A Review of Deep Learning Architectures for Speech and Audio Processing

Taiba Wani and Syed Asif Ahmad Qadri

wanitaiba1@gmail.com

syedasifahmadqadri@gmail.com

Abstract- While artificial neural systems have been in presence for over 50 years, it was not until year 2010 that they had made a huge effect on audio and speech single processing with a profound type of such networks. Belief reviews of earlier studies on neural networks and on (deep) generative models relevant to the introduction of deep neural networks (DNN) to speech and audio processing several years ago is given. Deep learning is machine learning subfield that addresses algorithm inspired by the structure and function of the brain cells called ANNs. To understand the effectiveness of various deep learning, different architectures like recurrent neural network (RNN), long short-term memory (LSTM) and convolution neural network (CNN) and Deep neural network (DNN) are reviewed in the study. Given the recent surge in developments of deep learning, this article provides a review of the state-of-the-art deep learning techniques for audio and speech signal processing.

Keywords- ANN, DNN, RNN, LSTM, MFCC, LPCC, Audio processing, Speech processing.

I. INTRODUCTION

Artificial neural systems have increased boundless consideration in three waves up until now, activated by 1) in 1957, the calculation of perceptron, 2) in 1986 the backpropagation calculation, lastly 3) the accomplishment of deep learning in speech acknowledgment and image classification.[1] In this "deep" worldview, structures with numerous parameters are prepared to gain from a monstrous measure of information turning late advances in machine parallelism. Deep learning is an expansion of artificial neural systems (ANNs) to find out about the example (for arrangement) and the highlights (for feature learning) by utilizing more than one shrouded layer. The utilization of numerous layers with nonlinear handling capacity for every node enable deep learning to get familiar with the example of numerous unpredictable signals, for example, speech, pictures, and recordings and the utilization of pre-preparing steps empower us to proficiently prepared a substantial system with millions of nodes. Deep learning algorithms have been for the most part used to additionally improve the capacities of PCs with the goal that it comprehends what people can do [2]. Deep learning permits computational models that are made from various handling layers to learn portrayals of information with different dimensions of reflection. These strategies have significantly improved the best in class speech recognition, visual object acknowledgment, object location and numerous different areas, for example, medicate revelation and genomics. Deep learning finds multifaceted structure in substantial informational collections by utilizing the backpropagation calculation to show how a machine should change its inner parameters that are utilized to register the portrayal in each layer from the portrayal in the past layer. Deep Learning is a subset of Machine Learning. One of the designs of Deep Learning is Deep Neural Networks [DNNs]. Deep learning has become increasingly popular since the introduction of an effective new way of learning deep neural networks in 2006. It has proved very successful for acoustic modeling in speech recognition especially for large-scale tasks, These DNNs are only a class of Artificial Neural Networks having many concealed layers when contrasted with primary neural systems, consequently the name Deep Neural Networks. Neural Networks have been around since numerous decades yet being a vast system, DNNs require more information to examine and thus increasingly amazing PCs. Accordingly, due to the abrupt ascent of incredible PCs utilizing GPUs, deep learning has picked up notoriety in numerous zones as of late.

Rest of the paper is organized as follows. Literature Analysis related to different Deep Learning Architectures is performed in Section 2, along with a comparison table of different articles about deep learning architectures. Section 3 briefly discusses the issues in speech and audio processing and summarizes different applications of architectures of Deep Learning. Section 4 concludes the paper along with the recommendation.

II.. LITERATURE REVIEW

2.1 DEEP LEARNING ARCHITECTURES

Deep Learning architectures have been connected in numerous zones, for example, speech coding, language identification, automatic speech recognition, audio compression, image processing and many more. The different architectures of deep learning and are shown in table 1.

Table 1. List of Deep Learning Architectures

No.	Deep Learning Architectures
1	Recurrent Neural Network (RNN)
2	Long Short-Term Memory (LSTM)
3	Convolutional Neural Network (CNN)
4	Deep Belief Network (DBN)
5	Deep Stacking Network (DSN) or Deep Convex Network (DCN)

2.1.1 Recurrent Neural Network (RNN)

The main engineering of deep learning is Recurrent Neural Networks [RNNs]. RNN has circle like structure because of which it can access past information. In RNN, recurrent associations are shaped in three different ways; between a neuron and a neuron itself or between a neuron and a neuron in a similar layer or with the neuron and a neuron in the past layer of a neural system design. These recurrent associations are framed with covered up and yield neurons just and not with information or inclination neurons. This kind of configuration makes it significant to continue past information to predict current information and to oversee unmistakable talking rates. RNNs are contemplated as a class deep system for the utilization in unsupervised learning in the situations where the profundity of the info information succession can be as enormous as the length since RNNs permit parameter sharing through the various layers of the system [1]. RNNs are created by the utilization of a similar arrangement of loads in a recursive way over a tree like structure, and the tree is crossed in topological request. The RNN is utilized predominantly to anticipate the future information arrangement using past information tests. The RNN is winning with regards to demonstrating grouping information, for example, speech or content.

2.1.2 Long Short-Term Memory (LSTM)

LSTM is an artificial recurrent neural system (RNN) engineering utilized in the field of deep learning. Not at all like standard feedforward neural systems, LSTM has input associations that make it a "broadly useful PC" (that is, it can process whatever a Turing machine can). It cannot just process single information focuses, (for example, pictures), yet in addition whole groupings of information, (for example, discourse or video). For instance, LSTM is material to undertakings, for example, unsegmented, associated penmanship acknowledgment or speech recognition. (LSTM) is a special type of Recurrent Neural Networks (RNNs). LSTM consists of state boxes receiving the inputs through time [3].

2.1.3 Convolutional Neural Network (CNN)

CNNs are viewed as a kind of differential deep engineering where each model contains a convolutional layer and a pooling layer and are stacked over one another [4]. Numerous loads are partaken in the convolutional layer, the pooling layer then again sub-tests the yield originating from the convolutional layer and diminishes the information rate of the underneath layer. The weight imparting together to appropriately picked pooling plans, results in invariance properties of the CNN. Some have contended that the restricted invariance found in CNN isn't acceptable for confounded example acknowledgment errands. Nonetheless, the CNNs have demonstrated adequacy when utilized in PC vision or picture recognition tasks [5]. Some fitting changes in the CNN for picture examination purposes with the end goal that it joins speech properties, the CNN can be used in discourse speech recognition also.

2.1.4 Deep belief network (DBN)

This deep learning architecture was firstly developed by Hinton[6] and has been executed in classification and feature learning successfully. Predominantly DBN consists an unsupervised learning subpart using restricted Boltzmann machines (RBMs) as its building blocks and a logistic regression layer for prediction.

2.1.5 Deep Stacking Network (DSN)

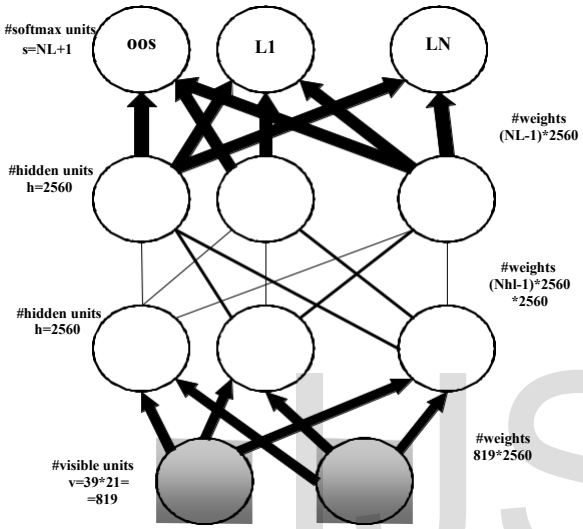
In[7], Deep stacking system (DSN) is indicated by the concept that complex function can be communicated by stacking various layers of shallow premise modules. It is known as deep convex network. Every premise module is comprised of an information layer, a hidden layer and an output layer. Both information layer and output layer are direct, and no mapping capacity is used in these two layers. Sigmoid capacity fills in as a basic nonlinear capacity in the concealed layer. To interface every premise module to develop a deep structure, output of a premise module, that is the anticipated labels by this module, and the first info information are consolidated as contribution to the following associating module.

2.2 RELATED WORK

Nowadays, we are at the dawn of Deep Learning (DL) because in a short time it has dramatically improved the state- of- the-art in many domains including SR. This approach allows us to use complex multi-layer models that learn representation of data with multiple levels of abstraction. The main advantage of DL is the fact that it requires very little engineering by hand, and it can benefit from today's increases of data amounts and computational power. Here, I provide the technical overview of 2 papers. The technical overview covers the better optimization and better types of neural activation function and better network architecture.

1. In [8]the authors have explored different avenues regarding the utilization of deep neural networks (DNNs) to automatic language identification (LID). Guided by the accomplishment of DNNs for acoustic demonstrating, they investigated their capacity to learn in partisan language data from speech signals. There is a correlation between the proposed DNNs

models to a few cutting-edge acoustic frameworks' dependent on I-vectors. Results on NIST LRE 2009 (8 dialects chose) and Google 5M LID datasets (9 dialects + 25 languages), show that DNNs beat, in a large portion of the cases, current condition of-workmanship approaches. This is particularly evident when enormous measure of information is accessible (> 20h), where not at all like I-vectors approaches, which appear to immerse, DNNs still gain from information. Then again, DNNs have a few downsides, including the preparation time, or the quantity of parameters to store as shown in figure 1. Likewise, changing the best possible number of hidden layers and units is an experimental exercise for each database. The authors gained that (for the datasets utilized) moving from 8 to 2 hidden layers, did not dramatically affect execution. Additionally, those changes should be possible disconnected, with testing time still sensible.



Fig, 1 DNN Network Topology

2. In [9], the authors have explored different avenues regarding two deep learning structures Convolutional Neural Networks (CNNs) which demonstrates astounding recognition execution for PC vision errands and Recurrent Neural Networks (RNNs) show impressive achievement in numerous successive data processing assignments as shown in figure 2. The examination is done over the consequence of the Speech Emotion Recognition (SER) calculation dependent on CNNs and RNNs prepared utilizing emotional speech database. The authors have proposed the change of speech signal to 2D portrayal utilizing STFT subsequent to pre-processing and 2D portrayal is broke down through CNNs and LSTM architectures.

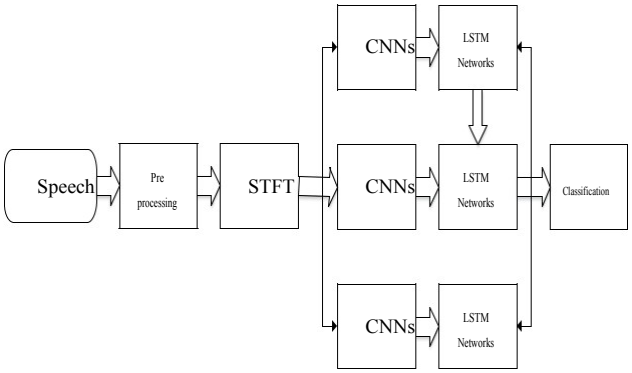


Fig. 2 Proposed block diagram for time distributed networks bases SER method

One of the fundamental points of their examination is to consolidate profound various levelled CNNs highlight extraction design with a model that can figure out how to perceive and orchestrate successive elements in speech signal as shown in figure 3.

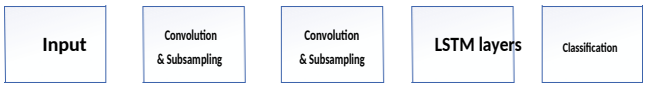


Fig. 3 The Proposed Time Distributed CNNs structure for emotion recognition in speech

The key thought whether CNN is to exploit the properties of sign: pooling, use of different layers, neighbourhood network and weight sharing as shown in figure 4.

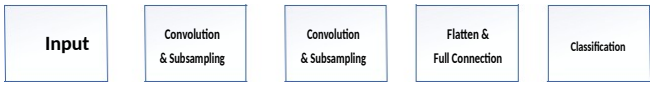


Fig. 4 The Proposed CNN structure for emotion recognition in speech

A few tests were performed for the fundamental CNN's and LSTM. The outcomes were better for CNN's - based time appropriated systems as shown in figure 5.

Fig. 5 The Proposed LSTM Structure for speech emotion recognition

2.3 COMPARISON TABLE

Authors (Year)	DL Architecture	Strengths	Limitations
R. Narasimhan, Xiaoli Z Fem, R Raich (2017) [10]	CNN	• New approaches for segmentation of bird recording to identify bioacoustics for birds.	• Did not take the temporal relation between bird syllables into account.
Zhaung, J. Tang, S. Xue, L. Dai (2016) [11]	RNN and BLSTM	• Can achieve over 10% of relative reduction in phone error rate.	• Not feasible with adaptation.
O. Pichot, L. Berget, H. Aronowitz (2016) [12]	DNN	• 50% of improvements for text dependent system and 48% for text independent system.	• Can't handle the distortion very well.
Liu, Shuchang, Li Guo and Geraint A. wiggins (2018) [13]	CNN	• Improvement of state of art approaches by applying several data set conditions for polyphonic transcription.	• Involves lots of steps.
Hennequin, Romain, Jimena Royo-letelier and M. Moussallam (2017) [14]	CNN	• Good performance on large scale database and robustness to codec type and re-sampling.	• Small scale databases are not used
Vijay Badrinarayanan ; Alex Kendall ; Roberto Cipolla (2018) [15]	CNN	• SegNet provides good performance with competitive inference time • significantly smaller in the number of trainable parameters than other competing architectures	• somewhat difficult to apply
Lam, M. W., Chen, X., Hu, S., Yu, J., Liu, X., & Meng, H. (2019) [16]	LSTM	• allows the optimal forms of gates being automatically learned for individual LSTM cells.	• Doesn't hold good for short-term information.
Qing Wang ; Jun Du ; Li Chai ; Li-Rong Dai ; Chin-Hui Lee [17]	DNN	• Less speech distortion	• Low frequency units are distorted.
Shiqing Zhang , Shiliang Zhang , Tiejun Huang , Wen Gao , Qi Tian (2018) [18]	DBN	• Validity of cross-media fine tuning scheme.	• Not good for audio-visual feature fusion.

III. DISCUSSION

• A new approach is proposed to deal with Automatic Language Identification (LID) in view of Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs). Persuaded by the ongoing achievement of Deep Neural Networks (DNNs) for LID, the authors investigated LSTM RNNs as a characteristic engineering to incorporate temporal contextual data inside a neural system framework. They have contrasted the proposed framework and an I-vector based framework and various designs of feed forward DNNs. Results on NIST LRE 2009 (8 dialects chose and 3s condition) demonstrate that LSTM RNN engineering accomplishes preferred execution over the best 4 layers DNN framework utilizing multiple times less parameters (∼1M versus ∼21M). Moreover, it is seen that LSTM RNN scores as preferable aligned over those created by the I-vector or the DNN frameworks. This work additionally demonstrates that both LSTM RNN and DNN frameworks amazingly outperform the exhibition of the individual I vector framework. Besides, both neural system methodologies can be joined prompting an improvement of >25% as far as Cavg concerning the best individual LSTM RNN framework. The best joined framework likewise consolidates the scores from the I-vector framework prompting an absolute improvement of 28%. However, LSTM RNNs can adequately exploit in temporal dependencies acoustic information, learning important features for language discrimination purposes. The work would have good better results if instead of contrasting with the feed forward DNN framework, the authors had distinguished with feedback DNN framework. The feedback DNN would have measured the output of the process, calculated the error and then adjusted one or more inputs to get the desired and more effective output value or resulting in more improvement i.e., greater than 28%.

• Most significant method for correspondence among people is language and essential medium utilized for the said is speech. The speech recognizers utilize a parametric type of a sign to acquire the most significant discernable features of speech signal for acknowledgment reason, In one of the works, Linear Prediction Cepstral Coefficient (LPCC), Mel Frequency Cepstral Coefficient (MFCC) and Bark recurrence Cepstral coefficient (BFCC) feature extraction methods for acknowledgment of Hindi Isolated, Paired and Hybrid words have been examined and the relating acknowledgment rates are thought about. Artificial Neural Network is utilized as back end processor. The exploratory outcomes demonstrate that the better acknowledgment rate is acquired for MFCC when contrasted with LPCC and BFCC for all the three sorts of words (Isolated, Paired and Hybrid). However, if the fusion of LPCC and BFCC is implemented the recognition rate can be improved up to mark of MFCC. These feature extraction techniques if used for other language databases like ENGLISH, CHINESE, SWEDISH etc., the recognition rate can be enhanced as the said languages have enough databases from where number of features can be extracted. The other feature extraction techniques like HOG, SURF and LBP can be implemented for the better results in place of LPCC and BFCC.

• The language model in conventional frameworks is a huge, smoothed n-gram model. Utilizing a Markov language model makes it a lot simpler to perform deciphering and empowers productive word cross sections. At the point when progressively complex language models are utilized, they for the most part pursue just a first deciphering pass is complete.

3.1 APPLICATIONS OF ARCHITECTURES OF DEEP LEARNING

3.1.1 Speech emotion recognition

Speech emotion recognition (SER) is by and by a standout amongst the most inclining subjects in the world. Numerous frameworks are being created to perceive various emotion from human speech. The advancement of SER appeared in 1920 when a celluloid toy, 'Radio Rex' was made. This toy used to take a shot at acoustic vitality discharged by the vowel 'Rex' for example 500Hz. 'Davis' in 1952 built up the principal speech feeling acknowledgment framework in Bell Laboratory, which used to perceive digits from 0-9 in male voice.

In paper[19] , an answer for the issue of 'context - aware ' emotional relevant feature extraction is proposed, by consolidating Convolutional Neural Networks (CNNs) with LSTM systems. This is known as start to finish SER. In this paper, it is likewise demonstrated that the proposed tropology fundamentally outflanks the customary methodologies dependent on signal processing strategies for the forecast of unconstrained and common emotions on the RECOLA database. The last commitment of this paper is that they have considered the door enactments of the repetitive layers and discover cells that are profoundly connected with prosodic features that were constantly accepted to cause excitement.

3.1.2 Audio Compression

The ordinary uses of data compression are omnipresent: spilling live recordings and music progressively over the planet, putting away many pictures and tunes on a solitary minor thumb drive and some more. Basically, all advanced pressure guidelines are hand-planned, including the most noticeable wideband speech coder: AMR-WB. It was made by eight speech coding scientists working at the Voice Age Corporation in Montreal.

In paper [20], a deep neural system model is exhibited which improves every one of the means of a wideband discourse coding pipeline (pressure, quantization, entropy coding and decompression) start to finish legitimately from crude discourse information, likewise no manual component building is fundamental and it prepares in hours. In this proposed framework, DNN-put together coder performs with respect to standard with the AMR-WB standard at an assortment of bitrates (9kbps up to 24kbps).

3.1.3 Speech Recognition

Speech recognition is the interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields. Some speech recognition systems require "training" (also called "enrollment") where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent". Sequence-to-sequence models only use shallow acoustic encoder networks. In paper [21], deep convolutional networks are trained to add more expressive power and better generalization for end-to-end ASR models. To build very deep recurrent and convolutional structures, network-in-network principles, batch normalization, residual connections and convolutional LSTMs are applied. Experiments have been performed with the WSJ ASR task and has achieved 10.5%-word error rate without any dictionary or language model using a 15-layer deep network.

3.1.4 Language Identification

Language identification (LI) is the issue of deciding the regular language that a record or part thereof is written in. Automatic LI has been widely explored for more than fifty years. Today, LI is a key piece of numerous content handling pipelines, as content preparing methods for the most part accept that the language of the info content is known. Research around there has as of late been particularly dynamic. LID is utilized in a few applications, for example, multilingual interpretation frameworks or crisis call directing, where the reaction time of a familiar local administrator may be basic. In paper [21], deep neural networks (DNNs) is utilized to address automatic language identification (LID). DNNs are adjusted to the issue of recognizing the language of a given expressed articulation from brief time acoustic highlights. This proposed methodology is additionally contrasted with best in class I-vector put together acoustic frameworks with respect to two diverse datasets: Google 5M LID corpus and NIST LRE 2009 respectivily. Exploratory outcomes have demonstrated that how LID can generally profit utilizing DNNs, particularly when enormous preparing information is accessible. Generally, 70% of upgrades were found in Cavg, over the gauge framework.

IV. CONCLUSION

The review deals with the critical analysis of two papers where deep learning architectures like DNN, RNN, CNN and LSTM have been applied to the areas such as speech emotion recognition and language identification. The architectures have generated results in a way, superior to Humans. The accuracy is the main point what separates deep learning from the other technologies. The table consists of different deep learning architectures which are applied in various applications are presented. The various application of audio and speech processing like speech emotion recognition, audio compression, language identification and speech recognition have been discussed. The work must be done on the training processing of deep learning architectures as they require unlabeled training data.

RECOMMENDATION

It is surprising to see that most of the researchers still use MFCCs as feature extraction for speech and audio signals in deep learning models. MFCCs were heavily used in classical classifiers such as HMM and GMM. It is worth trying when using deep learning models other feature extraction methods such as Linear Predictive Coding (LPC). Another observation is that there is little work on speech processing using Recurrent Neural Networks (RNN). Authors are highly recommended to conduct research using deep RNN in the future since RNN models, especially Long Short Time Memory (LSTM), are very powerful in speech processing.

REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, and T. N. Sainath, "Deep Learning for Audio Signal Processing," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[2] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

[3] E. Balouji, I. Y. Gu, M. H. Bollen, A. Bagheri, and M. Nazari, "A LSTM-based deep learning method with application to voltage dip classification," in *2018 18th International Conference on Harmonics and Quality of Power (ICHQP)*, 2018: IEEE, pp. 1-5.

[4] D. Guiming, W. Xia, W. Guangyan, Z. Yan, and L. Dan, "Speech recognition based on convolutional neural networks," in *2016 IEEE International Conference on Signal and Image Processing (ICSIP)*, 2016: IEEE, pp. 708-711.

[5] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533-1545, 2014.

[6] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected topics in applied earth observations and remote sensing*, vol. 7, no. 6, pp. 2094-2107, 2014.

[7] C. Sun, M. Ma, Z. Zhao, and X. Chen, "Sparse deep stacking network for fault diagnosis of motor," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3261-3270, 2018.

[8] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] T. Gulzar, A. Singh, and S. Sharma, "Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks," *International Journal of Computer Applications*, vol. 101, no. 12, pp. 22-27, 2014.

[10] R. Narasimhan, X. Z. Fern, and R. Raich, "Simultaneous segmentation and classification of bird song using CNN," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: IEEE, pp. 146-150.

[11] Z. Huang, J. Tang, S. Xue, and L. Dai, "Speaker adaptation of RNN-BLSTM for speech recognition based on speaker code," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: IEEE, pp. 5305-5309.

[12] O. Plchot, L. Burget, H. Aronowitz, and P. Matejka, "Audio enhancing with DNN autoencoder for speaker recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: IEEE, pp. 5090-5094.

[13] R. Kelz, S. Böck, and G. Widmer, "Deep polyphonic adsr piano note transcription," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: IEEE, pp. 246-250.

[14] R. Hennequin, J. Royo-Letelier, and M. Moussallam, "Codec independent lossy audio compression detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: IEEE, pp. 726-730.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481-2495, 2017.

[16] M. W. Lam, X. Chen, S. Hu, J. Yu, X. Liu, and H. Meng, "Gaussian process lstm recurrent neural network language models for speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: IEEE, pp. 7235-7239.

[17] Q. Wang, J. Du, L. Chai, L.-R. Dai, and C.-H. Lee, "A Maximum Likelihood Approach to Masking-based Speech Enhancement Using Deep Neural Network," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2018: IEEE, pp. 295-299.

[18] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030-3043, 2018.

[19] G. Trigeorgis *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2016: IEEE, pp. 5200-5204.

[20] S. Kankanahalli, "End-to-end optimized speech coding with deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018: IEEE, pp. 2521-2525.

[21] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: IEEE, pp. 4845-4849.