A Study Review of Neural Audio Speech Transposition over Language Processing

Chapter · April 2023

DOI: 10.1007/978-3-031-30396-8_1

4 authors:

Sharun Akter Khushbu
Daffodil International University

Moshfiqur Rahman Ajmain
Daffodil International University

Mahafozur Rahman Rudro
Daffodil International University

Sheak Rashed Haider Noori
Daffodil International University

A Study Review of Neural Audio Speech Transposition over Language Processing

Sharun Akter Khushbu, Moshfiqur Rahman Ajmain, Mahafozur Rahman, Sheak Rashed Haider Noori

Daffodil International University, 1341 Dhaka, Bangladesh
sharun.cse, moshfiqur15-14090, mahafozur15-2955 {@diu.edu.bd}
drnoori@daffodilvarsity.edu.bd

Abstract. Natural Language processing is the advancement of Artificial Intelligence in the modern technological era. Machine Translation this is the vast majority of languages transformation among human languages and computer interaction. NLP domain creates a sequential analysis of the path where the neural network basement is mathematically and theoretically strong enough. In unwritten language aim to multimodal language transformation. According to the spoken language there are several aspects of prosperity. Thereby coming up with the development of linguistic CNN models for all manpower who spoke in their mother tongue. Therefore, concern with english speech language processing advancement has a great impact on language transformation. In a sense other languages can be placed by the speech to language transformation computational period emerged on severally made corpus languages. Due to this implementation of model refer probabilistic model. Consequently, this is a study of a benchmark of recent happenings in unwritten language to language modeling in which summarization or transformation will be faster. With concern current research work that has described how performing best in CNN model and attention model, described statistical deployment. Finally, this paper is capable of inflicting the solution. Furthermore, studies have discussed the impact of result, observation, challenges and limitation with the respect of the solution. The challenge is voice identification without noise is challenging.

Keywords: NLP · CNN · AI · Attention Mechanism

1 Introduction

Machine Translation invented a Neural network possibility by creating a wide research field of Natural Language Processing. Furthermore NLP explores the area of visualization power across large amounts of data. Tropically, NLP has capability of processing various types of problems. NLP was introduced [1] many times ago but recently lots of repositories have been created throughout the knowledge invention. Moreover, researchers at present relate their languages to modify the language as instinctive modernized support. According to the field specification lot of sectors emerged on NLP such as Automatic Speech Recognition(ASR) [2-4], Speech to emotion recognition [15] [10], speech coding [5], NER [6-7] for text annotation, continuous sequences assistive technology [8-9]. Natural language processing study to help a sustainable productive research output. This study refers to a review on unwritten languages where study had been rendered as the current NLP path driven by the Convolutional Neural Network(CNN). The review analysis that focused on comparative study where specifically upholds how much machine translation is reliable and exists with audio speech synthesizing. This is the sequential model analysis of DL [10-15]. CNN layer is the expertise of the noise reduction onto any speech. Large integrated data in heterogeneous modules predict over the solution and limitation of speech recognition. Natural language processing attentively genre the computational resources where speech detect on its utterances. Moreover, input data remarks by hand correction before preprocessing the word phrase. In language processing determine the accurate speech output that totally divides the word and converts into text. Language processing replaces the machine readable power in order to change keyboard transformation through voice recognition. Definitely there are some limitations to gathering large volumes of data but more than advantages too. Throughout the voice synthesization method to text or something can reduce larger time boundaries. State of the art worked on different Corpora that predict language models which can suggest the same category of word. On the contrary, there are also difficulties in different corpora segmentation based on language utterances, word selection, length of sentences, relativity of words, speakers fluency, noiseful data. Furthermore, preprocessing is also challenging against different countries based on different languages. Due to this analysis the outcome mostly refers to several explorations such as emotion, hate words, sarcastic words, ASR and crime detection etc. According to this research initial input has a great part of sound activity in which tape to record voices using microphone, smartphone recorder and so on recording instruments. Language processing has been placed a tremendous avail for deaf those who do not hear by ear. In this discussion that reviewed those research which relates to unwritten languages and also comprehensive use of sequential CNN model learning in neural architectures. A machine translation model is an attention mechanism developed inside creating RNN models.

2 Brief Background

Speech transposition is the recent trend in various sub continental and other countries. In order to do so many researchers had worked on it. A brief background only can analyse, produce feedback, data set availability, model efficiency and comparative study deliver in the next advancement. Odette Scharenborg et al. [16] asserted by the exploration of deep learning throughout sequence to sequence language processing. Accepted audio file addressed by the unwritten language confessed about not only one but also three representations along with defeat necessary of language formulation difficulties. As mentioned with speech-translation [17, 18] required LSTM, speech-image besides retrieving image-speech required PyTorch. An unwritten language [19] needs to specify speech to meaning vice versa for utterances because of denoting by image, translation, documentation and auto text generation. Deep neural networks [17, 20, 21] modulate the signal [22] of any natural languages into text form [23,24]. Proceeded by speech translation paves that English-English, English-Japanese, Mboshi-Mboshi, Mboshi-french refers BLEU score later apart from speech-image stand by BLEU scores and PER(Phone Error Rate). In [25], The authors proclaimed the Speech emotion recognition replace the ASR technique where audio speech produces a signal after that applied by the attention mechanism it enhanced speech emotion recognition. In this research study developed a transfer learning scheme in which aligned between speech frames and speech text. Researchers had been establishing an attention mechanism model due to the RNN and LSTM model. According to the analysis showed comparative performance with LSTM+Attention, CNN+LSTM, TDNN+LSTM. The dataset of speech has been collected from the IEMOCAP source and trained by the Bidirectional LSTM model. With ideal parameter settings machines have learned a multimodal feature that produces a sequential model with fully connected hidden layers by the 0.001 learning rate and 16 kHz utterances pattern up to 20 seconds duration. According to the comparison rate speech emotion recognition states that Oracle text accuracy reminder value addition with other comparative analysis. In [26], the authors facilitated distorted speech signal processing that have used the Transformer MT system and LSTM. Removed noisy background found clean speech that are used as BPC phonetic class and (BPPG) posterior-gram developed SNR system. Nurture with TIMIT dataset was evaluated as BPSE rate ground truth rate, noise ratio and transformer, LSTM performance. The SNR system implies acoustic signals into symbolic sequences. The study of incorporating broad phonetic information for speech enhancement was outperformed by the overcome different SNR criteria. In [27], the authors employed an assistive technology that covers many sectors of COVID-19 with the audio speech synthesis analysis. It had been studied about speech transformation about Covid or not Covid cough sound samples, voice synthesizing for face mask wearing or not, breathing speed ups and down analysis, Covid speech to text analysis and Mental health sensitivity analysis from twitter, instagram sound clip. All are the unwritten languages detected Covid-19 situation that is customized by the attention mechanism and transformation of natural language processing. In [28], the authors elaborate between virtual speaker speech and real speaker speech that have shown the possibility of noise reduction. In [29], the authors had employed over voice recording in which (WCE) word count estimation using six numbers of various corpora of several languages. English languages such as French, Swidish, Canadian, Spanish are different languages covered by the daylong recordings from children. According to this audio speech produces consistent performance over model in all several corpora. Collected speech worked for two specifications one is speech activity detection where detect word depends on its utterance also deduct noise another work was syllabification of speech that check the phonetic models. Study had also illustrated the limitations of working on speech recognition synthesis. In [30], the authors demonstrate that emotion recognition over natural language processing. This is another speech transposition procedure towards enhancing LSTM based models. Also facilitate the preprocessing section where speech is processed by the word. According to the coefficient rate authors showed a visual representation of audio speech. The working section represents the functionality of CNN, LSTM that produce outperform results. In [31], the author's explanation states that LSTM 2D convolutional layers in order to perform speech transformation analysis according to the speech signal processing where signal can drive brain signals. Moreover, Graphical representations that reproduce clean signals from original signals. Separating clean signals after that using theoretical implementation noisy speech cleaned. In [32], the authors claimed a spectrogram convolutional neural networks towards a 2240 speech dataset where it combines with depressed and nondepressed related data. Due to speech signal processing developed a model that represents a convolutional 256 hidden layer with Dense layer containing max pooling and softmax. End to end convolution neural network model gained 80% highly accurate model check validity of F-score. This study also proclaimed for speech to depression detection behalf of the controversial LSTM model that justified all necessary parameters.

3 Methodology

3.1 Statistical Approach

Analysis is the data set that has the great effort in model end-to-end unwritten language transformation where neural networks recreate in sequence-to-sequence machine conversion. Regarding through the making larger dataset quite tuff in audio speech data collection. Due to audio speech or audio voice collection from different inputs there are more difficulties in the dataset. According to the unwritten languages collection phrase dataset included by noisy input. Therefore, SNR fixed the issues and reduced the noise ratio and reformed the speech as a clean speech that is standard for speech processing in model. This input signal converted into a wav signal afterwards wav signal put counting number head of word vectors. CNN multimodal language modeling trained large volumes of data so that states statistical of dataset very high volume. Sequence to sequence segmentation is the review of Dataset where models generate effective output. ASR technique using CNN model that experiment on publicly known larger dataset eg. LRS2 [33-34] around thousands of sentences on the news. According to the utterances separately another large dataset VoxCeleb2 [35] that had been made by the different spoken input among 6000 speakers. Tropically, the dataset is fed into the model by dividing two sizes and that is Training dataset and testing dataset.

Speech review work is not available. Here the discourse review work and discussion of conflicting results distinguish this work from other works.

3.2 Experimental Setup & Evaluation Protocol

CNN models transform in multimodal two languages output with the experimental setup. Thereby, input signals all types of difficulties have been measured by theoretical terms also collaboration with data preprocessing in which need to cover the challenges of spectacular noise removal signal. In the preprocessing method raw dataset delivers a significant word vector output by applying a lemmatization method afterwards parser parses the data with word level annotation. Any research study there has a justified protocol where machine translation in NLP by CNN represents an optimum, effective solution. Therefore, removing noise unavoidable input that raises speech quality where Interference Ratio of signal detect low voice, unclear speech. Furthermore, other staff such as SAR, SDR and STOI have been looking to wav signal quality ratio, intelligibility of transform sequential signal. Phase distortion calculated by matrix parameters what have reformed by character or word number [36]. Thereby, a great number of research domains apply ASR technique and enhanced WER in few studies.

3.3 CNN model

Due to the analysis of input framework CNN model is the right approach to centralized the actual accuracy path. CNN model contains by itself fully connected of 5 layers where each layer is centrally connected with hidden 8 layers [37,38]. Convolutional neural network developed with faster GPU module where highly powerful NVIDIA graphics that visionary signal processing output where CNN layer converted it into 512 unit or 256 unit. Nevertheless, CNN states the optimum output in each unit by adam optimizer. Relu functionality [39] that has pays render the activity of the Dense layer. Each unit maintains the hiding dense layer from starting to fully connected. Mostly lots of large volumes of data give compatible high order accuracy suggested in CNN. Popular dataset runs in CNN remain very high efficient like PASCAL VOC. Speech data requires a CNN model nust where recurrent neural network used in the model adjusting with LSTM model.

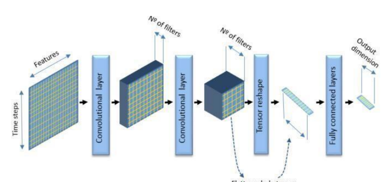

Fig. 1. ASR of language transformation acoustic model.

Paper Title	Speech technology for unwritten languages	Learning Alignment for Multimodal Emotion Recognition from Speech	Incorporating Broad Phonetic Information for Speech Enhancement	An Overview on Audio, Signal, Speech, & Language Processing for COVID-19
Author	Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa -Johnson, Florian Metze, Graham Neubig, Sebastian Stuker, Pierre Godard, Markus Muller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, Emmanuel Dupoux	Haiyang Xu, Hui Zhang, Kun Han, Yun Wang,Yiping Peng, Xiangang Li	Yen-Ju Lu, Chien-Feng Liao, Xugang Lu, Jeih-weih Hung, Yu Tsao	Gauri Deshpande, Bjorn W. Schuller
Method	Speech to Text conversion 3 speech translation approach for speech to image, image to speech	ASR system in Bidirectional method	Phonetic based acoustic model on speech to word LSTM two language modeling	Speech to text and emotion detection CNN method approached
Dataset	FlickRreal Dataset	IEMOCA P Dataset of speech files	TIMIT dataset	Self collected features of speech signal
Size	6000 images 30,000 speech files		3696 utterances of speech	NA
Accuracy	BLEU score 7.1%	PER score 14.7% WA-70.4 UA-69.5	LSTM score 78%, BPG score 0.824	Accuracy 0.69%
Year	2020	2020	2020	2020

Paper Title	The fifth 'CHIME' Speech Separation and Recognition Challenge: Dataset, task and baselines	The Conversation: Deep Audio-Visual Speech Enhancement	Phoneme-Specific Speech Separation
Author	Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal	Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman	Zhong-Qiu Wang, Yan Zhao and DeLiang Wang
Method	CNN encoder-decoder based LSTM network	Loss function estimation in SNR ratio measured PESQ on CNN based networks	ASR systems NMF method in DNN network model
Dataset	Self-collected features of speech signal	LRS2 Dataset on noisy audio speech	Self-collected dataset on ASR
Size	200k utterances	1000 sentences & 6000 different speakers	4026 several speech files
Accuracy	200k utterances (LF-MMI TDNN) 81.3 end-to-end 94.7	WER score is 98.9% accuracy, ground truth signal 8.8%	ASR performance in WER is 13.46%
Year	2020	2020	2020

3.4 LSTM

Language modeling difficulties occur by its language phonetics, morphology and anafora. Due to audio speech analysis LSTM [40] transfers a voice signal into sequential processing where the model can convert text or word or break into the sentence by labeling. LSTM gives the facility of regularizing, word embeddings [41] and optimizing. Recurrent neural networks developed its model on encode the input source and decode the reformed text then produce output with BLEU score. Encoded output new sequential text formed as metrics which is measured by categorical cross entropy entropy and adam optimizer. Sequential processing moderate the model setting parameters counting matrix weighted position. Tropically, LSTM is two languages modeling task implementation happening by the 3D shape matrix in which it is highly modulated [42-44].

3.5 Attention Mechanism

Many of the unwritten language translations that have involved another attention model analysis. This is another structured model enriched by the researcher where context follows multi head mechanisms that are double linked with self attention base multi head transfer. According to this transformer model developed feed forward networks with encoding signals and output decoding signals in word or character sequences or summary representation. Furthermore, attention mechanisms have a duty to positional encoding where encoding structured sequential sentences convert it into words. After that, positional encoding in terms of transfer the signal of word count as a head thisl head of word concat with each other and decode a sequential output. NMT requires output estimation of

BLEU scores by plotting n grams key. N grams key compute the final output sequence with BLEU score [45-52].

3.6 Findings and Limitation

In a sense, the reviews of the unwritten languages based on the neural network have lots of findings defining the phonetically, morphological, utterances high similarity of input with output. CNN Neural network, RNN, LSTM and attention mechanism derive rap id advancement. Findings that measure those model efficiency also enlightened the next upcoming related research where this model can contribute a big part for improving intelligibility and quality. These models handle noisy conditions that give noiseless output signals. Another study also has explained about LSTM networks good for phonetic module to speech capturing. SNR puts input from audio data after that fresh data ready for parsing, lemmatization. An acoustic model that has some validity addressing th e development of Corpora. NLP processes different languages and comes up with multi - modal transformation. In Table - 1 review that has some specification about CNN model higher accuracy. All popular dataset and few are self - collected data set responded to very high accuracy. According to this review, derived data-set volume and a healthy dataset will make an impact based on the model and also depending data-set on its own waveform high to low frequencies calculation. This paper's technical contribution is identifying the different types of statistical analysis on the other side exploratory analysis of speech's recent trend of work.

4 Impact of Result

To the best of our knowledge, a dataset plays a strong character when input source sounds good for volume and noiseless. By this observation, wav signal is capable of interpreting as much wrong and depending on it everything will spoil taking wrong input, preprocessing result i s not good, model unoccupied for transfer learning and so on problematic functionalities will arise. Thus all that staff misguide the multimodal language transformation process for sequential learning. To come up with the solution of limitation, the study make table where approached input the great compatibility of WER analysis, SNR technique making noiseless waveform, BPC technique in articulatory by place and manner. Furthermore, larger data set also requires effective accuracy and loss function calculation rate that can conduct or address functionality of computational error.

5 Conclusion

In this paper, which has reviewed the model flexibility over voice signal in benchmark of automatic speech recognition. According to the study of data-set it

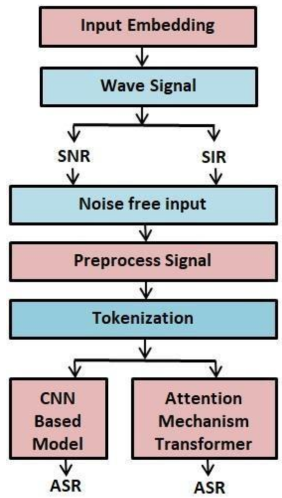

Fig. 2. ASR of language transformation acoustic model.

makes sense to have huge doman in unwritten languages depending on the several corpus of languages and volume of data. Concluding research summary finally after review we have got two known models is CNN and attention mechanism. These two models explain ASR technique possibilities with a very low loss function rate thereby accuracy in all data-set very effective. Therefore, Audio speech in which is collected by microphone and recorder after removing noise machine translation output estimate sequential output. This translation developed a multimodal language transformation. Lots of language domains related with unwritten languages are not involved in computational transformation multimodal languages. Various sub-continental languages or country languages have created the recent domain of research. In terms of small numbers of study also begin sequential analysis among unwritten languages. Due to this advancement poorly required to increase volume of the data set and figure out the corpus lackings also need to fix it and refill the corpus lackings which can surely reach a tremendous comprehensive module.

6 Acknowledgment

DIU NLP & Machine learning Research Lab give support to accomplish our research. We are thankful to give us support together with facilities and guidance.

References

1. Karen Sparck Jones, “Natural Language Processing: A Historical Review”, vol. 9-10, 1994.

2. J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, 2014.
3. H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase sensitive and recognition-boosted speech separation using deep recurrent neural networks," in Proc. ICASSP 2015.
4. F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in Proc. LVA/ICA, 2015.
5. A. J. Accardi and R. V. Cox, "A modular approach to speech enhancement with an application to speech coding," in Proc. ICASSP 1999.
6. Yonghao Jin, Fei Li and Hong Yu, "BENTO: A Visual Platform for Building Clinical NLP Pipelines Based on CodaLab", 2020.
7. Xiang Dai1, Sarvnaz Karimi, Ben Hachey, Cecile Paris, "An Effective Transition-based Model for Discontinuous NER", rXiv:2004.13454v1 [cs.CL], 2020.
8. D. Wang, "Deep learning reinvents the hearing aid," IEEE spectrum, vol. 54, no. 3, pp. 32–37, 2017.
9. Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.- H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," IEEE Transactions on Biomedical Engineering,vol.64,no.7,pp.1568– 1578, 2016.
10. Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, pp. 1849–1858, 2014.
11. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder."inProc.Interspeech2013.
12. Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "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.
13. M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 153–167, 2016.
14. K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual microphone mobile phones in close-talk scenarios," in Proc. ICASSP 2019.
15. J. Qi, J. Du, S. M. Siniscalchi, and C. Lee, "A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, 2019.
16. Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian St¨uker, Pierre Godard, Markus M¨uller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, Emmanuel Dupoux, "Speech technology for unwritten languages", 2020.
17. Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. "Listen and translate: A proof of concept for end-to-end speech to-text translation", In NIPS Workshop on end-to-end learning for speech and audio processing, 2016.
18. Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly transcribe foreign speech. arXiv preprint

arXiv:1703.08581, 2017. [19] Laurent Besacier, Bowen Zhou, and Yuqing Gao. Towards speech translation of non written languages. In Spoken Language Technology Workshop, 2006. IEEE, pages 222–225, 2006.
19. Laurent Besacier, Bowen Zhou, and Yuqing Gao. Towards speech translation of non written languages. In Spoken Language Technology Workshop, 2006. IEEE, pages 222–225, 2006.
20. Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird14, and Trevor Cohn. An attentional model for speech translation without transcription. In Proceedings of NAACL-HLT, pages 949–959, 2016.
21. Radek Fer, Pavel Matejka, Frantisek Grezl, Oldrich Plchot, Karel Vesely, and Jan Honza Cernocky. Multilingually trained bottleneck features in spoken language recognition. Computer Speech and Language, 46(Supplement C):252 – 267, 2017.
22. Fabrice Malfrere and Thierry Dutoit. High-quality speech synthesis for phonetic speech segmentation. In Proc. Eurospeech, pages 2631–2634, 1997.
23. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui.
24. Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In Proc. ICASSP, 2018.
25. W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell” CoRR, abs/1508.01211, 2015.
26. Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, Xiangang Li, “Learning Alignment for Multimodal Emotion Recognition from Speech”, arXiv:1909.05645v2 [cs.CL], 2020.
27. Yen-Ju Lu, Chien-Feng Liao, Xugang Lu, Jeih-weih Hung, Yu Tsao, “Incorporating Broad Phonetic Information for Speech Enhancement”, arXiv:2008.07618v1 [eess.AS], 2020.
28. Gauri Deshpande, Bjorn W. Schuller, “An Overview on Audio, Signal, Speech, & Language Processing for COVID 19”, arXiv:2005.08579v1 [cs.CY], 2020.
29. Jens Nirme, Birgitta Sahlen, Viveha Lyberg Ahlander, Jonnas Brannstrom, Magnus Haake, “Audio-Visual Speech Comprehension in noise with real and virtual speakers”, Elsevier Journal of Speech Communication, 116(2020) 40- 55.
30. Okko Rasanen, Shreyas Seshadri, Julien Karadayi, Eric Riebling, John Bunce, Alejandrina Cristia, Florian Metze, Marisa Casillas, Celia Rosemberg, Elika Bergelson, “Automatic word count estimation from daylong child centered recordings in various language environments using language-independent syllabification of speech”, Elsevier, 113(2019) 63–80.
31. Md. Zia Uddin, Erik G. Nilsson, “Emotion Recognition using Speech and Neural Structured Learning to Facilitate edge Intelligence”, 94(2020) 103775.
32. Enea Ceolini, Jens Hjortkjaer, Daniel D.E. Wong, James O’Sullivan, Vinay S. Raghavan, Jose Herrero, Ashesh D. Mehta, Shih-Chii Liu, Nima Mesgaran, “Brain informed speech separation (BISS) for enhancement of target speaker in multitalker Speech perception”, 223 (2020) 117282.
33. Srimadhur N.S, Lalitha S, “An End-to-End Model for Detection and Assessment of Depression Levels using Speech”, 171 (2020) 12–21.
34. J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, “Lip reading sentences in the world”, in Proc. CVPR, 2017.
35. J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, “Lip reading sentences in the world”, in Proc. BMVC., 2017.
36. “VoxCeleb2 : Deep Speaker Recognition”, arXiv Preprint arXiv: 1001.2267, 2018.

37. P. Mowlaee, "On Speech Intelligibility Estimation of Phase-aware Single-Channel Space Enhancement", ICASSP, 2015.
38. Ken Chatfield, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets", arXiv:1405.3531v4 [cs.CV], 2014.
39. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in NIPS, 2012, pp. 1106–1114.
40. Lu Yang, Qing Song, Zhihui Wang and Ming Jiang, "Parsing R-CNN for Instance-Level Human Analysis", 2019.
41. Stephen Merity, Nitish Shirish Keskar,Richard Socher, "An Analysis of Neural Language Modeling at Multiple Scales", arXiv:1803.08240v1 [cs.CL], 2018.
42. Hakan Inan, Richard Socher, "Tying Word Vectors And Word Classifiers : A Loss Framework For Language Modeling", arXiv:1611.01462v3 [cs.LG], 2017.
43. Martin Sundermeyer, Ralf Schluter, and Hermann Ney, "LSTM Neural Networks for Language Modeling", 2012.
44. Stephen Merity, Nitish Shirish Keskar, Richard Socher, "Regularizing and Optimizing LSTM Language Models", arXiv:1708.02182v1, 2017.
45. Hochreiter, S., Schmidhuber, J., "Long Short-Term Memory", Neural Computation 9 (8), 1997, pp. 1735–1780
46. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need", pages 5998– 6008, 2017.
47. Jan Niehues and Eunah Cho, "Exploiting linguistic resources for neural machine translation using multi-task learning", In Proceedings of the Second Conference on Machine Translation, pages 80–89, 2017.
48. Alessandro Raganato and Jorg Tiedemann, "An Analysis of Encoder Representations in Transformer-Based Machine Translation", 2018.
49. Ozan Caglayan, Loic Barrault, Fethi Bougares, "Multimodal Attention for Neural Machine Translation", arXiv:1609.03976v1 [cs.CL], 2016.
50. Qimin Zhou, Zhengxin Zhang, Hao Wu, "NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification", Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 189–194, 2018.
51. Kyu J. Han, Jing Huang, Yun Tang, Xiaodong He, Bowen Zhou, "Multi-Stride Self-Attention for Speech Recognition", 2019.
52. Bryan McCann, James Bradbury, Caiming Xiong, Richard Socher, "Learned in Translation: Contextualized Word Vectors", 31st Conference on Neural Information Processing Systems (NIPS), 2017.