Attention-Inspired Artificial Neural Networks for Speech Processing: A Systematic Review

Noel Zacarias-Morales 1, Pablo Pancardo 1,*, José Adán Hernández-Nolasco 1 and Matias Garcia-Constantino 2

Abstract: Artificial Neural Networks (ANNs) were created inspired by the neural networks in the human brain and have been widely applied in speech processing. The application areas of ANN include: Speech recognition, speech emotion recognition, language identification, speech enhancement, and speech separation, amongst others. Likewise, given that speech processing performed by humans involves complex cognitive processes known as auditory attention, there has been a growing amount of papers proposing ANNs supported by deep learning algorithms in conjunction with some mechanism to achieve symmetry with the human attention process. However, while these ANN approaches include attention, there is no categorization of attention integrated into the deep learning algorithms and their relation with human auditory attention. Therefore, we consider it necessary to have a review of the different ANN approaches inspired in attention to show both academic and industry experts the available models for a wide variety of applications. Based on the PRISMA methodology, we present a systematic review of the literature published since 2000, in which deep learning algorithms are applied to diverse problems related to speech processing. In this paper 133 research works are selected and the following aspects are described: (i) Most relevant features, (ii) ways in which attention has been implemented, (iii) their hypothetical relationship with human attention, and (iv) the evaluation metrics used. Additionally, the four publications most related with human attention were analyzed and their strengths and weaknesses were determined.

Keywords: artificial neural networks; deep learning; attention; speech; systematic review

1. Introduction

The analysis and processing of signals generated by the human speech consists in identifying and quantifying some physical features from the signals in such a way that they can be used for different speech related applications like identification, recognition and authentication. In that sense, Artificial Neural Networks (ANNs) have been a valuable computational tool because of their effectiveness in speech processing. Using deep learning algorithms, ANNs try to mimic the behaviour of the human brain to perform the functionalities involved in speech processing and, to improve the results, some algorithms implement some type of attention.

Given the above, it is of interest to know the diverse research works published between 2000 and 2020 that use ANNs and that implement attention for speech processing. While there are some systematic reviews related to speech processing using Artificial Intelligence techniques, to our best knowledge are no systematic reviews focused on attention such as the one presented in this paper.

Therefore, the literature search for this review was conducted on the ACM Digital Library, IEEE Explorer, Science Direct, Springer Link, and Web of Science databases to identify studies in the field of speech processing that reported the use of ANNs with some type of attention included in the title and/or abstract. We present a comprehensive...
and integrative update of the topic based on the main findings of 133 papers published between 2000 and 2020. This review aims to identify and analyze papers about the design and construction of neural networks that implement some speech processing attention mechanism. According to this objective, four research questions are presented:

- **RQ1**: In which way has attention been integrated in deep learning algorithms and its possible relationship with human auditory attention?
- **RQ2**: What are the features of the speech signals used?
- **RQ3**: What are the neural network models used in the research papers?
- **RQ4**: Which methods or metrics were used to evaluate the obtained results?

The main contributions of this systematic review are as follows: (i) to analyze neural network research works that have implemented attention for speech processing, and its hypothetical relation with human attention (cognitive processes), (ii) to identify the speech processing application areas that have been investigated more widely between 2000 and 2020, and (iii) to determine which are the main Artificial Intelligence algorithms that have been applied to speech processing.

This review was constructed following the steps of the PRISMA methodology [1] and it is organised as follows. Section 2 explains the background and related work. Section 3 presents in detail the implementation of the PRISMA methodology for the systematic review process. Section 4 reports the results obtained from the application of the PRISMA methodology and presents the answers to the research questions. Section 5 discusses the obtained results. Finally, conclusions and final remarks are presented in Section 6.

### 2. Background and Related Works

Audio analysis has been widely used to retrieve human speech for the purposes of identification or extraction. This process becomes more complex when there are other sounds included in addition to human speech, for example when there is more than one speech at a time. The audio analysis process becomes even more complex when noise is present. However, the human brain is capable of performing the task successfully, thanks to the attention process. On the other hand, in the area of Computer Science, Artificial Neural Networks that use deep learning algorithms have achieved outstanding results in speech processing.

#### 2.1. Related Works

To date, there are related systematic reviews, overviews, and surveys that collect information from different architectures and deep learning models. These publications are: (i) the publications that gather information from deep learning models with attention mechanisms, and (ii) the publications that collect the information from deep learning models applied to speech signal processing.

In the publications that gather information about deep learning models with attention mechanisms, we can mention the work of Galassi et al. [2]. This work presented a systematic overview to define a unified model for attention architectures in Natural Language Processing (NLP), focusing on those designed to work with vector representations of textual data. The publication provides an extensive categorization of the literature, presents examples of how attention models can utilize prior information, and discuss ongoing research efforts and open challenges. It also demonstrates how attention could be a key element in injecting knowledge into the neural model to represent specific features or to exploit previously acquired knowledge, as in transfer learning settings. This publication restricts their analysis to attentive architectures designed to work just with vector representation of textual data.

Lee et al. [3] conduct a survey on attention models in graphs and introduce three intuitive taxonomies to group the available work based on the problem setting (the type of input and output), the attention mechanism type used, and the task (e.g., graph classification, link prediction). They mention the main advantages of using attention on graphs, like that the attention allows the model: (i) to avoid or ignore noisy parts of the graph, thus
improving the signal-to-noise (SNR) ratio; (ii) to assign a relevance score to elements in the graph to highlight aspects with the most task-relevant information; and (iii) to provide a way to make the results of a model more interpretable. This publication restricts their analysis to examining and categorizing techniques that apply attention only to graphs (the methods that take graphs as input and solve some graph-based problem).

Within the works related to deep learning models applied to speech signal processing, the most recent are Nassif et al. [4], and Zhang et al. [5]. The first is a systematic literature review that identifies and examines the information from 174 articles that implement deep neural networks in speech-related applications like automatic speech recognition, emotional speech recognition, speaker identification, and speech enhancement [4]. Although several areas of application are involved, attention is not an issue.

The second work reviews recently developed and representative deep learning approaches for tackling non-stationary additive and convolutional degradation of speech to provide guidelines for those involved in developing environmentally robust speech recognition systems [5]. The authors focused their review only on models related to speech recognition and applied to noisy environments. Therefore, they do not consider other application areas.

Our systematic review differs from the existing studies because it identifies and analyzes publications about the design and construction of neural networks that implement some attention mechanism for speech processing.

2.2. Attention

According to cognitive psychology and neuroscience, attention can be identified as a cognitive activity that involves identifiable aspects of cognitive behavior [6,7]. In the literature, there are different definitions for the concept of attention, this is because it comprises several psychological and cognitive processes, which causes researchers from several fields to differ when it comes to having a definition that covers the different types of attention.

One of the definitions that possibly best describes attention is that of Richard Shiffrin [8], in which he mentions that attention refers to all those aspects of human cognition that the individual can control and to all those cognition aspects related to resource or ability limitations, including the methods to address such limitations. Thus, it is evident that the term attention is used to refer to different phenomena and processes, and not only among psychologists or neuroscientists but also in the everyday use of this term. Types of attention can be visual, auditory, and of sensory type; including conscious or unconscious attention.

Attention is not a single or unidirectional process, and it can be classified in terms of two different essential functions: (i) Top-Down attention, and (ii) Bottom-Up attention. Top-Down attention is a selective process that focuses cognitive resources on the most relevant sensory information to maintain a behavior directed to one or more objectives in the presence of multiple distractions. Top-Down attention implies the voluntary assignment of cognitive resources to an objective, while the other sensory stimuli are suppressed or ignored; this is why Top-Down attention is a process guided by objectives or expectations. Bottom-Up attention is a process triggered by unexpected or outstanding sensory stimuli, i.e., it refers to the orientation process of the attention guided purely by stimuli that are outstanding due to their inherent properties concerning the environment [9].

In the acoustic analysis, auditory attention is responsible for mediating perception and behavior, focusing sensory and cognitive resources on relevant information in the space of stimuli. Auditory attention is a selection process or processes that focuses the sensory and cognitive resources on the most relevant events in the soundscape. Stimulus-driven factors can modulate auditory attention in a Top-Down and Bottom-Up manner. Auditory attention samples sensory input and directs sensory and cognitive resources to the most relevant events in the soundscape [10].
2.3. Deep Learning and Neural Networks

Deep Learning is a subfield of Machine Learning that focuses on Artificial Neural Networks (ANNs) and the related algorithms to perform these networks' training. A deep learning model has at least two hidden layers of neurons (models that involve at least ten hidden layers are called Very Deep Neural Networks).

2.3.1. Artificial Neural Networks

Artificial Neural Networks (ANNs) are inspired by the functioning of neurons in the human brain. Inside the human brain each neuron receives stimuli and decides to activate itself or not. An activated neuron will send an electrical signal to other connected neurons, and then, if an extensive network of interconnected neurons is available, it is possible to learn to react to different inputs by adjusting the way they are connected and how sensitive they are to the stimuli [11].

While Artificial Neural Network models maintain the same principle of functioning of the human brain, they focus more on solving problems using data. A key component of a neural network is the neuron (also called a node). A node consists of one or more inputs \(X_i\), its weights \(W_l\), an input function \(Z_l\), an activation function \(A_l\), and an output \(Y\).

The input function takes the weighted sum of all the inputs, and the activation function uses the result to determine whether the node should be activated or not. The weights are adjusted during the learning process to amplify or reduce them according to the input data [11].

As a basis, the simplest structure is a single-layer neural network, and its main feature is that neurons belonging to the same layer cannot communicate. Next in complexity is the multi-layer neural network, where the first layer is called input layer, the last layer is called output layer, and the intermediate layers are called hidden layers.

The design and creation of deep neural networks involve the use of hyperparameters, which are parameters whose values are set and initialized prior to the training process of artificial neural network models, such as the number of layers in the neural network or the number of neurons in each layer. Some of the hyperparameters in deep neural network models are the following:

- Number of hidden layers
- Number of neurons in each layer
- Initialization weights
- The activation function
- The cost function
- An optimizer
- A learning rate

Deep learning comprises several types of artificial neural network architectures, including convolutional, recurrent, short-term and long-term memory, among others.

Convolutional Neural Networks (CNNs) is one of the most extensively used approaches for object recognition because their design is based on the visual cortex of animals. In convolutional neural networks, hidden layers of neurons are connected only to the previous layer containing the subset of neurons; this type of connectivity gives systems the ability to learn from the features implicitly [12].

Recurrent Neural Networks (RNNs) are ideal for processing tasks involving sequential inputs, such as Natural Language Processing (NLP) tasks (text and speech). In recurrent neural networks, the convolution layer is the most basic, but at the same time the most important layer; it convolves or multiplies a pixel array generated for the given image or object to produce an activation map for the given image [13]. The main advantage of the activation map is that it stores all the distinctive features of a given image and at the same time reduces the amount of data to be processed; unfortunately, there is also a problem in this neural network architecture: the storage of past information for a long time, i.e., long-term dependencies.
Long Short-Term Memory (LSTM) Neural Networks are a particular type of recurrent neural network that emerged to overcome the problem of recurrent neural networks with explicit memory since it uses special hidden nodes or units to remember the parameters in input form for a long time. In the literature, it is also possible to find a particular type of neural network called Bidirectional Long Short-Term Memory (Bi-LSTM) Neural Network, which consists of two regular long-short term memory networks: one with a forward direction and the other in the opposite direction.

In the current research in the literature it is common to find more complex neural networks; these make use of combinations of various neural network architectures, as some combinations are suitable to solve specific problems; the resulting architecture of the combinations is often called Deep Reinforcement Learning (DRL) [14].

2.3.2. Attention Mechanism in Neural Networks

Methods inspired by nature have been widely explored as efficient tools for solving real-world problems. In this sense, human attention mechanism could be ideally implemented through algorithms built from the synthesis of biological processes as a goal to reach a symmetry between attention inspired ANN and human auditory attention.

By the way, the attention mechanisms used in deep learning originated as an improvement to the encoder-decoder architecture used in natural language processing. Later, this mechanism and its variants were applied to other areas such as computer vision and speech processing. Before the attention mechanisms, the encoder-decoder architecture was based on stacked units of artificial neural networks of recurrent type and Long Short-Term Memory (LSTM).

The encoder (LSTM type neural network) is in charge of processing the input data and encoding them into a context vector (the last hidden state of the LSTM). It is expected this vector be a collection or summary of the input data since this vector is the initial hidden state of the decoder (intermediate encoder states are discarded); in other words, the encoder reads the input data and tries to make sense of it before summarizing them. The decoder (comprised of recurring units or LSTM) takes the context vector and produces the output data in sequential order.

As part of neural network architecture, attention mechanisms dynamically highlight the relevant features of the input data. The central idea behind the attention mechanism is not to discard the intermediate states of the encoder but to use them to build the context vectors required by the decoder to generate the output data, calculating a distribution of weights in the input sequence, and assigning higher values to the most relevant elements, and lower weights to the less relevant elements [2].

2.4. Speech

As human physiology allows for life in an air-based atmosphere, it was inevitable that humans would develop a form of communication-based on acoustic signals that support the movement of molecules in the air [15]. For humans, communication through speech implies:

- The physiological properties of sound generation in the vocal system.
- The mechanisms for processing speech in the auditory system.
- The configurations imposed by the various languages.

In today’s era, speech communication is no longer a process exclusive to humans. Advances in computerized speech processing allow for the continued development of technologies that attempt to improve the communication between humans and computer systems with ever-increasing performance. The challenges for speech processing in which the scientific community focuses its most significant dedication are: (i) speech recognition, (ii) language identification, (iii) emotion recognition, and (iv) speech enhancement.

Typically, these areas are studied separately; that is, researchers usually work on these specific areas to improve the performance of systems concerning systems that integrate the current state of the art, but in reality, the problem they face is the same: finding a way to
extract, represent and process the information contained in speech signals. Table 1 lists the objectives of the speech processing areas most studied by the scientific community.

Table 1. Objectives of the speech processing areas.

| Speech Processing Area        | Objective                                                                 |
|------------------------------|---------------------------------------------------------------------------|
| Speech Recognition            | Determine the content of the speech signals.                              |
| Speech Emotion Recognition    | Know the emotional state of a person.                                     |
| Language Identification       | Identify the language or dialect of a speech signal.                     |
| Speech Enhancement            | Remove background noise from the degraded speech without distorting the clean speech, thereby improving the speech quality and intelligibility. |
| Speaker Recognition           | Recognize the identity of a person from a speech signal.                 |
| Disease Detection             | Detect a specific disease from a speech signal.                           |

3. Methodology

We planned and conducted this study based on the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement [1] (we adapted the items in the checklist to research in Computer Science, which differs from medical research). It is important to note that the PRISMA statement involves systematic reviews and meta-analysis. This study only does a systematic review to provide a compilation of what is available in the literature. Before performing the systematic review, we conducted a pilot test with ten randomized publications to standardize the process and resolve doubts. We discussed and resolved the differences that arose.

3.1. Protocol And Registration

The objectives, methods, strategies and analysis applied in this systematic review were carried out according to the specifications of the systematic review protocol entitled: “Attention-Inspired Artificial Neural Networks for Speech Processing: Systematic Review Protocol” as established in PRISMA-P [16]. This protocol was written, validated and approved by all authors before the systematic review.

3.2. Eligibility Criteria

The inclusion and exclusion criteria used in this systematic review are as follows.

Inclusion criteria:
- Publications made between the years 2000 and 2020.
- Publications in English.
- Publications proposing models based on artificial neural networks.
- Publications using an attention-based approach.
- Publications that consider speech applications.

We selected the time range from 2000 to 2020 to have a historical context of the last two decades to cover all those papers that implement attention.

Exclusion criteria:
- Publications that use neural network models, but do not apply them to speech.
- Publications applied to speech, but not using neural network models.
- Publications that do not use attention-based approaches.
- Publications without evaluation methods or metrics.
- Publications without clear information about their origin (authors’ affiliation and name of the journal or conference where it was published).

3.3. Information Sources

In this systematic review, the following digital libraries were used to search for publications:
- ACM Digital Library
3.4. Search

The search strategy implemented in this systematic review consisted of two different steps: (i) the definition of the terms or keywords, and (ii) the definition of the search strings for each digital library.

First, we identified seven terms: comput*, model, neural network, speech, audi*, selecti* and attention; and 14 related words (words that share the same grammatical base, or synonyms): computer, computational, model, modeling, NN, deep learning, voice, speaker, audio, auditory, selective, selection, attention-based, and attention mechanism. After trying different structures, search strings for each digital library were generated, as shown in Table 2.

Table 2. Search strings.

| Digital Library   | Search String                                                                                                                                 |
|-------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| ACM               | Search items from: The ACM Guide to Computing Literature  
Title: attention OR speech  
Abstract: model AND attention AND 
(“neural network” OR “deep learning”) AND (speech OR voice)  
Publication Date: January 2000–October 2020 |
| IEEE Explorer     | Abstract: model AND attention AND (“neural network” OR “deep learning”) AND (speech OR voice)  
Filters Applied: 2000–2020 |
| Science Direct    | Find articles with these terms: model AND attention AND (“neural network” OR “deep learning”) AND (speech OR voice)  
Year(s): 2000–2020  
Title, abstract or author-specified keywords: model AND attention AND speech |
| Springer Link     | With all of the words: Model AND attention AND neural network AND speech  
With the exact phrase: neural network  
With at least one of the words: attention speech  
Where the title contains: attention  
Start year: 2000  
End year: 2020 |
| Web of Science    | AB = (model * AND attention AND (“neural network” OR “deep learning”) AND (speech OR voice))  
Year(s): 2000–2020 |

Some of the digital libraries allow using the asterisk (*) as a wildcard to search for words that have spelling variations or contain a specified pattern of characters. We used the asterisk (*) to find terms with the same beginning but different endings.

3.5. Study Selection

The search in the digital libraries generated a list of 902 publications. Subsequently, we carried out a filtering process to include only relevant publications in this systematic review. This process was carried out through scheduled meetings between the authors. The steps of the filtering process were as follows:

1. Remove all duplicate publications.
2. Review the title and abstract of each publication to apply the inclusion/exclusion criteria (when the information in the title and abstract was not sufficient to apply the inclusion/exclusion criteria, the full text of the publication was retrieved and reviewed).
3. Apply the quality assessment to identify publications that answered the research questions.

3.6. Data Collection Process

For the data extraction process, the researchers jointly developed a form to gather all the necessary information to answer the research questions. The form was applied separately by two of the authors, and it was reviewed by a third author. The differences of opinion that arose were discussed and resolved. It is important to mention that some publications included in the systematic review did not contain the necessary information to answer each of the items included in the form.

3.7. Data Items

The form used for data extraction contains a total of 21 items. The extracted data were divided into four general groups: (i) data on the source of the publication, (ii) data from the speech signal used, (iii) data from the deep learning models used, and (iv) details on the implementation of attention.

The individual items extracted were: digital library, type of publication, name of journal or conference, application area, publication date, publication title, names of authors, data source, features of the data used in the training, context of the original data, context of the data in the tests, language of the data, generation of the data, features extracted from the data, types of neural network used, other models used, details of the proposed model, evaluation metrics, method or process of implementing the correspondence between the model and the attention, contribution of the publication to science, and future work.

3.8. Risk of Bias in Individual Studies

In this systematic review it was considered critical to evaluate the quality of the publications to identify those that best answered the research questions. For this reason, an assessment of risk of bias (other authors refer to this study as: “quality assessment”) was applied.

For this process, 10 questions were defined to evaluate the publications; each question could obtain one of three possible answers with its respective score according to the following criteria: (i) question thoroughly answered = 1, (ii) question answered in a general way = 0.5, and (iii) question not answered = 0. The answer scores sum ranged from 1 to 10, and we selected only those publications that obtained a sum equal to or greater than 7 for the next stage of the systematic review. This evaluation was carried out by two of the authors separately and reviewed by a third researcher. The questions were:

- Q1: Is the source information clear?
- Q2: Does the publication have the primary sections of a scientific report?
- Q3: Do authors define the problem (or improvement) they address?
- Q4: Does the paper describe what the input (source) data are?
- Q5: Is the deep learning model (method) used clearly described?
- Q6: Do authors use metrics to evaluate the results?
- Q7: Is there mapping (correspondence) between the computational and biological/cognitive areas?
- Q8: Does the publication mention how attention is applied?
- Q9: Do the authors present the results in a clear way?
- Q10: In the discussion, are findings, implications, and relationship of results to other similar works considered?

The evaluation was developed based on the criteria used by the Center for Reviews and Dissemination from the University of York, published in [17].
3.9. Summary Measures

In this systematic review, we distinguished between two outcomes of interest, those considered primary (also known as primary outcomes), and those considered additional (known as secondary outcomes).

- **Primary outcome:** It identifies how researchers have implemented attention in neural network algorithms and the supposed correspondence between the proposal and human attention.
- **Secondary outcome:** It identifies the specific features extracted from the audio signals and how authors implemented them in the neural network models. Additionally, to know the areas of opportunity for future research.

4. Results

In this section are described the results obtained and the answers to the research questions of this systematic review.

4.1. Study Selection

The PRISMA-based flowchart in Figure 1 details how the review process was performed and the number of publications filtered at each stage for the final selection to be included.

![Flowchart](image)

**Figure 1.** Flowchart of the included eligible studies in the systematic review.

4.2. Study Characteristics

Appendix A lists the publications and includes the most important data related to the research questions, which are also considered significant for this systematic review.
4.3. Risk of Bias within Studies

Appendix B contains the results of the risk assessment for bias (quality assessment) for the publications.

4.4. Results of Individual Studies

Once the information from the 133 publications selected during the systematic review was organised, different research areas were identified (as shown in Table 3) and graphically illustrated (as presented in Figure 2). The 32.3% of the publications are journal papers, and the 67.7% are conference papers. The International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in its 2018, 2019, and 2020 editions were the conferences with the highest number of selected publications (36 out of 90 conference publications). Additionally, it was detected that 35.3% of the total number of publications did not include possible future work as a continuation to their research.

Table 3. Application areas identified in the publications.

| Application Area            | Number of Publications |
|------------------------------|------------------------|
| Speech Recognition           | 47                     |
| Speech Emotion Recognition   | 26                     |
| Language Identification      | 11                     |
| Speech Enhancement           | 8                      |
| Speech Separation            | 5                      |
| Speaker Recognition          | 4                      |
| Speaker Verification         | 4                      |
| Voice Conversion             | 4                      |
| Disease Detection            | 4                      |
| Voice Activity Detection     | 3                      |
| Others                       | 17                     |

Figure 2. Distribution of the identified application areas.

Speech recognition and emotion recognition are the areas where more than half of the publications are concentrated. The “disease detection” area included publications regarding depression severity detection, dysarthria, mood disorders, and SARS-CoV-2.

In the area of “Others”, there are applications with only one publication such as: adversarial examples generation, classification of phonation modes, classification of speech
utterances, cognitive load classification, detection of attacks, lyrics transcription, speaker adaptation, speech classification tasks, speech conflict estimation, speech dialect identification, speech disfluency detection, speech intelligibility estimation, speech pronunciation error detection, speech quality estimation, speech word rejection, speech-to-text translation, and word vectors generation.

Figure 3 shows the distribution of publications from 2000 to 2020. The oldest publications identified were published in 2000 and 2002 (one publication in each year). From 2003 to 2015, there were no publications identified that complied with all the requirements for inclusion. In 2016, the number of publications that met all the requirements increased substantially, being 2019 the year with the highest number of publications. Note that the number of publications in 2019 is higher than in 2020, which can be attributed to the fact that our search started in October 2020.

![Figure 3. Distribution of publications between 2000 and 2020.](image)

### 4.4.1. Answer to RQ1

After applying the inclusion/exclusion criteria and the risk assessment for bias, 133 publications were identified. Of these, 64.66% only introduce a mechanism of attention as an additional component within their neural network model. The proposed models used this mechanism to improve their performance since as mentioned by [18,19], it was found that the fusion of the neural network models and the mechanism of attention can help the models to learn where to “search” for the most significant information for the task. Thus, focusing on the relevant parts without considering the less relevant data (other terms that the authors refer to the attention mechanism are: module, layer, model, or block).

A 30.08% of the publications mention the use of an attention mechanism, but with more details or variations of this mechanism, as is the case of Bayesian attention layer [20], Multi-head Self-attention mechanism [21], or Monotonic attention mechanism [22]. In another 2.26% of the publications, it was found the application of the concept of attention in a different way than the publications that introduce a mechanism of attention. For example: in [23] they use an environment classification network as attention switch; in [24] they combine the benefits of several approaches using a language model based on attention, and in [25] they propose a selective attention strategy for the acceleration of learning in multi-layer perceptual neural networks.

The remaining 3% are publications that propose models based on neural networks with different approaches and degrees of correspondence to human attention. Specifically, Ref. [26] proposes an auditory attention model with two modules for the segregation and localization of the sound source. On the other hand, Ref. [27] proposes a selective attention algorithm based on Broadbent’s “early filtering” theory; Ref. [28] proposes a Top-Down auditory attention model. Finally, Ref. [29] improve the performance of its neural network
model for emotion recognition based on the mechanism of auditory signal processing and human attention.

4.4.2. Answer to RQ2

Training and testing of models based on artificial neural networks require sufficient and diverse data. In general, the most used datasets within the publications included in this systematic review are: (i) the Wall Street Journal corpus, (ii) the LibriSpeech corpus, and (iii) the TIMIT corpus; with presence in 11.3%, 10.5%, and 7.5% of the publications, respectively.

Regarding the features extracted from the audio files of the different datasets, the most used features are: (i) the Mel Frequency Cepstral Coefficients (MFCC), used in 25% of the publications; (ii) the Log-Mel filterbank, used in 16% of the publications; and (iii) the spectrograms, used in 13% of the publications. The sampling rate used in the audio files during the training was 16 kHz in 25.6% of the publications; 8 kHz in 4.5% and other sampling rates or multiple sampling rates in 4.5%. The most frequent languages used in the datasets are English, Mandarin, and Japanese; only 59.4% of the publications provide information about the language of the data used.

In terms of information that the authors did not find in all the publications reviewed, note the following with respect to features extracted, sampling rate and gender of the speech: (i) in 6.8% of the publications it was not found which were the features extracted from the data, (ii) in 65.4% of the publications there was no mention about the sampling rate used in models, and (iii) only 28.6% of the publications mention information about the gender of the speech in the datasets.

4.4.3. Answer to RQ3

Despite the different types of existing neural networks and the significant number of variations and combinations implemented in the publications, it was possible to identify the most used types of neural networks: (i) the neural network Bi-LSTM, (ii) the neural network LSTM, and (iii) the neural network CNN; used in 33.8%, 30.1%, and 25.6% of the publications, respectively.

The publications can use a single neural network or a combination of more than one model or neural network type. It was identified that 49.6% of the publications required only one type of neural network, 36.8% used at least two types, 9.8% used at least three types, and 3.8% used at least four types of neural network. Their combination is done by including layers of different types of neural networks or independent modules of a specific type of neural network that later are joined to create a more robust model.

Two interesting facts detected are: (i) that 12.8% of the publications do not mention information about the values of the hyper-parameters used in their neural network models, and (ii) that 12% of the publications used other additional models to complement the proposed neural network model, such as Gaussian Mixture Model (GMM), Convex Nonnegative Matrix Factorization (CNMF) and Hidden Markov Model (HMM).

4.4.4. Answer to RQ4

Among the techniques used to evaluate the performance of the diverse and different neural network models proposed in the publications, it was found that the most popular metric used was the Word Error Rate (WER) (used in 28.6% of the publications), followed by the Character Error Rate (CER) (used in 13.5% of the publications) and the Equal Error Rate (EER) (used in 12.8% of the publications). It was also found that 51.9% of the publications apply one metric, 37.6% use two metrics, 9.8% use three metrics, and only 0.8% use five metrics in their publication.

4.5. Synthesis of Results

It was found that 126 of the 133 publications introduce some mechanism, layer, or module of attention, which is added as an additional layer within their neural network model.
Only four publications implemented the combination of diverse techniques or algorithms to elaborate correspondence with human attention.

Regarding the data used in the research, it was found that the Wall Street Journal Corpus was the most used dataset, and MFCCs were the most commonly extracted features of the audio files. From what we observed in the publications, the sampling rates most used by the researchers are 16 kHz and 8 kHz, although more than half of the authors do not mention the sampling rate they used in their research. English, Mandarin, or Japanese are the most frequent languages in the datasets, except for language identification investigations, where the datasets contained data in at least four languages.

Despite the significant number of variations and combinations of the neural network models that implemented diverse attention mechanisms, it was possible to identify that the neural networks of Bi-LSTM type were the ones used, both as independent layers of the models or as independent modules. A point to consider is that we found publications that omitted information about the hyperparameters used, which makes it difficult to replicate the work for future comparisons.

Regarding the diverse metrics used to evaluate the performance of the proposed models, we found that the metrics vary even within each area of research in which the authors work; this makes it difficult to compare between works by having to find and implement some homologation of metrics that reflects the performance of each proposed model.

Table 4 summarizes the three most used datasets, features, models, and metrics by area of research or application.

| Application Area          | Datasets                              | Features                          | Models       | Metric                          |
|---------------------------|---------------------------------------|-----------------------------------|--------------|---------------------------------|
| **Speech Recognition**    | 1. WSJ dataset                        | 1. Log-Mel filterbank             | 1. Bi-LSTM   | 1. Word Error Rate              |
|                           | 2. LibriSpeech dataset                | 2. Mel-scale filterbank           | 2. LSTM      | 2. Character Error Rate         |
|                           | 3. CSJ corpus                         | 3. Pitch                          | 3. CNN       | 3. Phone Error Rate             |
|                           | 1. EMO-DB dataset                    | 1. MFCC                           | 1. CNN       | 1. Unweighted Accuracy          |
| Speech Emotion Recognition| 2. SAVEE dataset                      | 2. Spectrogram                    | 2. Bi-LSTM   | 2. Weighted Accuracy            |
|                           | 3. CASIA dataset                     | 3. Zero-Crossing Rate             | 3. DNN       | 3. Unweighted Average Recall    |
|                           | 1. API17-OLR database                 | 1. MFCC                           | 1. DNN       | 1. Equal Error Rate             |
| Language Identification   | 2. NIST LRE dataset                   | 2. Bottleneck features            | 2. Bi-LSTM   | 2. Average Detection Cost       |
|                           | 3. AP18-OLR database                  | 3. l-vector                       | 3. ResNet    | 3. Accuracy                     |
|                           | 1. Noises92 dataset                   | 1. Spectrogram                    | 1. CNN       | 1. Perceptual Evaluation of Speech Quality |
| Speech Enhancement        | 2. TIMT dataset                       | 2. MFCC                           | 2. DNN       | 2. Short-term Objective Intelligibility |
|                           | 3. CHiME dataset                     | 3. AMS                            | 3. LSTM      | 3. Log-Spectral Distance        |
|                           | 1. WSJ dataset                        | 1. Spectrogram                    | 1. Bi-LSTM   | 1. Signal to Distortion Ratio   |
| Speech Separation         | 2. AIR database                       | 2. AMS                            | 2. LSTM      | 2. Signal to Artifact Ratio     |
|                           | 3. MIR-1K dataset                     | 3. DRR                            | 3. CNN       | 3. Perceptual Evaluation of Speech Quality |
| Speaker Recognition       | 1. VoxCeleb dataset                   | 1. Spectrogram                    | 1. CNN       | 1. Equal Error Rate             |
|                           | 2. AllShell public dataset            | 2. Log-Mel filterbank             | 2. DNN       | 2. Top-1 and Top-5 accuracies   |
|                           | 3. Free S1 Chinese Corpus             | 3. MFCC                           | 3. ResNets   | 3. Word Error Rate              |
| Speaker Verification      | 1. VoxCeleb dataset                   | 1. Energy                         | 1. CNN       |                                 |
|                           | 2. ASVspoof dataset                   | 2. Linear filterbank              | 2. LSTM      | 1. Equal Error Rate             |
|                           | 3. BTA52016 dataset                   | 3. Log-Mel filterbank             | 3. Bi-LSTM   |                                 |
| Voice Conversion          | 1. CMU-ARCTIC dataset                 | 1. Mel-scale spectrograms         | 1. Bi-LSTM   | 1. Naturalness                  |
|                           | 2. VCC2016 dataset                    | 2. Phonetic posteriorgrams        | 2. CNN       | 2. Similarity                   |
|                           | 3. CHI-MEI mood database              | 3. Acoustic/raw spectral features  | 3. LSTM      | 3. Mel-Cepstral Distortion      |
| Disease detection         | 1. COVID19 dataset                    | 1. Fundamental frequency          | 1. LSTM      | 1. Mean Absolute Error          |
|                           | 2. DAKCW-OZ database                  | 2. Harmonic-Noise-Ratio           | 2. Bi-LSTM   | 2. Probability of False Alarm   |
|                           | 1. TIMT dataset                       | 3. Mel-filterbanks                | 3. CNN       | 3. Recall                       |
| Voice Activity Detection  | 2. HAVIC corpus                      | 1. MFCC                           | 1. Bi-LSTM   | 1. Accuracy                     |
|                           | 3. Noises92 dataset                   | 2. Log-Mel filterbank energies    | 2. LSTM      | 2. Area Under the Curve         |
|                           | 1. ASV spoof dataset                  | 3. Multiresolution cochleagram    | 3. FC-NN     | 3. Equal Error Rate             |
| Others                    | 2. BTEC corpus                        | 1. MFCC                           | 1. Bi-LSTM   | 1. Word Error Rate              |
|                           | 3. CCTV news corpus                   | 2. Mel-filterbank                 | 2. LSTM      | 2. Accuracy                     |

The publications that establish a more significant correspondence with human attention are analyzed in Table 5.
Table 5. Analysis of the publications that had correspondence with human attention.

| Item | [26] | [28] | [29] | [27] |
|------|------|------|------|------|
| Application area | Speech Separation | Speech Separation | Speech Emotion Recognition | Speech word rejection |
| Summary | Presents an auditory attention model for locating and extracting a target speech in a multi-source environment. It uses two modules: One module to extract features and segregate the speech, and another module for source location. | It presents a Top-Down auditory attention model to select and separate individual speech from an audio signal. The model consists of two modules: a Bottom-Up inference module, and a Top-Down attention module. First, it generates the spectrogram of the original mix, then it predicts the number of speeches in the mix with the bottom-up inference module, then it uses the Top-Down module to extract one of the speeches, and finally, the resulting spectrogram will replace the original mix. To extract another speech, the process is repeated, until there are no speeches left in the spectrogram. | It is based on the mechanism of processing auditory signals and human attention, and proposes a system of emotion recognition that combines a front-end based on auditory perception and a back-end based on attention. Use the back-end to extract features that include information on variations in intensity, duration, and period of the emotional regions, extracting features with a temporal attention model. | It proposes a selective attention algorithm based on Broadbent's "early filtering" theory, adding an attention layer in front of the input layer (of the multi-layer perception-type neural network) that works as a data filter. |
| Process | It uses two modules: One module for source location. First, it extracts the characteristics, then it separates the speech with a neural network that locates the source using the reverberation times, and finally, it identifies the nearby audio sources. | | | An attention filter layer is added before the input layer. |
| Details of the model | Module one is a 1DRNN. Module two is GMM-EM. Both modules (Bottom-Up inference and Top-Down attention) are Bi-LSTM-type neural networks. | | | The neural network used is a multi-layer perception. |
| Comparisons with human attention performance | (1) They propose a model of auditory attention. (2) The two modules attempt to imitate two of the functions of the human auditory system. (3) They use gamma filters and are proposed as a correspondence to the way the cochlea secretes acoustic signals based on their frequencies (in humans). | (1) They propose a model of auditory attention where they integrate the two modules that were created with correspondence to Top-Down and Bottom-Up attention. | (1) The auditory front-ends are used to functionally simulate the processing of signals in the auditory system from the cochlea to the thalamus. (2) They use the Gamma filterbank to imitate human hearing filters. (3) The back-ends of this system capture the emotional parts of the information of the temporal dynamics in the speech, similar to the human auditory system. | (1) They propose a model of selective attention. (2) They are based on a theory of psychological selective attention. (3) They used ZCPA characteristics motivated by the auditory periphery of mammals. |
| Strengths | (1) The research proposes two modules that attempt to perform two of the functions of the human auditory system: (a) segregate a source in complex environments and locate a source by estimating its distance. (2) By joining these modules, it is possible to reduce errors in selecting the best microphone (binaural scenario) and reduce ambiguities when identifying the desired target. (3) The characteristics and modules are completely described, as well as the results obtained with each module. | (1) The proposal seeks to imitate the human capacity to focus and separate a specific source in a complicated auditory environment. The two modules are: a Bottom-Up inference module that calculates the number of sources in the mix and extracts classification data, and a Bottom-Up attention module that is in charge of separating the signals. (2) The modules are based on the characteristics of human attention. (3) The modules are described in sufficient detail. (4) They mention that the model was based on cognitive science theories. (5) Its proposal can be used in other areas besides the separation of sources. | (1) This proposal is inspired by the human temporal attention mechanism. (2) The choice of features attempts to simulate the way the cochlea breaks down speech signals into acoustic frequency components. (3) The modules, the operating process, and the results are described in detail. | It is the oldest proposal, so it could be considered obsolete compared to the current research because the authors separate words, then it is not functional with phrases. |
| Weaknesses | The proposal imitates two of the abilities of the human auditory system, but not all the abilities of the human auditory system are considered. Its model is weak when there are similar speeches since this confuses the Bottom-Up inference module. | | | |

5. Discussion

As mentioned at the beginning of this document, this systematic review aimed to identify and analyze publications about the design and construction of neural networks that implement some mechanism of attention for speech processing (such as Top-Down and/or Bottom-Up attention) and its possible correspondence with human attention. Attention (from the human point of view) is seen as a process of allocation of cognitive resources, which respond to some priority according to events present in the environment. On the other hand, in deep learning the attention mechanisms in neural network models are designed to assign higher values of “weights” to relevant input information and ignore irrelevant information when the values of the “weights” are lower.

After conducting the systematic review, it was determined that most of the computer models based on the use of artificial neural networks (94.74%), implement only attention
mechanisms as an additional component within the architecture of their neural network models; and only 3% of the publications propose their neural network model with some degree of correspondence with human attention.

The current similarity (regarding attention functioning) between the deep learning models reviewed and the processes studied from the perspective of cognitive psychology are few and vague; which coincides with what is mentioned by [10,30]; the attention “mechanisms” currently used in artificial neural networks are an idea that can be implemented in different ways, more than an implementation of some models of the human attention [31]. This reflects the need to establish interdisciplinary collaborations to better understand the cognitive mechanisms of the human brain, as well as to explore human cognition processing from a computational perspective to develop bio-inspired computational models that have greater adaptive capabilities in uncertain and complex environments, such as acoustic environments.

Based on the evidence collected, it is not possible to establish superiority in terms of efficiency or performance between models of artificial neural networks with built-in attention mechanisms and those that attempt to establish a correspondence to attention, selective attention, or the human auditory attention system. The lack of publications that attempt to establish real correspondences with human auditory attention systems using artificial neural network models also reflects an opportunity for future research in the area of deep learning.

Regarding the features used for speech signals, it was found that 65% of the articles did not offer information about the sampling rate used for the training of the model, which implies that it is not possible to replicate the experiments, which is an essential characteristic in scientific research.

The same happens with the models of neural networks used since in some cases only the hyperparameters used are provided partially. The two situations mentioned above make it impossible to compare the results obtained in the articles analyzed with those obtained in new research.

When analyzing the metrics used in the research works it could be noticed that even in the same area of application, these evaluation methods are heterogeneous and therefore it is difficult to compare efficiencies in the results.

To our best knowledge, no systematic reviews have been conducted focusing on the different attention mechanisms implemented in deep learning algorithms for speech processing and their correspondence with human auditory attention. We only found two reviews related to attention models, the first for text processing [2] and the second for representing data as graphs [3], which confirms our assumption that there are no reviews about the inclusion of attention in deep learning algorithms for speech processing and whether there is a relationship with human auditory attention.

Difficulties in data collection due to missing information or the heterogeneity of the metrics used in the research limited comparisons between the efficiencies of the results when implementing the mechanisms of attention. Complete information would have made it possible to mention the strengths and weaknesses of each article analyzed for the others that address the same area of application.

This systematic review was limited to include proposals inspired by auditory attention, however, it is important to take into account that visual attention is a significant complement to speech processing [30]. Thus, a future systematic review will consider research works with both types of attention to analyze the efficiency of audiovisual models.

6. Conclusions

In this systematic review, we found that ANNs for speech processing have implemented some attention mechanism to improve results. We categorized the application areas, identified the most used datasets for the studies, the most used audio features, the neural network models, and the most-used metrics by the authors. We extracted some additional
data from the publications: sampling rate, language in the dataset, hyperparameters, and number of layers in ANNs.

However, the vast majority of publications that propose models of neural networks with some focus of attention for speech processing, in practice, make little correspondence with human cognitive processes of attention. This situation leads to proposals that are still far from the broad functionality and efficiency achieved by human auditory processing, therefore, the symmetry between human biological attention and attention-inspired ANNs is an utopia yet.

In many research works, the classical attention mechanism is only a part of the proposal and performs a specific function. At the same time, new research works are increasingly complex and require more elements to have better results.

The application areas of speech processing are very diverse. The classification presented in this paper may have a subclassification, and in many cases, authors addressed specific aspects (assigning weights, selecting features) of the application (speech recognition, speech separation).

We conclude that Neural Networks are essential or relevant for speech processing and therefore are the most used. Attention mechanisms have increased in a particular way in the last three years (2018–2020), and we observe an ascending behavior in terms of the number of publications. The recent boom in artificial intelligence, the advances in algorithms, and the new capabilities of hardware make it possible for areas studied for many years to regain relevance. Furthermore, given the new conditions, better results can be obtained.

We visualize a significant increase and greater relevance of computer science research inspired by nature for speech processing. In particular, proposals for neural systems with bio-inspired intelligence approaches for speech, biomedicine, biometrics, signals and images, and other applications [32].

Among the future works of speech processing, we consider that intelligent selective filtering based on previous and real-time generated knowledge will lead to proposals that are more related to how we apply auditory attention; that is, a bio-inspired proposal leads to better results.

**Author Contributions:** Conceptualization, P.P., N.Z.-M. and J.A.H.-N.; methodology, P.P., N.Z.-M.; investigation, N.Z.-M.; writing—original draft preparation, N.Z.-M., P.P., J.A.H.-N. and M.G.-C.; writing—review and editing, N.Z.-M., P.P., J.A.H.-N. and M.G.-C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by CONACYT.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is contained within the article.

**Acknowledgments:** We want to express our gratitude to the Consejo Nacional de Ciencia y Tecnologia (CONACyT) and the Juarez Autonomous University of Tabasco (UJAT) to support us with the necessary academic resources for this research.

**Conflicts of Interest:** The authors declare no conflict of interest.
Appendix A

Table A1. Publications characteristics.

| Pub   | Year | Area                        | Form of Implementation of Attention                                                                 | Neural Network Used | Network Used | Data                                              | Extracted Features                                                                 | Metrics Used                                                                 |
|-------|------|-----------------------------|------------------------------------------------------------------------------------------------------|--------------------|-------------|---------------------------------------------------|----------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| [26]  | 2018 | Speech Separation           | Combination of two modules (binaural source segregation and localization of a target speech signal) to make a auditory attention model | DRNN, LSTM        |             | TIMIT dataset, AIR database, NOIZEUS7 dataset    | MHEC, MFCC, RASTA-MFCC, GFCC, GBFB, PLP, RASTA-PLP, AMS, DRR                     | Source to interference ratio (SIR), Source to artifacts ratio (SAR), Source to distortion ratio (SDR). Unweighted Accuracy (UA), Weighted Accuracy (WA) |
| [33]  | 2018 | Speech Emotion Recognition  | Attention mechanism                                                                                    | Bi-LSTM            |             | IEMOCAP dataset                                   | Mel-Spectrogram                                                               | Word Error Rate (WER)                                                                 |
| [34]  | 2020 | Speech-to-text translation  | Multi-head Self-attention mechanism                                                                   | Encoder-decoder NN |             | BTEC corpus, Google synthesized speech             | Mel-Spectrogram                                                               | Phone Error Rate (PER), Character Error Rate (CER), Word Error Rate (WER) | Signal-to-Distortion Ratio (SDR), Mel-Cepstral Distortion (MCD), Root Mean Square Error (RMSE) |
| [35]  | 2018 | Speech Recognition          | Attention mechanism                                                                                    | DBN, BN-FEN        |             | TIMIT dataset, WSJ dataset                        | Mels filterbank                                                              |                                                                                                                                        |
| [28]  | 2018 | Speech Separation           | Top-Down Auditory Attention model                                                                     | Bi-LSTM            |             | WSJ dataset                                       | Spectrogram                                                                  |                                                                                                                                        |
| [36]  | 2019 | Voice Conversion            | Attention mechanism                                                                                    | Seq2seq ConvErson NeTwork (SCENT), WaveNet |             | CMU ARCTIC dataset                                | Mel-scale spectrograms                                                         |                                                                                                                                        |
| [37]  | 2018 | Speech Recognition          | Attention mechanism                                                                                    | Bi-LSTM            |             | TIMIT dataset, Voxforge dataset                  | MFCC                                                                           |                                                                                                                                        |
Table A1. Cont.

| Pub | Year | Area                          | Form of Implementation of Attention | Neural Network Used | Data                                      | Extracted Features                                      | Metrics Used                                      |
|-----|------|-------------------------------|-------------------------------------|---------------------|------------------------------------------|---------------------------------------------------------|---------------------------------------------------|
| [38] | 2020 | Language Identification       | Attention mechanism                 | Bi-LSTM             | NIST LRE dataset, RATS LID Dataset       | Short-term i-vectors/ x-vectors, Bottleneck features, MFCC | Equal Error Rate (EER), Accuracy, Average Detection Cost (Cavg) |
| [39] | 2018 | Speech Enhancement            | Local attention mechanism           | NS-LSTM             | Recordings in Chinese                    | Spectrogram, MFCC, LPC                                  | Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Log-Spectral Distance (LSD) |
| [40] | 2020 | Speech Emotion Recognition    | Attention mechanism                 | CNN, LSTM, GRU      | IEMOCAP dataset                          | Spectrogram                                             | Unweighted Accuracy (UA), Weighted Accuracy (WA) |
| [41] | 2019 | Disease detection (mood disorders) | Attention mechanism                 | CNN, LSTM           | CHI-MEI mood disorder database, MHMC emotion database | Zero-crossing rate, Root-mean-square, Fundamental frequency, Harmonic-Noise-Ratio, MFCC | Accuracy |
| [22] | 2020 | Disease detection (depression severity) | Soft attention mechanism (global attention approach) and Monotonic attention mechanism | Bi-LSTM, LSTM       | DAICW-OZ database                        | Spectrogram                                             | Root Mean Square Error (RMSE), Mean Absolute Error (MAE) |
| [18] | 2020 | Speech Enhancement            | Attention mechanism                 | CNN                 | TIMIT dataset, Noisex92 dataset          | Spectral vectors using STFT                              | Short-time objective intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ) |
| Pub  | Year | Area                      | Form of Implementation of Attention | Neural Used | Network Used | Data                                                                 | Extracted Features                                      | Metrics Used                                      |
|------|------|---------------------------|-------------------------------------|-------------|--------------|----------------------------------------------------------------------|---------------------------------------------------------|----------------------------------------------------|
| [42] | 2020 | Speech Emotion Recognition | Attention mechanism                | CNN         |              | IEMOCAP dataset, EMO-DB dataset, FAU-AIBO Corpus, EMOVO dataset, SAVEE dataset FAU-AIBO Corpus, EMO-DB dataset, Airplane Behavior Corpus | Spectrogram                                             | Mean Accuracy                                      |
| [43] | 2019 | Speech Emotion Recognition | Activation attention mechanism     | CNN         |              | FAU-AIBO Corpus, EMO-DB dataset, Airplane Behavior Corpus            | Spectrogram                                             | Unweighted Average Recall (UAR)                   |
| [44] | 2020 | Speech Enhancement        | Attention mechanism                | CNN, LSTM   |              | TIMIT dataset, Noisex92 dataset                                      | Spectrogram                                             | Short-term Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ), Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) Phone Error Rate (PER), Word Error Rate (WER), Accuracy Perceptual Evaluation of Speech Quality (PESQ), Short-term Objective Intelligibility (STOI) |
| [45] | 2020 | Speech pronunciation error detection | Attention mechanism              | Bi-LSTM     |              | CCTV news corpus, PSC-1176 corpus                                     | MFCC filterbank, 3-dimensional pitch                   | Character Error Rate (CER)                        |
| [23] | 2020 | Speech Enhancement        | Use of a classification neural network to act as a multidirectional attention switch | DNN         |              | TIMIT dataset, Noisex92 dataset                                      | Noise-aware features using STFT                        |                                                   |
| [46] | 2017 | Speech Recognition        | Attention mechanism                | Bi-LSTM, LSTM |              | WSJ dataset, CHiME dataset, HKUST dataset, CSJ corpus              | MFCC filterbank                                        |                                                   |
| Pub | Year | Area | Form of Implementation of Attention | Neural Network Used | Data | Extracted Features | Metrics Used |
|-----|------|------|--------------------------------------|---------------------|------|-------------------|--------------|
| [47] | 2019 | Speech Recognition | Attention mechanism | Bi-LSTM, LSTM | LIEPA corpus | Sequences of phonemes from raw audio files 3-D feature (the static, deltas and delta-deltas of Log-Mel spectrum filterbanks) | Accuracy, Word Error Rate (WER) |
| [48] | 2019 | Speech Emotion Recognition | Attention mechanism | Dilated CNN, Bi-LSTM | IEMOCAP dataset, EMO-DB dataset | MFCC, emobase2010, IS09, IS13 ComParE, MSF | Unweighted Accuracy (UA) |
| [29] | 2020 | Speech Emotion Recognition | Attention mechanism | 3D CNN, Bi-LSTM | IEMOCAP dataset, MSF-IMPROV dataset | RMSE, ZCR, fundamental frequency, HNR, MFCC | Unweighted Accuracy (UA), Weighted Accuracy (WA) |
| [49] | 2020 | Speech Emotion Recognition | Attention mechanism | HSF-DNN, MS-CNN, LLD-RNN | IEMOCAP dataset | Log-mel spectrogram | Average Processing Time, Average Accuracy, Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Artifact Ratio (SAR) |
| [50] | 2020 | Speech Emotion Recognition | Attention mechanism | CNN, Bi-LSTM | IEMOCAP dataset, RAVDESS dataset, SAVEE dataset | 3D scalogram | Unweighted Average Recall (UAR) |
| [51] | 2020 | Speech Emotion Recognition | Self-attention mechanism | 3D CNN LSTM | IEMOCAP dataset, EMO-DB dataset, SAVEE dataset | Log-mel spectrogram | Average Processing Time, Average Accuracy, Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), Signal to Artifact Ratio (SAR) |
| [31] | 2020 | Speech Separation | Attention mechanism | CNN, Bi-LSTM | MIR-1K dataset | Spectrogram | Accuracy Rate, Classification Rate |
| [52] | 2020 | Speech intelligibility estimation | Attention mechanism | LSTM | UA-Speech database | MFCC, energy of the modulation spectrum, LHMR, Three prosody-related features | Accuracy Rate, Classification Rate |
Table A1. Cont.

| Pub | Year | Area | Form of Implementation of Attention | Neural Used | Network | Data | Extracted Features | Metrics Used |
|-----|------|------|-------------------------------------|-------------|---------|------|-------------------|--------------|
| [53] | 2018 | Speech classification tasks | Attention mechanism | CNN | | UT-Podcast corpus, CHAINS corpus, eNTERFACE corpus, AP17-OLR database, NOISEX dataset, IIIT-H database | Spectrograms | Recall Score, Unweighted Average Recall (UAR) |
| [54] | 2020 | Language Identification | Attention mechanism | DNN, LSTM | | AP17-OLR database, NOISEX dataset | Shifted delta cepstral | Equal Error Rate (EER) |
| [55] | 2018 | Language Identification | Attention mechanism | DNN, DNN-WA | | AP17-OLR database, AP17-OLR database, IIIT-H database | MFCC | Equal Error Rate (EER) |
| [56] | 2020 | Speaker Verification | Attention mechanism | ResNet, SENet | | AP18-OLR database, NoCeleb dataset, NoCeleb dataset | Spectrograms | Equal Error Rate (EER) |
| [57] | 2019 | Language Identification | Self-attention mechanism | ResNet | | AP18-OLR database, NoCeleb dataset, NoCeleb dataset | MFCC | Equal Error Rate (EER) |
| [20] | 2019 | Speaker Recognition | Bayesian attention layer | DNN | | NIST dataset, OpenSLR corpus, NoCeleb dataset | NA | Equal Error Rate (EER) |
| [58] | 2019 | Voice Conversion | Multi-head Self-attention mechanism | Bi-LSTM, LSTM | | ARCTIC dataset, THCHS30 dataset, Free ST Chinese Mandarin Corpus, AIShell dataset | Phonetic posterior-grams | Similarity |
| [59] | 2019 | Speaker Recognition | Self-attention mechanism | CNN | | CMU ARCTIC dataset, Free ST Chinese Mandarin Corpus, AIShell public dataset | MFCC, Spectrogram | Word Error Rate (WER) |
| [60] | 2019 | Speech Recognition | Multi-headed additive attention mechanism | Bi-LSTM, LSTM | | LibriSpeech dataset | Log-mel filterbank | Word Error Rate (WER) |
| [61] | 2019 | Speech Separation | Additive attention mechanism | Bi-LSTM, LSTM | | WSJ0-2mix dataset | Magnitude spectrograms | Signal-to-distortion ratio (SDR) |
Table A1. Cont.

| Pub | Year | Area                        | Form of Implementation of Attention | Neural Network Used | Data                        | Extracted Features                                                                 | Metrics Used                  |
|-----|------|-----------------------------|------------------------------------|---------------------|-----------------------------|-----------------------------------------------------------------------------------|--------------------------------|
| [62] | 2018 | Voice Activity Detection    | Attention mechanism                | Bi-LSTM, LSTM       | TIMIT dataset               | MFCC                                                                              | Equal Error Rate (EER)        |
| [63] | 2018 | Speech Recognition          | Attention mechanism                | Bi-LSTM             | CSJ corpus, JNAS corpora    | Log Mel-scale filterbank, delta and acceleration coefficients                     | Word Error Rate (WER)         |
| [64] | 2018 | Speech Emotion Recognition  | Attention mechanism                | Bi-LSTM, LSTM       | IEMOCAP dataset             | MFCC, ZCR, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, 12D chroma vector, chroma deviation, harmonic ratio and pitch | Macro Average F-score (MAF), Macro Average Precision (MAP), Accuracy |
| [65] | 2019 | Adversarial examples generation | Attention mechanism                | RNN, GRU            | Data collected from a smart speaker | Mel-filterbank                                                                  | False Reject Rate (FRR), False Alarm Rate (FAR) |
| [66] | 2018 | Speech Recognition          | Attention mechanism                | CNN, LSTM           | Callcenter dataset, Reading dataset | NA                                                                                | Character Error Rate (CER)    |
| [67] | 2019 | Speech Enhancement          | Attention mechanism                | LSTM                | Musan corpus, CHIME3 dataset | Spectrograms, phase information                                                  | Perceptual Evaluation of Speech Quality (PESQ), Short-term Objective Intelligibility (STOI) |
| [68] | 2019 | Language Identification     | Multi-head attention mechanism     | RES-TDNN            | IIITH-ILSC database         | MFCC, SDC, i-vector, and phonetic                                                 | Equal Error Rate (EER)        |
Table A1. Cont.

| Pub | Year | Area                  | Form of Implementation of Attention | Neural Used | Network Used | Data                                                                 | Extracted Features | Metrics Used                                      |
|-----|------|-----------------------|-------------------------------------|-------------|--------------|----------------------------------------------------------------------|---------------------|---------------------------------------------------|
| [69] | 2019 | Speech Enhancement   | Self-attention mechanism            | Wave-U-Net  | CSTR VCTK Corpus, DEMAND Database                                   | NA                  | Perceptual Evaluation of Speech Quality (PESQ), Word Error Rate (WER) |
| [21] | 2020 | Speech Recognition   | Multi-head Self-attention mechanism | Dynamic convolution NN | CSJ corpus, Librispeech dataset, REVERVB dataset, CHiME dataset | NA                  | Character Error Rate (CER), Word Error Rate (WER) |
| [70] | 2017 | Speech Recognition   | Attention mechanism                 | Bi-LSTM, LSTM, NIN, CNN | WSJ dataset | MFCC, log Mel-spectrogram                                           | Character Error Rate (CER), Word Error Rate (WER) |
| [71] | 2019 | Speech Recognition   | Multi-head attention mechanism      | DNN         | CHiME dataset | Log-Mel filterbank, Acoustic and raw spectral features               | Naturalness, Similarity |
| [72] | 2019 | Voice Conversion     | Attention mechanism                 | Another author’s model (modified) | ASVspoof dataset, BTA52016 dataset | Equal Error Rate (EER)                                           |
| [73] | 2020 | Speaker Verification | Soft spatial attention module       | DenseNet-Bi-LSTM | ASVspoof dataset, BTA52016 dataset | NA                  | Word Error Rate Reduction (WERR)                     |
| [24] | 2020 | Speech Recognition   | Attention mechanism                 | LSTM        | Spoken dialog between users and digital assistants DAMP—Sing! 300 × 30 × 2 dataset | NA                  | Word Error Rate (WER)                                 |
| [74] | 2020 | Lyrics transcription | Self-attention mechanism            | CTDNN       | Mel-spectrogram filter banks                                       | Word Error Rate (WER) |
| [75] | 2019 | Speech Recognition   | Attention mechanism                 | RNN         | Microsoft Cortana dataset, LibriSpeech dataset, DEMAND database    | Rate of Succeed Attack (RoSA), Word Error Rate (WER) |
| [76] | 2020 | Speech Recognition   | Self-attention mechanism            | U-Net       | MFCC                                                                 |                      |                                                                 |
### Table A1. Cont.

| Pub  | Year | Area                        | Form of Implementation of Attention | Neural Used                | Network Used          | Data                        | Extracted Features                  | Metrics Used                      |
|------|------|-----------------------------|-------------------------------------|---------------------------|------------------------|-----------------------------|-------------------------------------|------------------------------------|
| [77] | 2019 | Speaker Verification        | Multi-head mechanism attention      | LSTM, cLSTM, CNN, DNN     | VoCeleb dataset        | Static log Mel filterbanks  | Equal Error Rate (EER)              | Unweighted Accuracy (UA)             |
| [78] | 2019 | Speech Emotion Recognition  | Self-attention mechanism            | CNN                       | IEMOCAP dataset        | Mel-spectrograms             | Weighted Accuracy (WA)              | Unweighted Average Recall (UAR)      |
| [79] | 2019 | Speech Conflict Estimation  | Global additive self-attention     | LSTM, CRNN                | SSPNet Conflict Corpus | Raw speech waveforms         | Pearson Correlation Coefficient (PCC) | Weighted Average Recall (WAR)          |
| [19] | 2018 | Speech Emotion Recognition  | Attention mechanism                | Bi-LSTM                   | IEMOCAP dataset        | Pitch, energy, zero-crossing rate, voicing probability, MFCC | Weighted Accuracy (WA), Unweighted Accuracy (UA) |
| [80] | 2018 | Speech Emotion Recognition  | Attention mechanism                | CNN                       | IEMOCAP dataset, Recola database | Log-Mel filterbanks         | Unweighted Average Recall (UAR)      | Top-1 and Top-5 accuracies           |
| [81] | 2019 | Speaker Recognition         | Self-attention mechanism           | VGG, CNN, ResNets         | VoCeleb dataset        | Log-Mel filterbanks         | Unweighted Accuracy (UA)             |                                    |
| [82] | 2017 | Speech Emotion Recognition  | Attention mechanism                | CNN-LSTM                  | eNTERFACE-05 corpus, MUSAN corpus | Log-Mel filterbanks         | MFCC, 1-dimensional logarithmic energy, voicing probability, HNR, logarithmic fundamental frequency, zero-crossing rate | Unweighted Accuracy (UA), F1 Scores |
| [83] | 2019 | Speech Emotion Recognition  | Multi-head Self-attention          | DRN, LSTM, DNN            | IEMOCAP dataset        |                            |                                     |                                    |
| Pub  | Year | Area                      | Form of Implementation of Attention | Neural Network Used | Network Used | Data                     | Extracted Features                                                                 | Metrics Used                                                                 |
|------|------|---------------------------|-------------------------------------|---------------------|--------------|--------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| [84] | 2019 | Speech disfluency detection | Attention mechanism                 | Bi-LSTM, LSTM       | CSJ corpus   | Mel-scale filterbank, delta and delta-delta, log-pitch               | F1 Scores, Word Fragments                                                             |
| [85] | 2020 | Speech Recognition         | Attention mechanism                 | Bi-LSTM, LSTM       | CSJ corpus   | Log-Mel filterbank, delta and acceleration coefficients              | Character Error Rate (CER), Kana Error Rate (KER), Word Error Rate (WER), Word Error Rate Reduction (WERR) |
| [86] | 2019 | Speech Recognition         | Attention mechanism                 | Bi-GRU, RNN         | Microsoft Cortana dataset | Log-Mel filterbank, MFCC, root-mean-square energy, zero-crossing rate, harmonics-to-noise ratio, fundamental frequency | Unweighted Averaged (UA)                                                            |
| [87] | 2018 | Speech Emotion Recognition | Attention mechanism                 | RNN, LSTM           | FAU-AIBO Corpus| IEMOCAP dataset, KSUEmotions database                               | F1 Scores, Overall Accuracy                                                           |
| [88] | 2019 | Speech Emotion Recognition | Attention mechanism                 | CNN, Bi-LSTM, DNN   | IEMOCAP dataset, KSUEmotions database                              | MFCC                                                                                 | Recognition Rate                                                                    |
| [89] | 2019 | Speech Emotion Recognition | Attention mechanism                 | CNN, Bi-LSTM        | FAU-AIBO Corpus, CASIA dataset English corpus, Chinese corpus, Amdo-Tibetan corpus | Log-Mel filter bank, Phoneme Error Rate (PER)                                          |
| [90] | 2020 | Speech Recognition         | Multi-head attention mechanism      | pBi-LSTM            | English corpus, Chinese corpus, Amdo-Tibetan corpus                | Mel-scale filterbank coefficients, energy (deltas and delta-deltas)                  | Character Error Rate (CER) and Word Error Rate (WER)                                  |
| [91] | 2016 | Speech Recognition         | Attention mechanism                 | Bi-RNN              | WSJ dataset                                                   | Mel-scale filterbank coefficients, energy (deltas and delta-deltas)                  | Character Error Rate (CER) and Word Error Rate (WER)                                  |
Table A1. Cont.

| Pub | Year | Area                  | Form of Implementation of Attention | Neural Network Used | Network Used | Data                                      | Extracted Features                                      | Metrics Used                                                                 |
|-----|------|-----------------------|-------------------------------------|---------------------|--------------|------------------------------------------|----------------------------------------------------------|-----------------------------------------------------------------------------|
| [92] | 2019 | Detection of attacks | Self-attention mechanism            | LCNN                |              | ASV spoof dataset                        | Spectral representations, Cepstral coefficients            | Equal Error Rate (EER), Tandem Decision Cost Function (T-DCF)            |
| [93] | 2019 | Language Identification | Attention mechanism             | GRU, RNN           |              | LRE2017 dataset                         | Bottleneck features                                       | Average Detection Cost (Cavg), Approximate Computational Time            |
| [94] | 2020 | Speech Recognition   | Attention mechanism               | Transformers       |              | WSJ dataset                              | Log-Mel filterbank coefficients (with pitch and their delta and delta delta), raw waveform audio signal | Word Error Rate (WER)                                                    |
| [95] | 2019 | Speech Recognition   | Attention mechanism               | Bi-LSTM            |              | Tibetan Ando dialect corpus (made by authors) | Mel-scale filterbank coefficients, pitch                  | Character Error Rate (CER)                                               |
| [96] | 2019 | Speech Recognition   | Attention mechanism               | Bi-LSTM, LSTM      |              | LibriSpeech dataset                      | Power-mel filterbank coefficients, Speech waveform        | Word Error Rate (WER)                                                    |
| [97] | 2019 | Classification of speech utterances | Attention mechanism             | DNN, CNN, Bi-LSTM  |              | Dataset made by authors                  | Mel-filterbank coefficients                              | Detection Error Tradeoff (DET), Equal Error Rate (EER)                |
| [98] | 2018 | Language identification | Attention mechanism             | Bi-GRU, CNN & GRU. |              | Dataset made by authors                  | Log-Mel filter bank                                       | Accuracy, Unweighted Average Recall (UAR)                              |
| [99] | 2020 | Speaker Recognition  | Attention mechanism               | CNN                 |              | VoxCeleb dataset                        | Spectrograms                                             | Equal Error Rate (EER)                                                  |
Table A1. Cont.

| Pub  | Year | Area                      | Form of Implementation of Attention | Neural Network Used | Data                          | Extracted Features      | Metrics Used                                      |
|------|------|---------------------------|-------------------------------------|---------------------|-------------------------------|-------------------------|---------------------------------------------------|
| [100]| 2020 | Speech quality estimation | Attention mechanism                 | CNN-LSTM            | Multiple datasets             | Log-mel spectrograms    | Root Mean Square Error (RMSE)                      |
|      |      |                           |                                     |                     |                               |                         | Unweighted Average Recall (UAR), Weighted Average Recall (WAR) |
| [101]| 2020 | Speech Emotion Recognition| Attention mechanism                 | Bi-LSTM             | IEMOCAP dataset               | Spectrogram             |                                                   |
|      |      |                           |                                     |                     |                               |                         |                                                   |
| [102]| 2020 | Speech Emotion Recognition| Multi-head Self-attention mechanism | CNN                 | IEMOCAP dataset               | MFCC                    |                                                   |
|      |      |                           |                                     |                     |                               |                         |                                                   |
| [103]| 2018 | Speech Recognition        | Attention mechanism                 | VResTDCTC           | CSJ corpus                    | Non-spliced filterbank features | Word Error Rate (WER) |
| [104]| 2018 | Speech Recognition        | Attention mechanism                 | RNN                 | Dataset made by authors       | NA                      | Word Error Rate (WER) |
| [105]| 2019 | Speech Recognition        | Attention mechanism                 | Another author’s model | LibriSpeech dataset          | NA                      | Word Error Rate (WER) |
| [106]| 2019 | Speech Recognition        | Attention mechanism                 | BiRNN, LSTM         | Dataset made by authors       | Pitch, delta, pitch     | Character Error Rate (CER) |
| [107]| 2017 | Speech Recognition        | Attention mechanism                 | Bi-LSTM, LSTM, RNN, CNN | WSI dataset, CSJ Corpus, HKUST dataset, VoxForge dataset | Filterbank, pitch      | Character Error Rates (CER), Accuracies/Error Rates |
| [108]| 2019 | Disease detection (dysarthria) | Attention mechanism                 | LSTM                | TORGO database                | Mel-filterbanks, Time-Domain filterbanks | Unweighted Average Recall (UAR) |
| [109]| 2016 | Speech Recognition        | Attention mechanism                 | pBi-LSTM            | Google voice search utterances | Log-mel filterbank      | Word Error Rate (WER) |
| [110]| 2019 | Word vectors generation  | Attention mechanism                 | RNN, Bi-LSTM        | LibriSpeech dataset           | MFCC                    | Word Similarity                                    |
| [111]| 2020 | Speech Emotion Recognition| Multi-head mechanism                | Transformer          | IEMOCAP dataset               | Log-Mel filterbank Energies | Weighted Average (WA), Unweighted Average (UA) |
| Pub | Year | Area                | Form of Implementation of Attention | Neural Network Used | Data                          | Extracted Features                                                                 | Metrics Used                                                                 |
|-----|------|---------------------|-------------------------------------|---------------------|-------------------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| [112] | 2019 | Language Identification | Multi-head Self-attention mechanism | DNN                  | AP17-OLR database             | Shifted delta cepstral features (computed using MFCC)                            | Equal Error Rate (EER)                                                         |
| [113] | 2019 | Speech Separation   | Attention mechanism                  | Bi-LSTM             | TSP corpus, THCHS-30 dataset | Amplitude spectrum                                                             | Perceptual Evaluation of Speech Quality (PESQ), Short-term Objective Intelligibility (STOI) |
| [114] | 2020 | Voice Activity Detection | Attention mechanism                  | Bi-LSTM             | Dataset made by authors       | Log-Mel filterbank energies Waveforms (mean-variance normalized)                | F1 Scores, Accuracy                                                            |
| [115] | 2019 | Language Identification | Attention mechanism                  | CNN, VD-CNN, ResNet | Dataset made by authors       | Mel-scale filterbank Coefficients, logarithmic fundamental frequency, aperiodicity | F1 Scores, Accuracy                                                            |
| [116] | 2018 | Speech Recognition   | Attention mechanism                  | DNN, CNN-LSTM       | CSJ corpus                    | Mel-Cepstral Coefficients, logistic fundamental frequency, aperiodicity         | Character Error Rate (CER)                                                      |
| [117] | 2019 | Voice Conversion    | Self-attention mechanism             | CNN                  | VCC2016 dataset, Data collect of internet | Word Error Rate (WER), Word Error Rate Reduction (WERR) Perceptual Evaluation of Speech Quality (PESQ), CSIG, CBAK, COVL, Segmented SNR | Speaker/Singer Identity, Naturalness                                          |
| [118] | 2018 | Speaker adaptation  | Attention mechanism                  | RNN                  | Switchboard (SWB) task        | Word Error Rate (WER), Word Error Rate Reduction (WERR) Perceptual Evaluation of Speech Quality (PESQ), CSIG, CBAK, COVL, Segmented SNR | Word Error Rate (WER), Word Error Rate Reduction (WERR) Perceptual Evaluation of Speech Quality (PESQ), CSIG, CBAK, COVL, Segmented SNR |
| [119] | 2020 | Speech Recognition   | Attention mechanism                  | Bi-LSTM, CNN        | Switchboard (SWB) task        | Word Error Rate (WER), Word Error Rate Reduction (WERR) Perceptual Evaluation of Speech Quality (PESQ), CSIG, CBAK, COVL, Segmented SNR | Word Error Rate (WER), Word Error Rate Reduction (WERR) Perceptual Evaluation of Speech Quality (PESQ), CSIG, CBAK, COVL, Segmented SNR |
| [120] | 2019 | Speech Enhancement  | Self-attention mechanism             | FCNN                 | VCTK speech dataset            | NA                                                                               | Perceptual Evaluation of Speech Quality (PESQ), CSIG, CBAK, COVL, Segmented SNR |
| Pub     | Year | Area                              | Form of Implementation of Attention | Neural Network Used | Data                                      | Extracted Features | Metrics Used                                      |
|---------|------|-----------------------------------|--------------------------------------|---------------------|-------------------------------------------|-------------------|-------------------------------------------------|
| [121]   | 2020 | Classification of phonation modes | Attention mechanism                   | CNN, RAN            | Four different datasets                   | Mel-scaled magnitude spectrum | F1 Scores, Accuracy                               |
|         |      |                                   |                                      |                     |                                           |                   | F1 Scores, Recall (sensitivity), Probability of False Alarm (PFA) |
| [122]   | 2020 | Disease detection (SARS-CoV-2)    | A self-supervised attention-based transformer | Transformer        | COVID19 dataset, Librispeech dataset     | Mel-scaled frequencies |                                                       |
| [123]   | 2019 | Speech Recognition                | Self-attention mechanism             | Self-attention network | HKUST dataset, CasiaMTS dataset          | Filterbanks (with delta and delta-delta transforms), Log-Mels (with delta and delta-delta transforms), i-vectors | Character Error Rate (CER) |
| [124]   | 2019 | Speech Recognition                | Attention mechanism                  | CNN, RNN            | BN-6000 Corpus                          | Log-Mel spectrogram, Zero-Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Rolloff, MFCC, Chroma Vector, Chroma Deviation | Word Error Rate (WER) |
| [125]   | 2019 | Speaker Verification              | Attention mechanism                  | CNN, GRU            | Tencent wake-up word dataset             | Log-Mel spectrogram, MFCC, Chroma Vector, Chroma Deviation, Spectral Flux, Spectral Rolloff, MFCC (and energy augmented by delta and delta-delta), Log-spectrogram, eGeMAPS | Equal Error Rate (EER) |
| [126]   | 2020 | Speech Emotion Recognition        | Attention mechanism                  | CNN, Bi-LSTM        | Berlin dataset, DaFEx dataset, CASIA dataset |                                                             | Emotion-Wise Accuracy |
| [127]   | 2019 | Speech Emotion Recognition        | Self-attention mechanism             | DNN, CNN            | IEMOCAP dataset, RAVDESS dataset         | IEMOCAP dataset, RAVDESS dataset | Unweighted Accuracy (UA)                          |
| Pub  | Year | Area               | Form of Implementation of Attention | Neural Network Used | Network Used | Data                          | Extracted Features | Metrics Used                                                                 |
|------|------|--------------------|-------------------------------------|---------------------|--------------|-------------------------------|-------------------|------------------------------------------------------------------------------|
| [128] | 2018 | Speech Recognition | Attention mechanism                | RNN                 |              | CHiME dataset, WSJ dataset    | log-Mel filterbank | Word Error Rate (WER), Signal-to-Distortion Ratio (SDR), Perceptual Evaluation of Speech Quality (PESQ) |
| [129] | 2019 | Speech Recognition | Attention mechanism                | DNN, LSTM, Bi-LSTMP |              | LibriSpeech dataset, TED-LIUM dataset, WSJ dataset | Mel scale filterbank, 3 pitch features | Word Error Rate (WER)                                                          |
| [130] | 2019 | Speech Emotion Recognition | Attention mechanism                | CNN, LSTM          |              | CASIA dataset                 | Spectrograms      | Precision, Recall, F1 Scores, Word Error Rate (WER)                        |
| [131] | 2019 | Speech Recognition | Self-attention mechanism           | TD-NN               |              | LibriSpeech dataset           | High resolution MFCC, i-vectors | Word Error Rate (WER), Word Error Rate (WER), Word Error Rate Reduction (WER) |
| [132] | 2018 | Speech Recognition | Multi-head mechanism               | LSTM                |              | Google voice search traffic   | Log-Mel features   | Word Error Rates (WER)                                                       |
| [133] | 2020 | Speech Recognition | Self-attention mechanism           | CNN, RNN, Transformer |              | LibriSpeech dataset           | Log-mel spectral energies, pitch information | Word Error Rates (WER)                                                       |
| [134] | 2019 | Speech Recognition | Attention mechanism                | LSTM, TDLSTM       |              | WSJ dataset, LibriSpeech corpus, HKUST dataset | Log-Mel spectral energies, pitch feature | Word Error Rates (WER)                                                       |
| [135] | 2020 | Speech Recognition | Self-attention mechanism           | Transformer, RNN-T  |              | LibriSpeech dataset           | Log-Mel energy values | Word Error Rates (WER)                                                       |
| [136] | 2019 | Speech Recognition | Attention mechanism                | CNN, Bi-LSTM, LSTM |              | WSJ dataset, LibriSpeech corpus, HKUST dataset | Log-Mel spectral energies | Character Error Rate (CER), Word Error Rates (WER)                           |
| Pub  | Year | Area                  | Form of Implementation of Attention | Neural Used | Network Used | Data                          | Extracted Features                      | Metrics Used                          |
|------|------|-----------------------|--------------------------------------|-------------|--------------|-------------------------------|----------------------------------------|----------------------------------------|
| [137] | 2019 | Language Identification | Self-attention mechanism              | DCNN, Bi-LSTM | NIST LRE dataset | Log-Mel filterbank energies | Average Detection Cost (Cavg), Equal Error Rate (EER) |
| [138] | 2018 | Voice Activity Detection | Attention mechanism                   | FC-NN, LSTM  | TIMIT dataset, Noisex92 dataset, HAVIC corpus, WSJCAM0 corpus, MC-WSJ-AV corpus | Multiresolution cochleagram features (MRCG) | Area Under the Curve (AUC)         |
| [139] | 2019 | Speech Recognition     | Attention mechanism                   | DCNN        | TIMIT dataset, WSJ dataset, LibriSpeech dataset | Mel filterbanks (with delta and delta-delta components) | Word Error Rate (WER)         |
| [140] | 2019 | Speech Recognition     | Attention mechanism                   | DNN, LSTM   | NIST LRE dataset, MUSAN dataset, RIR dataset | MFCC                          | Average Detection Cost (Cavg), Word Error Rate (WER) |
| [141] | 2020 | Language Identification | Attention mechanism                   | DNN         | NIST LRE dataset, MUSAN dataset, RIR dataset | MFCC                          | Average Detection Cost (Cavg), Equal Error Rate (EER) |
| [142] | 2020 | Speech Emotion Recognition | Attention mechanism                   | RNN, DNN    | IEMOCAP dataset, EmotAsS dataset | Set of prosodic features (Duration, Energy, F0 and its dynamics, Voice quality), MFCC | Weighted Average Recall (WAR), Unweighted Average Recall (UAR) |
| [143] | 2020 | Speech Recognition     | Attention mechanism                   | LSTM, BLSTM | TED-LIUM dataset | Log-Mel f-bank features       | Character Error Rate (CER), Word Error Rate (WER) |
| [144] | 2019 | Speech dialect Identification | Attention mechanism                   | DNN         | Chinese dialects speech database | Prosodic features (F0, Energy, Loudness, Pitch), I-vector Coefficients derived from mel-scale filter banks | Equal Error Rate (EER)         |
| [25]  | 2002 | Speech Recognition     | Performing partial computation guided by attention criterions | MLP         | Speech isolated-words | Speech isolated-words | Learning Time (sec) |
| Pub  | Year | Area                     | Form of Implementation of Attention | Neural Used | Network Used | Data                                                                 | Extracted Features                  | Metrics Used                                                                 |
|------|------|--------------------------|--------------------------------------|-------------|--------------|----------------------------------------------------------------------|--------------------------------------|--------------------------------------------------------------------------------|
| [145]| 2020 | Speech enhancement       | Self-attention mechanism             | DNN         |              | Voice Bank Corpus database, Chinese Mandarin Test CD, Noisex92 dataset, PNL 100 Non-speech database HAVRUS corpus, VoxForge dataset, M-AILABS corpus | MFCC, AMS, RASTA-PLP, cochleagram, PNCC | Perceptual Evaluation of Speech Quality (PESQ), Short-term Objective Intelligibility (STOI) |
| [146]| 2020 | Speech Recognition       | Attention mechanism                  | LSTM, Bi-LSTM, ResNet                  |              | HAVRUS corpus, VoxForge dataset, M-AILABS corpus                     | NA                                   | Character Error Rate (CER), Word Error Rate (WER)                                |
| [147]| 2019 | Cognitive Load Classification | Attention mechanism                  | LSTM                          |              | CSLE database                                                        | Log-Mel filterbank energies         | Unweighted Average Recall (UAR)                                                  |
| [148]| 2019 | Speech Recognition       | Attention mechanism                  | Bi-LSTM, LST, VggCNN                |              | ATC corpus                                                           | Mel-scale filterbank coefficients, pitch features | Character Error Rate (CER), Sentence Error Rate (SER)                            |
| [149]| 2019 | Speech Recognition       | Attention mechanism                  | CNN, LSTM, Bi-LSTM, MLP             |              | VoxForge dataset, M-AILABS corpus, SPIIRAS corpus                    | Spectrogram, filterbank, deltas features | Character Error Rate (CER), Word Error Rate (WER), Real-Time Factor (RTF)          |
| [150]| 2020 | Speech Recognition       | Attention mechanism                  | Bi-LSTM, LSTM                       |              | VoxForge dataset                                                     | MFCC, pitch features                | Character Error Rate (CER)                                                        |
| [27] | 2000 | Speech word rejection    | Inclusion of an attention layer       | MLP                                   |              | A isolated-word database                                              | Zero Crossing with Peak Amplitude   | In-vocabulary Rejection Rate, Out-of-vocabulary Rejection Rate                   |
| [151]| 2020 | Speech Recognition       | Attention mechanism                  | RNN, GRU                             |              | TIMIT dataset, WSJ dataset                                            | Mel scale filterbank, energy        | Word Error Rate (WER), Phone Error Rate (PER)                                     |
| [152]| 2019 | Speech Emotion Recognition | Self-attention mechanism             | DNN, CNN, Bi-LSTM, ELM              |              | IEMOCAP dataset                                                      | Spectrogram                         | Accuracy                                                                       |
Appendix B

Table A2. Assessment of risk of bias.

| Publication | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Final Score |
|-------------|----|----|----|----|----|----|----|----|----|-----|-------------|
| [26]        | 1  | 1  | 1  | 1  | 1  | 0  | 0.5| 1  | 0.5 | 1   | 9           |
| [33]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 1   | 8.5         |
| [34]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 8           |
| [35]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 7.5         |
| [28]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 0   | 1   | 8.5         |
| [36]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 0   | 1   | 8           |
| [37]        | 1  | 1  | 0.5| 1  | 1  | 0  | 1  | 1  | 0.5 | 1   | 8           |
| [38]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0.5| 8.5         |
| [39]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0.5| 8           |
| [40]        | 1  | 0.5| 1  | 0.5| 0  | 0  | 1  | 1  | 1   | 0.5| 7           |
| [41]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 8           |
| [42]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0.5| 8.5         |
| [43]        | 1  | 0.5| 1  | 0  | 0  | 1  | 1  | 1  | 0   | 1   | 7.5         |
| [44]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 7.5         |
| [45]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 9             |
| [46]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0.5| 8.5         |
| [47]        | 1  | 1  | 1  | 0.5| 0  | 1  | 1  | 1  | 0   | 1   | 0.5         |
| [48]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 0.5| 8.5         |
| [49]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 1   | 10          |
| [50]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 7.5         |
| [51]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 9           |
| [52]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 0.5         |
| [53]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 9           |
| [54]        | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1   | 0.5| 8.5         |
| [55]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0.5| 8           |
| [56]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 7.5         |
| [57]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 8           |
| [58]        | 1  | 1  | 0.5| 0.5| 0.5| 1  | 0  | 1  | 1   | 0.5| 7           |
| [59]        | 1  | 0.5| 1  | 1  | 1  | 1  | 1  | 0  | 1  | 0   | 7.5         |
| [60]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 7.5         |
| [61]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 7.5         |
| [62]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 7           |
| [63]        | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0   | 7.5         |
| [64]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 1   | 8           |
| [65]        | 1  | 0.5| 1  | 1  | 0.5| 0  | 1  | 1  | 0   | 1   | 8           |
| [66]        | 1  | 1  | 1  | 0.5| 1  | 0  | 1  | 1  | 0.5 | 1   | 8           |
| [67]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 0.5         |
| [68]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 1   | 8.5         |
| [69]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 7.5         |
| [70]        | 1  | 1  | 1  | 0.5| 0.5| 1  | 0  | 1  | 1   | 0.5| 7.5         |
| [71]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 7.5         |
| [72]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 1   | 7           |
| [73]        | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1   | 0.5| 8.5         |
Table A2. Cont.

| Publication | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Final Score |
|-------------|----|----|----|----|----|----|----|----|----|-----|-------------|
| [24]        | 1  | 0.5| 1  | 0.5| 1  | 1  | 0  | 1  | 1  | 0   | 7           |
| [74]        | 1  | 1  | 1  | 0.5| 1  | 0  | 1  | 1  | 0   | 7.5  |
| [75]        | 1  | 0.5| 1  | 1  | 0.5| 0  | 1  | 1  | 0   | 7   |
| [76]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8.5  |
| [77]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8   |
| [78]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8.5  |
| [79]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8   |
| [80]        | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8   |
| [81]        | 0.5| 1  | 1  | 1  | 1  | 0.5| 0  | 1  | 1  | 0   | 7.5  |
| [82]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8.5  |
| [83]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8   |
| [84]        | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [85]        | 0.5| 1  | 1  | 0.5| 1  | 0  | 1  | 1  | 0   | 7   |
| [86]        | 0.5| 1  | 1  | 1  | 1  | 0  | 0.5| 1  | 0   | 7   |
| [87]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [88]        | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [89]        | 0.5| 1  | 1  | 1  | 1  | 0.5| 0  | 1  | 1  | 0   | 7   |
| [90]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8.5  |
| [91]        | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [92]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8.5  |
| [93]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8.5  |
| [94]        | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [95]        | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [96]        | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [97]        | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [98]        | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8.5  |
| [99]        | 0.5| 1  | 1  | 1  | 1  | 0.5| 0  | 1  | 1  | 0   | 7   |
| [100]       | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 8   |
| [101]       | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 8.5  |
| [102]       | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 8   |
| [103]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [104]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [105]       | 0.5| 1  | 1  | 0.5| 1  | 0  | 1  | 1  | 0.5 | 7.5  |
| [106]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [107]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [108]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [109]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [110]       | 0.5| 1  | 1  | 1  | 1  | 0.5| 0  | 1  | 1  | 0   | 7   |
| [111]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [112]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [113]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5 | 8   |
| [114]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [115]       | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 9   |
| [116]       | 1  | 0.5| 1  | 1  | 0.5| 1  | 0  | 0.5| 0.5| 0.5| 7   |
| [117]       | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [118]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [119]       | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [120]       | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5  |
| [121]       | 0.5| 1  | 1  | 1  | 0.5| 0  | 1  | 1  | 0   | 7   |
| [122]       | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 9   |
Table A2. Cont.

| Publication | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Final Score |
|-------------|----|----|----|----|----|----|----|----|----|-----|-------------|
| [123]       | 1  | 0.5| 1  | 1  | 1  | 1  | 1  | 0  | 1  | 0   | 7.5         |
| [124]       | 1  | 0.5| 1  | 1  | 0.5| 1  | 1  | 0  | 1  | 0   | 7           |
| [125]       | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0  | 0   | 7.5         |
| [126]       | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 0.5| 8.5         |
| [127]       | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 9           |
| [128]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [129]       | 1  | 1  | 1  | 0.5| 1  | 1  | 0  | 1  | 1  | 1   | 8           |
| [130]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [131]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 7.5         |
| [132]       | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 7.5         |
| [133]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5| 7.5         |
| [134]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [135]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [136]       | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 8           |
| [137]       | 1  | 0.5| 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1   | 7.5         |
| [138]       | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5| 8.5         |
| [139]       | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0.5| 8.5         |
| [140]       | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 1  | 1  | 1   | 9           |
| [141]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [142]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [143]       | 1  | 0.5| 1  | 1  | 0.5| 1  | 0  | 1  | 1  | 0   | 7           |
| [144]       | 1  | 0.5| 1  | 1  | 0.5| 1  | 0  | 1  | 1  | 0   | 7           |
| [145]       | 1  | 0.5| 1  | 0.5| 1  | 1  | 0  | 1  | 1  | 0.5| 8           |
| [146]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [147]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [148]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [149]       | 1  | 0.5| 1  | 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [150]       | 1  | 0.5| 1  | 0.5| 1  | 1  | 0  | 1  | 1  | 0   | 7           |
| [151]       | 1  | 0.5| 1  | 1  | 0.5| 1  | 1  | 1  | 0.5| 8.5| 7.5         |
| [152]       | 1  | 1  | 1  | 0.5| 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [153]       | 1  | 0.5| 1  | 1  | 1  | 0  | 1  | 1  | 0   | 7.5         |
| [154]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 0   | 6           |
| [155]       | 1  | 0.5| 1  | 0  | 1  | 0  | 1  | 1  | 0   | 6           |
| [156]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 1   | 6.5         |
| [157]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 1   | 6.5         |
| [158]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 0.5| 7.5         |
| [159]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 1   | 6           |
| [160]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 0   | 6.5         |
| [161]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 0.5| 6.5         |
| [162]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 1   | 6.5         |
| [163]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 1   | 6.5         |
| [164]       | 1  | 0.5| 1  | 0.5| 0.5| 0.5| 0  | 1  | 0.5| 6           |
| [165]       | 1  | 0.5| 0.5| 0.5| 0.5| 0.5| 0  | 1  | 1   | 6           |

References

1. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* 2009, 6, e1000097, doi:10.1371/journal.pmed.1000097.
2. Galassi, A.; Lippi, M.; Torroni, P. Attention in Natural Language Processing. *IEEE Trans. Neural Netw. Learn. Syst.* 2020, doi:10.1109/TNNLS.2020.3019893.
3. Lee, J.B.; Rossi, R.A.; Kim, S.; Ahmed, N.K.; Koh, E. Attention Models in Graphs: A Survey. *ACM Trans. Knowl. Discov. Data* 2019, 13, doi:10.1145/3363574.
Symmetry 2021, 13, 214

29. Peng, Z.; Li, X.; Zhu, Z.; Unoki, M.; Dang, J.; Akagi, M. Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends. IEEE Access 2020, 8, 16560–16572, doi:10.1109/ACCESS.2020.2967791.

30. Fu, D.; Weber, C.; Yang, G.; Kerzel, M.; Nan, W.; Barros, P.; Wu, H.; Liu, X.; Wermler, S. What Can Computational Models Learn From Human Selective Attention? A Review From an Audiovisual Unimodal and Crossmodal Perspective. Front. Integr. Neurosci. 2020, 14, doi:10.3389/fint.2020.00010.

31. Yuan, C.M.; Sun, X.M.; Zhao, H. Speech Separation Using Convolutional Neural Network and Attention Mechanism. Discret. Dyn. Nat. Soc. 2020, 2020, 2196893, doi:10.1155/2020/2196893.

32. Travisio-González, C.M.; Alonso-Hernández, J.B. Special issue on developing nature-inspired intelligence by neural systems. Neural Comput. Appl. 2020, 32, 17823–17824, doi:10.1007/s00521-020-05454-w.

33. Zhao, Z.; Zhao, Y.; Bao, Z.; Wang, H.; Zhang, Z.; Li, C. Deep Spectrum Feature Representations for Speech Emotion Recognition. In Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data, ASMMC-MMAC’18, Seoul, Korea, 26 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 27–33, doi:10.1145/3267935.3267948.

34. Kano, T.; Sakti, S.; Nakamura, S. Study of robust language identification techniques for future smart cities. Adv. Ubiquitous Comput. 2020, 163–183, doi:10.1016/B978-0-12-816801-1.00005-0.

35. Vuddagiri, R.K.; Gurugubelli, K.; Thirumuru, R.; Vuppala, A.K. Study of robust language identification techniques for future smart cities. Adv. Ubiquitous Comput. 2020, 163–183, doi:10.1016/B978-0-12-816801-1.00005-0.

36. Ocquaye, E.N.N.; Mao, Q.; Xue, Y.; Song, H.; Xu, G.; Xue, Y. Dual Exclusive Attentive Transfer for Unsupervised Deep Convolutional Neural Networks. Knowl. Based Syst. 2018, 161, 90–100, doi:10.1016/j.knosys.2018.07.033.

37. Vuddagiri, R.K.; Vydana, H.K.; Vuppala, A.K. Curriculum learning based approach for noisy robust language identification using DNN with attention. Expert Syst. Appl. 2018, 110, 290–297, doi:10.1016/j.eswa.2018.06.004.
56. Xu, J.; Wang, X.; Feng, B.; Liu, W. Deep multi-metric learning for text-independent speaker verification. *Neurocomputing* 2020, 410, 394–400, doi:10.1016/j.neucom.2020.06.045.

57. Monteiro, J.; Alam, J.; Falk, T.H. Residual convolutional neural network with attentive feature pooling for end-to-end language identification from short-duration speech. *Comput. Speech Lang.* 2019, 58, 364–376, doi:10.1016/j.csl.2019.05.006.

58. Lu, H.; Wu, Z.; Li, R.; Kang, S.; Jia, J.; Meng, H. A Compact Framework for Voice Conversion Using WaveNet Conditioned on Phonetic Posteriorgrams. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6810–6814, doi:10.1109/ICASSP.2019.8682938.

59. Gong, S.; Chang, R.; Hao, T.; Wu, G.; Wang, Y. A Convenient and Extensible Offline Chinese Speech Recognition System Based on Convolutional CTC Networks. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; pp. 7606–7611, doi:10.23919/ChiCC.2019.8865580.

60. Guo, J.; Sainath, T.N.; Weiss, R.J. A Spelling Correction Model for End-to-end Speech Recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5651–5655, doi:10.1109/ICASSP.2019.8683745.

61. Ochiai, T.; Delcroix, M.; Kinoshiha, K.; Ogawa, A.; Nakatani, T. A Unified Framework for Neural Speech Separation and Extraction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6975–6979, doi:10.1109/ICASSP.2019.8683448.

62. Yu, Y.; Kim, Y. A Voice Activity Detection Model Composed of Bidirectional LSTM and Attention Mechanism. In Proceedings of the 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 29 November–2 December 2018; pp. 1–5, doi:10.1109/HNICEM.2018.8666342.

63. Ueno, S.; Inaguma, H.; Mimura, M.; Kawahara, T. Acoustic-to-Word Attention-Based Model Complemented with Character-Level CTC-Based Model. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5804–5808, doi:10.1109/ICASSP.2018.8462576.

64. Tao, F.; Liu, G. Advanced LSTM: A Study About Better Time Dependency Modeling in Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2906–2910, doi:10.1109/ICASSP.2018.8461750.

65. Wang, X.; Sun, S.; Shan, C.; Hou, J.; Xie, L.; Li, S.; Lei, X. Adversarial Examples for Improving End-to-end Attention-based Small-footprint Keyword Spotting. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6366–6370, doi:10.1109/ICASSP.2019.8683479.

66. Jiang, D.; Zou, W.; Zhao, S.; Yang, G.; Li, X. An Analysis of Decoding for Attention-Based End-to-End Mandarin Speech Recognition. In Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, Taiwan, 26–29 November 2018; pp. 384–388, doi:10.1109/ISCSLP.2018.8706686.

67. Hao, X.; Shan, C.; Xu, Y.; Sun, S.; Xie, L. An Attention-based Neural Network Approach for Single Channel Speech Enhancement. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6895–6899, doi:10.1109/ICASSP.2019.8683169.

68. Mandava, T.; Vuppala, A.K. Attention based Residual-Time Delay Neural Network for Indian Language Identification. In Proceedings of the 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2019; pp. 1–5, doi:10.1109/IC3.2019.8844889.

69. Giri, R.; Isik, U.; Krishnaswamy, A. Attention Wave-U-Net for Speech Enhancement. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 249–253, doi:10.1109/WASPA.2019.8937186.

70. Tjandra, A.; Sakti, S.; Nakamura, S. Attention-based Wav2Text with feature transfer learning. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 309–315, doi:10.1109/ASRU.2017.8268951.

71. Meng, Z.; Li, J.; Gong, Y. Attentive Adversarial Learning for Domain-invariant Training. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6740–6744, doi:10.1109/ICASSP.2019.8683486.

72. Tanaka, K.; Kameoka, H.; Kaneko, T.; Hojo, N. ATTTS2-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6805–6809, doi:10.1109/ICASSP.2019.8683282.

73. Huang, L.; Pun, C. Audio Replay Spoofer Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network. *IEEE/ACM Trans. Audio Speech Lang. Process.* 2020, 28, 1813–1825, doi:10.1109/TASLP.2020.2998870.

74. Demirel, E.; Ablübeck, S.; Dixon, S. Automatic Lyrics Transcription using Dilated Convolutional Neural Networks with Self-Attention. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8, doi:10.1109/IJCNN48605.2020.9207052.

75. Meng, Z.; Gaur, Y.; Li, J.; Gong, Y. Character-Aware Attention-Based End-to-End Speech Recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 949–955, doi:10.1109/ASRU46091.2019.9004018.
76. Yang, C.; Qi, J.; Chen, P.; Ma, X.; Lee, C. Characterizing Speech Adversarial Examples with Self-Attention U-Net Enhancement. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3107–3111, doi:10.1109/ICASSP40776.2020.9053288.

77. Zhou, T.; Zhao, Y.; Li, J.; Gong, Y.; Wu, J. CNN with Phonetic Attention for Text-Independent Speaker Verification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 718–725, doi:10.1109/ASRU46091.2019.9003826.

78. Li, C.; Jiao, J.; Zhao, Y.; Zhao, Z. Combining Gated Convolutional Networks and Self-Attention Mechanism for Speech Emotion Recognition. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, UK, 3–6 September 2019; pp. 105–109, doi:10.1109/ACIIW.2019.8925283.

79. Rajan, V.; Brutti, A.; Cavallaro, A. ConflictNET: End-to-End Learning for Speech-Based Conflict Intensity Estimation. IEEE Signal Process. Lett. 2019, 26, 1668–1672, doi:10.1109/LSP.2019.2944004.

80. Neumann, M.; Thang Vu, N.G. CROSS-lingual and Multilingual Speech Emotion Recognition on English and French. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5769–5773, doi:10.1109/ICASSP.2018.8462162.

81. An, N.N.; Thanh, N.Q.; Liu, Y. Deep CNNs With Self-Attention for Speaker Identification. IEEE Access 2019, 7, 85327–85337, doi:10.1109/ACCESS.2019.2917470.

82. Huang, C.; Narayanan, S.S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 583–588, doi:10.1109/ICME.2017.8019296.

83. Li, R.; Wu, Z.; Jia, J.; Zhao, S.; Meng, H. Dilated Residual Network with Multi-head Self-attention for Speech Emotion Recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6675–6679, doi:10.1109/ICASSP.2019.8682154.

84. Tanaka, T.; Masumura, R.; Moriya, T.; ObA, T.; Aono, Y. Disfluency Detection Based on Speech-Aware Token-by-Token Sequence Labeling with BLSTM-CRFs and Attention Mechanisms. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1009–1013, doi:10.1109/APSIPAASC47483.2019.9023119.

85. Moriya, T.; Sato, H.; Tanaka, T.; Ashihara, T.; Masumura, R.; Shinohara, Y. Distilling Attention Weights for CTC-Based ASR Systems. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6894–6898, doi:10.1109/ICASSP40776.2020.9053578.

86. Meng, Z.; Li, J.; Gaur, Y.; Gong, Y. Domain Adaptation via Teacher-Student Learning for End-to-End Speech Recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 268–275, doi:10.1109/ASRU46091.2019.9003776.

87. Hsiao, P.; Chen, C. Effective Attention Mechanism in Dynamic Models for Speech Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2526–2530, doi:10.1109/ICASSP.2018.8461431.

88. Hifny, Y.; Ali, A. Efficient Arabic Emotion Recognition Using Deep Neural Networks. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6710–6714, doi:10.1109/ICASSP.2019.8683632.

89. Cao, G.; Tang, Y.; Sheng, J.; Cao, W. Emotion Recognition from Children Speech Signals Using Attention Based Time Series Deep Learning. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 1296–1300, doi:10.1109/BIBM47256.2019.8982992.

90. Zhu, X.; Huang, H. End-to-End Amrdo-Tibetan Speech Recognition Based on Knowledge Transfer. IEEE Access 2020, 8, 170991–171000, doi:10.1109/ACCESS.2020.3023783.

91. Bahdanau, D.;chorowski, J.; serdyuk, D.; brakel, P.; bengio, Y. End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4945–4949, doi:10.1109/ICASSP.2016.7472618.

92. Monteiro, J.; Alam, J.; Falk, T.H. End-To-End Detection Of Attacks To Automatic Speaker Recognizers With Time-Attentive Light Convolutional Neural Networks. In Proceedings of the 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), Pittsburgh, PA, USA, 13–16 October 2019; pp. 1–6, doi:10.1109/MLSP.2019.8918703.

93. Padi, B.; Mohan, A.; Ganapathy, S. End-to-End Language Recognition Using Attention Based Hierarchical Gated Recurrent Unit Models. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5966–5970, doi:10.1109/ICASSP.2019.8683895.

94. Chang, X.; Zhang, W.; Qian, Y.; Roux, J.L.; Watanabe, S. End-To-End Multi-Speaker Speech Recognition With Transformer. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6134–6138, doi:10.1109/ICASSP40776.2020.9054029.

95. Sun, J.; Zhou, G.; Yang, H.; Wang, M. End-to-end Tibetan Ando dialect speech recognition based on hybrid CTC/attention architecture. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 628–632, doi:10.1109/APSIPAASC47483.2019.9023130.
96. Kim, C.; Kim, S.; Kim, K.; Kumar, M.; Kim, J.; Lee, K.; Han, C.; Garg, A.; Kim, E.; Shin, M.; et al. End-to-End Training of a Large Vocabulary End-to-End Speech Recognition System. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 562–569, doi:10.1109/ASRU46091.2019.9003976.

97. Norouzian, A.; Mazoure, B.; Connolly, D.; Willett, D. Exploring Attention Mechanism for Acoustic-based Classification of Speech Utterances into System-directed and Non-system-directed. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7310–7314, doi:10.1109/ICASSP.2019.8683565.

98. Ubale, R.; Qian, Y.; Eivanini, K. Exploring End-To-End Attention-Based Neural Networks For Native Language Identification. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 84–91, doi:10.1109/SLT.2018.8639689.

99. Yadav, S.; Rai, A. Frequency and Temporal Convolutional Attention for Text-Independent Speaker Recognition. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6794–6798, doi:10.1109/ICASSP40776.2020.9054440.

100. Mittag, G.; Möller, S. Full-Reference Speech Quality Estimation with Attentional Siamese Neural Networks. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 346–350, doi:10.1109/ICASSP40776.2020.9053951.

101. Liu, S.; Jiao, J.; Zhao, Z.; Dineley, J.; Cummins, N.; Schuller, B. Hierarchical Component-attention Based Speaker Turn Embedding for Emotion Recognition. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 June 2020; pp. 1–7, doi:10.1109/IJCNN48605.2020.9207374.

102. Xu, M.; Zhang, F.; Khan, S.U. Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion. In Proceedings of the 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 1058–1064, doi:10.1109/CCWC47524.2020.9031207.

103. Li, S.; Lu, X.; Takashima, R.; Shen, P.; Kawahara, T.; Kawai, H. Improving Very Deep Time-Delay Neural Network With Vertical-Attention For Effectively Training CTC-Based ASR Systems. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 77–83, doi:10.1109/SLT.2018.8639675.

104. Schumann, R.; Angkititrakul, P. Incorporating ASR Errors with Attention-Based, Jointly Trained RNN for Intent Detection and Slot Filling. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6059–6063, doi:10.1109/ICASSP.2018.8461598.

105. Kim, H.; Na, H.; Lee, H.; Lee, J.; Kang, T.G.; Lee, M.; Choi, Y.S. Knowledge Distillation Using Output Errors for Self-attention End-to-end Models. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6181–6185, doi:10.1109/ICASSP.2019.8682775.

106. Park, H.; Seo, S.; Rim, D.J.; Kim, C.; Son, H.; Park, J.; Kim, J. Korean Grapheme Unit-based Speech Recognition Using Attention-CTC Ensemble Network. In Proceedings of the 2019 International Symposium on Multimedia and Communication Technology (ISMAC), Quezon City, Philippines, 19–21 August 2019; pp. 1–5, doi:10.1109/ISMAC.2019.8836146.

107. Watanabe, S.; Hori, T.; Hershey, J.R. Language Independent end-to-end architecture for joint language identification and speech recognition. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 265–271, doi:10.1109/ASRU.2017.8268945.

108. Millet, J.; Zeghidour, N. Learning to Detect Dysarthria from Raw Speech. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5831–5835, doi:10.1109/ICASSP.2019.8682324.

109. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–23 March 2016; pp. 4960–4964, doi:10.1109/ICASSP.2016.7472621.

110. Cui, D.; Yin, S.; Gu, J.; Liu, L.; Wei, S. MSAM: A Multi-Layer Bi-LSTM Based Speech to Vector Model with Residual Attention Mechanism. In Proceedings of the 2019 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC), Xi’an, China, 12–14 June 2019; pp. 1–3, doi:10.1109/EDSSC.2019.8753946.

111. Nediyanchath, A.; Paramasivam, P.; Yenigalla, P. Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7179–7183, doi:10.1109/ICASSP40776.2020.9054073.

112. Vuddagiri, R.K.; Mandava, T.; Vydana, H.K.; Vuppala, A.K. Multi-Head Self-Attention Networks for Language Identification. In Proceedings of the 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2019; pp. 1–5, doi:10.1109/IC3.2019.8844925.

113. Li, M.; Lan, T.; Peng, C.; Qian, Y.; Liu, Q. Multi-layer Attention Mechanism Based Speech Separation Model. In Proceedings of the 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi’an, China, 16–19 October 2019; pp. 506–509, doi:10.1109/ICCT46805.2019.8947242.

114. Li, H.; Kang, Y.; Ding, W.; Yang, S.; Yang, S.; Huang, G.Y.; Liu, Z. Multimodal Learning for Classroom Activity Detection. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 9234–9238, doi:10.1109/ICASSP40776.2020.9054407.
115. Ubale, R.; Ramanarayanan, V.; Qian, Y.; Evanini, K.; Leong, C.W.; Lee, C.M. Native Language Identification from Raw Waveforms Using Deep Convolutional Neural Networks with Attentive Pooling. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 403–410. doi:10.1109/ASRU46091.2019.9003872.

116. Tanaka, T.; Masumura, R.; Moriya, T.; Aono, Y. Neural Speech-to-Text Language Models for Rescoring Hypotheses of DNN-HMM Hybrid Automatic Speech Recognition Systems. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 196–200. doi:10.23919/APSIPA.2018.8659622.

117. Hu, J.; Yu, C.; Guan. F. Non-parallel Many-to-many Singing Voice Conversion by Adversarial Learning. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 125–132, doi:10.1109/APSIPAASC47483.2019.9023357.

118. Pan, J.; Liu, D.; Wan, G.; Du, J.; Liu, Q.; Ye, Z. Online Speaker Adaptation for LVCSR Based on Attention Mechanism. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 183–186, doi:10.23919/APSIPA.2018.8659609.

119. Pan, J.; Wan, G.; Du, J.; Ye, Z. Online Speaker Adaptation Using Memory-Aware Networks for Speech Recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 2020, 28, 1025–1037, doi:10.1109/TASLP.2020.2980372.

120. Zhang, Y.; Duan, Q.; Liao, Y.; Liu, J.; Wu, R.; Xie, B. Research on Speech Enhancement Algorithm Based on SA-Unet. In Proceedings of the 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCCE), Hohhot, China, 25–27 October 2019; pp. 818–813, doi:10.1109/ICMCCE48743.2019.00187.

121. Sun, X.; Jiang, Y.; Li, W. Residual Attention Based Network for Automatic Classification of Phonation Modes. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6, doi:10.1109/ICME46284.2020.9102952.

122. Pinkas, G.; Karny, Y.; Malachi, A.; Barkai, G.; Bachar, G.; Aharonson, V. SARS-CoV-2 Detection from Voice. IEEE Open J. Eng. Med. Biol. 2020, 1, doi:10.1109/OJEMB.2020.3026468.

123. Dong, L.; Wang, F.; Xu, B. Self-attention Aligner: A Latency-control End-to-end Model for ASR Using Self-attention Network and Chunk-hopping. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5656–5660, doi:10.1109/ICASSP.2019.8682954.

124. Huang, Y.; Thomas, S.; Suzuki, M.; Tüske, Z.; Sansone, L.; Picheny, M. Semi-Supervised Training and Data Augmentation for Adaptation of Automatic Broadcast News Captioning Systems. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 867–874, doi:10.1109/ASRU46091.2019.9003943.

125. Zhang, Y.; Yu, M.; Li, N.; Yu, C.; Cui, J.; Yu, D. Seq2Seq Attentional Siamese Neural Networks for Text-dependent Speaker Verification. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6131–6135, doi:10.1109/ICASSP.2019.8682676.

126. Fu, C.; Dissanayake, T.; Hosoda, K.; Maekawa, T.; Ishiguro, H. Similarity of Speech Emotion in Different Languages Revealed by a Neural Network with Attention. In Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 3–5 February 2020; pp. 381–386, doi:10.1109/ICSC.2020.00076.

127. Jalal, M.A.; Moore, R.K.; Hain, T. Spatio-Temporal Context Modelling for Speech Emotion Classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 853–859, doi:10.1109/ASRU46091.2019.9004037.

128. Ochiai, T.; Watanabe, S.; Katagiri, S.; Horii, T.; Hershey, J. Speaker Adaptation for Multichannel End-to-End Speech Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6707–6711, doi:10.1109/ICASSP.2018.8462161.

129. Bansal, S.; Malhotra, K.; Ganapathy, S. Speaker and Language Aware Training for End-to-End ASR. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 494–501, doi:10.1109/ICASSP.2019.9004000.

130. Wei, C.; Sun, X.; Tian, F.; Ren, F. Speech Emotion Recognition with Hybrid Neural Network. In Proceedings of the 2019 5th International Conference on Big Data Computing and Communications (BIGCOM), Qingdao, China, 9–11 August 2019; pp. 298–302, doi:10.1109/BIGCOM.2019.00051.

131. Han, K.J.; Prieto, R.; Ma, T. State-of-the-Art Speech Recognition Using Multi-Stream Self-Attention with Dilated 1D Convolutions. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 54–61, doi:10.1109/ASRU46091.2019.9003730.

132. Chiu, C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778, doi:10.1109/ICASSP.2018.8462105.

133. Moritz, N.; Horii, T.; Le, J. Streaming Automatic Speech Recognition with the Transformer Model. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6074–6078, doi:10.1109/ICASSP40776.2020.9054476.
134. Moritz, N.; Hori, T.; Roux, J.L. Streaming End-to-End Speech Recognition with Joint CTC-Attention Based Models. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 936–943, doi:10.1109/ASRU46091.2019.903920.

135. 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. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7829–7833, doi:10.1109/ICASSP40776.2020.9053896.

136. Moritz, N.; Hori, T.; Roux, J.L. Triggered Attention for End-to-end Speech Recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5666–5670, doi:10.1109/ICASSP.2019.8683510.

137. Cui, W.; Cai, D.; Huang, S.; Li, M. Utterance-level End-to-end Language Identification Using Attention-based CNN-BLSTM. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5991–5995, doi:10.1109/ICASSP.2019.8682386.

138. Kim, J.; Hahn, M. Voice Activity Detection Using an Adaptive Context Attention Model. IEEE Signal Process. Lett. 2018, 25, 1181–1185, doi:10.1109/LSP.2018.2811740.

139. Li, N.; Ge, M.; Wang, L.; Dang, J. A Fast Convolutional Self-attention Based Speech Dereverberation Method for Robust Speech Recognition. In Neural Information Processing; Gedeon, T., Wong, K.W., Lee, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; Volume 11955, pp. 295–305, doi:10.1007/978-3-030-36718-3_25.

140. Qin, C.X.; Zhang, W.L.; Qu, D. A new joint CTC-attention-based speech recognition model with multi-level multi-head attention. EURASIP J. Audio Speech Music Process. 2019, 18, 18, doi:10.1186/s13636-019-0161-0.

141. Miao, X.; McLaughlin, I.; Yan, Y. A New Time-Frequency Attention Tensor Network for Language Identification. Circuits Syst. Signal Process. 2020, 39, 2744–2758, doi:10.1007/s11034-019-01286-9.

142. Alex, S.B.; Mary, L.; Babu, B.P. Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. Circuits Syst. Signal Process. 2020, 39, 5681–5709, doi:10.1007/s11034-020-01429-3.

143. Kürzinger, L.; Chavez Rosas, E.R.; Li, L.; Watzel, T.; Rigoll, G. Audio Adversarial Examples for Robust Hybrid CTC/Attention Recognition Model with Transformer Encoders and RNN-T Loss. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7829–7833, doi:10.1109/ICASSP40776.2020.9053896.

144. Markovnikov, N.; Kipyatkova, I. Markovnikov, N. Experimenting with Attention Mechanisms in Joint CTC-Attention Models for Russian Speech Recognition. In Speech and Computer; Karpov, A., Potapova, R., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12335, pp. 214–222, doi:10.1007/978-3-030-60276-5_22.

145. Gallardo-Antolín, A.; Montero, J.M. External Attention LSTM Models for Cognitive Load Classification from Speech. In Statistical Language and Speech Processing; Martín-Vide, C., Purver, M., Pollak, S., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12335, pp. 255–266, doi:10.1007/978-3-030-60276-5_26.

146. Qiu, Y.; Ma, Y.; Jin, Y.; Li, S.; Gu, M. Chinese Dialects Identification Using Attention-Based Deep Neural Networks. In Communications, Signal Processing, and Systems; Liang, Q.; Mu, J., Jia, M., Wang, W., Feng, X., Zhang, B., Eds.; Lecture Notes in Electrical Engineering; Springer: Singapore, 2019; Volume 463, pp. 2051–2058, doi:10.1007/978-981-10-6571-2_250.

147. Cheng, J.; Liang, R.; Zhao, L. DNN-based speech enhancement with self-attention on feature dimension. Multimed. Tools Appl. 2020, doi:10.1007/s11042-020-09345-z.

148. Cui, Z., Pan, J., Zhang, S., Xiao, L., Yang, J., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; Volume 11816, pp. 139–150, doi:10.1007/978-3-030-31372-2_12.

149. Zhou, K.; Yang, Q.; Sun, X.; Liu, S.; Lu, J. Improved CTC-Attention Based End-to-End Speech Recognition on Air Traffic Control. In Proceedings of the 34th International Conference on Machine Learning, ICML’17, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 2837–2846.
155. Chen, J.-N.; Gao, S.; Sun, H.-Z.; Liu, X.-H.; Wang, Z.-N.; Zheng, Y. An End-to-end Speech Recognition Algorithm based on Attention Mechanism. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Hefei, China 27–29 July 2020; pp. 2935–2940, doi:10.23919/CCC50068.2020.9189026.

156. Battenberg, E.; Chen, J.; Child, R.; Coates, A.; Li, Y.G.Y.; Liu, H.; Satheesh, S.; Sriram, A.; Zhu, Z. Exploring neural transducers for end-to-end speech recognition. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 206–213, doi:10.1109/ASRU.2017.8268937.

157. San, L.; Moritz, N.; Hori, T.; Roux, J.L. Unsupervised Speaker Adaptation Using Attention-Based Speaker Memory for End-to-End ASR. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7384–7388, doi:10.1109/ICASSP40776.2020.9054249.

158. Chazan, S.E.; Gannot, S.; Goldberger, J. Attention-Based Neural Network for Joint Diarization and Speaker Extraction. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 301–305, doi:10.1109/IWAENC.2018.8521259.

159. Shan, C.; Weng, C.; Wang, G.; Su, D.; Luo, M.; Yu, D.; Xie, L. Component Fusion: Learning Replaceable Language Model Component for End-to-end Speech Recognition System. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5361–5635, doi:10.1109/ICASSP.2019.8682490.

160. Koizumi, Y.; Yatabe, K.; Delcroix, M.; Masuyama, Y.; Takeuchi, D. Speech Enhancement Using Self-Adaptation and Multi-Head Self-Attention. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 181–185, doi:10.1109/ICASSP40776.2020.9053214.

161. Xia, T.; Rui, X.; Huang, C.; Chu, I.H.; Wang, S.; Han, M. An Attention Based Deep Neural Network for Automatic Lexical Stress Detection. In Proceedings of the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Ottawa, ON, Canada, 11–14 November 2019; pp. 1–5, doi:10.1109/GlobalSIP45357.2019.8969232.

162. Chiu, C.; Han, W.; Zhang, Y.; Pang, R.; Kishchenko, S.; Nguyen, P.; Narayanan, A.; Liao, H.; Zhang, S.; Kannan, A.; et al. A Comparison of End-to-End Models for Long-Form Speech Recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 889–896, doi:10.1109/ASRU46091.2019.9003854.

163. Miao, H.; Cheng, G.; Zhang, P.; Yan, Y. Online Hybrid CTC/Attention End-to-End Automatic Speech Recognition Architecture. IEEE Acm Trans. Audio Speech Lang. Process. 2020, 28, 1452–1465, doi:10.1109/TASLP.2020.2987752.

164. Doetsch, P.; Hannemann, M.; Schluter, R.; Ney, H. Inverted Alignments for End-to-End Automatic Speech Recognition. IEEE J. Sel. Top. Signal Process. 2017, 11, 1265–1273, doi:10.1109/JSTSP.2017.2752691.

165. Audhkhasi, K.; Rosenberg, A.; Saon, G.; Sethy, A.; Ramabhadran, B.; Chen, S.; Picheny, M. Recent Progress in Deep End-to-End Models for Spoken Language Processing. IBM J. Res. Dev. 2017, 61, 2:1–2:10, doi:10.1147/JRD.2017.2701207.