Transformers in Speech Processing: A Survey

Siddique Latif[1], Aun Zaidi[2], Heriberto Cuayáhuitl[3], Fahad Shamshad[4], Moazzam Shoukat[5], and Junaid Qadir[6]

[1]Queensland University of Technology (QUT), Australia
[2]Information Technology University, Pakistan
[3]University of Lincoln, UK
[4]Mohamed bin Zayed University of Artificial Intelligence
[5]EmulationAI
[6]Qatar University, Doha

Abstract—The remarkable success of transformers in the field of natural language processing has sparked the interest of the speech-processing community, leading to an exploration of their potential for modeling long-range dependencies within speech sequences. Recently, transformers have gained prominence across various speech-related domains, including automatic speech recognition, speech synthesis, speech translation, speech para-linguistics, speech enhancement, spoken dialogue systems, and numerous multimodal applications. In this paper, we present a comprehensive survey that aims to bridge research studies from diverse subfields within speech technology. By consolidating findings from across the speech technology landscape, we provide a valuable resource for researchers interested in harnessing the power of transformers to advance the field. We identify the challenges encountered by transformers in speech processing while also offering insights into potential solutions to address these issues.

I. INTRODUCTION

Transformers have garnered significant attention in the speech processing and natural language processing (NLP) communities [1]–[6] owing to their remarkable performance across a spectrum of applications, including machine translation [7], automatic speech recognition (ASR) [8], [9], question answering [10], speech enhancement [11], speech emotion recognition [12], and speech separation [13], to name a few. These models have even surpassed traditional recurrent neural networks (RNNs), that struggle with long sequences and the vanishing gradient problem on sequence-to-sequence tasks [1]. The rapid development and popularity of transformer-based models in speech processing have generated a wealth of literature investigating the unique features that underlie their superior performance.

Transformers have an advantage in comprehending speech, as they analyze the entire sentence at once, whereas RNNs can only process smaller sections at a time. This is made possible by the unique self-attention-based architecture of transformers [14], which enables them to learn long-term dependencies, which is critical for speech processing tasks. Moreover, the multi-head attention mechanism [15]—a specialized feature in transformers—allows for more efficient parallelization during training, making them ideal for handling large datasets, which is a common challenge in speech processing tasks. This

unique combination of self-attention and multi-head attention empowers transformers to achieve exceptional performance in sequence-to-sequence modeling, making them an indispensable tool for researchers and practitioners in the field of speech processing.

As the use of transformers for speech processing research community is gaining popularity, it is timely to review the existing literature and present a comprehensive overview of the field. In this regard, we provide a comprehensive overview of transformer model applications in the speech processing domain. Our aim is to assist researchers and practitioners in grasping the major trends and recent advancements in the field. Specifically, the main contributions of this survey are as follows:

- We present the first comprehensive survey of the application of transformer models in the speech processing field. Our survey covers more than 100 papers to cover the recent progress.
- We provide detailed coverage of this rapidly evolving field by categorizing the papers based on their applications Specifically, these applications include automatic speech recognition, neural speech synthesis, speech translation, speech enhancement, multi-modal applications, and spoken dialogue systems.
- Finally, based on our thorough analysis, we identify various challenges and propose future research directions. Moreover, we provide valuable insights into potential solutions based on the literature reviewed.

We compare our paper with recent surveys on transformers and speech processing in Table I. It can be found that most of the transformer-related survey papers are focused on computer vision and natural language processing (NLP). The articles focused on speech processing do not cover transformers. Here, we focus on recent development in speech technology using transformers. Although other recent surveys have focused on deep learning techniques for SR [16], ASR [17], [18], and SER [19], [20], none of these has focused on transformers for speech processing. This study bridges this gap by presenting an up-to-date survey of research that focused on speech processing using transformers. The paper is organised as follows. Section II provides an overview of the applications of seq2seq models in SP and introduces the salient concepts underlying transformers.

Email: siddique.latif@usq.edu.au

Section III presents a comprehensive review of the applications of transformer models in SP. Section IV discusses open problems and future research directions. Finally, in Section V, we summarize and conclude the paper.

II. Background

In this section, we will provide a comprehensive overview of transformer architecture, starting with a brief overview of sequential models and their limitations in handling sequential data. We will then delve into the key concepts behind the transformer's operation, highlighting the unique features that enable it to outperform traditional recurrent neural networks. Lastly, we will discuss popular transformers for speech processing.

A. Sequential Models for Speech Processing

Early deep learning approaches in the SP domain typically employed variants of convolutional neural networks (CNNs) [37], [38]. However, a drawback of these CNN-based approaches is their inability to capture the sequential nature of speech data. This limitation of CNNs led to the development of sequence-to-sequence (seq2seq) architectures, such as RNNs and long short-term memory networks (LSTMs), which are specifically designed for sequential data. RNNs are well-suited for sequential data because they can process long sequences step-by-step with limited memory of previous sequence elements. More recently, researchers have also combined the strengths of CNNs and RNNs by using CNNs to extract audio features and using these features as input to train RNNs. However, RNNs have been shown to have issues with the vanishing or exploding gradient problem. To address this issue, LSTMs use a gating mechanism and memory cells to control the flow of information and alleviate gradient problems. Many LSTM variations—such as Frequency-LSTM, Time-Frequency LSTMs, Bi-directional LSTMs, ConvLSTMs, and Stacked LSTMs—have been proposed for SP tasks. Despite their effectiveness, seq2seq models are limited in important ways: they cannot take advantage of parallel computing hardware and have difficulties modeling long-term context.

B. Overview of Transformers

Transformers were first introduced in the seminal work of Vaswani et al. titled "Attention Is All You Need" [14] for machine translation tasks in Natural Language Processing (NLP) and have recently shown impressive performance in other domains, including computer vision, medical imaging, and remote sensing. Transformer models use self-attention layers to effectively capture long-range dependencies among the input sequence, which is in contrast to traditional recurrent neural networks that struggle to capture such interaction. Furthermore, self-attention allows for more parallelization compared to recurrent neural networks, as these can process the speech sequence as a whole without relying on past states to capture dependencies. More specifically, two types of attention were introduced by Vaswani et al. including (1) scaled dot-product attention, and (2) multi-head attention. In addition, positional encoding is also used to inject information about the relative or absolute position of the tokens in the sequence. Due to these desirable properties, transformers have garnered immense interest in the speech community, and several approaches have been proposed that build upon transformers. We provide next a brief overview of the core components of transformers, which are multi-head self-attention layers and a position-wise feed-forward network (positional encoding).

1) Self-Attention Layer: Self-attention (SA) layer aims to capture the internal correlation of sequence or features by aggregating global information from the entire input sequence, a task that is challenging for conventional recurrent models. Specifically, the input $X \in \mathbb{R}^{N \times D}$ consisting of N entities each having dimension D is transformed into query ($Q \in \mathbb{R}^{N \times D_k}$), key ($K \in \mathbb{R}^{N \times D_k}$) and value ($V \in \mathbb{R}^{N \times D_k}$) matrices via learnable weight matrices $W^Q \in \mathbb{R}^{D \times D_k}$, $W^K \in \mathbb{R}^{D \times D_k}$, and $W^V \in \mathbb{R}^{D \times D_k}$ respectively. Mathematically,

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V. \quad (1)$$

Then the dot-product of the query matrix Q with all the keys Q in a given sequence is computed and the resulting matrix is normalized using the softmax operator to get the attention matrix $A \in \mathbb{R}^{N \times N}$ as

$$A = softmax\left(\frac{QK^T}{\sqrt{D_k}}\right)V. \quad (2)$$

The output of the SA layer Z is the attention matrix A multiplied by the value matrix V

$$Z = AV. \quad (3)$$

2) Masked Self-Attention: In the original transformers paper [14], the SA blocks used in the decoder are masked to prevent attending to the subsequent future entities by element-wise multiplication of the mask $M \in \mathbb{R}^{N \times N}$ as:

$$Z = softmax\left(\frac{QK^T}{\sqrt{d_k}} \circ M\right)V, \quad (4)$$

where M is the upper triangular matrix and $\circ$ denotes the Hadamard product. This is called masked self-attention.

3) Multi-Head Attention: Rather than only computing the attention once, Multi-Head Self-Attention (MHSA) consists of multiple SA blocks. These SA blocks are concatenated together channel-wise to model dependencies among different elements in the input sequence. Each head in MHSA has its own learnable weight matrices denoted by $\{\mathbf{W}^{Q_i}, \mathbf{W}^{K_i}, \mathbf{W}^{V_i}\}$, where $i = 0 \cdots (h-1)$ and h denotes total number of heads in MHSA block. Specifically,

$$\text{MHSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{Z}_0, \mathbf{Z}_1, ..., \mathbf{Z}_{h-1}]\mathbf{W}^O,$$

whereas $\mathbf{W}^O \in \mathbb{R}^{h.D_k \times N}$ computes linear transformation of heads and $\mathbf{Z}_i$ can be written as,

$$\mathbf{Z}_i = softmax\left(\frac{\mathbf{Q}\mathbf{W}^{Q_i}(\mathbf{K}\mathbf{W}^{K_i})^T}{\sqrt{D_k/h}}\right)\mathbf{V}\mathbf{W}^{V_i}.$$

TABLE I: Comparison of this paper with other recent survey papers. Where Speech Processing = SP, Speech Emotion Recognition (SER), Natural Language Processing = NLP, Computer Vision = CV, Action Recognition = AR, and Reinforcement Learning = RL

Author (year)	Speech Focused	Transformer Focused	Domain	Details
Deng et al. 2016 [21]	✓	×	SP	The paper reviews self-supervised representation learning methods in speech processing, discussing generative, contrastive, and predictive methods and multi-modal data approaches.
Khalil et al. 2019 [19]	✓	×	SER	This paper provides an overview of deep learning techniques for speech based emotion recognition, covering databases used, emotions extracted, contributions made, and limitations related to it.
Nassif et al. 2019 [18]	✓	×	SP	This survey paper reviews the use of deep learning for speech-related applications, providing a statistical analysis of 174 papers published between 2006 and 2018.
Alam et. al. 2020 [22]	×	×	Multimodal	The survey covers DNN architectures, algorithms, and systems for speech and vision applications, not limited to transformers or speech.
Bracsoveanu et al. 2020 [23]	×	✓	NLP	The survey paper focuses on explaining Transformer architectures through visualizations to provide better understanding and proposes a set of requirements for future Transformer visualization frameworks.
Tan et al. 2021 [24]	✓	×	SP	This paper provides a comprehensive survey on neural text-to-speech, covering key components such as text analysis, acoustic models, as well as advanced topics like fast, low-resource, robust, expressive, and adaptive TTS, and it also discusses future research directions.
Alharbi et al. 2021 [25]	✓	×	ASR	This survey paper provides a systematic review of automatic speech recognition (ASR) technology, covering significant topics and recent challenges published in the last six years.
Malik et al. 2021 [26]	✓	×	ASR	The survey paper compares various deep learning techniques and feature extraction methods for ASR and discusses the impact of different speech datasets on ASR performance, providing online resources and language models for ASR formulation.
Liu et al. 2021 [27]	×	✓	CV	The survey explores Transformer-based architectures in CV tasks, proposing a taxonomy and evaluating and comparing existing methods. It suggests three research directions for future investment.
Xu et al. 2022 [28]	×	✓	Multimodal	The paper surveys Transformer techniques in multimodal learning, including theoretical reviews and applications. It aims to provide insights for researchers and practitioners.
Lin et al. 2022 [2]	×	✓	Multimodal	The survey reviews various Transformer variants in AI fields and proposes a new taxonomy. It covers architectural modifications, pre-training, applications, and potential directions for future research.
Shamshad et al. 2022 [29]	×	✓	Medical Imaging	The paper reviews Transformer models in medical imaging, discussing their applications and identifying open problems and future directions.
Acheampong et al. 2022 [30]	×	✓	NLP	The paper reviews Transformer-based models used for NLP tasks in emotion recognition, highlighting their strengths and limitations, and providing future research directions.
Khan et al. 2022 [31]	×	✓	CV	The paper reviews Transformer models in CV tasks, covering a wide range of tasks and comparing the advantages and limitations of popular techniques. It also discusses research directions and future works.
Tay et al. 2022 [3]	×	✓	Multimodal	The article provides an overview of "X-former" models in multiple domains, aimed at improving efficiency and helping researchers navigate the evolving field.
Aleissaee et al. 2022 [32]	×	✓	Remote Sensing	The paper reviews transformers-based methods for remote sensing problems. It also discusses different challenges and open issues.
Ulhaq et al. 2022 [33]	×	✓	AR	It reviews literature on vision transformer techniques for action recognition, providing taxonomies, network learning strategies, and evaluation metrics, while discussing challenges and future research directions.
Bhangale et al. 2022 [34]	✓	×	SP	The paper presents a survey of deep learning techniques for various speech processing applications. It covers various deep learning models such as Auto-Encoder, GAN, RBN, DBN, DNN, CNN, RNN, and DRL, along with speech databases and evaluation metrics.
Lahoud et al. 2023 [35]	×	✓	3D Vision	It reviews over 100 transformer methods on various 3D CV tasks, comparing their performance to common non-transformer methods on 3D benchmarks. It also discusses open directions and challenges.
Li et al. 2023 [36]	×	✓	RL	The paper reviews recent advances and applications of transformers in reinforcement learning, providing a taxonomy of existing works in the field and summarizing future prospects.
This paper	✓	✓	SP	The paper reviews applications of transformers and challenges faced in various speech processing applications, like speech recognition, synthesis, translation, and enhancement, and suggests future research directions for improving speech technology with transformers.

4) Positional Encoding: The self-attention mechanism in transformer models allows for input speech frames to be processed in no particular order or position. To account for this, positional encoding is used to provide the transformer model with information about the order of the input sequence. This is done by associating each position in the input sequence with a vector that helps the transformer learn positional relationships. Positional encoding can be learned during training or pre-defined and can be encoded in relative or absolute ways for SP tasks.

C. Popular Transformers for Speech

Transformers are a novel neural network architecture that relies solely on attention mechanisms to handle sequential
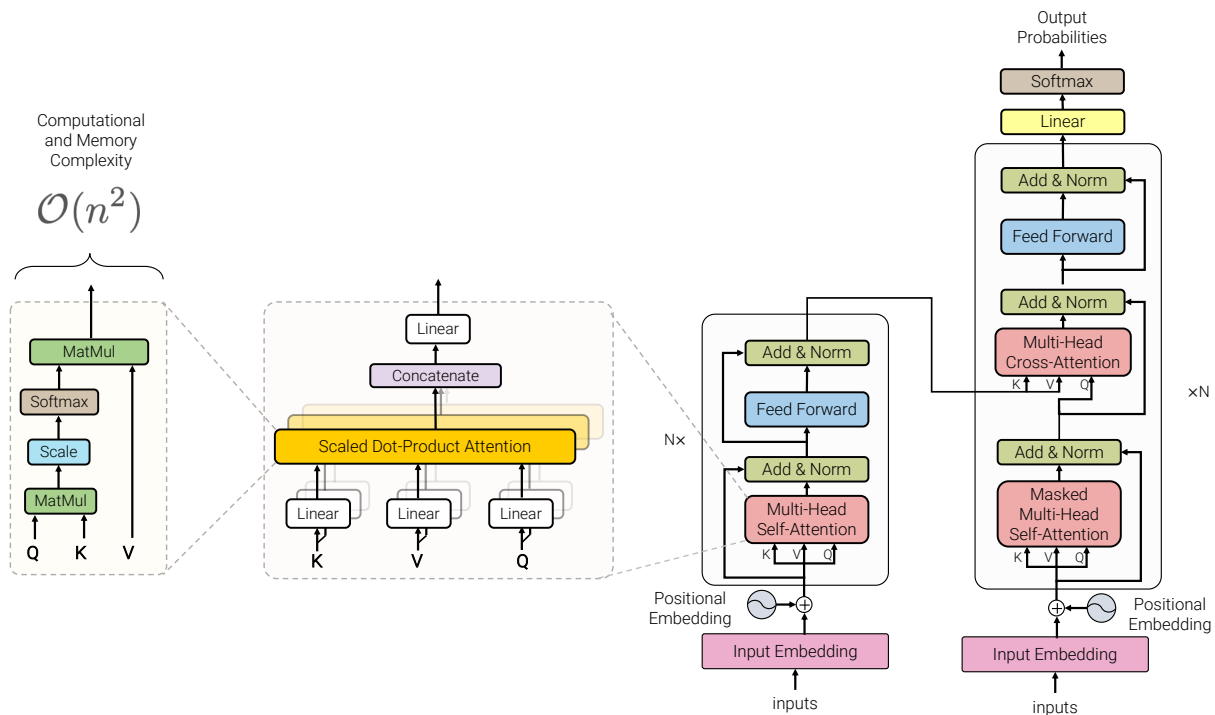
Fig. 1: Architecture of the standard transformer (Adapted from Vaswani et al., 2017 [14] and Tay et al., 2020 [3]). *The model comprises encoder and decoder layers, each with stacked self-attention and feed-forward sub-layers. The encoder produces hidden states from an input token sequence, while the decoder generates predictions from an output token sequence and attends to the encoder's states for input information.*

data, such as natural language and speech, for various tasks. Since the seminal work of Vaswani et al. (2017) [14], many extensions and applications of transformers have been developed for natural language processing (NLP) tasks, such as language modeling, question answering, sentiment analysis, text generation, etc. Transformers are also becoming increasingly popular in the speech community due to their suitability for various tasks including speech recognition, enhancement, text-to-speech synthesis, speaker recognition, and multi-microphone processing. Various open-source libraries including Hugging Face, SpeechBrain, and torch audio are accelerating the research in the speech domain. Big tech companies like Google, Meta, Amazon, etc., are building large speech domain-related transformer models. While transformers were initially developed for NLP tasks, they have since been adapted for other data types, including speech. BERT [39] is a language model that uses masked-language modeling (MLM) as its pre-training objective. BERT consists of two modules: an embedding layer that maps tokens to vectors, and an encoder layer that applies self-attention and feed-forward networks to learn contextualized representations. While BERT and similar text-based large language models (e.g., GPT [40]), XLNet [41], T5 [42]), etc.) have been successful in various NLP tasks, their application to speech processing is limited due to several shortcomings. For instance, they require discrete tokens as input, which means it needs a tokenizer or a speech recognition system to convert raw audio signals into text, introducing errors and noise in the process that can negatively impact the performance

[43]. Additionally, these models are pre-trained on large-scale text corpora, which may not match the domain or style of speech data, leading to domain mismatch issues.

1) wav2vec: To overcome these limitations, dedicated frameworks for learning speech representations, such as wav2vec, have been developed. wav2vec uses a self-supervised training approach that leverages the contrastive predictive coding (CPC) loss function to learn speech representations without the need for transcription or segmentation [44], [45]. This approach allows wav2vec to achieve state-of-the-art performance on several speech processing tasks, including speech recognition, speaker recognition, and spoken language understanding, among others [44], [45]. w2v-BERT is a framework that combines contrastive learning and MLM for self-supervised speech pre-training and builds on the success of wav2vec. w2v-BERT consists of three modules: a feature encoder, a quantization module, and a masked prediction module. The feature encoder is similar to wav2vec, but the quantization module discretizes the continuous speech representations into a finite set of speech units using Gumbel-softmax. The main differences between wav2vec and w2v-BERT lie in their pre-training objectives and the data types they operate on. wav2vec focuses on raw audio signals and uses contrastive learning to learn speech representations. In contrast, w2v-BERT operates on discrete tokens and it uses both MLM and contrastive learning as pre-training objectives. After the success of the wav2vec models, Baevski et al. demonstrated that it could be fine-tuned with a small amount of labeled data to achieve state-of-the-art results

on speech recognition tasks [46], [47]. This breakthrough led to the development of a series of models aimed at building cross-lingual or multilingual speech recognition systems using pre-trained transformer models. Another model is XLS-R [48], a large-scale model for cross-lingual speech representation learning based on wav2vec 2.0. XLS-R is trained on nearly half a million hours of publicly available speech audio in 128 languages and achieves state-of-the-art results on a wide range of tasks, domains, data regimes, and languages. These models leverage large-scale multilingual data augmentation and contrastive learning techniques to learn universal speech representations that can be transferred across languages and domains.

2) data2vec: While the success of wav2vec was inspired by the achievements of BERT, it paved the way for the development of other dedicated frameworks that use transformers to learn representations from multi-modal data. One such example is data2vec, which aims to learn multi-modal representations of data, including speech, images, and text, using a contrastive learning objective [49]. Similar to wav2vec, data2vec uses a self-supervised training approach that does not require labels or annotations and learns representations by maximizing agreement between differently augmented views of the same data sample. However, unlike wav2vec, which focuses solely on speech signals, data2vec can operate on various types of data and can learn joint representations that capture cross-modal correlations and transfer knowledge across modalities. Data2vec's self-supervised training approach allows it to learn representations without the need for labeled data, making it a scalable and cost-effective solution for many applications. Data2vec has been shown to outperform other unsupervised approaches for learning multimodal representations, such as Skip-thought [50] and Paragraph Vector [51], on several benchmark datasets [49]. However, it should be noted that while data2vec is suitable for learning representations that generalize across domains and modalities, it may not perform as well as domain-specific models for certain tasks, such as speech recognition or speaker identification, where the data has a specific domain or language.

3) Whisper: Whisper [52] is a general-purpose model designed for speech recognition in noisy or low-resource settings, and is capable of performing multiple speech-related tasks. Whisper uses weak supervision and a minimalist approach to data pre-processing. It achieves state-of-the-art results, showcasing the potential of using advanced machine-learning techniques in speech processing. Whisper is capable of performing multilingual speech recognition, speech translation, and language identification. It is trained on a large dataset of diverse audio and is a multitasking model that can handle various speech-related tasks, such as transcription, voice assistants, education, entertainment, and accessibility. The Whisper model is unique in that it uses a minimalist approach to data pre-processing, allowing models to predict the raw text of transcripts without significant standardization. This eliminates the need for a separate inverse text normalization step to produce naturalistic transcriptions, simplifying the speech recognition pipeline. The resulting models can generalize well to standard benchmarks and are competitive with prior fully supervised results without fine-tuning.

4) Tacotron: Transformer-based models have also gained popularity in speech synthesis tasks. A prime example of such a model is Tacotron [53], which uses a sequence-to-sequence architecture with attention mechanisms to generate high-quality speech from text input. The limitations of the Griffin-Lim algorithm used for audio signal generation led to the development of Tacotron 2 [54] by Google AI in 2018, which used WaveNet to generate raw audio waveforms directly from mel-spectrograms, resulting in more natural-sounding speech. Furthermore, Microsoft introduced Transformer TTS [55] in 2019, which employs a transformer network instead of the convolutional and recurrent networks used in Tacotron 2, along with a duration predictor and a new training method that uses teacher forcing for faster convergence and better performance. Despite these advancements, current systems still have limitations in generating natural-sounding speech for non-English languages, handling complex intonations and accents, and real-time speech synthesis for applications such as voice assistants and automated phone systems.

5) VALL-E: VALL-E [78] is another model that has gained attention, a zero-shot text-to-speech synthesis system that uses a language modeling approach, treating TTS as a conditional language modeling task rather than continuous signal regression. It is trained using discrete codes from an off-the-shelf neural audio codec model and pre-trained on 60,000 hours of English speech data, providing strong in-context learning capabilities. Unlike previous TTS systems, VALL-E does not require additional structure engineering, pre-designed acoustic features, or fine-tuning. It can synthesize high-quality personalized speech with only a 3-second acoustic prompt from an unseen speaker. The model also provides diverse outputs with the same input text and can preserve the acoustic environment and the speaker's emotion of the acoustic prompt. VALL-E's speaker dimension is built on a generalized TTS system, leveraging a large amount of semi-supervised data. This approach is significant, as scaling up semi-supervised data has been underestimated for TTS. Evaluation results show that VALL-E significantly outperforms the state-of-the-art zero-shot TTS system on LibriSpeech and VCTK datasets in terms of speech naturalness and speaker similarity. VALL-E X [79] was developed as a natural extension of VALL-E to address the challenge of cross-lingual speech synthesis. Cross-lingual speech synthesis involves generating speech in a target language, using a source language speech prompt and a target language text prompt. While VALL-E was designed for zero-shot TTS in English, there was a need for a model that could handle cross-lingual speech synthesis in multiple languages. VALL-E X, therefore, extends VALL-E to support cross-lingual synthesis by training a multi-lingual model to predict acoustic token sequences in the target language using prompts in both source and target languages. This enables the model to generate high-quality speech in the target language while preserving the voice, emotion, and acoustic environment of the unseen speaker and effectively alleviates the foreign accent problem, which can be controlled by a language ID.

6) Conformer: Recent advances in transformers for speech processing have also seen the emergence of the Conformer

TABLE II: Transformer Speech Models: Release Year and Parameter Count with Task Compatibility. * means that the parameters for this model cannot be found or are proprietary.

Model Name	Release Year	Number of Parameters	Tasks			Multimodal
			Speech Synthesis (TTS)	Speech Translation (ST)	Automatic Speech Recognition (ASR)	
Tacotron [53]	2017	13 million [56]	✓	×	×	×
Tacotron 2 [54]	2017	28.2 million [57]	✓	×	×	×
Transformer-TTS [55]	2018	30.7 million [58]	✓	×	×	×
vq-wav2vec [44]	2019	34 million [59]	×	×	✓	×
Mockingjay [60]	2019	85 million [59]	×	×	✓	×
FastSpeech [58]	2019	23 million [61]	✓	×	×	×
wav2vec [62]	2019	16 million [63]	×	×	✓	×
wav2vec 2.0 [45]	2020	317 million [59]	×	×	✓	×
FastSpeech 2 [61]	2020	27 million [61]	✓	×	×	×
FastPitch [64]	2020	26.8 million	✓	×	×	×
Conformer [65]	2020	1 billion [59]	×	✓	✓	×
DeCoAR 2.0 [66]	2020	317 million [59]	×	×	✓	×
w2v-Conformer	2021	1 billion [59]	×	×	✓	×
w2v-BERT [67]	2021	1 billion [59]	×	×	✓	×
HuBERT [68]	2021	317 million [59]	×	×	✓	×
XLS-R [48]	2021	2 billion [59]	×	✓	✓	×
UniSpeech [69]	2021	317 million [59]	×	×	✓	×
UniSpeech-SAT [70]	2021	317 million [59]	×	×	✓	×
BigSSL [71]	2021	8 billion [59]	×	×	✓	×
WavLM [72]	2021	317 million [59]	×	×	✓	×
DeltaLM [73]	2021	360 million [74]	✓	✓	×	✓
SpeechT5 [75]	2021	11 billion [75]	✓	✓	✓	✓
data2vec [49]	2022	*	×	×	✓	✓
data2vec 2.0 [76]	2022	*	×	×	✓	✓
SpeechFormer [63]	2022	3.5 million [63]	×	×	✓	✓
Whisper [52]	2022	1.6 billion [77]	×	✓	✓	×
VALL-E [78]	2023	*	✓	×	×	×
VALL-E X [79]	2023	*	✓	×	×	×

models [65]. The Conformer architecture combines convolutional and transformer layers, enabling it to capture both local and global context information. This makes Conformer models well-suited for speech-processing tasks such as speech recognition and speaker identification, where capturing long-range dependencies is crucial. Conformer achieved state-of-the-art performance on benchmarks such as LibriSpeech and AISHELL-1. However, previous limitations in speech synthesis and recognition, such as the struggle to produce natural-sounding speech in languages other than English and generate speech in real-time, remained. In response, Wang et al. [80] presented an ASR model that uses a combination of noisy student training with SpecAugment and giant Conformer models pre-trained using the wav2vec 2.0 pre-training method on the Libri-Light dataset. They achieved state-of-the-art word error rates on the LibriSpeech dataset. In 2021, Wang et al. [81] extended Conformer and developed Conformer-LHUC, which utilized learning hidden unit contribution (LHUC) for speaker adaptation. Conformer-LHUC showed superior performance in elderly speech recognition and has potential implications for clinical diagnosis and treatment of Alzheimer's disease.

7) UniSpeech: There are other emerging models that are gaining traction in speech processing, such as the UniSpeech model, which focuses on developing models that can handle low-resource and cross-lingual speech tasks. Microsoft's UniSpeech approach [82] proposes a unified pre-training method that combines supervised and unsupervised learning for speech representation learning. The approach uses phonetic CTC learning and phonetically-aware contrastive self-supervised learning to capture more phonetic information and generalize better across languages and domains. The authors evaluate UniSpeech on cross-lingual representation learning and achieve state-of-the-art results on low-resource speech recognition tasks. In a related paper [83], the authors propose UniSpeech-SAT, a universal speech representation learning method with speaker-aware pre-training. The method improves existing self-supervised learning for speaker representation learning by using utterance-wise contrastive learning and utterance mixing augmentation. The method achieves state-of-the-art performance in universal representation learning, especially for speaker identification tasks, and can be easily adapted to downstream tasks with minimal fine-tuning.

8) Speechformer: After the success of UniSpeech models, the field of end-to-end speech recognition has continued to advance. In June 2021, Speechformer [84], a self-supervised pre-trained model for end-to-end speech recognition that leverages masked acoustic modeling and contrastive predictive coding was proposed. Unlike previous models that used either convolutional or recurrent neural networks, Speechformer uses a transformer-based encoder-decoder architecture with relative position encoding and layer normalization. It is pre-trained on 53k hours of unlabeled speech data and achieves competitive

results on several ASR benchmarks.

9) WavLM: Microsoft Research Asia released WavLM [85], a large-scale self-supervised pre-trained model that can solve full-stack downstream speech tasks such as ASR, TTS, and speaker verification. WavLM jointly learns masked speech prediction and denoising in pre-training and employs gated relative position bias for the Transformer structure to better capture the sequence ordering of the input speech. It is trained on a massive dataset of 94k hours of speech and achieves state-of-the-art results on several downstream speech tasks for 10 languages.

There are various other transformers for speech-related tasks that we present in Table II.

III. LITERATURE REVIEW

A. Automatic Speech Recognition (ASR)

ASR enables machines to recognize uttered speech and transform it into the corresponding sequence of text (words or sub-words). State-of-the-art ASR systems achieved improved performance by using RNNs with long short-term memory (LSTM) [86] units as their backbone networks. Recently, there has been an increasing interest in the exploitation of transformers [14] for ASR, inspired by their success in different NLP tasks such as language modeling [87] and machine translation [88]. RNNs process the input signal in a sequential manner by utilizing expensive back-propagation through time (BPTT) [89] algorithm to learn temporal dependencies. Transformers circumvent this with a self-attention mechanism to capture the temporal correlations among the sequential data. This enables transformers to capture longer temporal correlations with less computation complexity. Another advantage of using a transformer is the ability to parallelize the computations in transformers, which can reduce the time for training deeper models on larger datasets.

In ASR, transformers achieved a competitive recognition rate compared to RNN-based baseline models. For instance, Karita et al. [1] experimentally compared transformers with conventional RNNs. Based on the results, they showed various training and performance benefits achieved with transformers in comparison to RNNs. In [90], Zeyer et al. performed a comparison of the transformer encoder decoder-attention model with RNNs and found that transformers are more stable compared to LSTM, however, they face the problem of overfitting and generalization. They also found that the pretraining leads to faster convergence and performance boost. Li et al. [91] performed a comparison between RNN and transformer-based end-to-end models using the 65 thousand hours of Microsoft anonymized training data. They found that transformer-based attention encoder-decoder architecture achieved the best accuracy. Similarly, studies [92], [93] also performed a comparison of transformers with different ASR systems and highlights the benefits of transformers and pointers for future research.

In hybrid ASR, an acoustic encoder is used to encode an input sequence to high-level embedding vectors that are exploited to generate the posterior distribution of tied states of the hidden Markov model (HMM). Combined with other knowledge sources, these posterior distributions are used to construct a search graph. A decoder network is then used to determine the best hypothesis. Different deep models can be used as acoustic encoders in hybrid ASR. Recently, studies started using transformers for improving hybrid acoustic modeling. Wang et al. [92] evaluated a transformer-based acoustic model for hybrid speech recognition. They explored multiple modeling choices and losses for training deep transformers. Based on the results, they showed that the proposed hybrid ASR can achieve significantly improved WER compared to the very strong bi-directional LSTM (BLSTM) baselines.

For streaming applications of ASR, Wu et al. [94] presented an acoustic model based on an augmented memory self-attention transformer for hybrid ASR. The proposed model attends a short segment of the input sequence and accumulates information into memory banks. This makes the segment information equally accessible. Evaluations were performed on Librispeech data, which showed that the proposed model achieves a 15% error reduction in contrast to the widely used LC-BLSTM baseline.

Recurrent sequence-to-sequence models have achieved great progress in ASR. These models are based on the encoder-decoder architecture, where the encoder transforms the speech feature sequence into hidden representations and generates an output sequence. Conventional RNNs-based sequence-to-sequence models suffer from slow training and training parallelization issues. In a transformer-based sequence-to-sequence model, the encoder and decoder network are composed of multi-head attention and position-wise feed-forward networks rather than RNNs. Also, the encoder outputs are attended by each decoder block respectively. This makes training transformer-based sequence-to-sequence models faster and allows for parallel training. Dong et al. [8] presented a Speech-Transformer with no recurrence to learn positional dependencies in speech signals entirely relying on attention mechanisms. Evaluations were performed on Wall Street Journal (WSJ) dataset, which showed that transformers can achieve a competitive word error rate (WER), significantly faster than the published results using RNN-based sequence-to-sequence models. Zhou et al. [93] explored the modeling units in ASR using transformer-based sequence-to-sequence models on Mandarin Chinese speech. They performed a comparison among five modeling units including context-independent phonemes, syllables, words, sub-words, and characters. Based on the results, they found that the character-based model performs best and achieves state-of-the-art (SoTA) CER on the HKUST dataset.

In a study by Zhou et al. [95], the authors compared the performance of a context-independent (CI)-phoneme-based model and a syllable-based model using a transformer on the HKUST dataset. The results showed that the syllable-based model performed better than the CI-phoneme-based model. In Hrinchuk et al. [96], a transformer-based sequence-to-sequence model was proposed to improve the performance of automatic speech recognition (ASR) by correcting the ASR system output. The proposed model was able to correct erroneous outputs into semantically and grammatically correct text, which helped improve the performance. To address the issue of asynchronous encoding and decoding in sequence-to-sequence models, Tain

et al. [97] presented a synchronous transformer that can predict the output in chunks. The experiments showed that the proposed model was able to encode and decode synchronously, which led to an improved character error rate (CER) rate.

Transformers have also shown promising results in large-scale ASR. Lu et al. [98] explored transformers for large-scale ASR with 65,000 hours of training data. They investigated different aspects such as warm-up training, model initialisation, and layer normalization techniques on scaling up transformers for ASR. Chen et al. [99] evaluated the potential of transformer Transducer models for the first pass decoding with low latency and fast speed on a large-scale ASR dataset. Based on the experiments, they showed that the Transformer Transducer model outperforms RNN Transducer (RNN-T) [100], streamable transformer, and hybrid model in the streaming scenario. In [101], Li et al. focused on a large-scale Mandarin ASR and propose three optimization strategies to improve the efficiency and performance of SpeechTransformer [8]. Wang et al. [92] performed a comparative study on the transformer-based acoustic model on large-scale ASR. They found that the transformer-based ASR model achieves better performance compared to LSTM for voice assistant tasks. The aforementioned studies show the effectiveness of transformers for ASR. We summarise recent studies on transformers for ASR in Table III.

B. Neural Speech Synthesis

Neural speech synthesis, or Neural text-to-speech (TTS), is an important field of research that aims to synthesize speech from text input. Traditional TTS systems are composed of complex components including acoustic frontends, duration model, acoustic prediction model, and vocoder models [102]. The complexity of the TTS has been recently overcome with deep end-to-end TTS architectures [53], [103]. These systems can synthesize realistic-sounding speech by training on <text,audio> pairs, and eliminate the need for complex sub-components and their separate training. Prominent models includes Tacotron [53], Tacotron 2 [54], Deep Voice 3 [104], and Clarinet [105]. These models generate Mel-spectrogram from text input, which is then used to synthesize speech by vocoder such as Griffin-Lim [106], WaveNet [107], and Waveglow [108].

Recently, transformers are becoming popular to generate Mel-spectrogram in TTS systems. Particularly, they replace RNN structures in end-to-end TTS to improve training and inference efficiency. In [55], Li et al. attempted to utilize the multi-head attention mechanism to replace RNN structures as well as the vanilla attention mechanism in Tacotron 2. This helps in improving pluralization by solving the long-distance dependency problem. They generated the Mel-spectrogram using the phoneme sequences as input and exploited WaveNet as a vocoder to synthesize speech samples. Based on the results, they showed that transformer TTS was able to speed up training 4.25 times compared to Tacotron 2 and achieve a similar MOS performance.

In order to improve the inference speed, FastSpeech [58] used a feed-forward network based on 1D convolution [117], [118]

and the self-attention mechanism in transformers to generate Mel-spectrogram in parallel. It utilizes the length regulator based on duration predictor to solve the issue of sequence length mismatch between the Mel-spectrogram sequence and its corresponding phoneme sequence. Fastspeech was evaluated on the LJSpeech dataset and results showed that it can significantly speed up the generation of Mel-spectrogram while achieving comparable performance to the autoregressive transformer model. FastPitch [109] improves FastSpeech by conditioning the TTS model on fundamental frequency or pitch contour. Pitch conditioning improved the convergence and removed the requirement for knowledge distillation of Mel-spectrogram targets in FastSpeech. Fastspeech 2 [61] is another transformer-based TTS system that resolved the issues in Fastspeech and better addressed the one-to-many mapping problem in TTS. It uses more diverse information of speech (e.g., energy, pitch, and more accurate duration) as conditional inputs and directly train the system on a ground-truth target. Fastspeech 2s is another variant proposed in [61], which further simplifies the speech synthesis pipeline by directly generating speech from the text in inference without using Mel-spectrograms as intermediate output. Experiments on the LJSpeech data showed that FastSpeech 2 and FastSpeech 2s present a simplified training pipeline with fast, robust, and controllable speech synthesis compared to FastSpeech.

End-to-end TTS systems, such as FastSpeech [58] and Durian [119], utilize a duration model to align output acoustic features with the input text. However, the multi-stage training pipeline used in these systems can be slow. To address this issue, Lim et al. proposed a jointly trained duration-informed transformer (JDI-T) that uses a feed-forward transformer with a duration predictor to generate acoustic feature sequences without explicit alignments. JDI-T achieved state-of-the-art performance on the Korean Single Speaker Speech (KSS) dataset and synthesized high-quality speech compared to other popular TTS models.

However, neural TTS models can suffer from robustness issues and generate poor audio samples for unseen or unusual text. To overcome these issues, Li et al. proposed RobuTrans, a robust transformer that converts input texts to linguistic features before feeding them to the encoder [120]. They also modified the attention mechanism and position embedding to improve the learning of holistic information from the input, resulting in improved MOS scores compared to other popular TTS models. Another approach to achieving robustness in TTS systems is the segment-transformer (s-Transformer) [121] proposed by Wang et al. The s-Transformer is capable of modeling speech at the segment level, allowing it to capture long-term dependencies and use segment-level encoder-decoder attention to handle long sequence pairs. This approach enables the s-Transformer to achieve similar performance to the standard transformer while also exhibiting robustness on extra-long sentences. Lastly, Zheng et al. [122] proposed an approach that incorporates a local recurrent neural network into the transformer to capture both sequential and local information in sequences. Evaluation on a 20-hour Mandarin speech corpus demonstrated that this model outperforms the transformer alone in terms of performance.

Speech synthesis using multi-speaker voices is another inter-

TABLE III: Recent studies on transformers for **Automatic Speech Recognition (ASR)**.

Author (year)	Dataset	Performance	Architecture
Zeyer et al. [90]	LibriSpeech (1000 hr), TED-LIUM 2 (200 hr) and Switchboard (300 hr)	WER LibriSpeech: 2.81% TED LIUM: 12.0% Switchboard: 10.6 %	Transformer encoder-decoder-attention model and LSTM encoder-decoder-attention model.
Li et al. [91]	Microsoft transcribed data (65000 hr)	WER: 9.16%	Recurrent neural network transducer (RNN-T), RNN attention-based encoder-decoder (AED), and Transformer-AED
Wang et al. [92]	Personal Assistant Dataset	WER: 3.94	Transformer, Emformer (streamable variant of Transformer), latency-controlled BLSTM (LCBLSTM), and LSTM
Wu et al [94]	LibriSpeech and German and Russian video dataset	WER LibriSpeech: 2.8% Russian: 18.0% German: 17.4%	Streaming Transformer with self-attention with augmented memory (SAAM) module.
Dong et al. [8]	WSJ dataset	WER: 10.9%	Transformer encoder-decoder-attention model with 2D-Attention mechanism.
Zhou et al. [95]	HKUST Dataset	CER: 28.77%	Transformer encoder-decoder-attention model with syllable-based input and output.
Hrinchuk et al. [96]	Prepared own dataset 101K sample for first name and 580K sample for last name	WER: 9.2% on first and 6.6% on last name	Transformer encoder-decoder-attention model for ASR post-processing.
Tian et al. [97]	AIShell	CER: 8.91%	Synchronous Transformer (a variant of Transformer with chunk-by-chunk prediction)
Lu et al. [98]	Microsoft data (65000 hr)	WER: 12.2%	Transformer encoder-decoder
Chen et al. [99]	Microsoft data (65000 hr)	WER : 8.19%	Transformer transducer
Li et al. [101]	AiShell 1(165 hr) HKUST(156 hr)	CER AiShell-1: 13.09% HKUST: 28.95%	SpeechTransformer
Lancucki et al. [109]	LJSpeech-1.1 Dataset	MOS Values 4.071±0.164	FastPitch, a variant of FastSpeech
Mohammed et al. [110]	LJSpeech-1.1 Dataset	WER: 4.7%	Transformer encoder-decoder with convolutional context modules
Zhang et al. [111]	SWBD AMI AISHELL	WER: SWBD: 7.1% AMI: 24.1% AISHELL: 4.7%	TransMask, a transformer encoder-decoder with mask prediction heads
Moriya et al. [112]	WER: WSJ, Switchboard, Librispeech, CSJ and NTT Japanese dataset.	WER: SWBD: 8.9% LibriSpeech: 4.4% CER: CSJ: 3.9% NTT: 4.2%	CTC-Transformer, a transformer encoder-decoder with connectionist temporal classification (CTC) loss
Cao et al. [113]	LibriSpeech	WER: 3.5%	Streaming Transformer, a transformer encoder-decoder with block processing and latency control mechanisms
Tsunoo et al. [114]	WSJ, Librispeech, VoxForge Italian, and AISHELL-1	WER: Librispeech: 4.6% WSJ: 5.7% AISHELL-1: 7.6% VoxForge: 10.3%	Transformer encoder-decoder with contextual block processing (CBP), a technique to improve streaming ASR performance by using past and future context information
Jain et al. [115]	YLE news dataset	WER: 17.71 %	Transformer encoder-decoder with deep self-attention layers
Yu et al. [116]	LibriSpeech and MultiDomain	WER: LibriSpeech: 2.5% MultiDomain: 6.0%	Dual-mode ASR (DM-ASR), a hybrid model that combines streaming ASR (S-ASR) and full-context ASR (F-ASR) using two parallel transformer encoders and one shared decoder.

esting field of research. Chen et al. presented a MultiSpeech model, based on transformer TTS, that can synthesize high-quality speech in multi-speaker voices with fast inference speed. To achieve this, they designed a special component in the transformer to preserve positional information and prevent copy between consecutive speech frames. The MultiSpeech model was evaluated on VCTK and LibriTTS datasets, and the results demonstrated superior performance compared to existing models. Voice conversion, on the other hand, focuses on altering the source speaker's voice to match the target voice without changing the linguistic content. While various studies have explored RNN-based sequence-to-sequence models for voice conversion, these models require extensive training data and often suffer from mispronunciation issues. To address these challenges, Huang et al. presented a voice transformer network that utilized pre-training to improve data-efficient

training and achieve better results compared to RNN-based models. Recent studies have continued to push the boundaries of speech synthesis systems, exploring various approaches to improve performance. The summary of recent studies on speech synthesis is presented in Table IV.

C. Speech Translation (ST)

Speech Translation (ST) is the process of translating the spoken speech in the source language into the target language. ST systems are typically divided into two categories: cascaded systems and end-to-end systems. Cascaded ST systems consist of an automatic speech recognition (ASR) system and a machine translation (MT) system. ASR system generates text from the spoken sentence, which is then used by a machine translation system to translate it into the target language. Cascaded ST systems face the problem of errors compounding

TABLE IV: Recent studies on transformers for **Speech Synthesis**.

Author (year)	Datasets	Performance	Architecture
Ren et al. [58]	LJSpeech dataset	MOS: 3.84 ± 0.08	FastSpeech, a feed-forward Transformer TTS
Ren et al. [61]	LJSpeech dataset	MAE: FastSpeech2:0.131 FastSpeech2s:0.133	FastSpeech 2, an improved FastSpeech with more variance information
Chen et al. [123]	VCTK and LibriTTS datasets	MOS: VCTK: 3.65 ± 0.14 LibriTTS: 2.95 ± 0.14	MultiSpeech, a multi-speaker Transformer TTS with speaker embeddings and classifier loss
Łańcucki et al. [109]	LJSpeech Dataset	MOS: 3.707 ± 0.218	FastPitch, a FastSpeech variant with pitch prediction and control
Gehringi et al. [117]	WMT'14 English-French WMT'14 English-German. WMT'14 English-Romanian.	WMT'14 English-French: 20.51 WMT'14 English-German: 26.43 WMT'14 English-Romanian: 41.62	ConvS2S, a CNN-based sequence-to-sequence model.
Yu et al. [119]	Hours of Speech: Male speaker: 18 hours Female speaker: 7 hours	Male: 4.11 Female: 4.26	DurIAN, a multimodal TTS model with duration-informed attention network (DIAN)
Lim et al. [124]	Internal Speaker Dataset Korean Speaker Dataset	MOS Values on a scale of 5 Internal: 3.77 KSS: 3.52	JDI-T, a feed-forward Transformer TTS with duration predictor
Wang et al. [121]	Professional enUS A speaker dataset (46 hour)	MOS Values on a scale of 5 Short: 4.29 Long: 4.2 Extra Long: 3.99	s-Transformer, a segment-wise Transformer TTS
Zheng et al. [122]	Mandarin speech corpus	MOS Values 4.34±0.05	LRN-Transformer, a Transformer TTS with LRNs
Huang et al. [125]	CMU Arctic dataset	WER: 7.8% CER: 4.8%	VTN, a seq2seq voice conversion model with TTS pretraining
Hu et al. [126]	LibriTTS VCTK	WER: LibriTTS: 33.3 ± 1.2 VCTK: 20.3 ± 1.2	MI-TTS, an unsupervised TTS model with VAEs and mutual information minimization
Chen et al. [127]	LJSpeech VCTK	WER: LJSpeech: 9.5% VCTK: 12.5 %	TransformerTTS with local style tokens (LST) and cross-attention blocks
Liu et al. [128]	LJSpeech database	RMSE: 1.625%	GraphSpeech, a graph neural network (GNN) based TTS that encodes syntactic information as a dependency graph
Wang et al. [129]	LJSpeech database	CER: 1.7%	PatNet, a phoneme-level autoregressive Transformer (TTS) that predicts mel-spectrograms from phoneme sequences

TABLE V: Recent studies on transformers for **Speech Translation**.

Author (year)	Datasets	Performance	Architecture(s)
Kano et al. 2021 [130]	Fisher Spanish-English, LibriSpeech English-French, LibriSpeech English-German	BLEU: 16.9 (es-en), 15.8 (en-fr), 10.1 (en-de); MOS: 3.87 ± 0.08 (es-en), 3.81 ± 0.09 (en-fr), 3.77 ± 0.09 (en-de); MAE: 0.131 (es-en), 0.132 (en-fr), 0.133 (en-de)	Encoder-Decoder with attention and transcoder module
Zhang et al. 2019 [131]	IWSLT 2017 en-de ST task, MuST-C en-de ST task	BLEU: 17.9 on IWSLT 2017 en-de; BLEU: 20.8 on MuST-C en-de	A novel controllable lattice attention mechanism that leverages the extra information from the lattice structure of ASR output
Huawei et al. 2022 [132]	WMT 2014 English-German5, WMT 2014 English-French5, WMT 2016 Romanian-English5, WMT 2016 English-Czech5, WMT 2017 English-Turkish	BLEU score: 28.4 (en-de), 41.0 (en-fr), 32.8 (ro-en), 26.7 (en-cs), 24.9 (en-tr)	Encoder-decoder with self-attention layers
Ao et al. 2021 [75]	LibriSpeech, LibriTTS, Common Voice, LJSpeech, AISHELL-1/2/3, VCTK-Corpus, VoxCeleb1/2	WER: 2.9% (LibriSpeech test-clean), 7.0% (LibriSpeech test-other), 6.8% (AISHELL-1), 6.0% (AISHELL-2), 7.9% (AISHELL-3); MOS: 4.06 (LibriTTS), 4.01 (LJSpeech); BLEU: 17.8 (English-to-Chinese ST); MCD: 6.38 dB (voice conversion); PESQ: 2.97 (speech enhancement); EER: 0.83% (speaker identification)	Shared encoder-decoder network with six modal-specific pre/post-nets, pre-trained with contrastive loss, masked speech/text modeling loss and cross-modal alignment loss

between components, e.g., recognition errors leading to larger translation errors. In contrast, end-to-end ST systems optimize a single model that directly translates the spoken utterance into the target language. Various studies explored methods and techniques to improve the performance of both cascaded ST systems [133]–[135] as well as end-to-end ST systems [136]–[138].

Presently, ST research explores transformers for solving different issues. Vila et al. [139] used a transformer for end-to-end speech translation. Evaluations were performed on the Spanish-to-English translation task and a bilingual evaluation understudy score was computed, which showed that the end-

to-end architecture was able to outperform the concatenated systems. Zhang et al. [131] presented a lattice transformer for speech translation, which also uses lattice representation in addition to the traditional sequential input. They evaluated the proposed model on Spanish-English Speech Translation Corpus and achieved improvements over strong baseline results. In [140], the authors present an adaptation of the transformer to end-to-end ST. They performed down-sampling of input with convolutional neural networks to i) make the training process feasible on GPUs, ii) model the bi-dimensional nature of a spectrogram, and iii) add a distance penalty to the attention, so as to bias it towards the local context. Furthermore, several distinct studies (Jia et al., 2021 [141]; Zhang et al., 2023 [79]; Huang et al., 2022 [142]; Li et al., 2020 [143]; Wang et al., 2020 [144]; Zeng et al., 2021 [145]) explore various Speech Translation model implementations or models designed for Speech Translation tasks.

D. Speech Paralinguistics

Speech paralinguistics is a term that refers to the non-verbal aspects of speech communication, such as tone, pitch, volume, speed, emotion, and accent [151]. In terms of NLP, speech paralinguistics is an area that aims to analyse and synthesize speech signals with paralinguistic features. These features can convey important information about the speaker's identity, intention, attitude, and mood, and can enhance the performance and naturalness of various speech applications. In this section, we discuss how transformers can be used for speech paralinguistic tasks, which involve analyzing and synthesizing speech signals with non-verbal features such as emotion, speaker identity, and accent. We focus on a recent paper by Chen et al. (2021) [85], which proposes a new pre-trained model called WavLM for full-stack speech processing. WavLM uses a masked speech prediction and a speech-denoising objective to learn universal speech representations from large-scale unlabeled data. WavLM also employs a gated relative position bias mechanism to capture the sequence order of speech signals. The paper shows that WavLM achieves state-of-the-art results on the SUPERB benchmark [152] and improves performance on several other speech benchmarks for tasks such as speech emotion recognition (SER), speaker verification (SV), speaker diarization (SD), and speech separation (SS). We review the main contributions of WavLM and compare it with other transformer-based models for speech paralinguistic tasks.

In speech paralinguistics, Xu et al. (2021) proposed a new attention mechanism called local dense synthesizer attention (LDSA) [148]. The mechanism restricts attention scope to a local range around the current frame and eliminates dot products and pairwise interactions to improve the performance of end-to-end speech recognition models while reducing computational complexity. The study also combines LDSA with self-attention to extract both local and global information from speech signals. Shor et al. (2022) proposed a new pre-trained model, Conformer-HuBERT, that combines conformers and HuBERT to learn universal speech representations for paralinguistic tasks such as SER, SV, SD, and SS [147]. Conformers are hybrid architectures that integrate CNNs and transformers, while

HuBERT is a self-supervised learning framework that learns from large-scale unlabeled data. The study demonstrates that Conformer-HuBERT outperforms existing models on several benchmarks for paralinguistic tasks, achieving state-of-the-art results.

Shor and Venugopalan (2022) [146] propose a collection of small and performant models called TRILLsson, which are distilled from a large self-supervised model called CAP12 [147]. TRILLsson uses knowledge distillation on public data to reduce the size of CAP12 by up to 100x while retaining 90-96 percent of its performance. The paper demonstrates that TRILLsson outperforms previous models on the Non-Semantic Speech (NOSS) benchmark [153]. Additionally, the paper releases the TRILLsson models publicly. Another attempt by Chen et al. (2022) [63] propose a novel framework, SpeechFormer, that incorporates the unique characteristics of speech signals into transformer models. The framework comprises three components: a hierarchical encoder that reduces the input sequence length using convolutional and pooling layers, a local self-attention module that captures dependencies within a fixed window size, and a global self-attention module that captures dependencies across different windows. The paper demonstrates that SpeechFormer achieves competitive results on several speech benchmarks for tasks such as automatic speech recognition (ASR), speaker verification (SV), speaker diarization (SD), and emotion recognition. Another recent paper, SpeechFormer++: by Chen et al. (2023) [150] builds on the previous work of SpeechFormer [63] and incorporates the unique characteristics of speech signals into transformer models. The framework includes a unit encoder that models the intra- and inter-unit information, a merging block that generates features at different granularities based on the hierarchical relationship in speech signals, and a word encoder that integrates word-grained features into each unit encoder. The paper demonstrates that SpeechFormer++ outperforms the standard transformer on various paralinguistic tasks, such as speech emotion recognition (SER), depression classification (DC), and Alzheimer's disease detection (ADD).

Gao et al. (2022) introduced Paraformer, a new model for non-autoregressive end-to-end speech recognition that uses parallel attention and parallel decoder techniques [149]. Paraformer's encoder-decoder architecture allows each decoder layer to attend to all encoder outputs simultaneously without waiting for previous decoder outputs, and each output token to be predicted independently without depending on previous output tokens. The paper shows that Paraformer outperforms existing non-autoregressive models on several ASR datasets, achieving faster inference speed and higher accuracy. These recent innovations in transformer-based models such as WavLM, Conformer-HuBERT, TRILLsson, SpeechFormer, and Paraformer have shown promising results for speech paralinguistic tasks, paving the way for more natural and efficient speech applications.

E. Speech Enhancement and Separation

The area of speech enhancement involves the application of various algorithms for enhancing the quality of speech. Speech

TABLE VI: Recent studies on transformers for **Speech Paralinguistics**.

Author (year)	Datasets	Performance	Architecture(s)
Chen et. al. 2022 [85]	SUPERB benchmark, LibriSpeech (ASR), VoxCeleb1/2 (SV), AMI (SD), CommonVoice (LID), CREMA-D (SER)	SUPERB: 0.9231234, LibriSpeech WER: 1.9/4.81, VoxCeleb1/2 EER: 0.69/0.831, AMI DER: 7.51, CommonVoice LID ACC: 0.9781, CREMA-D SER UAR: 0.7221	Transformer with gated relative position bias and utterance mixing.
Shor et al. 2022 [146]	SUPERB benchmark, LibriSpeech (ASR), VoxCeleb1/2 (SV), AMI (SD), CommonVoice (LID), CREMA-D (SER)	SUPERB: 0.906123, LibriSpeech WER: 2.4/5.81, VoxCeleb1/2 EER: 0.75/0.911, AMI DER: 8.71, CommonVoice LID ACC: 0.9761, CREMA-D SER UAR: 0.713	EfficientNet-B0 with Audio Spectrogram Transformer encoder and ResNet-50 encoder for fixed-length and arbitrary-length inputs respectively.
Shor et al. 2022 [147]	SUPERB benchmark, LibriSpeech (ASR), VoxCeleb1/2 (SV), AMI (SD), CommonVoice (LID), CREMA-D (SER)	SUPERB: 0.9511234, LibriSpeech WER: 2.3/5.61, VoxCeleb1/2 EER: 0.67/0.831, AMI DER: 8.31, CommonVoice LID ACC: 0.9781, CREMA-D SER UAR: 0.726	A stack of convolution-augmented transformer blocks known as Conformers with a total of 600M parameters trained on YT-U dataset using self-supervision.
Xu et al. 2021 [148]	IEMOCAP, MSP-IMPROV, RAVDESS	IEMOCAP: WA: 0.661, UA: 0.633; MSP-IMPROV: WA: 0.651, UA: 0.644; RAVDESS: WA: 0.713, UA: 0.712	A transformer-based end-to-end model that consists of a CNN encoder, a transformer encoder, a temporal attention layer, and a softmax classifier.
Gao et al. 2022 [149]	ComParE 2019-2021, IEMOCAP, MSP-IMPROV, RAVDESS, SAVEE, EmoDB, CREMA-D, EMOVO-Corpus, ShEMO-Corpus	ComParE 2019: UAR: 0.72 ComParE 2020: UAR: 0.67 ComParE 2021: UAR: 0.65 IEMOCAP: F1: 0.71 MSP-IMPROV: F1: 0.69 RAVDESS: UAR: 0.76 SAVEE: UAR: 0.77 EmoDB: UAR: 0.80 CREMA-D: UAR: 0.81 EMOVO-Corpus: UAR: 0.79 ShEMO-Corpus: UAR: 0.78	A hierarchical framework that consists of a transformer-based encoder-decoder model and a multi-task learning module.
Chen et al. 2023 [150]	LibriSpeech test-clean/test-other.	WER: 2.3%/5.4%	A parallel transformer that utilizes a continuous integrate-and-fire based predictor to predict the number of tokens and generate hidden variables.

TABLE VII: Recent studies on transformers for **Speech Enhancement**.

Author (year)	Datasets	Performance	Architecture(s)
Yu et al. 2022 [154]	VCTK, CHiME-3	PESQ: 2.97 ± 0.01, STOI: 0.94 ± 0.00, SI-SNRi: 16.9 ± 0.1 dB	Encoder-LSTM-Multi-head attention-Decoder
Kim et al. 2020 [155]	VCTK	PESQ: 3.06 ± 0.01, STOI: 0.95 ± 0.00, SI-SNRi: 17.8 ± 0.1 dB	Encoder-Self-attention with Gaussian weights-Decoder
Wang et al. 2021 [156]	VCTK, DNSCL, DNSCL-R	PESQ: Voice Bank + DEMAND: 3.08, DNSCL: 3.02, DNSCL-R: 2.93; STOI: Voice Bank + DEMAND: 0.95, DNSCL: 0.94, DNSCL-R: 0.92	Encoder-Two-stage transformer module-Masking module-Decoder
Subakan et al. 2021 [13]	WSJ0-2mix, WSJ0-3mix	WSJ0-2mix: SDR: 20.8 dB, SI-SNR: 21.9 dB; WSJ0-3mix: SDR: 17.6 dB, SI-SNR: 18.7 dB	Multi-scale transformer with multi-head attention and feed-forward layers
Zhang et al. 2022 [157]	LibriSpeech, LibriMix, WHAMR, WHAM	LibriMix: PESQ: 3.25, STOI: 0.95, ESTOI: 0.93; WHAMR: PESQ: 2.98, STOI: 0.94, ESTOI: 0.91; WHAM PESQ: 3.04, STOI: 0.95, ESTOI: 0.92;	Streaming transformer with cross-attention between encoder and decoder layers
Zhao et al. 2021 [158]	WSJ0-2mix, WSJ0-3mix	WSJ0-2mix: SDR: 20.6 dB, SI-SNR: 21.7 dB; WSJ0-3mix9: SDR: 17.5 dB, SI-SNR: 18.6 dB	Multi-scale group transformer with dense-fusion or light-fusion and time-domain audio separation network (TasNet)
Jiang et al. 2023 [159]	LibriSpeech	PESQ: 3.25, STOI: 0.95, ESTOI: 0.93, SI-SNRi: 19.1 dB	Hierarchical frame-level Swin Transformer with adaptive windowing and convolutional layers

enhancement (SE) aims to isolate the speech of a targeted user from a group of others. Previously, neural networks have been employed in an attempt to achieve this goal. One such implementation utilized an audio embedding network to extract the audio embedding of different speakers, which was then utilized in a spectrogram masking network to produce an output

with masks [160]. This approach yielded a more efficient and faster model as the embedding for each speaker was computed in advance. The features extracted were subsequently employed in the PSE network for enhancing the speech signals of specific users, enabling the segregation of separate networks and allowing for further individual improvements [160], [161]. At inference, the speaker embedding is concatenated with the intermediate features of the PSE network for conditionality purposes. In another approach, audio signal embedding vectors representing the desired speaker were utilized to improve noise and echo cancellation and speech enhancement [162]. Another model, known as the Sound-Filter model, employed unlabeled data for speech enhancement. This wave-to-wave convolutional neural network was trained using mixtures generated from a collection of unlabeled audio recordings. It was assumed that speech was from a single source for the entire duration of a clip, based on the use of short intervals of audio signals for the same type of sound. This problem was approached as a one-shot learning challenge, resulting in models with conditioning encoder clusters that mapped acoustically similar sounds together in the embedding space [163]. Additionally, activation functions were learned to personalise the output [164].

The application of speech separation is a crucial aspect of speech processing, and Recurrent Neural Networks (RNNs) were predominantly utilized for this purpose since their introduction. However, the trend has shifted towards the usage of Transformers as they enable parallel computation through the attention mechanism. The sequence-modeling capabilities of Transformers have the potential to enhance speech separation. An example of this approach is the proposed SepFormer model in [13]. The authors leveraged the parallel computation benefits of Transformers in the implementation of their model. Furthermore, the utilization of Transformers in speech separation has been observed in other studies as well, such as the work of Zhang et al. [111]. The issue of increased computational complexity with longer sequences of sentences in the field of speech separation has been addressed through the utilization of the multi-scale group transformer (MSTG) approach. As reported by Zhao et al. [158], this approach leverages the self-attention capabilities of transformers and incorporates multi-scale fusion to capture long-term dependencies, thereby reducing computation complexity. The size of state-of-the-art models for speech separation tasks, however, often reaches hundreds of gigabytes, presenting a common challenge in the field.

To mitigate this challenge, various approaches have been employed including Knowledge Distillation [165]. The Teacher-Student model [166] has been utilized in an approach aimed at reducing the size and complexity of models for speech separation. Another approach reported in [167] achieved this through the utilization of non-overlapping blocks in latent space and compact latent summaries calculated for each chunk. The authors developed the RE-SepFormer and achieved performance on par with existing state-of-the-art models. Another innovative approach, Tiny-Sepformer [168], uses a time-domain transformer neural network and achieved this by splitting Convolution Attention (CA) blocks and implementing

parameter sharing within CA blocks.

F. Spoken Dialogue Systems

Table VIII shows a list of related works on spoken dialogue systems using Transformer networks. But it should be noted that most of those neural architectures have been originally applied to text-based language processing tasks, not to speech data with some exceptions [169], [170]. From the popularity chart in Fig. 2, it can be noted that the most popular neural architecture is BERT (Bidirectional Encoder Transformer) [171] and the second most popular choice of architecture is either the original/vanilla transformer architecture [14] or GPT-2 (Generative Pre-Trained Transformer) [40]. Whilst BERT and GPT-2 are generalizations of the vanilla transformer networks, BERT uses encoder blocks (no decoder blocks) and bidirectional representations whilst GPT-2 uses decoder blocks (no encoder blocks) and left-to-right representations. Other generalizations of Transformers include the following: XLM (Cross-lingual Language Model) [172] to benefit from data in different languages; DistillBERT (Distilled version of BERT) [173] to train more compact models via knowledge distillation; ConvERT [174] to perform faster training via pre-trained response selection; BART (Bidirectional Auto-Regressive Transformers) [175] to pre-train sequence-to-sequence models via a denoising autoencoder (by predicting outputs without noise from noisy inputs); T5 (Text-to-Text Transfer Transformer) [176] to learn from data in multiple language tasks by converting it to text-to-text format and then carrying out transfer learning to a specific language task; Conformer (Convolution-augmented Transformer) [169] to bring the advantages of Transformer and Convolutional neural nets into a single architecture suitable for audio sequences; DIET (Dual Intent and Entity Transformer) [170] to perform language understanding of intents and entities in utterances without pretraining; GPT-3 (large Generative Pre-Trained Transformer) [177] to learn large language models without the need of fine-tuning; and RoBERTa (Robustly optimised BERT approach) [178] to learn language models with an improved methodology over BERT. Some other related architectures have been proposed and trained with text instead of speech data [179]–[194].

From Fig. 3, it can be noted that Transformer networks have been mostly applied to language understanding tasks (52% of related works) including intent recognition and semantic tagging (a.k.a. slot filling). They are followed by turn-taking prediction and dialogue generation[1] (25% of related works). Other less popular application areas (23% of related works) include emotion recognition, performance prediction (of language understanding or user satisfaction), dialogue state tracking, punctuation prediction, disfluency detection, text normalization, and speech recognition for conversational data. The full set of tasks gives us an idea of the range of skills involved in a spoken dialogue system.

Regarding the performance of Transformers, BERT models have been shown to outperform their predecessor recurrent/convolutional neural nets in language understanding tasks

---

[1]Dialogue generation in this paper refers to generating a system response either word by word or by selecting a sentence from a pool of candidates).

TABLE VIII: Representative list of publications in Transformer-based **Spoken Dialogue Systems** in the period 2019-2023.

| Author (year)	Architecture(s)	Task(s)	Dataset(s)
Takatsu et al. 2019 [195]	BERT	Intent recognition	Majority Vote
Chen et al. 2019 [196]	Vanilla Transformer	Dialog state tracking	DSTC2
Liu et al. 2019 [197]	BERT, BiLSTM	Intent recognition	CamRest
Korpusik et al. 2019 [198]	BERT	Semantic tagging	ATIS, Restaurants
Zhao H. Wang. 2019 [199]	Vanilla Transformer, CRF	Intent recognition, semantic tagging	ATIS, SNIPS
Quian et al. 2020 [200]	GPT-2	Semantic tagging	DSTC2
Hong et al. 2020 [201]	BERT	Semantic tagging	CamRest
Ekstedt et al. 2020 [202]	GPT-2	Turn-taking prediction	Maptask, Switchboard
Gopalakrishnan et al. 2020 [203]	GPT-2	Dialogue generation: writen vs. spoken	Topical-chat
Chen et al. 2020 [204]	Vanilla/CT Transformer	Punctuation/disfluency prediction	IWSLT2011, in-house Chinese data
Hori et al. 2020 [205]	Vanilla Transformer	ASR specialized for conversational data	Switchboard, HKUST
Lian et al. 2021 [206]	Vanilla Transformer	Emotion recognition in dialogue	IEMOCAP, MELD
Andreas et al. 2021 [207]	BART, GPT, BERT	Dialogue generation: model comparison	CamRest
Chapuis et al. 2021 [208]	BERT, hierarchical arch.	Emotion recognition, dialogue act recog.	SILICONE
Lai et al. 2021 [209]	BERT, RoBERTa	(Inverse) Text normalization	Google TN dataset
Kim et al. 2021 [210]	BERT	Intent recognition/classification	FSC, SNIPS, SmartLights
Lopez-Zorrilla et al. 2021 [211]	GPT-2, RL	Dialogue generation: policy learning	DSTC2
Bechet et al. 2021 [212]	BERT	SLU performance prediction	ATIS, MEDIA, SNIPS, M2M
Okur et al. 2022 [213]	BERT, ConveRT, DIET	Intent recognition	Math-Game
Sakuma et al. 2022 [214]	Conformer	End-of-utterance prediction	Japanese Restaurant Data
Lin et al. 2022 [215]	Vanilla Transformer	User-state/barge-in/backchannel detection	Chinese dialogues
Ahmed 2022 [216]	BERT	Intent recognition, semantic tagging	ATIS, SNIPS
Bekal et al. 2022 [217]	BERT, HuBERT	Turn-taking prediction: barge-in detection	12k chat-bot prompts
Dong et al. 2022 [218]	BERT	Intent recognition/classification	FSC, SmartLights
Svec et al. 2022 [219]	T5	Intent recognition/classification	HHTT, TIA (in Czech)
Shen et al. 2022 [220]	Vanilla Transformer, TBM	User satisfaction prediction	Industry data: 18M/32M/5M samples
Yang et al. 2022 [221]	Vanilla Transformer	Turn-taking prediction	Commercial phone-dialogues
Sunder et al. 2022 [222]	BERT	Dialogue act recognition	HarperValleyBank
Lopez-Zorrilla et al. 2022 [223]	GPT-2, RL	Dialogue generation: policy learning	DSTC2
Firdaus et al. 2023 [224]	mBERT, RoBERTa, XML	Intent recognition, semantic tagging	ATIS, TRAINS, SNIPs, FRAMES
Mei et al. 2023 [225]	BERT	Intent recognition, semantic tagging	ATIS, SNIPS
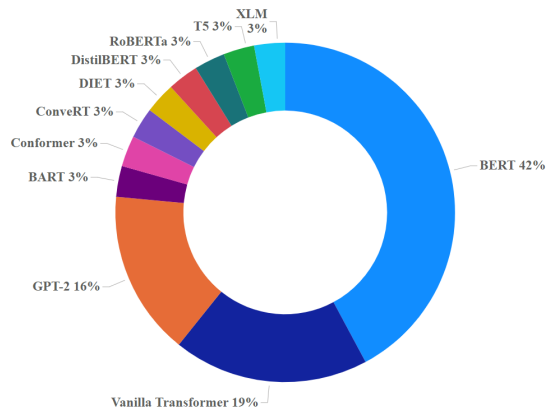
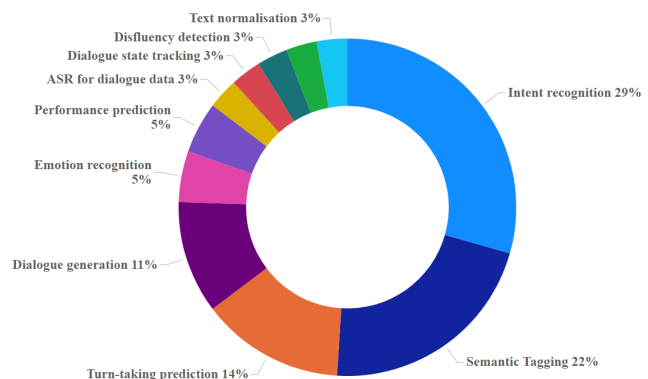Fig. 2: Transformer-based architectures in Spoken Dialogue Systems.

Fig. 3: Tasks of Transformer-based Spoken Dialogue Systems.

(semantic tagging and intent recognition) [197], [198], [201], [224], [225], text normalization [209], performance prediction of language understanding [212], and turn-taking prediction [217] – though not always clearly and sometimes with only small differences. Using the popular stages of pre-training and fine-tuning has been reported to achieve better performance over using only one stage avoiding pre-training [195], [210], [218]. When extended in a hierarchical way to model word sequences at the low level, utterances at the mid-level, and dialogues at the high level, a hierarchical BERT is able to outperform its single model counterpart [208]. Hierarchical models are also trained with fewer parameters than their counterpart non-hierarchical models [222]. Similarly, when combining

specialised architectures such as DIET and ConVERT has shown improved results over using standard Transformers with pre-trained BERT embeddings [213]. Even when combining more standard architectures such as BERT with biLSTMs (bidirectional Long-Short Term Memory Networks) has shown improved results over BERT-like baselines [216]. Moreover, BERT models have also shown to benefit from using data augmentation to obtain further gains over only pre-training and fine-tuning [210]. Furthermore, BERT models can also benefit from combining multiple modalities (such as audio and text) and using a cross-modal contrastive loss over using a single modality [218].

Similarly, GPT-2 models have shown to outperform their predecessor recurrent neural nets in turn-taking prediction

[202]. GPT-2 models have also shown promising results by training a semantic tagger jointly with a speech recognizer and outperforming independent models [200]. When training GPT-2-based models on text data and testing on noisy data derived from speech recognition, the selected responses may be out of context—suggesting the need for training Transformer-based spoken dialogue systems using noisy data [203]. The latter is supported by the work of [211], [223], who extended GPT-2 pre-trained language models with audio embeddings in order to train models with improved performance over using only text-based representations. When extending GPT-2 models (among other Transformer-based architectures) with knowledge embeddings (KE) to port a knowledge base into the model parameters [184], the KE-enhanced models outperform the KE-unaware models [207]. In a similar vein, vanilla transformers have shown to outperform their predecessor recurrent neural nets (RNN) in tasks such as dialogue state tracking [196], intent recognition and semantic tagging [199], punctuation prediction and disfluency detection [204], speech recognition for conversational data [205], emotion recognition [206], user satisfaction prediction [220], and turn-taking prediction [221]. Whilst transformers have mostly reported positive improvements over RNN baselines across different spoken language processing tasks, [215] found no significant differences between them in the task of turn-taking prediction using a dataset of 10K labeled instances.

Some recent models have shown further improvements over vanilla, BERT, or GPT-2 Transformers. The following are some examples. [207] have reported BART and T5 models to outperform vanilla and GPT-2 Transformers in the task of dialogue generation. [213] have shown that models combining DIET and ConVERT are able to outperform vanilla transformers with BERT embeddings in the task of intent recognition. [224] have found that multilingual multi-task BERT models are able to outperform strong baselines such as RoBERTa and XLM in the tasks of intent recognition and semantic tagging.

Other recent works in Transformer-based dialogue systems but using text data are also mostly based on BERT-based architectures [181], [186]–[191] or GPT-based architectures [179], [180], [182], [184], [192]–[194]. Notable large-scale efforts include GODEL [192] and InstructGPT [226]. GODEL (Grounded Open Dialogue Language Model) is framed as suitable for open-ended goal-directed dialogue, in its base and large versions containing 220M and 770M parameters, has been shown to outperform BART, T5, DialoGPT and GPT-3 across different benchmarks. Those results are based on automatic and human evaluations. InstructGPT, a fine-tuned GPT-3 model using reinforcement learning with human feedback, outperforms GPT-3 baselines that do not learn from human feedback. The latter results in a more compact model of 1.3 billion parameters being preferred over a larger model of 175 billion parameters. InstructGPT is the neural architecture of the widely known dialogue system ChatGPT[2].

Although previous works have shown remarkable progress in spoken and text-based language processing using Transformer

networks, however, it is unclear which is the best neural architecture for a particular task or across tasks. Some works include a few baselines and some others have different baselines, and there is a wide variety of datasets being used—due to the number of tasks involved in dialogue systems. While performing such comparisons requires significant resources in terms of data and computing power, they would be highly valuable for establishing a clearer understanding of the current state-of-the-art in the field. Efforts such as GLUE in NLP [227] are needed in spoken dialogue systems. Nonetheless, the battle so far seems to be between BERT and GPT—and novel architectures performing even better remain to be discovered. While there have been some notable successes in conversational AI, the average citizen has yet to fully benefit from them. Many services, whether accessed over the phone or through the web, still rely on rudimentary methods, such as requiring people to fill out long web-based forms or endure long wait times to speak with customer representatives. However, this is likely to change as technologies such as Transformer-based text and spoken dialogue systems become more accessible and easier to deploy. Additionally, given that experimental results in this field continue to show room for improvement in the correctness and safety of generated responses, there is a need for more effective methods to be developed.

G. Multi-Modal Applications

In human-computer interaction, a modality refers to the representation of a human sense using an individual channel of sensory input/output. The modalities in computing encompass vision, audition, reaction, gustation, and olfaction, among others. Multi-modal learning encompasses the use of multiple modalities in conjunction to solve various real-world applications, in much the same way as humans use their multiple senses to complete tasks. For instance, an image of a road sign board can provide an understanding of the type of sign being displayed, while the text on the board adds further context. To mimic this process in computers, data must be solved as separate problems, such as NLP and computer vision. Multi-Modal Learning (MML) provides a general approach to constructing robust models utilizing information from multi-modal data [234]. In order to achieve generalized models in the real world, it is necessary to have excellent models of certain modalities and to use them in conjunction with one another, as some modalities are interdependent in context and deeper understanding, especially in speech processing and NLP-related problems and tasks [235]. Transformers have demonstrated promise in achieving good generalization for multi-modal applications, as evidenced by the utilization of multi-model transformers in building AI models for classification [236], [237], segmentation [238], and cross-modal retrieval [239].

Recently, there has been a shift towards developing innovative fusion models and architectures. In 2019, the Multimodal Transformer (MulT) was proposed [240] to address the issues of long-range dependencies across modalities and non-alignment of data with different sampling rates through the utilization of the attention mechanism in transformers. Another approach was presented in [241] where the authors aimed to learn low-level

---

[2]https://openai.com/blog/chatgpt

TABLE IX: Recent studies on transformers for **Multi-Modal Applications**.

Author (year)	Datasets	Performance	Architecture(s)
Chuang et al. 2019 [228]	LibriSpeech, TIMIT, SQuAD v2.01, SWBD-Fisher	LibriSpeech: WER 9.8% (dev-clean), 23.4% (dev-other); TIMIT: PER 14.7% (test); SQuAD v2.0: F1 score 76.6%, EM score 69.8%; SWBD-Fisher: WER 10.9%	BERT-base model with a speech encoder consisting of two convolutional layers and four self-attention layers
Song et al. 2019 [229]	TIMIT; WSJ	TIMIT: PER 12.5% (test); WSJ: WER 11.4% (dev93), 10.4% (eval92)	XLNet-base model with a speech encoder consisting of two convolutional layers and six self-attention layers.
Ao et al. 2021 [75]	LibriSpeech; LibriTTS; Common Voice; TIMIT; WSJ; LJSpeech; VCTK	LibriSpeech: WER 4.0% (test-clean), 10.9% (test-other)1; LibriTTS: MOS 4.011; Common Voice: WER 6.8% (en); TIMIT: PER 9.7% (test); WSJ: WER 6.0% (dev93), 5.7% (eval92)1; LJSpeech: MOS 4.021; VCTK: MOS 3.97	T5-base model with a speech encoder consisting of two convolutional layers and six self-attention layers.
Arjmand et al. 2021 [230]	CMU-MOSI	Accuracy score of 76%, F-score of 0.75, MAE score of 0.94 on test set	A pre-trained language model such as BERT or RoBERTa with a speech prefix consisting of two convolutional layers and one self-attention layer.
Sant et al. 2022 [231]	MuST-C, CoVoST v2	MuST-C: BLEU 25.9 (en-de), 28.8 (en-es), 29.1 (en-fr), 20.3 (en-it), 21.0 (en-nl), 22.8 (en-pt), 18.9 (en-ro); CoVoST v2: BLEU 24.1 (es-en), 23.3 (fr-en)	A Transformer model with a head-configurable self-attention module that allows the use of different attention mechanisms in each head.
Lin et al. 2022 [232]	LibriSpeech, VoxCeleb, VoxCeleb	LibriSpeech: WER 3.0% (test-clean), 7.6% (test-other); VoxCeleb1: EER 1.67%; VoxCeleb2: EER 2.12%	A simplified version of HuBERT with a convolutional encoder and a Transformer decoder
Chung et al. 2018 [233]	TIMIT, Buckeye Corpus	TIMIT: Accuracy 0.81 (skip-gram), 0.80 (cbow); Buckeye Corpus: Accuracy 0.83 (skip-gram), 0.82 (cbow)	A sequence-to-sequence model with an RNN encoder and decoder that learns fixed-length vector representations of speech segments.
Li et al. 2019 [55]	LJSpeech, Blizzard2012, Blizzard2011	LJSpeech: MOS 4.13 ± 0.08, MAE:0.13110; Blizzard2012: MOS 4.03 ± 0.07, MAE:0.13610; Blizzard2011: MOS 4.01 ± 0.07, MAE:0.138	A non-autoregressive Transformer model with a multi-head self-attention network and a feed-forward network for both encoder and decoder.

representations by utilizing both visual and linguistic modalities to generate high-level features without explicit supervision. This model, built on top of BERT, utilized bidirectional joint distributions over sequences of visual and linguistic tokens, derived from vector quantization of video data and speech recognition outputs respectively, producing state-of-the-art results. A similar approach was presented in [242] with the proposal of a unified image-and-text generative framework based on a single multi-modal model to jointly study bi-directional tasks. The use of an Encoder architecture for learning generalized representations has gained significant popularity in recent times. A notable effort was made by Nagrani et al. [237] to implement this architecture for multi-modal applications, utilizing fusion bottlenecks for the integration of data from various modalities at multiple layers. This fusion process enabled the bottleneck layers to learn more comprehensive representations of the data, leading to improved performance and reduced computational expenses.

Self-supervised learning has also been utilized as a solution to multi-modal problems, with a majority of such work being focused on video and image applications. Zellers et al. [243] introduced a model, MERLOT, which leveraged self-supervised learning to acquire multi-modal script knowledge through observation of videos transcribed with speech. The model was pre-trained with both frame-level and video-level targets, allowing it to contextualize the data globally. This approach has gained widespread popularity in the video domain due to the abundance of video data available on the internet. Gabeur et al. [244] proposed a multi-modal transformer, which jointly encodes different modalities in video and facilitates their mutual attention, enabling the encoding and mapping of temporal information. A novel modification to the Encoder architecture was presented in the work of Chen et al. [245]. The authors proposed a joint random masking technique applied to two modalities and utilized conditional masking for pre-training tasks, resulting in the creation of the UNITER model. Another self-supervised approach was introduced in the study by Akbari et al. [246], where their model, VATT, took raw input and extracted multi-modal representations through a tokenizer layer, embedding layer, and transformer. This approach leveraged the attention mechanism of transformers to learn the representations of data, producing a model that was more robust for visual and language tasks than prior modality-specific models. The popularity of adversarial learning in creating robust models was also applied, as demonstrated in the work of Li et al. [247], where an adversarial learning model was implemented

on noise input to produce the improved MANGO model on top of UNITER. This approach achieved state-of-the-art results in terms of robustness benchmarks.

The application of Transformers in multi-modal systems is made feasible by their non-recurrent architecture, which enables sequential modeling. The attention mechanism of Transformers allows for learning across a sequence. The Factorised Multi-modal Transformer (FMT) [248] is a new model that makes use of Transformers in multi-modal applications and offers an improvement over existing state-of-the-art models through asynchronous modeling of both intra-modal and inter-modal dynamics. The input in the architecture is first passed through an embedding layer, followed by Multi-modal Transformer Layers (MTL), where each MTL comprises multiple Factorised Multimodal Self-attentions (FMS) that factor in inter-modal and intra-modal aspects of the multi-modal input. The authors of FMT conducted evaluations of its zero-shot task performance and examined if the model learns general representations from pre-trained models. Moreover, they demonstrated that a reduction in the model's losses does not always translate to expected performance gains in multi-modal Transformers. Research has shown that multi-modal Transformer models outperform deeper models with modality-specific attention mechanisms when compared with modality-specific models [249].

IV. CHALLENGES AND FUTURE WORK

A. Training Challenges

Transformers have been proven very effective in speech-related tasks as presented in Section III. However, transformers' training is complex and requires non-trivial efforts regarding carefully designing cutting-edge optimizers and learning rate schedulers [250]. The challenge in terms of applying self-attention to speech recognition is that individual speech frames are not like lexical units such as words. Speech frames do not convey distinct meanings or perform unique functions, which makes it hard for the self-attention mechanism to compute proper attentive weights on speech frames. Considering that adjacent speech frames could form a chunk to represent more meaningful units like phonemes, some sort of pre-processing mechanisms such as convolutions to capture an embedding for a group of nearby speech frames would be helpful for self-attention. Transformers were originally proposed for machine translation, where sequence lengths are short in contrast to speech technology. For instance, sequence lengths in SER are larger and contain a few thousand frames. Self-attention encoders in transformers have quadratic computational complexity and computation of self-attention between all the pairs of frames is expensive to compute. In addition, speech sequences are less informationally dense compared to the word sequences in textual data. Therefore, researchers exploit tricks including time-restricted self-attention [251], truncated self-attention [252], down-sampling [253], sub-sampling [254] and pooling [255] are being used as transformers in speech technology to tackle sequence length problems. Positional encoding is another main component in transformers to include a piece of positional information about each word about its position in the sentence.

This helps the transformers to capture longer dependencies in sequential data. The original paper utilized sinusoidal position encoding, which can hurt performance in speech-based systems due to longer sequences [256] and generalize poorly in certain conditions [257]. Different approaches [110], [257] have been explored to address this issue. However, these approaches are exploited in ASR and further research is required in other speech-related domains.

B. Computational Cost and Efficiency

Recently, transformer-based end-to-end models have achieved great success in many speech-related areas. However, compared to LSTM models, the heavy computational cost of the transformer during inference is a key issue to prevent their applications [98]. The computational cost of the Transformer Transducer grows significantly with respect to the input sequence length, which obstacles the practical use of T-T. Recently conformer Transducer (C-T) [65] was proposed to further improve T-T, but it is not streamable because its encoder has attention on full sequence [99]. However, it requires access to the full sequence, and the computational cost grows quadratically with respect to the input sequence length. These factors limit its adoption for streaming applications [94]. Additionally, transformers' high memory consumption and inference time pose practical difficulties for deploying and updating large-scale models.

Self-attention mechanism in transformers has quadratic complexity with respect to sequence length, limiting scalability for long sequences. To address this issue, several solutions, such as sparse attention patterns [258]–[260], low-rank factorization [261], random feature maps [262], and locality-sensitive hashing [259], [263], [264], have been proposed [3]. The memory consumption of transformers grows linearly with sequence length and quadratically with hidden dimension size, creating challenges for large-scale data training and inference. Some solutions to this problem include reversible residual connections [265], [266], gradient checkpointing [267], weight sharing [268], [269], or parameter pruning [270], [271] to save memory. Chen et al. [272] also show the use of streaming processing and early stopping to reduce latency and run-time cost in speech models. Lin et al. [232] used weight pruning, head pruning, low-rank approximation, and knowledge distillation to reduce parameters. Parallelizing and accelerating transformer models on different hardware platforms may encounter challenges such as load imbalance, communication overhead, or memory fragmentation [273]. Several solutions have been proposed to improve hardware utilization, including tensor decomposition [274]–[277], kernel fusion [278], mixed precision arithmetic [279], or hardware-aware optimization [280], [281].

Efficiency is another major concern for Transformers due to their large and complex architectures. When these models are pre-trained, they may not be efficient for all downstream tasks due to different data distributions. To improve efficiency, recent efforts have attempted to find solutions to use fewer training data and/or parameters. These solutions include knowledge distillation, simplifying and compressing the model, using asymmetrical network structures, improving utilization of training

samples, compressing and pruning the model, optimizing the complexity of self-attention, optimizing the complexity of self-attention-based multimodal interaction/fusion, and optimizing other strategies. Several specific methods have been proposed to address these issues, including knowledge distillation by Miech et al. [282] and Touvron et al. [283], model simplification by Xu et al. [284], Kim et al. [239], and Akbari et al. [246], weight-sharing by Wen et al. [285] and Lee et al. [236], training with fewer samples by Li et al. [286], compressing and pruning the model by Gan et al. [287], optimizing the complexity of self-attention by Child et al. [288] and Transformer-LS [289], optimizing the complexity of self-attention based multimodal interaction/fusion by Nagrani et al. [237] and optimizing other strategies by Yan et al. [290]. These efforts demonstrate the importance of addressing efficiency in the development of Transformers.

C. Large Data Requirements

One significant challenge faced by transformers-based speech models is the requirement for a large amount of data for effective training. While recent text-based conversational AI models have been trained on large amounts of data (e.g., GODEL, which used around 800GB of data, and GPT-3, which was pretrained on 570GB of data), the amount of speech data available for training models for spoken technology is more limited. For instance, the works in Table VIII (with some exceptions) use datasets consisting of several thousand spoken dialogues for various tasks, which is far less compared to the hundreds of gigabytes of text data used by text-based models. This highlights the need to either develop more efficient ways of creating datasets to train speech-related systems more efficiently.

One approach to enhancing the performance of transformer-based models for speech recognition is to collect a large-scale dataset of multilingual and multitask audio data from various sources [52]. This dataset can then be used to train a transformer-based model with self-attention layers using weak supervision. Evaluating the model on various downstream tasks and benchmarks can further improve its performance. Additionally, data augmentation and transfer learning can also be used to improve model performance [85]. Data augmentation techniques such as pitch shifting, time stretching, noise injection, SpecAugment, etc., can be employed to increase the diversity and robustness of the training data. Another solution is to use pre-trained models by training them on learning to learn generalised representation from large-scale unlabeled data [291] [292]. These models can be fine-tuned using few-shot learning to get better performance on downstream tasks. Multi-task learning approaches can also be utilized to enhance the performance of transformer-based models for speech and language processing with smaller datasets [85]. This includes masked acoustic modeling, contrastive predictive coding, speaker classification, emotion recognition, and sentiment analysis.

D. Generalization and Transferability

Transformers face challenges with generalization and transferability that can affect their ability to handle a broader range of tasks and scenarios. One of the main issues with generalization is the absence of inductive biases in pure transformers, which makes them heavily reliant on large-scale training data for optimal performance [293]. This can lead to poor performance on downstream tasks if the training data quality is poor. Additionally, transformers lack built-in biases, unlike convolutional neural networks, which makes it more difficult for them to generalize to new tasks or scenarios [293]. To address the challenge of poor generalization on new domains, Xue et al. [294] proposed the Bayesian Transformer Language Model (BTLM), which integrates a Bayesian framework to enhance the model's ability to handle out-of-domain data. Bayesian Transformer [294] uses variational inference to estimate the latent parameter posterior distributions and account for model uncertainty while another model [232] resolves generalization issue by applying several compression techniques, such as weight pruning, head pruning, low-rank approximation, and knowledge distillation, to a 12-layer Transformer model trained with contrastive predictive coding (CPC). This approach delivers improved performance on out-of-domain data compared to traditional language models. Other proposed solutions include Parallel Scheduled Sampling (PSS) and Relative Positional Embedding (RPE) [295]. PSS improves robustness and reduces exposure bias by randomly sampling tokens from either ground truth or predicted sequences during training. RPE encodes relative distances between tokens rather than absolute positions, which enhances the modeling capability for long sequences. Various other papers also discuss the issues and solutions of generalisation in transformer-based speech models along with various solutions. Zhou et al. [295] propose a text-to-speech model, GenerSpeech that transfer unseen style features from an acoustic reference to a target text. It improves generalization by decomposing the speech variation into style-agnostic and style-specific parts, and by using a content adaptor with Mix-Style Layer Normalization to eliminate style information in the linguistic content representation.

Transferability is also a significant challenge for transformers due to domain gaps. Several methods have been proposed to enhance the transferability of transformers, including data augmentation, adversarial perturbation strategies, learning a shared embedding space, and knowledge distillation [296]–[298]. While these methods have shown promising results, there are still some obstacles to transferability for multimodal applications. One of the challenges is the distribution gap between training data and practical data, as shown by Zhan et al. [299]. They demonstrate that transferring supervised multimodal transformers pre-trained on well-aligned cross-modal pairs/tuples to weakly aligned test data is challenging. Rahman et al. [300] and Xia et al. [301] show that transferring multimodal transformers across different tasks requires careful adaptation and fine-tuning. In multi-language data settings, transformers also face transferability challenges as demonstrated by Zhou et al. [302] and Ni et al. [303].

E. Multimodal Training

In Multimodal Learning (MML) Transformers, a fusion of information across multiple modalities is typically achieved at

three conventional levels: input (i.e., early fusion), intermediate representation (i.e., middle fusion), and prediction (i.e., late fusion) [304]. Middle fusion can be achieved by directly feeding the representations of two modalities into the standard attention module, which is followed by latent adaptation and ends up with a late fusion of final bimodality representations [240], [305]. This idea can be extended by alternating or compounding with unimodal attention, or token exchange across modalities [306]–[308]. On the other hand, inspired by the success of BERT, different modalities can integrate as early as at the input stage [237], [241], [245], [309]–[315]. These models are known as one-stream architecture, which allows the adoption of the merits of BERT with minimal architectural modification. However, a major difference with these one-stream models is the usage of problem-specific modalities with variant masking techniques. A noticeable fusion scheme is introduced based on a notion of bottleneck tokens, which applies for both early and middle fusion by simply choosing to-be-fused layers [237]. Late fusion based on prediction is less adopted in MML Transformers as the focus is on learning stronger multimodal contextual representations [304], [316]. The interaction between modalities can be explored further for enhancing and interpreting the fusion of MML [317].

Another issue with multimodal transformer models in real-world scenarios is that the data often exist in multiple modalities with intrinsic synchronization, such as audio-visual correspondence [318], which forms the basis for cross-modal alignment. Recently, there has been a surge of interest in leveraging large quantities of web data (e.g., image-text pairs) for vision and language tasks using Transformers-based alignment [297], [319]–[321]. The approach involves mapping two modalities into a common representation space with contrastive learning over paired samples and is typically implemented using massive multimodal models (MML) that are expensive to train. As a result, subsequent works have focused on utilizing pre-trained models for various downstream tasks [322]–[326]. These alignment models are capable of zero-shot transfer, particularly for image classification via prompt engineering, which is remarkable given that image classification is traditionally considered a uni-modal learning problem, and zero-shot classification remains a difficult challenge despite extensive research [327]. The approach has also been extended to more challenging and fine-grained tasks, such as object detection [328], visual question answering [245], [311], [329], [330], and instance retrieval [330], [331], by introducing region-level alignment, which incurs additional computational costs from explicit region detection. TEASEL [230] uses a speech-prefixed language model that takes speech features as input and predicts masked tokens in text and resolves issues with multimodal training.

F. Robustness

Despite their widespread adoption in speech processing applications, transformers exhibit sensitivity to domain shifts and noise in speech data leading to a sub-optimal performance in downstream tasks. Additionally, these models may not generalize well to other languages when trained solely on monolingual data [332]. This issue has been identified as one of the causes of performance degradation in transformer-based ASR system [333], speech-to-animation models [47], speaker recognition and speech-to-speech translation models [141] as raw speech features used as input render these models sensitive to noise and speaker variations. Moreover, the performance of these models is negatively impacted by the lack of prosody information, as they do not explicitly model it.

Several solutions have been proposed to overcome these challenges. One approach is to learn robust and multilingual speech representations using contrastive learning, which is a self-supervised technique that encourages the model to distinguish between similar and dissimilar speech segments. A multilingual phonetic vocabulary is used to capture cross-lingual similarities and enable transfer learning across languages. Additionally, self-training and semi-supervised learning are applied to leverage unlabeled data and enhance the quality of the representations. These approaches have been shown to outperform previous methods on various benchmarks and achieve state-of-the-art results on low-resource speech recognition [332]. To address the challenges associated with noisy or distorted speech signals, a novel audio-visual ASR model has been proposed, leveraging both speech and lip movement information to improve recognition accuracy and robustness. The model is based on the Efficient Conformer architecture, which combines convolutional neural networks. To overcome the challenges faced by speaker recognition models, an unsupervised approach that uses a contrastive loss to learn speaker embeddings directly from raw speech signals has also been proposed. The model uses a multi-task learning approach, where both speaker recognition and ASR tasks are learned jointly. This approach can help to overcome the variability in speech signals caused by different speakers and speaking styles, thereby improving the robustness of the ASR system.

V. SUMMARY AND CONCLUSIONS

The transformer architecture has emerged as a highly effective neural network architecture in the field of speech processing due to its ability to handle sequential data for various speech-related tasks. The popularity of transformers has been further accelerated by the availability of specialized libraries for transformer-based speech-processing tasks. The key innovation of transformers lies in their ability to capture long-range dependencies among input sequences using self-attention layers, and its effectiveness has been demonstrated in various speech processing tasks such as automatic speech recognition, text-to-speech synthesis, speaker recognition, multi-microphone processing, and speech translation.

This pioneering paper presents the first detailed and comprehensive survey of the applications of transformers in the audio domain. Our review shows that transformers are increasingly becoming popular in the speech-processing community. The main findings of this paper are summarized below.

- Transformers provide a competitive alternative to Recurrent Neural Network (RNN)-based models in Speech Processing tasks and have shown promising results in Automatic Speech Recognition (ASR) and Text-to-Speech

(TTS). Transformers use self-attention layers to capture long-range dependencies among input sequences, allowing more parallelization than RNNs.

- Pre-training with self-supervised learning techniques like wav2vec and data2vec can lead to faster convergence and performance boost in transformers.
- Hybrid models that combine transformers with conventional acoustic modeling techniques can improve word error rates and reduce computational complexity.
- Attention should be paid to overfitting and generalization problems in transformer models in ASR and TTS.
- Different modeling units such as syllables and phonemes should be compared in ASR using transformers.
- Multi-head attention mechanisms in transformers can improve parallelization by solving the long-distance dependency problem in end-to-end TTS architectures.
- Length regulator based on duration predictor can be used to solve the issue of sequence length mismatch in TTS.
- Data augmentation and contrastive learning techniques can be used to build cross-lingual or multilingual speech recognition systems using pre-trained transformer models.
- Ethical concerns around privacy, bias, and fairness should be considered when developing and deploying speech processing systems using transformers.
- Robustness of transformer models to adversarial attacks and out-of-distribution inputs should be carefully evaluated and addressed in ASR and TTS applications.

Future research on cross-lingual/multilingual systems is needed to address the performance issues highlighted in this review. These recommendations are intended for researchers and developers in the field of speech processing.

VI. Acknowledgements

References

[1] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs RNN in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[2] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, 2022.

[3] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.

[4] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[6] ——, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[7] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, "Learning deep transformer models for machine translation," *arXiv preprint arXiv:1906.01787*, 2019.

[8] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[9] Q. Song, B. Sun, and S. Li, "Multimodal sparse transformer network for audio-visual speech recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[10] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1686–1697.

[11] F. Dang, H. Chen, and P. Zhang, "Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6857–6861.

[12] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[15] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *arXiv preprint arXiv:1905.09418*, 2019.

[16] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2193–2240, 2023.

[17] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, pp. 1–28, 2018.

[18] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE access*, vol. 7, pp. 19 143–19 165, 2019.

[19] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[20] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[21] L. Deng, "Deep learning: from speech recognition to language and multimodal processing," *APSIPA Transactions on Signal and Information Processing*, vol. 5, p. e1, 2016.

[22] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on deep neural networks in speech and vision systems," *Neurocomputing*, vol. 417, pp. 302–321, 2020.

[23] A. M. Brașoveanu and R. Andonie, "Visualizing transformers for nlp: a brief survey," in *2020 24th International Conference Information Visualisation (IV)*. IEEE, 2020, pp. 270–279.

[24] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[25] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri, and M. Almojil, "Automatic speech recognition: Systematic literature review," *IEEE Access*, vol. 9, pp. 131 858–131 876, 2021.

[26] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021.

[27] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *arXiv preprint arXiv:2111.06091*, 2021.

[28] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *arXiv preprint arXiv:2206.06488*, 2022.

[29] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[30] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

[31] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[32] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia *et al.*, "Transformers in remote sensing: A survey," *arXiv preprint arXiv:2209.01206*, 2022.

[33] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian, "Vision transformers for action recognition: A survey," *arXiv preprint arXiv:2209.05700*, 2022.

[34] K. B. Bhangale and M. Kothandaraman, "Survey of deep learning paradigms for speech processing," *Wireless Personal Communications*, vol. 125, no. 2, pp. 1913–1949, 2022.

[35] J. Lahoud, J. Cao, F. S. Khan, H. Cholakkal, R. M. Anwer, S. Khan, and M.-H. Yang, "3D vision with transformers: A survey," *arXiv preprint arXiv:2208.04309*, 2022.

[36] W. Li, H. Luo, Z. Lin, C. Zhang, Z. Lu, and D. Ye, "A survey on transformers in reinforcement learning," *arXiv preprint arXiv:2301.03044*, 2023.

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

[38] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv preprint arXiv:1701.02720*, 2017.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[40] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[41] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[42] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[43] N. R. Wu and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-networks," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, p. 3982–3992.

[44] A. Baevski, M. A. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=r1gEjCNYwS

[45] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[46] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[47] S. Novoselov, G. Lavrentyeva, A. Avdeeva, V. Volokhov, and A. Gusev, "Robust speaker recognition with transformers using wav2vec 2.0," *arXiv preprint arXiv:2203.15095*, 2022.

[48] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[49] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[50] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," https://arxiv.org/abs/1506.06726, 2015.

[51] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, p. 1188–1196.

[52] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[53] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[54] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[55] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[56] G. Wang, "Deep text-to-speech system with seq2seq model," *arXiv preprint arXiv:1903.07398*, 2019.

[57] S. Beliaev and B. Ginsburg, "Talknet 2: Non-autoregressive depth-wise separable convolutional model for speech synthesis with explicit pitch and duration prediction," *arXiv preprint arXiv:2104.08189*, 2021.

[58] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[59] J. Bgn, "Timeline of transformers for speech," https://jonathanbgn.com/2021/12/31/timeline-transformers-speech.html, 2021, accessed: 2023-03-07.

[60] A. T.-Y. Liu, S.-w. Yang, and C.-S. F. J. C.-H. H. H. yi Lee Lin-shan Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *arXiv preprint arXiv:1910.12638*, 2019.

[61] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[62] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[63] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "Speechformer: A hierarchical efficient framework incorporating the characteristics of speech," *arXiv preprint arXiv:2203.03812*, 2022.

[64] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[65] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[66] S.-w. Chang, P.-H. Hsu, Y.-A. Tsai, S.-L. Chen, and H.-y. Lee, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[67] Z. Wang and S. Zhang, "Bridging commonsense reasoning and probabilistic planning via a differentiable neural logic framework," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019, p. 1091–1097.

[68] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[69] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 937–10 947.

[70] Z. Liu, Y. Chen, X. Liang, and J. Zhang, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," *arXiv preprint arXiv:2110.05752*, 2021.

[71] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang *et al.*, "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, 2022.

[72] Z. Liu, Y. Chen, X. Liang, and J. Zhang, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[73] J. Liu, Y. Hou, and W. Che, "Deltalm: Encoder-decoder pre-training for language generation and understanding," 2021.

[74] S. Ma, L. Dong, S. Huang, D. Zhang, A. Muzio, S. Singhal, H. H. Awadalla, X. Song, and F. Wei, "Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders," *arXiv preprint arXiv:2106.13736*, 2021.

[75] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," *arXiv preprint arXiv:2110.07205*, 2021.

[76] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," *arXiv preprint arXiv:2212.07525*, 2022.

[77] OpenAI, "Whisper: Robust speech recognition via large-scale weak supervision," https://github.com/openai/whisper.

[78] Z. Chen, J. Zhang, Z. Liu, X. Chen, and X. Liang, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2022.

[79] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," 2023. [Online]. Available: https://arxiv.org/abs/2303.03926

[80] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[81] T. Wang, J. Deng, M. Geng, Z. Ye, S. Hu, Y. Wang, M. Cui, Z. Jin, X. Liu, and H. Meng, "Conformer based elderly speech recognition system for alzheimer's disease detection," *arXiv preprint arXiv:2206.13232*, 2022.

[82] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "UniSpeech: Unified speech representation learning with labeled and unlabeled data," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 937–10 947.

[83] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[84] S. Papi, M. Gaido, M. Negri, and M. Turchi, "Speechformer: Reducing information loss in direct speech translation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1698–1706.

[85] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[86] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[87] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988.

[88] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. Gomez, S. Gouws, L. Jones, Ł. Kaiser, N. Kalchbrenner, N. Parmar *et al.*, "Tensor2tensor for neural machine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 2018, pp. 193–199.

[89] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[90] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and LSTM encoder decoder models for ASR," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 8–15.

[91] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao, and S. Liu, "On the comparison of popular end-to-end models for large scale speech recognition," *Proc. Interspeech 2020*, pp. 1–5, 2020.

[92] Y. Wang, Y. Shi, F. Zhang, C. Wu, J. Chan, C.-F. Yeh, and A. Xiao, "Transformer in action: a comparative study of transformer-based acoustic models for large scale speech recognition applications," *arXiv preprint arXiv:2010.14665*, 2020.

[93] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 210–220.

[94] C. Wu, Y. Wang, Y. Shi, C.-F. Yeh, and F. Zhang, "Streaming transformer-based acoustic models using self-attention with augmented memory," *Proc. Interspeech 2020*, pp. 2132–2136, 2020.

[95] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," *Proc. Interspeech 2018*, pp. 791–795, 2018.

[96] O. Hrinchuk, M. Popova, and B. Ginsburg, "Correction of automatic speech recognition with transformer sequence-to-sequence model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7074–7078.

[97] Z. Tian, J. Yi, Y. Bai, J. Tao, S. Zhang, and Z. Wen, "Synchronous transformers for end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7884–7888.

[98] L. Lu, C. Liu, J. Li, and Y. Gong, "Exploring transformers for large-scale speech recognition," *arXiv preprint arXiv:2005.09684*, 2020.

[99] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset," *arXiv preprint arXiv:2010.11395*, 2020.

[100] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[101] J. Li, X. Wang, Y. Li *et al.*, "The speechtransformer for large-scale Mandarin Chinese speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7095–7099.

[102] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[103] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *International Conference on Machine Learning*. PMLR, 2017, pp. 195–204.

[104] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *Proc. ICLR*, pp. 214–217, 2018.

[105] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2018.

[106] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[107] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," https://arxiv.org/abs/1609.03499, 2016.

[108] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[109] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," *arXiv preprint arXiv:2006.06873*, 2020.

[110] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for asr," *arXiv preprint arXiv:1904.11660*, 2019.

[111] Z. Zhang, B. He, and Z. Zhang, "Transmask: A compact and fast speech separation model based on transformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5764–5768.

[112] T. Moriya, T. Ochiai, S. Karita, H. Sato, T. Tanaka, T. Ashihara, R. Masumura, Y. Shinohara, and M. Delcroix, "Self-distillation for improving CTC-Transformer-Based ASR Systems," in *INTERSPEECH*, 2020, pp. 546–550.

[113] S. Cao, Y. Kang, Y. Fu, X. Xu, S. Sun, Y. Zhang, and L. Ma, "Improving streaming transformer based ASR under a framework of self-supervised learning," *arXiv preprint arXiv:2109.07327*, 2021.

[114] E. Tsunoo, Y. Kashiwagi, T. Kumakura, and S. Watanabe, "Transformer ASR with contextual block processing," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 427–433.

[115] A. Jain, A. Rouhe, S.-A. Grönroos, M. Kurimo *et al.*, "Finnish ASR with deep transformer models." in *Interspeech*, 2020, pp. 3630–3634.

[116] J. Yu, W. Han, A. Gulati, C.-C. Chiu, B. Li, T. N. Sainath, Y. Wu, and R. Pang, "Dual-mode ASR: Unify and improve streaming ASR with full-context modeling," *arXiv preprint arXiv:2010.06030*, 2020.

[117] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1243–1252.

[118] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "Fftnet: A real-time speaker-dependent neural vocoder," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2251–2255.

[119] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[120] N. Li, Y. Liu, Y. Wu, S. Liu, S. Zhao, and M. Liu, "Robutrans: A robust transformer-based text-to-speech model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8228–8235.

[121] X. Wang, H. Ming, L. He, and F. K. Soong, "s-transformer: Segment-transformer for robust neural speech synthesis," *arXiv preprint arXiv:2011.08480*, 2020.

[122] Y. Zheng, X. Li, F. Xie, and L. Lu, "Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2020, pp. 6734–6738.

[123] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "Multispeech: Multi-speaker text to speech with transformer," *Proc. Interspeech 2020*, pp. 4024–4028, 2020.

[124] D. Lim, W. Jang, O. Gyeonghwan, H. Park, B. Kim, and J. Yoon, "JDI-T: Jointly Trained Duration Informed Transformer for Text-To-Speech without Explicit Alignment," *Proc. Interspeech 2020*, pp. 4004–4008, 2020.

[125] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," *Proc. Interspeech 2020*, pp. 4676–4680, 2020.

[126] T.-Y. Hu, A. Shrivastava, O. Tuzel, and C. Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2020, pp. 3267–3271.

[127] L.-W. Chen and A. Rudnicky, "Fine-grained style control in transformer-based text-to-speech synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7907–7911.

[128] R. Liu, B. Sisman, and H. Li, "Graphspeech: Syntax-aware graph attention network for neural speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2021, pp. 6059–6063.

[129] S. Wang, Z. Ling, R. Fu, J. Yi, and J. Tao, "Patnet: A phoneme-level autoregressive transformer network for speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2021, pp. 5684–5688.

[130] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *2021 IEEE Spoken Language Technology Workshop (SLT)*.   IEEE, 2021, pp. 958–965.

[131] P. Zhang, N. Ge, B. Chen, and K. Fan, "Lattice transformer for speech translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6475–6484.

[132] Huawei. (2022) Speaking your language: The transformer in machine translation. [Online]. Available: https://blog.huawei.com/2022/02/01/speaking-your-language-transformer-machine-translation/

[133] H. Ney, "Speech translation: Coupling of recognition and translation," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 1. IEEE, 1999, pp. 517–520.

[134] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[135] Q. T. Do, S. Sakti, and S. Nakamura, "Toward expressive speech translation: A unified sequence-to-sequence lstms approach for translating words and emphasis." in *INTERSPEECH*, 2017, pp. 2640–2644.

[136] A. Bérard, O. Pietquin, L. Besacier, and C. Servan, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *NIPS Workshop on end-to-end learning for speech and audio processing*, 2016.

[137] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2018, pp. 6224–6228.

[138] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," *Proc. Interspeech 2017*, pp. 2625–2629, 2017.

[139] L.-C. Vila, C. Escolano, J. A. Fonollosa, and M.-R. Costa-Jussà, "End-to-end speech translation with the transformer," *Proc. IberSPEECH 2018*, pp. 60–63, 2018.

[140] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting transformer to end-to-end spoken language translation," in *INTERSPEECH 2019*. International Speech Communication Association (ISCA), 2019, pp. 1133–1137.

[141] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," *arXiv preprint arXiv:2107.08661*, 2021.

[142] R. Huang, Z. Zhao, J. Liu, H. Liu, Y. Ren, L. Zhang, and J. He, "Transpeech: Speech-to-speech translation with bilateral perturbation," *arXiv preprint arXiv:2205.12523*, 2022.

[143] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.

[144] C. Wang, Y. Tang, X. Ma, A. Wu, S. Popuri, D. Okhonko, and J. Pino, "fairseq s2t: Fast speech-to-text modeling with fairseq," *arXiv preprint arXiv:2010.05171*, 2020.

[145] X. Zeng, L. Li, and Q. Liu, "Realtrans: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer," *arXiv preprint arXiv:2106.04833*, 2021.

[146] J. Shor and S. Venugopalan, "TRILLsson: Distilling universal paralinguistic speech representations," 2022.

[147] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2022, pp. 3169–3173.

[148] M. Xu, S. Li, and X.-L. Zhang, "Transformer-based end-to-end speech recognition with local dense synthesizer attention," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2021, pp. 5899–5903.

[149] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[150] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "Speechformer++: A hierarchical efficient framework for paralinguistic speech processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[151] S. Schötz, "Paralinguistic phonetics in nlp models & methods," *NLP term paper*, 2002.

[152] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[153] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," *arXiv preprint arXiv:2002.12764*, 2020.

[154] W. Yu, J. Zhou, H. Wang, and L. Tao, "SETransformer: speech enhancement transformer," *Cognitive Computation*, pp. 1–7, 2022.

[155] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: transformer with Gaussian-weighted self-attention for speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2020, pp. 6649–6653.

[156] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.   IEEE, 2021, pp. 7098–7102.

[157] S. Zhang, M. Chadwick, A. G. C. Ramos, and S. Bhattacharya, "Cross-attention is all you need: Real-time streaming transformers for personalised speech enhancement," *arXiv preprint arXiv:2211.04346*, 2022.

[158] Y. Zhao, C. Luo, Z.-J. Zha, and W. Zeng, "Multi-scale group transformer for long sequence modeling in speech separation," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3251–3257.

[159] W. Jiang, C. Sun, F. Chen, Y. Leng, Q. Guo, J. Sun, and J. Peng, "Low complexity speech enhancement network based on frame-level Swin transformer," *Electronics*, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/6/1330

[160] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[161] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika *et al.*, "Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition," *arXiv preprint arXiv:2009.04323*, 2020.

[162] T. O'Malley, A. Narayanan, Q. Wang, A. Park, J. Walker, and N. Howard, "A conformer-based ASR frontend for joint acoustic echo cancellation, speech enhancement and speech separation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.   IEEE, 2021, pp. 304–311.

[163] B. Gfeller, D. Roblek, and M. Tagliasacchi, "One-shot conditional audio filtering of arbitrary sounds," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 501–505.

[164] A. G. C. P. Ramos, A. Mehrotra, N. D. Lane, and S. Bhattacharya, "Conditioning sequence-to-sequence networks with learned activations," in *International Conference on Learning Representations*, 2022.

[165] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[166] S. Chen, Y. Wu, Z. Chen, J. Wu, T. Yoshioka, S. Liu, J. Li, and X. Yu, "Ultra fast speech separation model with teacher student learning," *arXiv preprint arXiv:2204.12777*, 2022.

[167] C. Subakan, M. Ravanelli, S. Cornell, F. Lepoutre, and F. Grondin, "Resource-efficient separation transformer," *arXiv preprint arXiv:2206.09507*, 2022.

[168] J. Luo, J. Wang, N. Cheng, E. Xiao, X. Zhang, and J. Xiao, "Tiny-sepformer: A tiny time-domain transformer network for speech separation," *arXiv preprint arXiv:2206.13689*, 2022.

[169] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, and Y. Wu, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*. ISCA, 2020, p. 5036–5040.

[170] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol, "DIET: lightweight language understanding for dialogue systems," *CoRR*, vol. abs/2004.09936, 2020.

[171] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019.

[172] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *NeurIPS*, 2019.

[173] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019.

[174] M. Henderson, I. Casanueva, N. Mrksic, P. Su, T. Wen, and I. Vulic, "Convert: Efficient and accurate conversational representations from transformers," in *Findings of the ACL: EMNLP 2020*, vol. EMNLP 2020. Association for Computational Linguistics, 2020.

[175] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*. Association for Computational Linguistics, 2020.

[176] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, 2020.

[177] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *NeurIPS*, 2020.

[178] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A robustly optimized BERT pre-training approach with post-training," in *Chinese Computational Linguistics CCL*, ser. Lecture Notes in Computer Science, vol. 12869. Springer, 2021.

[179] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *CoRR*, vol. abs/1901.08149, 2019.

[180] P. Budzianowski and I. Vulic, "Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems," in *3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP*. Association for Computational Linguistics, 2019.

[181] Z. Liu, G. I. Winata, Z. Lin, P. Xu, and P. Fung, "Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems," in *AAAI*. AAAI Press, 2020.

[182] Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *ACL System Demonstrations*. Association for Computational Linguistics, 2020.

[183] D. Adiwardana, M. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," *CoRR*, vol. abs/2001.09977, 2020.

[184] A. Madotto, S. Cahyawijaya, G. I. Winata, Y. Xu, Z. Liu, Z. Lin, and P. Fung, "Learning knowledge bases with parameters for task-oriented dialogue systems," in *Findings of the ACL: EMNLP*, vol. EMNLP. Association for Computational Linguistics, 2020.

[185] H. Lin, N. Lubis, S. Hu, C. van Niekerk, C. Geishauser, M. Heck, S. Feng, and M. Gasic, "Domain-independent user simulation with transformers for task-oriented dialogue systems," in *SIGdial*. Association for Computational Linguistics, 2021.

[186] M. Rohmatillah and J. Chien, "Causal confusion reduction for robust multi-domain dialogue policy," in *Interspeech*. ISCA, 2021.

[187] M. Epps, J. Uribe, and M. Korpusik, "A new dataset for natural language understanding of exercise logs in a food and fitness spoken dialogue system," in *2021 IEEE Spoken Language Technology Workshop SLT*, 2021.

[188] W. Sun, S. Zhang, K. Balog, Z. Ren, P. Ren, Z. Chen, and M. de Rijke, "Simulating user satisfaction for the evaluation of task-oriented dialogue systems," in *SIGIR*. ACM, 2021.

[189] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *EACL*. Association for Computational Linguistics, 2021.

[190] T. Wu and B. Juang, "Induce spoken dialog intents via deep unsupervised context contrastive clustering," in *Interspeech*. ISCA, 2022.

[191] W. A. Abro, A. Aicher, N. Rach, S. Ultes, W. Minker, and G. Qi, "Natural language understanding for argumentative dialogue systems in the opinion building domain," *Knowl. Based Syst.*, vol. 242, 2022.

[192] B. Peng, M. Galley, P. He, C. Brockett, L. Liden, E. Nouri, Z. Yu, B. Dolan, and J. Gao, "GODEL: large-scale pre-training for goal-directed dialog," *CoRR*, vol. abs/2206.11309, 2022.

[193] Y. Jang, J. Lee, and K. Kim, "GPT-Critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems," in *ICLR*. OpenReview.net, 2022.

[194] A. Srivastava, I. Pandey, M. S. Akhtar, and T. Chakraborty, "Response-act guided reinforced dialogue generation for mental health counseling," *CoRR*, vol. abs/2301.12729, 2023.

[195] H. Takatsu, K. Yokoyama, Y. Matsuyama, H. Honda, S. Fujie, and T. Kobayashi, "Recognition of intentions of users' short responses for conversational news delivery system," in *Interspeech*. ISCA, 2019.

[196] T. Chen, C. Naik, H. He, P. Rastogi, and L. Mathias, "Improving long distance slot carryover in spoken dialogue systems," *CoRR*, vol. abs/1906.01149, 2019.

[197] D. Liu, Z. Zhao, and L.-D. Gan, "Intention detection based on bert-bilstm in taskoriented dialogue system," in *International Computer Conference on Wavelet Active Media Technology and Information Processing*, 2019.

[198] M. Korpusik, Z. Liu, and J. R. Glass, "A comparison of deep learning methods for language understanding," in *Interspeech*. ISCA, 2019.

[199] L. Zhang and H. Wang, "Using bidirectional transformer-crf for spoken language understanding," in *International Conference on Natural Language Processing and Chinese Computing NLPCC*, ser. Lecture Notes in Computer Science, vol. 11838. Springer, 2019.

[200] Y. Qian, Y. Shi, and M. Zeng, "Discriminative transfer learning for optimizing ASR and semantic labeling in task-oriented spoken dialog," in *Interspeech*. ISCA, 2020.

[201] T. Hong, O. Kwon, and Y. Kim, "End-to-end task-oriented dialog system through template slot value generation," in *Interspeech*. ISCA, 2020.

[202] E. Ekstedt and G. Skantze, "TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog," in *Findings of the ACL: EMNLP*, vol. EMNLP 2020. Association for Computational Linguistics, 2020.

[203] K. Gopalakrishnan, B. Hedayatnia, L. Wang, Y. Liu, and D. Hakkani-Tür, "Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study," in *Interspeech*. ISCA, 2020.

[204] Q. Chen, M. Chen, B. Li, and W. Wang, "Controllable time-delay transformer for real-time punctuation prediction and disfluency detection," in *ICASSP*. IEEE, 2020.

[205] T. Hori, N. Moritz, C. Hori, and J. L. Roux, "Transformer-based long-context end-to-end speech recognition," in *Interspeech*. ISCA, 2020.

[206] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, 2021.

[207] V. M. Andreas, G. I. Winata, and A. Purwarianti, "A comparative study on language models for task-oriented dialogue systems," *CoRR*, vol. abs/2201.08687, 2022.

[208] E. Chapuis, P. Colombo, M. Manica, M. Labeau, and C. Clavel, "Hierarchical pre-training for sequence labelling in spoken dialog," in *Findings of the ACL: EMNLP*, ser. Findings of ACL, vol. EMNLP 2020. Association for Computational Linguistics, 2020.

[209] T. M. Lai, Y. Zhang, E. Bakhturina, B. Ginsburg, and H. Ji, "A unified transformer-based framework for duplex text normalization," *CoRR*, vol. abs/2108.09889, 2021.

[210] S. Kim, G. Kim, S. Shin, and S. Lee, "Two-stage textual knowledge distillation for end-to-end spoken language understanding," in *ICASSP*. IEEE, 2021.

[211] A. López-Zorrilla, M. I. Torres, and H. Cuayáhuitl, "Audio embeddings help to learn better dialogue policies," in *ASRU*. IEEE, 2021.

[212] F. Béchet, C. Raymond, A. Hamane, R. Abrougui, G. Marzinotto, and G. Damnati, "Can we predict how challenging spoken language understanding corpora are across sources, languages, and domains?" in *Conversational AI for Natural Human-Centric Interaction - International Workshop on Spoken Dialogue System Technology IWSDS*, ser. Lecture Notes in Electrical Engineering, vol. 943. Springer, 2021.

[213] E. Okur, S. Sahay, R. Fuentes-Alba, and L. Nachman, "End-to-end evaluation of a spoken dialogue system for learning basic mathematics," *CoRR*, vol. abs/2211.03511, 2022.

[214] J. Sakuma, S. Fujie, and T. Kobayashi, "Response timing estimation for spoken dialog systems based on syntactic completeness prediction," in *SLT*. IEEE, 2022.

[215] T. Lin, Y. Wu, F. Huang, L. Si, J. Sun, and Y. Li, "Duplex conversation: Towards human-like interaction in spoken dialogue systems," in *KDD*. ACM, 2022.

[216] A. Waheed, "Combining neural networks with knowledge for spoken dialogue systems," Ph.D. dissertation, Ulm University, 2022.

[217] D. Bekal, S. Srinivasan, S. Bodapati, S. Ronanki, and K. Kirchhoff, "Device directedness with contextual cues for spoken dialog systems," *CoRR*, vol. abs/2211.13280, 2022.

[218] J. Dong, J. Fu, P. Zhou, H. Li, and X. Wang, "Improving spoken language understanding with cross-modal contrastive learning," in *Interspeech*. ISCA, 2022.

[219] J. Svec, A. Frémund, M. Bulín, and J. Lehecka, "Transfer learning of transformers for spoken language understanding," in *International Conference on Text, Speech, and Dialogue (TSD)*, ser. Lecture Notes in Computer Science, vol. 13502. Springer, 2022.

[220] W. Shen, X. He, C. Zhang, X. Zhang, and J. Xie, "A transformer-based user satisfaction prediction for proactive interaction mechanism in dueros," in *International Conference on Information & Knowledge Management (CIKM)*, M. A. Hasan and L. Xiong, Eds. ACM, 2022.

[221] J. Yang, P. Wang, Y. Zhu, M. Feng, M. Chen, and X. He, "Gated multimodal fusion with contrastive learning for turn-taking prediction in human-robot dialogue," in *ICASSP*. IEEE, 2022.

[222] V. Sunder, S. Thomas, H. J. Kuo, J. Ganhotra, B. Kingsbury, and E. Fosler-Lussier, "Towards end-to-end integration of dialog history for improved spoken language understanding," in *ICASSP*. IEEE, 2022.

[223] A. López-Zorrilla, M. I. Torres, and H. Cuayáhuitl, "Audio embedding-aware dialogue policy learning," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, 2023.

[224] M. Firdaus, A. Ekbal, and E. Cambria, "Multitask learning for multilingual intent detection and slot filling in dialogue systems," *Inf. Fusion*, vol. 91, 2023.

[225] J. Mei, Y. Wang, X. Tu, M. Dong, and T. He, "Incorporating BERT with probability-aware gate for spoken language understanding," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, 2023.

[226] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," *CoRR*, vol. abs/2203.02155, 2022.

[227] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *International Conference on Learning Representations ICLR*. OpenReview.net, 2019.

[228] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L.-s. Lee, "Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering," *arXiv preprint arXiv:1910.11559*, 2019.

[229] X. Song, G. Wang, Z. Wu, Y. Huang, D. Su, D. Yu, and H. Meng, "Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks," *arXiv preprint arXiv:1910.10387*, 2019.

[230] M. Arjmand, M. J. Dousti, and H. Moradi, "Teasel: a transformer-based speech-prefixed language model," *arXiv preprint arXiv:2109.05522*, 2021.

[231] G. Sant, G. I. Gállego, B. Alastruey, and M. R. Costa-Jussà, "Multi-former: A head-configurable transformer-based model for direct speech translation," *arXiv preprint arXiv:2205.07100*, 2022.

[232] T.-Q. Lin, T.-H. Yang, C.-Y. Chang, K.-M. Chen, T.-h. Feng, H.-y. Lee, and H. Tang, "Compressing transformer-based self-supervised models for speech processing," *arXiv preprint arXiv:2211.09949*, 2022.

[233] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *arXiv preprint arXiv:1803.08976*, 2018.

[234] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[235] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[236] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, and Y. Song, "Parameter efficient multimodal transformers for video representation learning," *arXiv preprint arXiv:2012.04124*, 2020.

[237] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 200–14 213, 2021.

[238] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Trans-former for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[239] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[240] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[241] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7464–7473.

[242] Y. Huang, H. Xue, B. Liu, and Y. Lu, "Unifying multimodal transformer for bi-directional image and text generation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1138–1147.

[243] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 634–23 651, 2021.

[244] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *European Conference on Computer Vision*. Springer, 2020, pp. 214–229.

[245] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[246] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 206–24 221, 2021.

[247] L. Li, Z. Gan, and J. Liu, "A closer look at the robustness of vision-and-language pre-trained models," *arXiv preprint arXiv:2012.08673*, 2020.

[248] A. Zadeh, C. Mao, K. Shi, Y. Zhang, P. P. Liang, S. Poria, and L.-P. Morency, "Factorized multimodal transformer for multimodal sequential learning," *arXiv preprint arXiv:1911.09826*, 2019.

[249] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Ne-matzadeh, "Decoupling the role of data, attention, and losses in multi-modal transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 570–585, 2021.

[250] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the difficulty of training transformers," in *Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*. Association for Computational Linguistics (ACL), 2020.

[251] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5874–5878.

[252] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, "Transformer-transducer: End-to-end speech recognition with self-attention," *arXiv preprint arXiv:1910.12977*, 2019.

[253] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," *Proc. Interspeech 2018*, pp. 3723–3727, 2018.

[254] J. Salazar, K. Kirchhoff, and Z. Huang, "Self-attention networks for connectionist temporal classification in speech recognition," in *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2019, pp. 7115–7119.

[255] L. Dong, F. Wang, and B. Xu, "Self-attention aligner: A latency-control end-to-end model for ASR using self-attention network and chunk-hopping," in *ICASSP 2019-2019 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5656–5660.

[256] A. Bie, B. Venkitesh, J. Monteiro, M. Haidar, M. Rezagholizadeh *et al.*, "A simplified fully quantized transformer for end-to-end speech recognition," *arXiv preprint arXiv:1911.03604*, 2019.

[257] T. Likhomanenko, Q. Xu, G. Synnaeve, R. Collobert, and A. Rogozhnikov, "CAPE: Encoding relative positions with continuous augmented positional embeddings," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 079–16 092, 2021.

[258] C. Zhao, J. Wang, X. Qu, H. Wang, J. Xiao *et al.*, "Adaptive sparse and monotonic attention for transformer-based automatic speech recognition," *arXiv preprint arXiv:2209.15176*, 2022.

[259] B. J. Woo, H. Y. Kim, J. Kim, and N. S. Kim, "Speech separation based on dptnet with sparse attention," in *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC).* IEEE, 2021, pp. 339–343.

[260] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Constructing transformers for longer sequences with sparse attention methods," *Google AI Blog*, 2021.

[261] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, and P. Fung, "Lightweight and efficient end-to-end speech recognition using low-rank transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 6144–6148.

[262] K. Choromanski and L. Colwell, "Rethinking attention with performers," *Google AI Blog*, 2020.

[263] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.

[264] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 53–68, 2021.

[265] Y. Yu, D. Park, and H. K. Kim, "Auxiliary loss of transformer with residual connection for end-to-end speaker diarization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 8377–8381.

[266] Z. Duan, G. Gao, J. Chen, S. Li, J. Ruan, G. Yang, and X. Yu, "Dual-residual transformer network for speech recognition," *Journal of the Audio Engineering Society*, vol. 70, no. 10, pp. 871–881, 2022.

[267] W. G. Tech, "Sparse transformers and longformers: A comprehensive summary of space and time optimizations on transformer architectures," https://medium.com/walmartglobaltech/sparse-transformers-and-longformers-a-comprehensive-summary-of-space-and-time-optimizations-on-4caa5c388693, 2021, accessed: 2022-01-28.

[268] T. Xiao, Y. Li, J. Zhu, Z. Yu, and T. Liu, "Sharing attention weights for fast transformer," *arXiv preprint arXiv:1906.11024*, 2019.

[269] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, and J. Wang, "On the connection between local attention and dynamic depth-wise convolution," *arXiv preprint arXiv:2106.04263*, 2021.

[270] J. Li, R. Cotterell, and M. Sachan, "Differentiable subset pruning of transformer heads," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1442–1459, 2021.

[271] S. Li, R. Dabre, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation." in *Interspeech*, 2019, pp. 4400–4404.

[272] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 5904–5908.

[273] Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 6783–6787.

[274] X. Ma, P. Zhang, S. Zhang, N. Duan, Y. Hou, M. Zhou, and D. Song, "A tensorized transformer for language modeling," *Advances in neural information processing systems*, vol. 32, 2019.

[275] J. Gu, B. Keller, J. Kossaifi, A. Anandkumar, B. Khailany, and D. Z. Pan, "Heat: Hardware-efficient automatic tensor decomposition for transformer compression," *arXiv preprint arXiv:2211.16749*, 2022.

[276] H. Pham Minh, N. Nguyen Xuan, and S. Tran Thai, "Tt-vit: Vision transformer compression using tensor-train decomposition," in *Computational Collective Intelligence: 14th International Conference, ICCCI 2022, Hammamet, Tunisia, September 28–30, 2022, Proceedings.* Springer, 2022, pp. 755–767.

[277] S. Li, P. Zhang, G. Gan, X. Lv, B. Wang, J. Wei, and X. Jiang, "Hypoformer: Hybrid decomposition transformer for edge-friendly neural machine translation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7056–7068.

[278] D. Timonin, B. Y. Hsueh, and V. Nguyen, "Accelerated inference for large transformer models using nvidia triton inference server," https://developer.nvidia.com/blog/accelerated-inference-for-large-transformer-models-using-nvidia-fastertransformer-and- Aug 2022.

[279] J. Xu, S. Hu, J. Yu, X. Liu, and H. Meng, "Mixed precision quantization of transformer language models for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 7383–7387.

[280] H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, and S. Han, "Hat: Hardware-aware transformers for efficient natural language processing," *arXiv preprint arXiv:2005.14187*, 2020.

[281] O. Kuchaiev, B. Ginsburg, I. Gitman, V. Lavrukhin, J. Li, H. Nguyen, C. Case, and P. Micikevicius, "Mixed-precision training for nlp and speech recognition with openseq2seq," *arXiv preprint arXiv:1805.10387*, 2018.

[282] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking fast and slow: Efficient text-to-visual retrieval with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9826–9836.

[283] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning.* PMLR, 2021, pp. 10 347–10 357.

[284] H. Xu, M. Yan, C. Li, B. Bi, S. Huang, W. Xiao, and F. Huang, "E2e-vlp: end-to-end vision-language pre-training enhanced by visual learning," *arXiv preprint arXiv:2106.01804*, 2021.

[285] K. Wen, J. Xia, Y. Huang, L. Li, J. Xu, and J. Shao, "Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2208–2217.

[286] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," *arXiv preprint arXiv:2110.05208*, 2021.

[287] Z. Gan, Y.-C. Chen, L. Li, T. Chen, Y. Cheng, S. Wang, J. Liu, L. Wang, and Z. Liu, "Playing lottery tickets with vision and language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 652–660.

[288] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[289] Z. Gan, Y.-C. Li, Y. Cheng, J. Liu, and J. Gao, "Long-short transformer: Efficient transformers for language and vision," *arXiv preprint arXiv:2107.02192*, 2021.

[290] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, and C. Schmid, "Multiview transformers for video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3333–3343.

[291] X. Ma, W.-N. Zhang, and C. Wang, "Data augmentation for end-to-end speech recognition," *arXiv preprint arXiv:2301.02111*, 2020.

[292] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[293] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[294] B. Xue, J. Yu, J. Xu, S. Liu, S. Hu, Z. Ye, M. Geng, X. Liu, and H. Meng, "Bayesian transformer language models for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 7378–7382.

[295] P. Zhou, R. Fan, W. Chen, and J. Jia, "Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding," *arXiv preprint arXiv:1911.00203*, 2019.

[296] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6616–6628, 2020.

[297] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning.* PMLR, 2021, pp. 8748–8763.

[298] C. Kervadec, C. Wolf, G. Antipov, M. Baccouche, and M. Nadri, "Supervising the transfer of reasoning patterns in vqa," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 256–18 267, 2021.

[299] X. Zhan, Y. Wu, X. Dong, Y. Wei, M. Lu, Y. Zhang, H. Xu, and X. Liang, "Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 782–11 791.

[300] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020. NIH Public Access, 2020, p. 2359.

[301] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, and M. Zhou, "XGPT: Cross-modal generative pre-training for image captioning," in *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*. Springer, 2021, pp. 786–797.

[302] M. Zhou, L. Zhou, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, "Uc2: Universal cross-lingual cross-modal vision-and-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4155–4165.

[303] M. Ni, H. Huang, L. Su, E. Cui, T. Bharti, L. Wang, D. Zhang, and N. Duan, "M3p: Learning universal representations via multitask multilingual multimodal pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3977–3986.

[304] T. Chen and R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998.

[305] A. K. Sahu, L.-P. Morency, and T. Baltrušaitis, "Low rank fusion based transformers for multimodal sequences," in *Proceedings of The Third Works*.

[306] P. Gao, Z. Jiang, H. You, Z. Lu, S. C. Hoi, and X. Wang, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 6639–6648.

[307] Z. Li, Z. Liu, and X. Zhang, "Causal attention for vision-language tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 10533–10542.

[308] Y. Chen, Y. Wang, X. Liu, C. Qian, L. Lin, and C. C. Loy, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, p. 10442–10451.

[309] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[310] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," *arXiv preprint arXiv:1908.06066*, 2020.

[311] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.

[312] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.

[313] L. Zhu, Z. Xu, and Y. Yang, "Actbert: Learning global-local video-text representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9697–9706.

[314] W. Qi, Y. Su, Y. Zhu, Q. Huang, L. Li, and G. Wang, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," *arXiv preprint arXiv:2001.07966*, 2020.

[315] C. Zhuge, H. Zhang, J. Liang, X. Zhang, and Z. Luo, "Kaleido-bert: Vision-language pre-training on fashion domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, p. 14529–14538.

[316] A. Owens, A. A. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Audio-visual scene analysis with self-supervised multisensory features," *arXiv preprint arXiv:1804.03641*, 2018.

[317] Y. Liu, H. Zhang, X. Liang, P. Liang, and J. Sun, "Probing inter-modality: Visual parsing with self-attention for vision-language pre-training," *arXiv preprint arXiv:2106.13488*, 2021.

[318] P. Morgado, Y. Li, and N. Nvasconcelos, "Learning representations from audio-visual spatial alignment," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4733–4744, 2020.

[319] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[320] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021.

[321] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7331–7341.

[322] Z. Wang, N. Codella, Y.-C. Chen, L. Zhou, J. Yang, X. Dai, B. Xiao, H. You, S.-F. Chang, and L. Yuan, "Clip-td: Clip targeted distillation for vision-language tasks," *arXiv preprint arXiv:2201.05729*, 2022.

[323] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963.

[324] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[325] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2video: Mastering video-text retrieval via image clip," *arXiv preprint arXiv:2106.11097*, 2021.

[326] M. Narasimhan, A. Rohrbach, and T. Darrell, "Clip-it! language-guided video summarization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 988–14 000, 2021.

[327] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[328] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[329] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, p. 5103–5114.

[330] L. H. Liu, H. Shen, and A. L. Yuille, "Align and prompt: Video-and-language pre-training with entity prompts," *arXiv preprint arXiv:2112.09583*, 2021.

[331] ——, "Gilbert: Generative vision-language pre-training for image-text retrieval," in *Proceedings of the Web Conference 2021*, 2021, p. 1143–1154.

[332] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord, "Learning robust and multilingual speech representations," *arXiv preprint arXiv:2001.11128*, 2020.

[333] M. Burchi and R. Timofte, "Audio-visual efficient conformer for robust speech recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2258–2267.