the text corresponding to the input speech signal. The classification model takes the extracted features from the previous stage and generates the output text.
- ***Language Model:*** consists of various types of grammatical rules and semantics of a language. Language models are necessary for recognizing the output token from the classifier and is also used to make corrections on the output text.

Artificial Intelligence & Information Analysis Lab

ASR categories



[Malik2021] ASR categories

Datasets

CallHome English, Spanish and German databases.

- They contain conversational data, high number of out-of-vocabulary words
- Challenging databases with foreign words and telephone channel distortion

Datasets

TIMIT

- broadband recordings from American English, where each speaker reads 10 phonetically rich sentences.
- Time-aligned orthographic, phonetic and word transcriptions
- 16kHz speech waveform file for each utterance.
- Training set of audios from 462 speakers
- Validation set of 50 speakers
- Test set of 24 speakers

Datasets

Wall Street Journal

- contains audio from speakers that read texts from Wall Street Journal newspapers
- Subsets of 5000 and 20000 words

Datasets

LibriSpeech

- A corpus of approximately 1000 hours of 16kHz speech of English language
- The dataset is derived from read audio-books from the LibriVox project

Feature extraction

- ***Mel-frequency Cepstral coefficients***
- A human ear is a non-linear system concerning how it perceives the audio signal
- To cope with the change in frequency Mel-scale makes a linear model of the human auditory system
- Only frequencies f_{Hz} in the range of [0,1] kHz can be transformed to the Mel-scale, while the rest frequencies are considered to be logarithmic

$$f_{mel} = \frac{1000}{\log(2)}\left[1 + \frac{f_{Hz}}{1000}\right]$$

Recurrent Neural Networks

- RNN methods
  - Speech recognition with deep recurrent neural networks
  - Encoder-Decoder RNN-Transducer
  - Streaming end-2-end speech recognition for mobile devices
- Attention RNN methods
  - Attention-based recurrent sequence generator (ARSG)
  - Listen-Attend-Spell (LAS)
  - Hybrid CTC and attention

Recurrent Neural Networks

Recurrent Neural Networks (***RNN***).

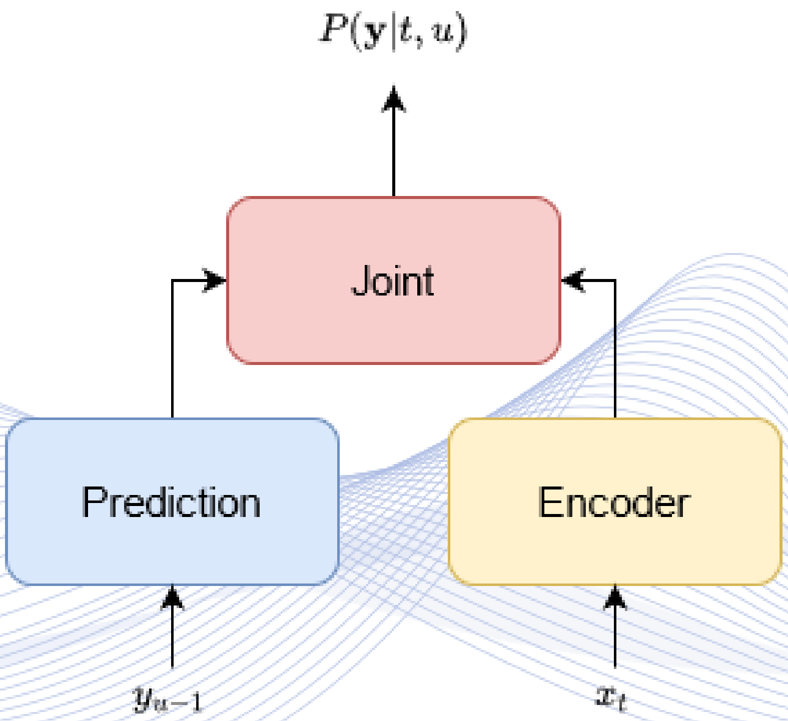
- RNNs have good performance on sequential data, as they exploit temporal data relations.
- They are based on the concept of ***state variables***, which store system status.
- Output(hidden state) is fed back into next timestep
- RNNs model time-series signals and capture model dependencies between different time-steps of the input

Speech recognition with deep recurrent neural networks

NETWORK	WEIGHTS	EPOCHS	PER
CTC-3L-500H-TANH	3.7M	107	37.6%
CTC-1L-250H	0.8M	82	23.9%
CTC-1L-622H	3.8M	87	23.0%
CTC-2L-250H	2.3M	55	21.0%
CTC-3L-421H-UNI	3.8M	115	19.6%
CTC-3L-250H	3.8M	124	18.6%
CTC-5L-250H	6.8M	150	18.4%
TRANS-3L-250H	4.3M	112	18.3%
PRETRANS-3L-250H	**4.3M**	**144**	**17.7%**

[Graves2013] Results on TIMIT dataset with different settings

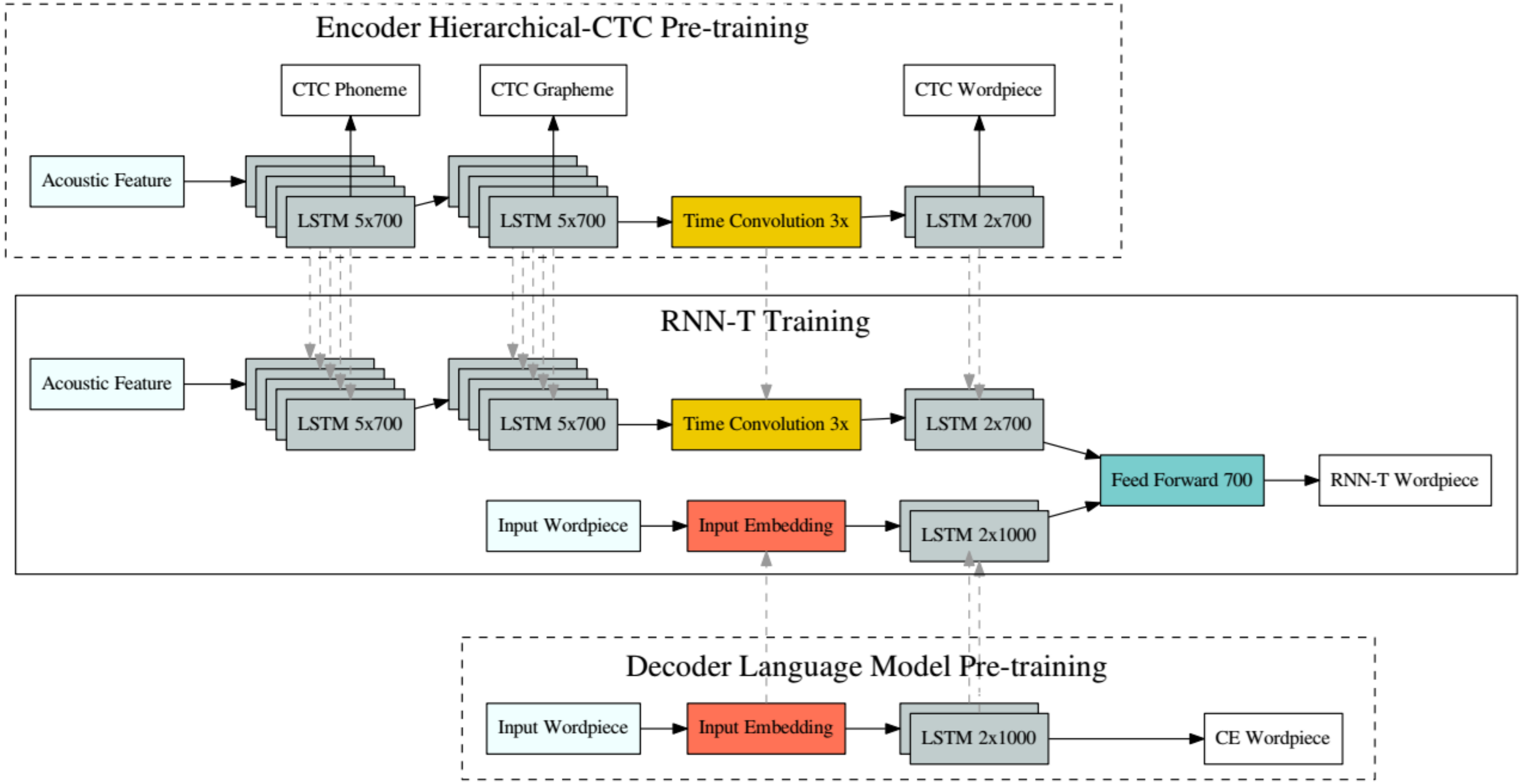
RNN-Transducer



Overview of RNN-Transducer

Artificial Intelligence & Information Analysis Lab

Encoder-Decoder RNN-Transducer



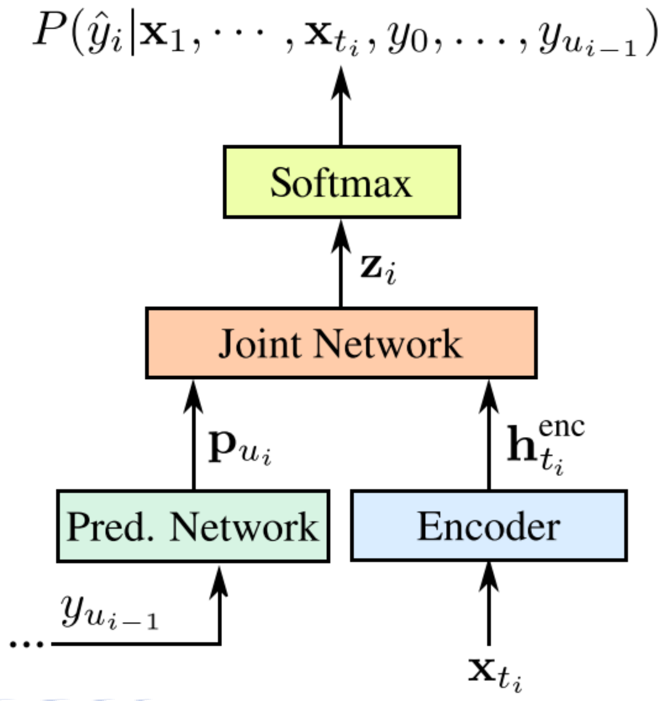
[Rao2017] Overview of encoder-decoder RNN-T

Artificial Intelligence &
Information Analysis Lab

Streaming e2e speech recognition for mobile devices

- RNN-T with 8 layers of uni-directional LSTM cells
- Time-reduction layer to speed up training and inference.
- Memory caching techniques to save about 50 – 60% of the prediction network computations.
- Multithreading has a speedup of 28% compared against single-threaded execution.
- Parameters are quantized from 32-bit floating-point precision into 8-bit to reduce memory consumption and operate in real-time.

Streaming e2e speech recognition for mobile devices

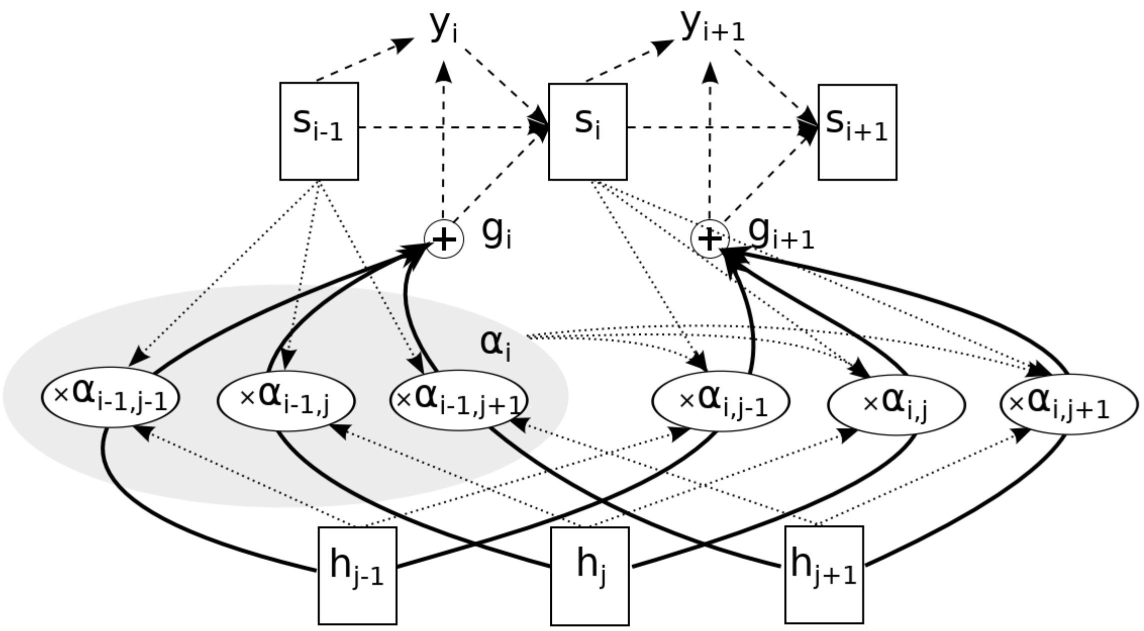


[He2019] Network for ASR on mobile devices

Attention RNNs

- Encoder-Decoder architecture as in machine translation
- Encoder transforms input text into a sequence of vectors (rather than a single vector)
- Decoder use an attention method at each output step to assign different weights to each vector in this sequence
- Does not require alignment of data

Attention-based recurrent sequence generator (ARSG)

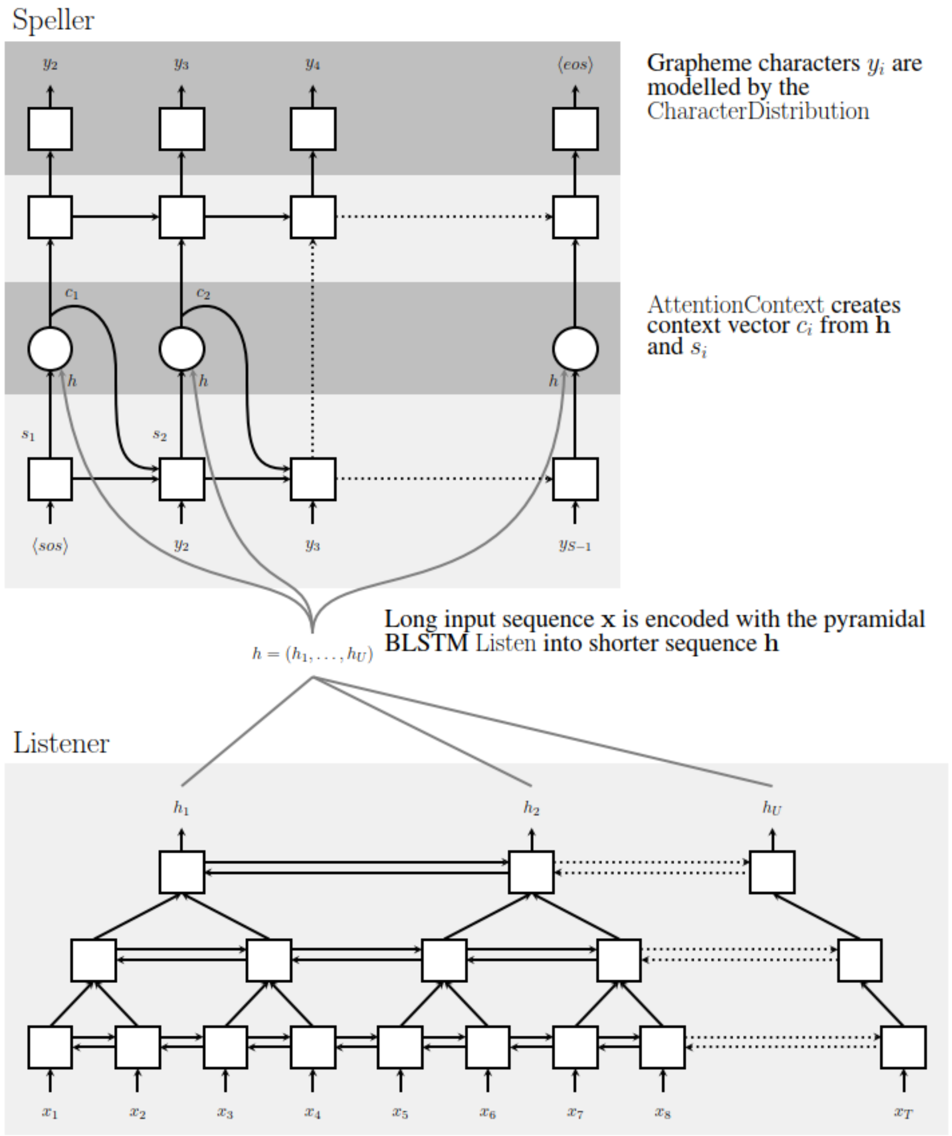


[Chorowski2015] ARSG Method

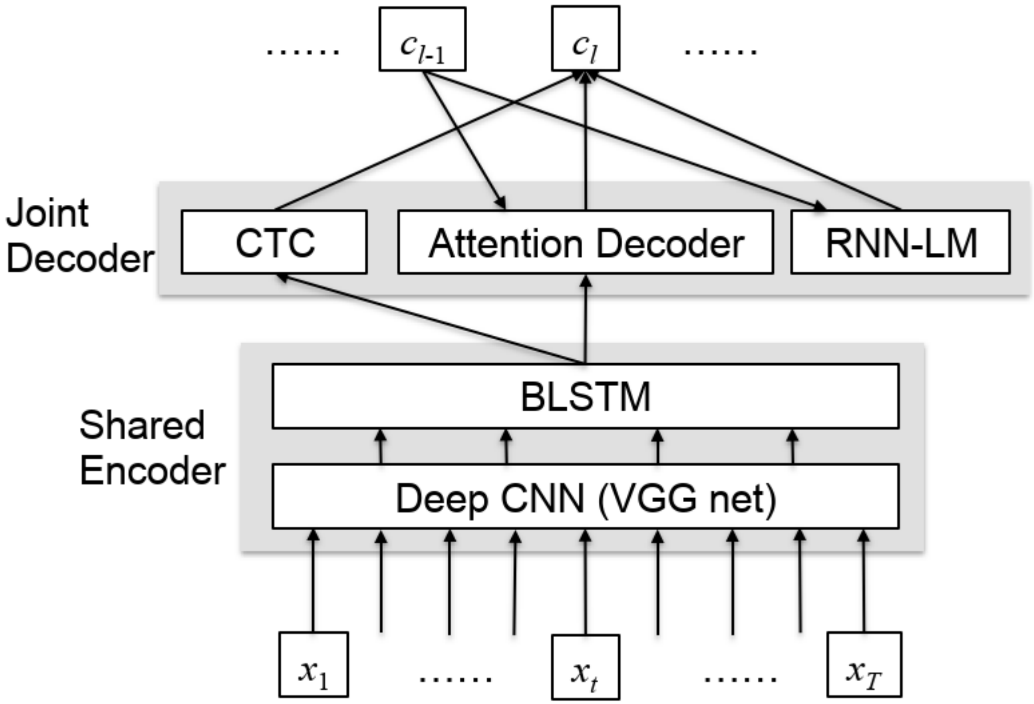
Artificial Intelligence & Information Analysis Lab

Listen-Attend-Spell (LAS)

[Chan2016] Listen-Attend-Spell (LAS) overview



Hybrid CTC and attention model



[Hori2018] Overview of the method

Artificial Intelligence & Information Analysis Lab

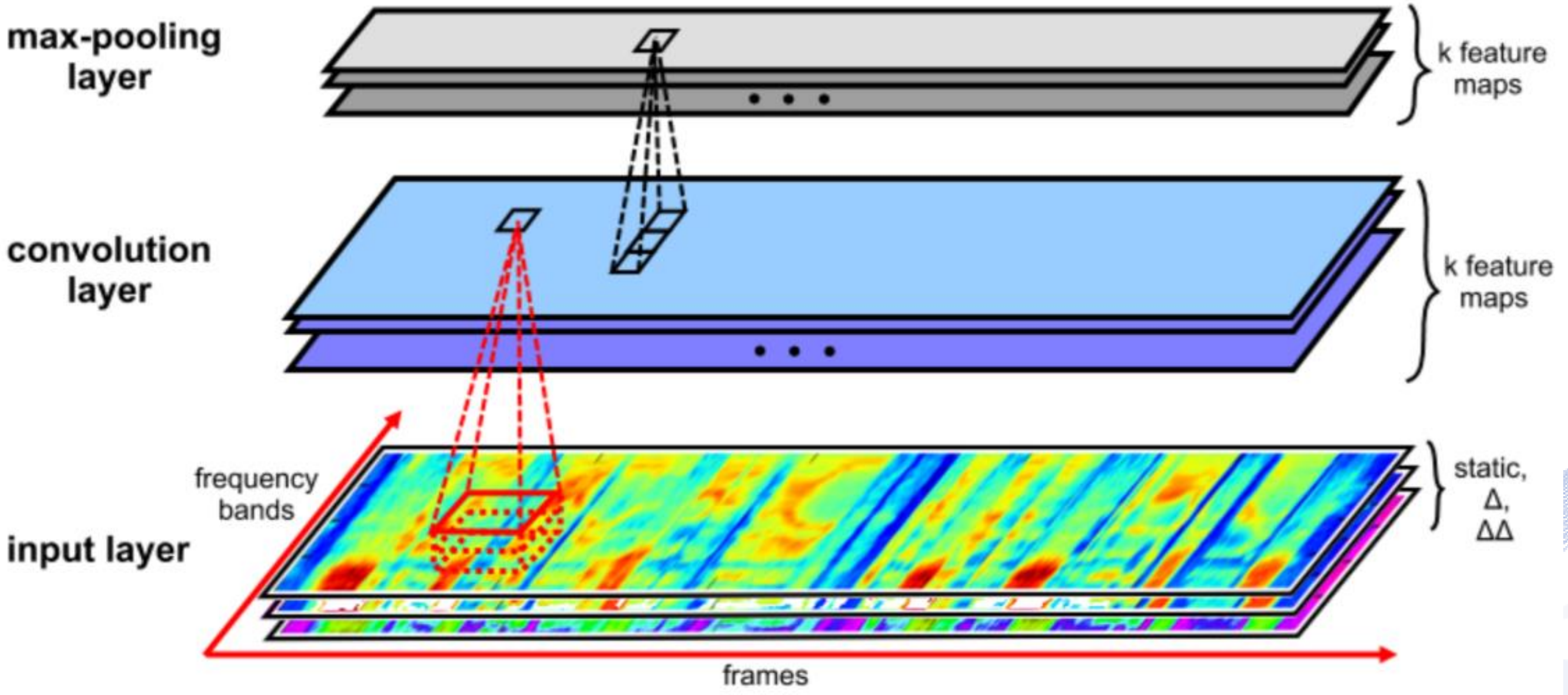
Convolutional Neural Networks

- Methods
  - 1D-CNN for speech recognition
  - Fully Convolutional method for speech recognition
  - Residual Convolutional CTC Networks for Automatic Speech Recognition
  - Jasper: An End-to-End Convolutional Neural Acoustic Model
  - Sequence-to-Sequence Speech Recognition with Time-Depth Separable Convolutions
  - ContextNet

Convolutional Neural Networks

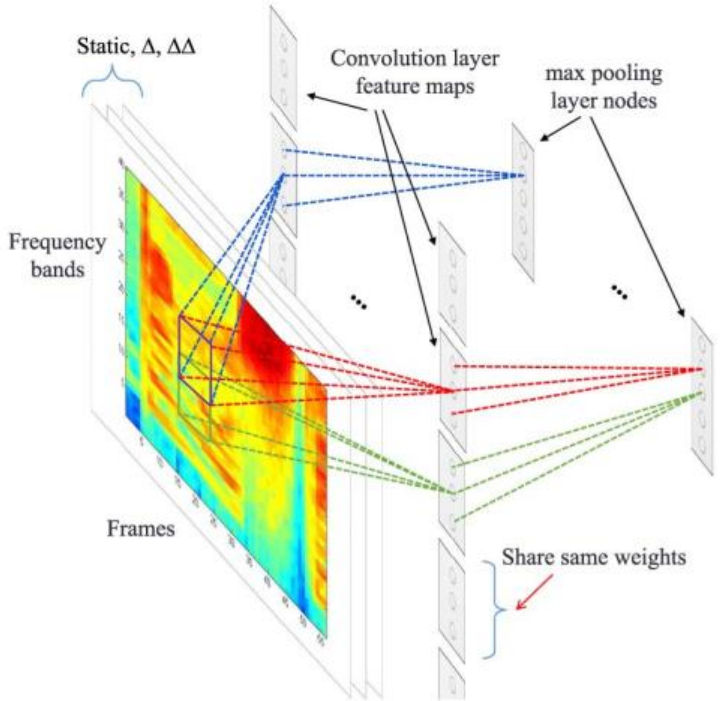
- Common method for computer vision
- Also adopted for speech recognition
- Usually have alternative pooling and convolutional layers, with fully connected layers in the end
- 1D CNNs: Speech signal as input
- 2D CNNs: Input signal is transformed to 2D similar to images

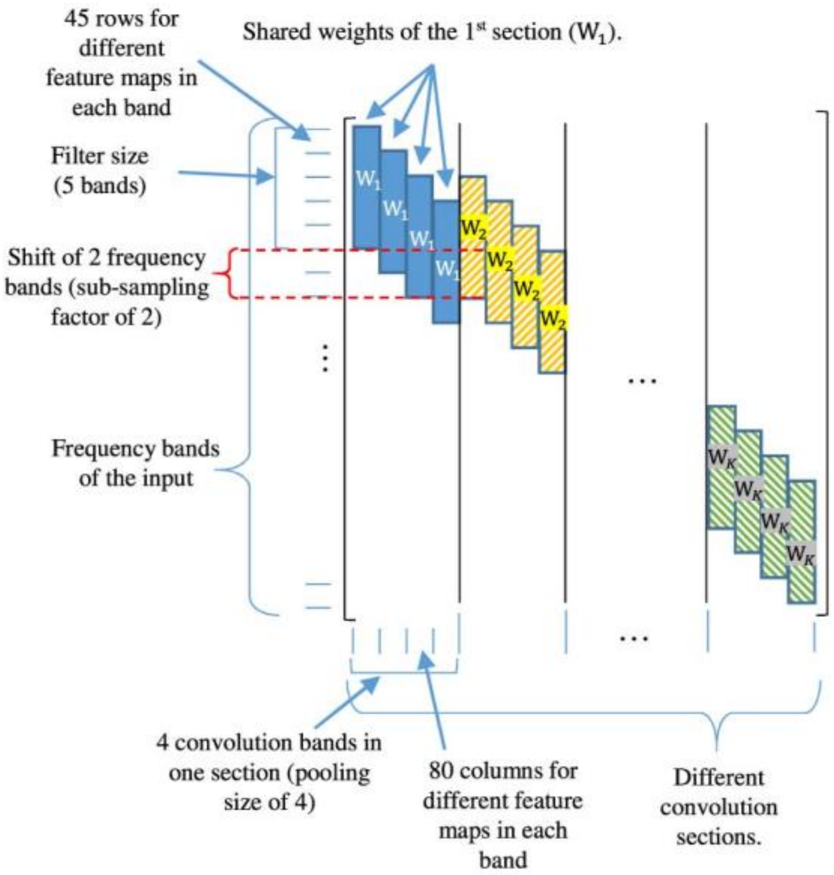
Convolutional Neural Networks



[Zhang2019]Example of 2D CNN and input of time-frequency signal

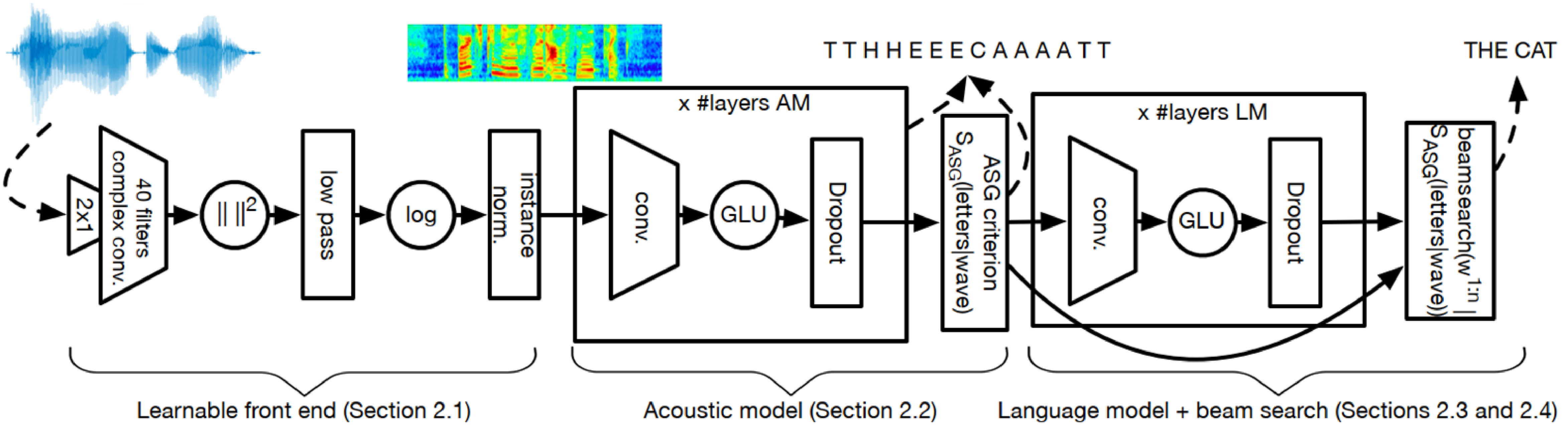
1D-CNN for speech recognition





[Abdel2014] LWS illustration

Fully Convolutional method for speech recognition

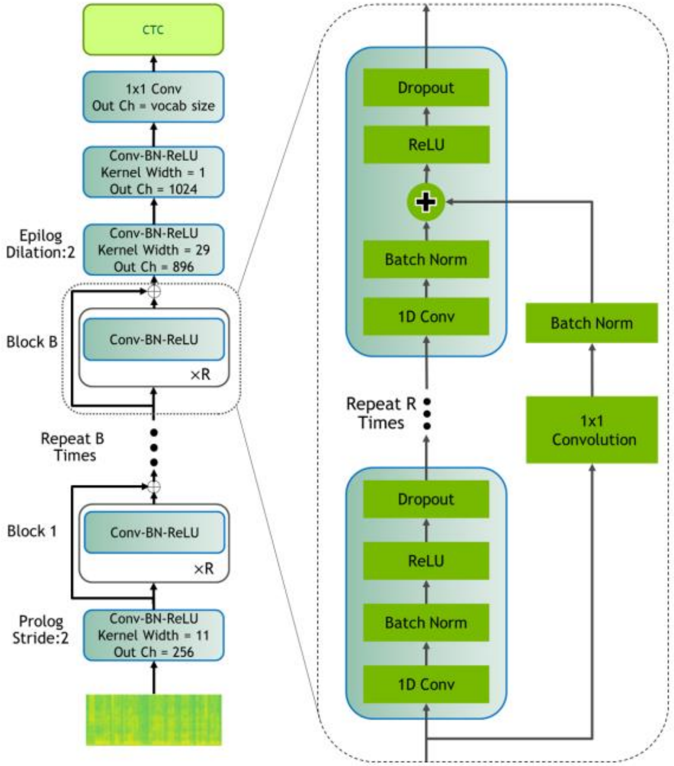


[Zeghidour2018] Illustration of fully convolutional network

Jasper: An End-to-End CNN Acoustic Model

- End-to-end ASR system with convolutional layers
- Input: mel-filterbank features obtained from 20 msec windows with a 10msec overlapping
- CNN has residual and dense blocks
- Tested with different types of normalization and activation functions
- Each block is optimized to fit on a single GPU kernel for faster inference

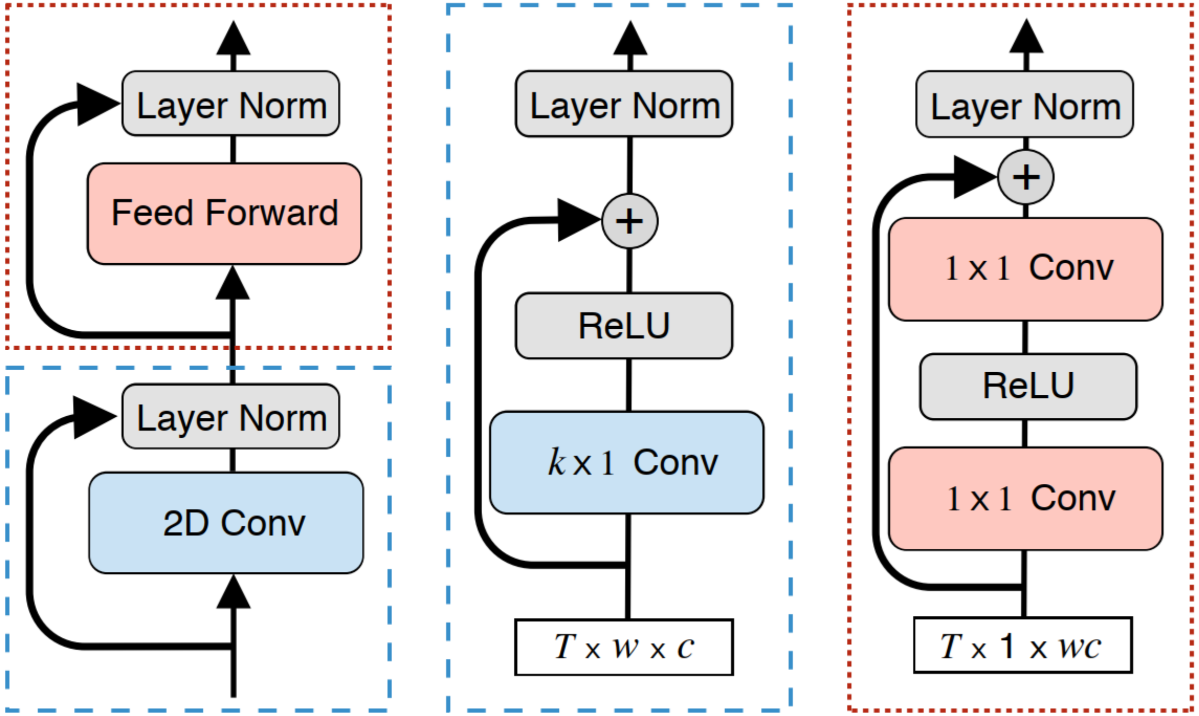
Jasper: An End-to-End CNN Acoustic Model



[Li2019] Overview of Jasper method

Artificial Intelligence &
Information Analysis Lab

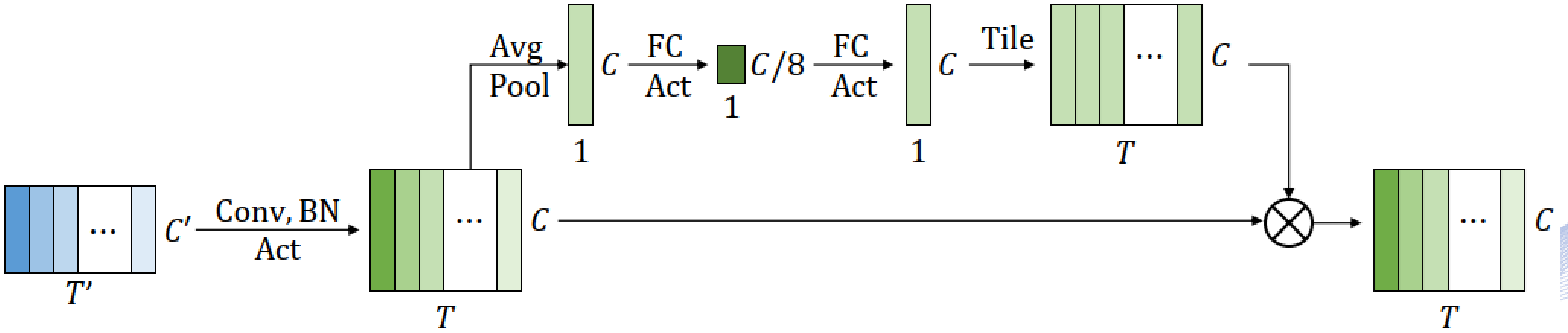
Speech Recognition with Time-Depth Separable Convolutions



[Hannun2019] Time-separable CNN

Artificial Intelligence & Information Analysis Lab

ContextNet



[Han2020] SE module

Transformers

- Methods
  - Speech Transformer
  - Transformer Transducer
  - Conformer
  - Semantic Masked Transformer

Transformers

- With the introduction of Transformer networks machine translation and speech recognition have seen significant improvements.
- Transformer models that are designed for speech recognition are usually based on the encoder-decoder architecture similarly to seq2seq models.
- They are based on the self-attention mechanism instead of recurrence that is adopted by RNNs.

Speech Transformer

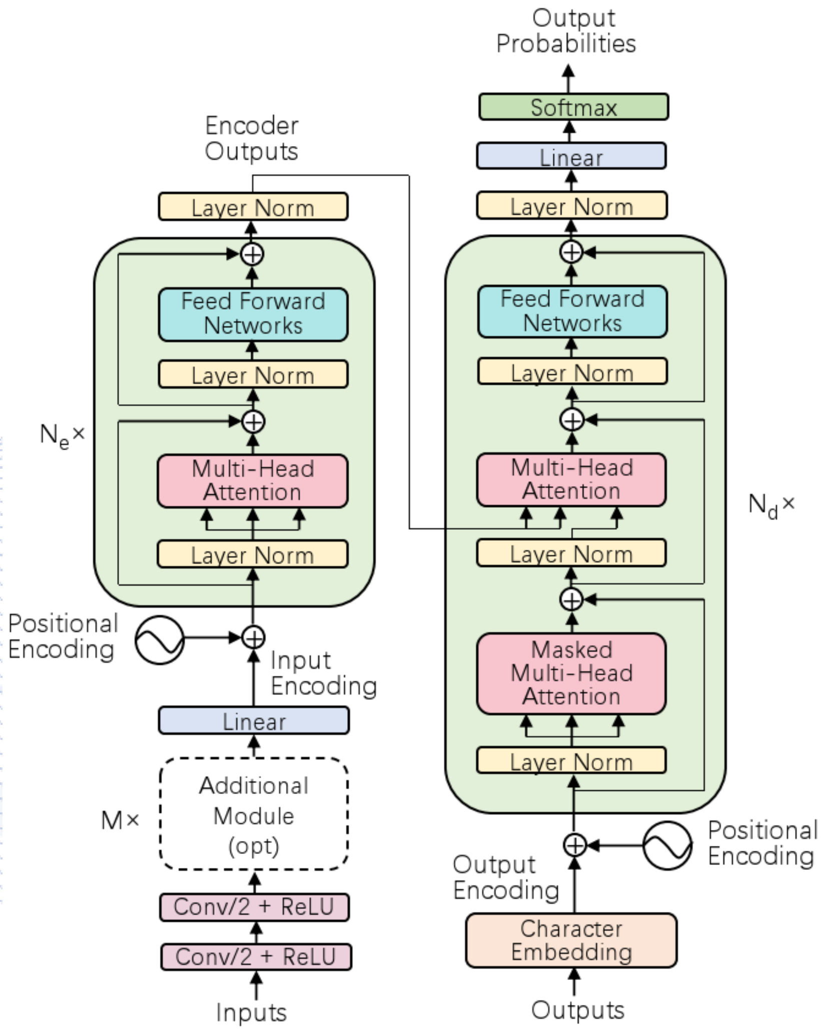
- Transforms the speech feature sequence to the corresponding character sequence.
- The feature sequence which is longer from the output character sequence is constructed from 2-dimensional spectrograms with time and frequency dimensions.
- CNNs are used in the input to exploit the structure locality of spectrograms and mitigate the length mismatch by striding along time.

Speech Transformer

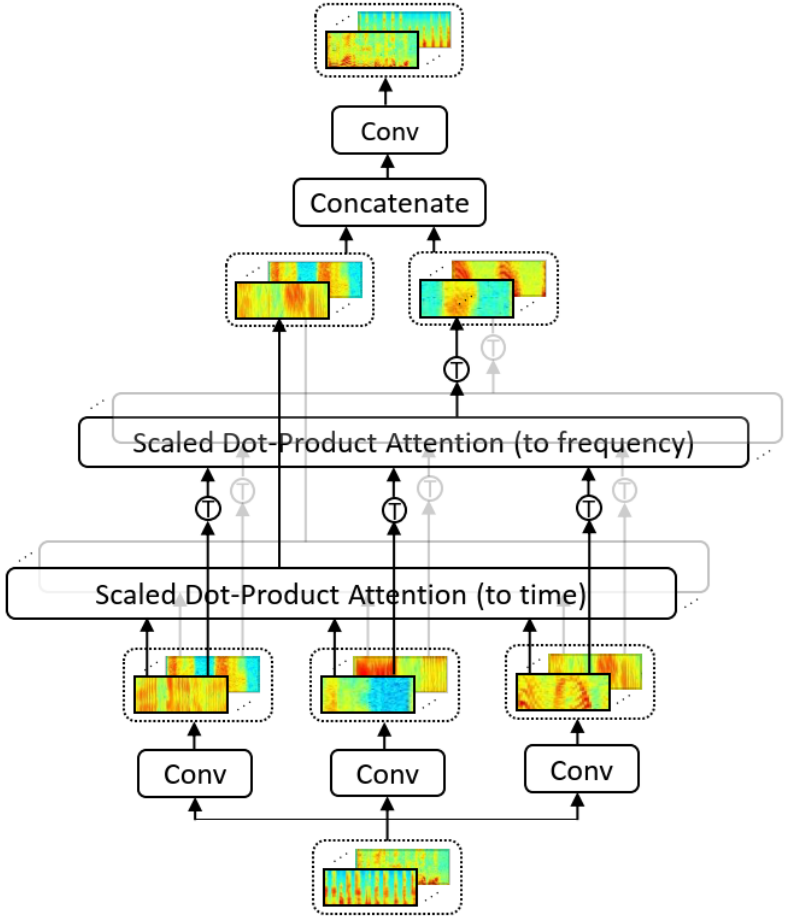
- In the Speech Transformer, 2D attention is used in order to attend at both the frequency and the time dimensions.
- The queries, keys and values are extracted from CNNs and fed to the 2 self-attention modules.

Speech Transformer

[Dong2018] Speech Transformer



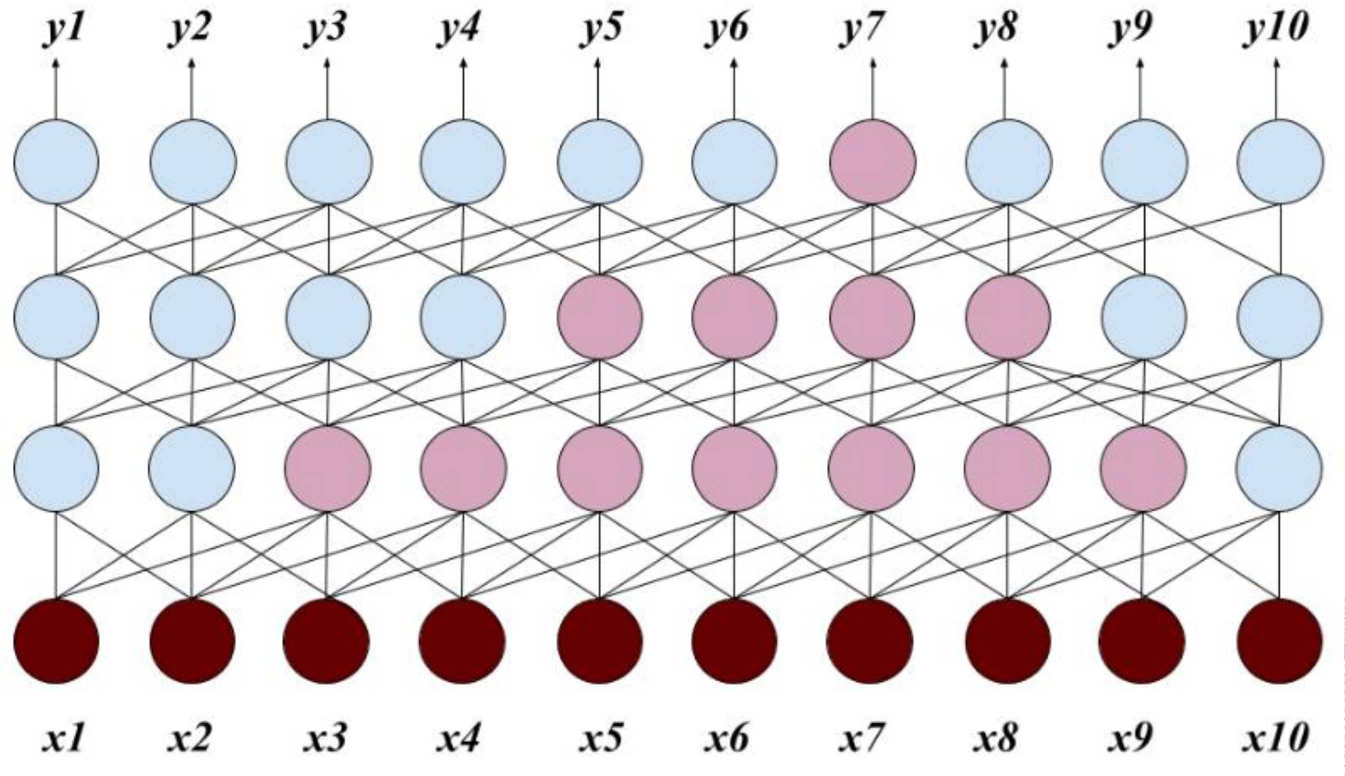
Speech Transformer



[Dong2018] 2D attention module from Speech-Transformer

Artificial Intelligence & Information Analysis Lab

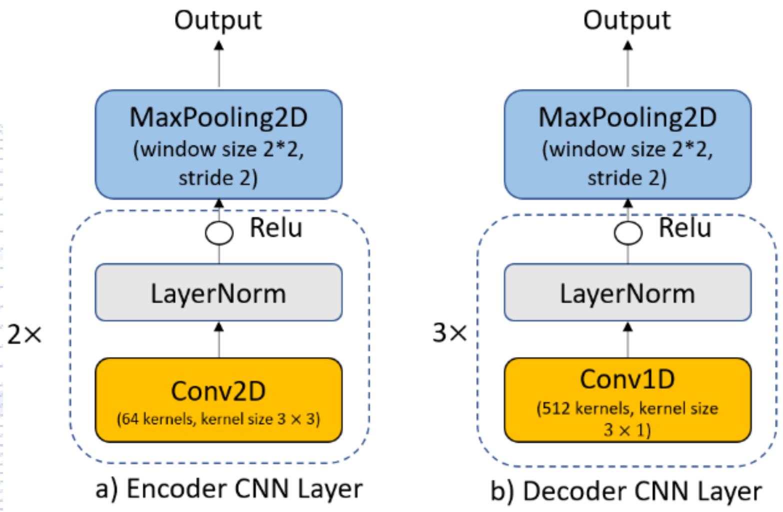
Transformer Transducer

- 

[Zhang2020] Example of context masking for label y_7

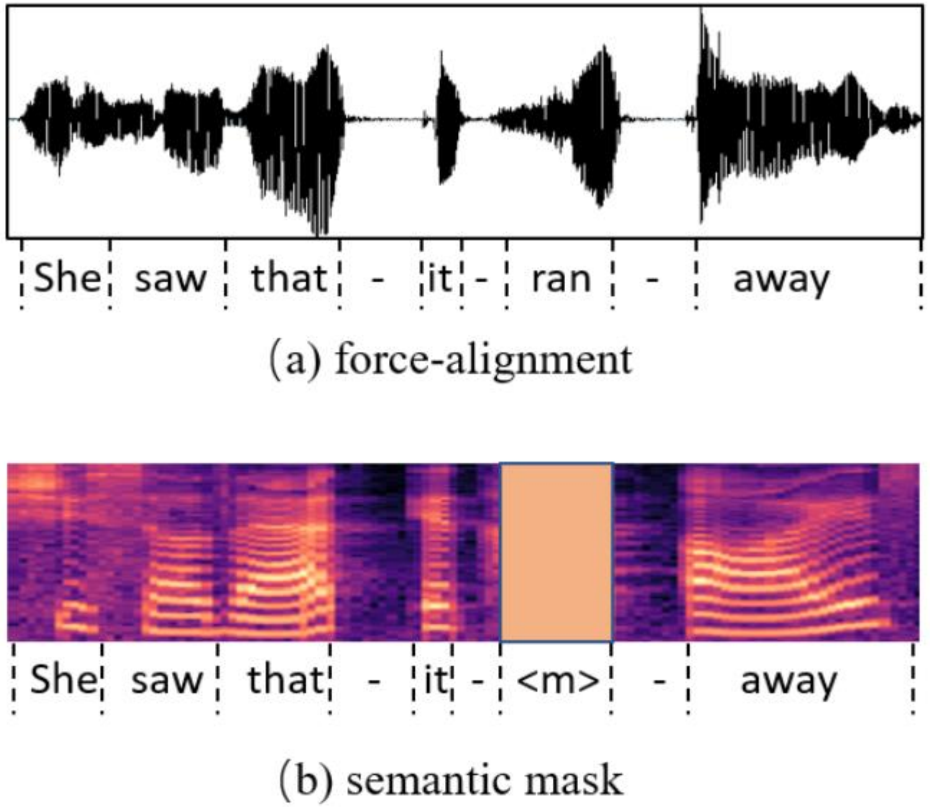
Semantic Mask for Transformer for ASR

- Transformer encoder –decoder architecture
- Mel-scale features with dimension equal to 83
- VGG-like input convolutions for local relationships



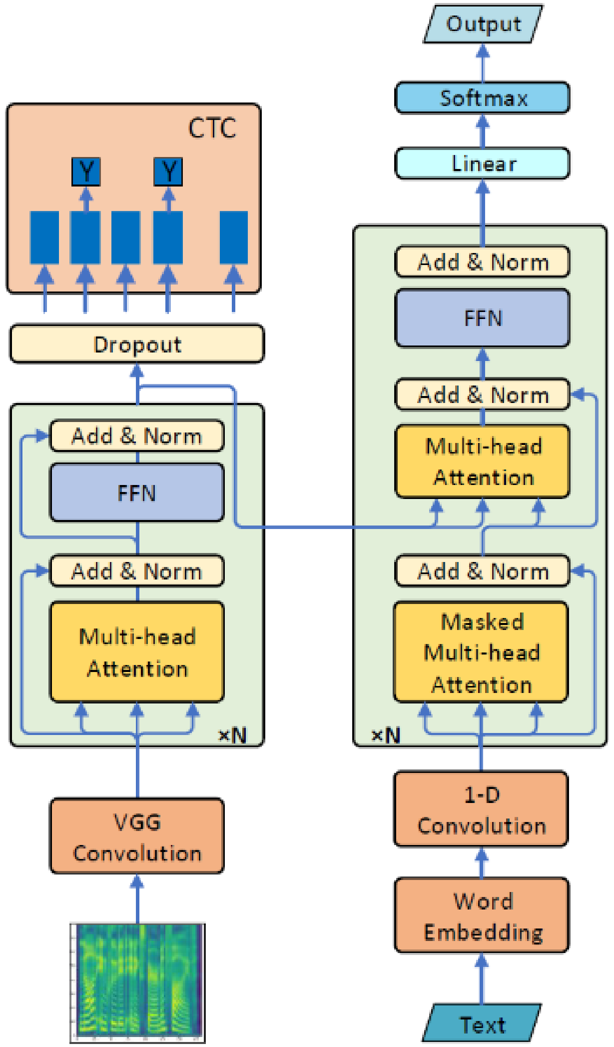
[WangC2020] CNN layer

Semantic Mask for Transformer for ASR



[WangC2020] Semantic mask

Semantic Mask for Transformer for ASR



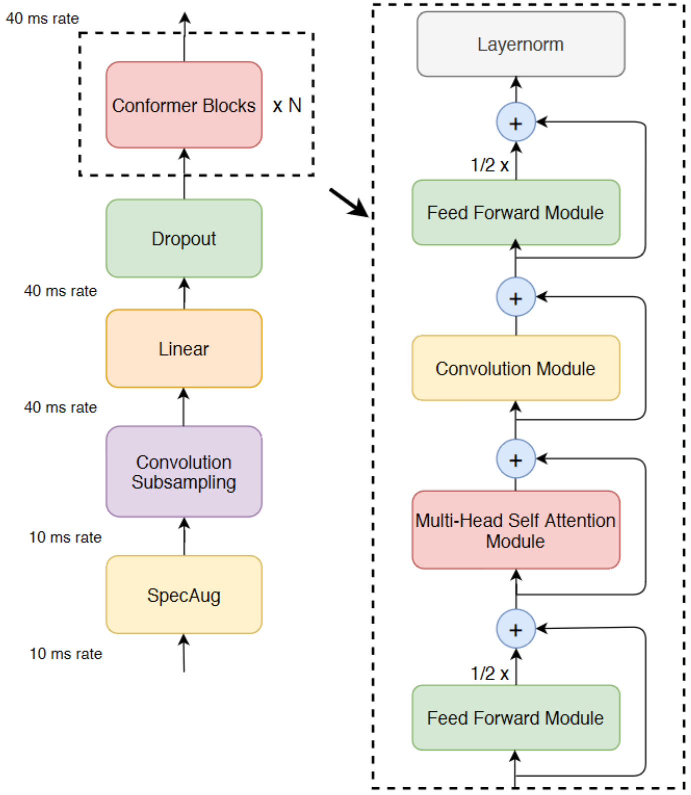
[WangC2020] Model architecture

Artificial Intelligence & Information Analysis Lab

Conformer

- is a variant of the original Transformer that combines CNNs and transformers in order to model both local and global speech dependencies by using a more efficient architecture and fewer parameters.
- The main module of the Conformer contains two feedforward layers (FFN), one convolutional layer (CNN) and a multi head attention module (MHA).

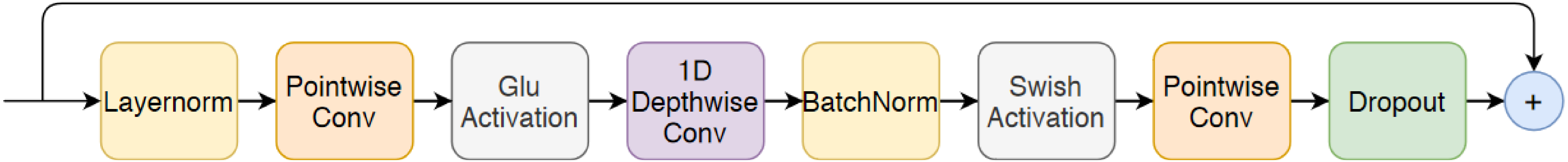
Conformer



[Gulati2020] Conformer method

Artificial Intelligence & Information Analysis Lab

Conformer



[Gulati2020] CNN block of the Conformer

Bibliography

- [Chorowski2015] Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. arXiv preprint arXiv:1506.07503. 2015 Jun 24.
- [Zhang2019] Zhang W, Zhai M, Huang Z, Liu C, Li W, Cao Y. Towards end-to-end speech recognition with deep multipath convolutional neural networks. InInternational Conference on Intelligent Robotics and Applications 2019 Aug 8 (pp. 332-341). Springer, Cham.
- [Hori2018] Hori T, Cho J, Watanabe S. End-to-end speech recognition with word-based RNN language models. In2018 IEEE Spoken Language Technology Workshop (SLT) 2018 Dec 18 (pp. 389-396). IEEE
- [wang2017] Wang Y, Deng X, Pu S, Huang Z. Residual convolutional CTC networks for automatic speech recognition. arXiv preprint arXiv:1702.07793. 2017 Feb 24.
- [Zeghidour2018] Zeghidour N, Xu Q, Liptchinsky V, Usunier N, Synnaeve G, Collobert R. Fully convolutional speech recognition. arXiv preprint arXiv:1812.06864. 2018 Dec 17.
- [Malik2021] Malik M, Malik MK, Mehmood K, Makhdoom I. Automatic speech recognition: a survey. Multimedia Tools and Applications. 2021 Mar;80(6):9411-57.

Bibliography

- [Graves2013] Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In2013 IEEE international conference on acoustics, speech and signal processing 2013 May 26 (pp. 6645-6649). Ieee.
- [Rao2017] Rao K, Sak H, Prabhavalkar R. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2017 Dec 16 (pp. 193-199). IEEE.
- [Abdel2014] Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing. 2014 Jul 16;22(10):1533-45.
- [Chan2016] Chan W, Jaitly N, Le Q, Vinyals O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016 Mar 20 (pp. 4960-4964). IEEE.

Bibliography

- [He2019] He Y, Sainath TN, Prabhavalkar R, McGraw I, Alvarez R, Zhao D, Rybach D, Kannan A, Wu Y, Pang R, Liang Q. Streaming end-to-end speech recognition for mobile devices. InICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019 May 12 (pp. 6381-6385). IEEE.
- [Wang2017] Wang Y, Deng X, Pu S, Huang Z. Residual convolutional CTC networks for automatic speech recognition. arXiv preprint arXiv:1702.07793. 2017 Feb 24.
- [Li2019] Li J, Lavrukhin V, Ginsburg B, Leary R, Kuchaiev O, Cohen JM, Nguyen H, Gadde RT. Jasper: An End-to-End Convolutional Neural Acoustic Model}}. Proc. Interspeech 2019. 2019:71-5.
- [Hannun2019] Hannun A, Lee A, Xu Q, Collobert R. Sequence-to-sequence speech recognition with time-depth separable convolutions. arXiv preprint arXiv:1904.02619. 2019 Apr 4.
- [Han2020] Han W, Zhang Z, Zhang Y, Yu J, Chiu CC, Qin J, Gulati A, Pang R, Wu Y. ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context.

Bibliography

- [Dong2018] Dong L, Xu S, Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018 Apr 15 (pp. 5884-5888). IEEE.
- [Zhang2020] Zhang Q, Lu H, Sak H, Tripathi A, McDermott E, Koo S, Kumar S. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. InICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020 May 4 (pp. 7829-7833). IEEE.
- [Gulati2020] Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y, Pang R. Conformer: Convolution-augmented Transformer for Speech Recognition}}. Proc. Interspeech 2020. 2020:5036-40.
- [Povey2011] Povey D, Burget L, Agarwal M, Akyazi P, Kai F, Ghoshal A, Glembek O, Goel N, Karafiát M, Rastrow A, Rose RC. The subspace Gaussian mixture model—A structured model for speech recognition. Computer Speech & Language. 2011 Apr 1;25(2):404-39.
- [GMM2020] https://www.mathworks.com/help/audio/ug/speaker-verification-using-gaussian-mixture-model.html

Bibliography

- [WangC2020] Wang C, Wu Y, Du Y, Li J, Liu S, Lu L, Ren S, Ye G, Zhao S, Zhou M. Semantic Mask for Transformer Based End-to-End Speech Recognition}}. Proc. Interspeech 2020. 2020:971-5.

Artificial Intelligence &
Information Analysis Lab

Bibliography

[OPP2013] A. Oppenheim, A. Willsky, Signals and Systems, Pearson New International, 2013.

[MIT1997] S. K. Mitra, Digital Signal Processing, McGraw-Hill, 1997.

[OPP1999] A.V. Oppenheim, Discrete-time signal processing, Pearson Education India, 1999.

[HAY2007] S. Haykin, B. Van Veen, Signals and systems, John Wiley, 2007.

[LAT2005] B. P. Lathi, Linear Systems and Signals, Oxford University Press, 2005.

[HWE2013] H. Hwei. Schaum's Outline of Signals and Systems, McGraw-Hill, 2013.

[MCC2003] J. McClellan, R. W. Schafer, and M. A. Yoder, Signal Processing, Pearson Education Prentice Hall, 2003.

Bibliography

[PHI2008] C. L. Phillips, J. M. Parr, and E. A. Riskin, Signals, Systems, and Transforms, Pearson Education, 2008.

[PRO2007] J.G. Proakis, D.G. Manolakis, Digital signal processing. PHI Publication, 2007.

[DUT2009] T. Dutoit and F. Marques, Applied Signal Processing. A MATLAB-Based Proof of Concept. New York, N.Y.: Springer, 2009

Bibliography

[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.
[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr