Evaluation Of LSTM, Transformer and TCN In the Field of Auditory Research

Ran Xue *

Department of Electrical Engineering, Universiti Sains Malaysia, State of Penang,14300, Malaysia

* Corresponding Author Email: 748470789xueran@student.usm.my

Abstract. Capturing the temporal correlations inherent in auditory signals and the interconnections among complex features accurately stands as the core technical challenge currently. This paper conducts a systematic review focusing on three mainstream deep learning architectures: Long Short-Term Memory (LSTM) networks, Transformers, and Temporal Convolutional Networks (TCNs). First, it elaborates on the core mechanisms of each model: LSTMs rely on gating control, Transformers leverage the self-attention mechanism, and TCNs are built on causal/dilated convolutions. Second, it summarizes their typical applications in core auditory tasks—including speech recognition, speech emotion recognition, and audio classification—and analyzes their adaptive strategies for special scenarios such as low-resource environments and noisy conditions. Finally, the paper evaluates the strengths and weaknesses of each model across three dimensions and puts forward scenario-specific selection recommendations. Key findings highlight the complementary advantages of the three architectures: LSTMs, with their lightweight design, are well-suited for edge computing under resource-constrained environments; Transformers excel at high-precision, large-scale tasks due to their superior global feature capture; and TCNs excel at tasks requiring local feature sensitivity and real-time processing. This work offers a comprehensive reference for both auditory research and engineering practice.

Keywords: auditory domain; deep learning; LSTM; Transformer; TCN.

1. Introduction

With the deep integration of artificial intelligence and hearing, the field of hearing research will undergo a change, moving from the traditional signal processing paradigm to a deep learning driven paradigm. Core tasks - from speech recognition to audio event detection and even making hearing aids better - need models that can use very tricky time connections and how features match up. So as a result, searching for correct deep learning architecture that can conform to characteristics of auditory signals is now a main prerequisite to making any sort of technological improvements.

Signals heard have changing time properties which have both close and far away information spread all over them, like quick small parts and longer big parts. The early CNNs were able to extract local information from an image, but failed to capture longer term temporal information. traditional RNNs were inherently designed for time modeling, but suffered from the gradient vanishing /exploding problems – which made it impractical for long sequence processing. Against this technological backdrop emerged three architectural paradigms that would power forward movement through the auditory domain: Long Short-Term Memory (LSTM) networks, TransFormers, Temporal convolutional networks.

LSTMs overcome the shortcoming present with conventional RNNs when it comes to working with the passage of time by using a collaborative gating mechanism that brings together forget, input, and an output gate which makes it possible to accurately filter out and keep information tied to important times. such designs provide great success for tasks like Speech Emotion Recognition and Speech Synthesis [1] with this unique design. This kind of design has achieved breakthrough results on tasks that require strong long-term understanding of context. Transformers use self-attention, so they can break the rules of time being connected; they do things at the same time, see all around them at once, and they've been better than LSTMs with big things like speaking out and sorting sounds [2]. TCN, based on causal and dilated convection, can further prolong the field of view without losing timing, so it is more convenient for integrating local details and global information. The Architectural

benefits have introduced innovative Real-time centric applications like Speech improvement, audio-event recognition etc [3].

Although they belong to the mainstream models, these three architectures still have different application performance, which is mainly determined by their own differences in the architectural structure, time modeling logic and computational speed. Thus, this paper reviews LSTMs, Transformers, and TCNs with respect to their overall research evolution, typical applications in auditory domain(s) and their strengths and weakness across three criteria: temporal model, computational cost of model, and data dependency, and it summarizes the current research bottlenecks and the future development directions of this paper as well, trying to give a hint to those academics and professionals working on the auditory field.

2. Auditory Science Research

2.1. Core task application breakdown

2.1.1 Speech Recognition

The development of speech recognition technology is an iterative process, alternating between end-to-end type and hybrid architecture, LSTM, Transformer, and TCN models complement each other technically. Open AI's Whisper model is a major innovation due to its use of a purely self-attention-based structure, breaking away from the old hybrid model architecture "acoustic model-language model – decoding module" And it creates an end-to-end system that does not require the manual extraction of the feature extraction module. It will map the raw audio signals directly to text. [4] Its main strength comes from the self-attention mechanism being able to calculate correlation between any position in the speech sequence, which can correctly capture long-range meaning dependencies. Furthermore, it makes use of parallel computation to relinquish the sequence computation that LSTM always sticks to during model training and inference, which can make model training and inference more effective, improve the efficiency of the entire system's deployment and deployment. The development of speech recognition is constantly refined by iteration based on a combined model and an end-to-end model; LSTM, Transformer, and TCN models are all important technical foundations complementing each other. OpenAI's whisper model is a landmark, since it adopts pure self-attention, unlike most other models that are based on "acoustic model - language model - decoder". It forms an endtoend system doing away with manually engineered feature extraction blocks making it possible to take raw audio signals and turn them straight into text [4-6]. A big strength with this model is that the self-attention mechanism is able to calculate the inter-positional correlation over the whole length of the speech sequence, which means we can get a good long-range dependency in the speech. And also, via making use of parallel computing, Whisper rejects the sequential calculation limitation that is an attribute of LSTMs, thus improving both the training and reasoning efficiency of the model, and making the deployment process more streamlined.

2.1.2 Speech Emotion Recognition

The essence of SER is to extract emotional information from two categories of information: prosodic feature information such as speech rate, tone, etc., and spectral feature information such as MFCC, mel spectrogram, etc. Based on Transformer architecture of wac2vec 2.0 and HuBERT variants, the core distinction lies in taking advantage of attention for modeling global prosodic trends and local spectral details at the same time. And it surpasses the old models where the capturing can only be one-dimensional, making manual feature extraction and segmentation unnecessary and using the pre-training paradigm to create cross-scenario transferal ability as well as the end-to-end fusion in learning the emotion information [7].

Long Short-term Memory (LSTM) networks were as well largely utilized in SER tasks. just like Zhang et al.'s fusion of LSTM -CINN model, extract complementary characteristics of time and locality, and obtain stable results. Therefore, it proves the important role of lstm in emotion modeling research [8]. Also, CNN-LSTM + bayes optimization achieves superior recognition resilience within

noisy environments owing to the implementation of automatic parameterization via Bayesian optimization [9-13].

This transformer-based model has excellent practical performance. MSP-Podcast dataset on the three-dimensional sentiment recognition task, it attained a valence prediction CCC accuracy level of 0.638, setting a new record for the dataset. In cross corpus validation for IEMOCAP& MOSI, the model could maintain stable performance, and in English to Chinese bilingual transfer learning, the model got 76% accuracy - 68% of LSTM's accuracy was way higher. It mainly has the following advantages. The self-attention mechanism can flexibly assign weights, which helps in fully capturing correlations and differences in affective features; The model is more adaptable to disturbances like background noise and speech rate changes; performance is still excellent under complex environmental conditions; It demonstrates better robustness compared to CNN based baseline models.

It has some shortcomings in application. First is the quadratic complexity, a lot of hardware required, hard to put on devices that have less computational power. second is that it heavily relies on a large amount of labeled data, thus it has a limit in small sample cases. Third it is inexact about getting at weak sentiment trait s [7]. But it is still effective at resolving inherent shortages of traditional kinds – hard to balance global and local features, poor generalizability, environmental sensitiveness, and complicated process of making models – and makes the model-making process simpler, improves recognition's efficiency and stability

2.1.3 Audio Classification

Audio classification—ambient sound, music, and speech—calls for a balance: in feature extraction efficiency and classification accuracy, there's three kinds of model types, each taking over certain beneficial circumstances. An important development in this area is the Audio Spectrogram Transformer (AST) proposed by the MIT team, which mainly combines Mel-spectrogram segmentation with Visual Transformer (ViT) transfer learning. and it can overcome the lack of global semantic correlation of the traditional approach. The audio Mel-spectrogram is treated as a 2-D image-like feature, so the Transformer's self-attention can be used to model local spectral information and global structure information simultaneously. For that matter it eliminates the requirement of manual feature extraction module(s). Resultantly we arrive at a clean end-to-end classification framework [5].

The AST performs excellently in practical evaluations. On the GTZAN Music dataset involving 10 music genres, the model reaches a classification accuracy of up to 94.2%, which represents an improvement over the 83.5% accuracy of LSTM models and establishes a new state-of-the-art record on this dataset. And it also can have very stable performance on different dataset like AudioSet(largesample general audio) and ESC-50(environment sound) which can proof it's very good at all kinds of situation[5]. The model's key strength is two-fold - first, the self-attention mechanism allows for capturing both global structures e.g. musical melodies, as well as local fine-grained details e.g. instrument timbres at the same time which leads to more effective classification, second the model reuses pre-trained ViT parameters which means that less large-scale annotated data is needed for training the model as well as shorter training cycles. There are quite some limitations of using the AST in real world. At first, it is computational complex with quadratic time complexity, so it can't be deployed on low power edges because it needs special hardware. The second, its inference speed is slow, cannot satisfy the millisecond-level real-time requirement. Thirdly, there is limited improvement when dealing with smaller sized dataset, so this model has worse cost effectiveness than LSTM model [5]. Despite these issues, the AST can efficiently overcome the shortcomings of the LSTM method: weak global correlation, cumbersome and complex manual feature design, and poor generalization to different tasks, so as to achieve "all-in-one" high-precision audio classification.

2.2. Application Scenarios: Extended Analysis

2.2.1 Small samples and dialects

The core problem in low resource setting is mainly due to insufficient feature learning caused by lack of data and difficulty in captureing unique prosodic charateristics of rare languages and dialects using general model. Three kinds of models, LSTM, Transformer and TCN, have developed distinctive adaptation pathways corresponding to their inherent characteristics, which have effectively resolved resource limitations.

In low-resource settings, the Transformer stands as the “benchmark of accuracy,” mainly leaning on “pretrain-finetune” approach and PEFT. Taking advantage of large scale multilingual pre trained models like XLSR - 53, HuBERT, accumulating cross - lingual knowledge by adapting to target language with minor param fine - tuning, decreasing data depend. Take Tsinghua University’s zero-resource speech recognition solution as an example, which reached an average WER of 33% across 8 minor languages, surpassing supervised models trained with just 10 hours of label data. In small sample audio classification, it got to 89% of the full dataset accuracy while only using 5% of data, very data efficient [10]. Transformer’s self-attention can accurately grasp the global semantic connections so as to alleviate the shortcomings of feature sparsity under the condition of limited resources. However, it has a lot of computational challenges, so it can’t be used on this type of edge type of devices, and it’s better suited for some cloud high-precision type environments.

LSTM is the best choice for “cheap adaptation” because of it lightweight and can-do transfer learning. By making the gating structure simpler and reducing the dimensions of features to make the model simpler, and by inheriting and pretraining the parameters of the general language model, the LSTM structure can avoid the problem of overfitting caused by training from scratch. When doing dialect recognition work with only 10 hours of labeled data, the LSTM model cut the WER down to 12.3%, which was better than the starting point by 35%. Furthermore, another method of hybrid CNN-LSTM and it is using Bayesian optimization, which help with stable recognition on the case of noisy data by an automatic parameter adjustment [13]. LSTM-LRASR was proposed by Shufan et al, it reached a relative WER of 29.9% on comparison with baselines through optimal feature extraction methods, data augmentation and discriminative training, which proves its practicality for resource-limited situations. The main advantage of LSTM is that it is less costly as it can run on an Edge Device without too much computing power. Nevertheless, its cross-lingual generalization ability is still inferior to that of Transformer models.

TCN provides a new solution that focuses on "Local Feature Extraction and Robustness Improvement". TCN has achieved more efficient and effective feature extraction through optimizing convolution kernels and increasing dilation factors, in order to improve local feature extraction. At the same time, the improvement of robustness is achieved by using data augmentation methods such as spectral masking to alleviate the problem of feature sparsity. This method got an 18% better accuracy on small-sample speech emotion recognition and could really catch dialect-specific intonation patterns in dialect emotion recognition situations [12]. In terms of the improvement made, the ATCN-GRU model proposed by Fan Yonghong et al., combining TCN with attention and GRU alleviated the issue of sample imbalance leading to recognition bias in the EMODB, IEMOCAP dataset—this proved that TCN was effective in local features. TCN has no sensitivity to language similarity, the memory grows linearly over time, but it has a small computational cost, these compensate for the loss of LSTM in capturing local information. These characteristics make TCN especially suitable for niche tasks like dialect phoneme recognition, but it’s not great at modeling global semantics.

And these three model types together form an integrated technical ecosystem: transformers dominate high-precision cloud-based scenarios, LSTMs fit the low-computation edge tasks, and TCNs fill the hole of local feature-specific apps. both drive the varied use of speech processing tech in scarce-resource places.

2.2.2 Noise Environments (Indoor/Outdoor Noise)

The main requirement of speech technology in the presence of noise is to suppress the interference while retaining the target speech signal. LSTM and Transformer arch structures have formed unique optimization routes due to the inherent difference in structure and thus show different robust advantages and practical value in stable-state and unsteady-state noise situations.

Transformer is the "benchmark of high accuracy" of the noisy environment task through the way of multimodal fusion and cross-layer attention. On indoor steady-state noise scenarios, it is based on multimodal attention to fuse speech and noise features. The model learns the noise distribution in advance by incorporating ambient noise data (such as air conditioner hum, computer clicking, etc.). For example, at SNR = 5dB, the Whisper model can achieve WER as low as 10.2%, compared with the WER of 15.6% of LSTMs [4], which is a huge improvement. When it comes to outdoors non-steady-state noises such as traffic noise, crowd sound noise, Transformer maintains target speech temporary continuity through Cross-layer Attentions Fusion and better positional encoding. Even with the unfavorable SNR = 0 dB, its WER only increases to 18.7%, significantly better than the 28.3% WER of the LSTM model [7]. The main advantage of the Transformer is its self-attention method that can capture global semantic information precisely for a good distinction of the target speech features over the noise ones. And because it takes advantage of the enlarged generalization ability, thanks to lots of pre-training, it can address different types of noise. But it has a very high demand for computing, so it is very hard for it to be applied in situations with low computation power and real time

LSTM has become a "feasible selection" for noisy environment tasks that require low-computational-power usage. This has been made possible by lightweight optimization and noise adaptation. For indoor steady-state noise, NAT optimizes the gating threshold, with or without combining with spectral subtraction as a preprocessing step, the result WER is 15.6% at SNR = 5 dB which balances accuracy and resource cost [8]. In outdoor non-steady-state noise conditions, the reset gate of the Gated Recurrent Unit (GRU) quickly forgets impulsive noise. Although its WER (28.3%) is larger than transformers, its training cost is greatly reduced. Furthermore, hybrid model based on the combination of LSTM with Kalman filtering can improve the recognition stability through dynamic state estimation and is a feasible method for low-computational-power devices [8]. In terms of gating mechanism, it can help the LM do dynamic filtering of temporal signals so as to adapt to noisy circumstances;LSTMs after noise-adaptive training shows excellent results in the acoustic echo cancellation task, effectively eliminating speech interference [11]- -showing its practical application value.

These two kinds of models go together: Transformers excel at high-precision, non-real-time scenarios, and LSTMs are better for low-computational-power, real-time scenarios. Their unique optimization paths give different examples about how to make speeches work better in noisy places when trying new things, and this pushes strong speeches to get better at being right but still not using too much money.

3. Model Strengths and Weaknesses System Evaluation

3.1. Multi-dimensional Evaluation Framework

Focusing on the central processing power, long/short-range temporal dependence coverage, features acquisition accuracy, ability for new task/adaptive generalization are the 3 main ones. The first metric requires the acquisition of both long sentence logical relationships and short-term phonetic details at the same time, while the last two respectively control the accuracy of key features and the performance of cross-task transfer. commonly used quantitiative ones are speech recognition accuracy and sentiment classification F1-score.

With respect to computing resource expenses and operational efficiency, the principal indicators encompass training/inference pace, computational force costs and store size footprint. Short-sequence

speech takes precedence on real-time response performance; long-sequence ones focus mostly on efficient use of computational resources and memory. They collectively indicate whether an actual working model is possible based on different resources:

Taking engineering implementation and scenario adaptability as the focus of evaluation, the main assessment dimensions are deployment complexity, small-sample/noisy-environment adaptability dimension, and hyperparameter tuning dimension. Deployment difficulty directly ties to the application scope for a model, how well a scenario can be used determines if lab-bred techs can be truly used in life, how complex the tuning is also is connected to enginering maintenance costs, an important factor that is very different from plain performance metrics in practical task.

3.2. Comparative Analysis of Three Model Categories

3.2.1 LSTM vs. Transformer

In long-range temporal dependency capture, the Transformer leverages self-attention mechanisms to establish direct connections between arbitrary positions in the sequence. Its information propagation path length follows a constant order, eliminating reliance on chained state transmission. On long-sequence speech datasets such as TED-LIUM-v2, the Transformer demonstrates significantly higher recognition accuracy than LSTM [14]. In contrast, LSTMs rely on chained hidden state propagation; although their gating mechanism mitigates gradient vanishing issues, substantial information decay still occurs in ultra-long speech sequences exceeding 300-time steps, resulting in an accuracy that is 15–20% lower than that of Transformers [2].

In terms of computational efficiency, Transformers support highly parallelized computation, processing all elements of the speech sequence simultaneously during training. Under the same number of training iterations, their efficiency in handling long sequences is far superior to that of LSTMs. LSTMs inherently perform sequential computation, where the state at time step t strictly depends on the result at time step t-1. Even with GPU-based parallel optimization, their training and inference speeds for long sequences remain significantly slower, and they only maintain competitiveness in short-sequence tasks [14].

There are other distinctions as well: LSTMs exhibit better robustness in small-sample speech scenarios, simple hyperparameter tuning is required, and less dependency on data augmentation techniques, a 15% performance improvement can be achieved just by applying basic augmentation. On the contrary, Transformers require massive speech datasets, they’re very likely to overfit in case of small samples, dealing with this kind of problem needs complex data increasing approaches such as SpecAugment, which could improve the result by at most 33%. Also, self-attention makes memory use go up by the square of the length of the sequence, which means there are really bad memory limits when dealing with big blocks of words to hear [14].

3.2.2 LSTM vs. TCN

In terms of feature extraction and dependency extraction, TCN used Causal Conv with dilated Conv. This building can make TCN be good at learning local pronunciation characteristics and also expand the receptive field flexibly with dilation factors, so it can get effective long-term dependence modeling. thus, TCN has better noise robustness for feature extraction of noisy speech recognition tasks Different from LSTMs, they are more natural in long term semantic logic for speech but with the structural characteristics leading to low precision for local features, compared to TCNs. Also, the vanishing gradient problem in long sequence is not solved and the performance is prone to degeneration [12].

On a more efficient note, TCN can carry out computations that run in parallel very efficiently as it has a lot quicker training speeds than LSTMs run through sequentially, so when it comes to dealing with lengthy speech sequences, this becomes very apparent. On the contrary, the LSTM has a simpler design and usually has a smaller size than the TCN. When it comes to short-sequence speech tasks, LSTMs typically consume less memory than TCNs and have only slightly worse inference latency, which makes them more appropriate for short, strictly real-time interactive situations [3].

And the other difference is that LSTM has less difficulty of adopting engineering, higher compatibility and can adapt to all different kinds of edges and clouds very quickly. The TCN model design comes with many hyperparameters, such as the convolution kernel's size, dilation and the network's depth. So, making it difficult to adjust. However, in speech related temporal tasks (e.g., fault prediction of Industrial IoT devices), TCNs have better performance stability because they have good local features' capturing ability [3].

Noise Environment – indoor noise / outdoor noise

Speech technology requirements for noisy environments mainly refer to noise suppression technology, which requires noise to be suppressed while retaining the target speech. LSTM and Transformer architectures develop different routes to optimization depending on their structure, and they display different levels of robustness advantage and real-world value in steady-state vs non-steady-state noise scenarios.

Transformer is now the "benchmark for high precision on noisy environments" due to its use of multimodal fusion, cross-layer attention. In an indoor steady-state noise environment, it combines the speech features with noise by using a multimodal attention approach; The model learns the noise distribution beforehand by using ambient noises such as air condition hum, keyboard clicks during Pretraining. To show the superiority, for example, Whisper model can reach an ultra-low Word Error Rate (WER) of only 10.2% at SNR=5dB, which is much less than the 15.6% WER of LSTM models. In the case of facing outdoor nonsteady - state noise, its traffic noise, people talking, the transformer preserves the temporal coherence of the target speech through cross - layer attention fusion and an enhanced position encoding and even at an extremely unfavorable SNR = 0dB, it's WER only climbs up to a very impressive 18.7%, which is many times better than the 28.3% WER reported for LSTM models [7]. The most direct reason for this is that the core advantage of Transformer is self-attention, which can precisely capture the global semantic information, so as to distinguish the difference between target speech and noise characteristics. Furthermore, the broadening from large-scale pre-training allows it to cope with all sorts of noise. But it has high computational consumption and delay, so it can't be used in real-time, low computational consumption scene;

LSTM has also become practical in environments where the tasks and the computation power are smaller due to being a lighter task as it also adapts for noise. Indoor steady state noise, for Noise Adaptive Training (NAT), optimizing of gating threshold takes place when utilized with spectral subtration pre-processing, WER of 15.6% is kept at SNR=5dB, and a compromise on accuracy and resource consumption is achieved [8]; When it comes to outdoor non-steady state noise environment, we can see that the reset gate of GRU, a variant of LSTM, allows for quick elimination of impulsive noise. Even though its WER (28.3%) is more, the training cost is lower. And also, A hybrid model that is a combination of LSTM with Kalman filter improves on stability through the use of dynamic state estimation, and provides a good solution for low computational devices (8). The Gating Mechanism of LSTMs enables dynamic filter suitable temporal signal processing for noisy signal not restricted to only noise adaptable training. The performance of LSTM network based on Dual-Signal Conversion was also well enough for the Acoustic Echo Cancellation tasks in terms of suppressing of human speaking interference [11] further demonstrating the real scenario environment use of LSTM application.

These two kinds of models are complementary: Transformers are used in high-precision, non-real-time scenarios. LSTMs handle low-computation-low-real-time scenarios. And its different optimized paths give more and more specific situations of speech technology applications on noisiness environment's multi-references and lead a strong speech recognition towards accuracy and cost's compromise.

3.2.3 Transformer vs. TCN

In terms of dependency capture and generalization, the Transformer's self-attention is good at finding out global speech correlation such as the semantic correspondence between paragraphs in a long utterance which can prove its better task adaptiveness and generalization. Complex tasks like multilingual speech synthesis and cross-regional speech data analysis performance is also very good.

On the contrary, TCN likes to strike a balance between capturing enough local information and capturing long term dependency to be good at basic things like recognizing someone talking and knowing what's going on at different times. Although, it can't capture global correlations as well as Transformers, and also has weaker cross-transfer across tasks [5].

To put it in terms of efficiency and resource consumption, TCN can be comparable to Transformers on the efficiency of parallel processing and relatively controllable memory usage. TCN avoids the quadratic memory rise brought about by self-attention systems and shows better resource efficiency for ultra-long speech sequences. When it comes to tasks with equivalent accuracy requirements, the Transformers incur larger computing burdens and do not exhibit good resource efficiency in the case of a few samples. Their performance benefits come from big size pre-training and strong computational resources [15].

Other notable differences are that the transformer is easy to scale because it has a module-style design, and hybrid models combining transformers and LSTMs have achieved very good results on certain complex temporal tasks, becoming an important research theme in deep learning. The opposite is true with TCNs having better robustness to noise. Moreover, modern TCN models can achieve a better trade-off between efficiency and capability due to structure optimization compared to some variants of transformers--these relevant findings often appear in top speech processing conferences and journals [3].

3.3. Comprehensive Scenario Selection Recommendations

Resource-rich, high-precision long-sequence scenarios: For applications such as cloud-based long-form speech transcription, multilingual speech synthesis, and large-scale speech sentiment analysis, Transformer models are recommended as the priority. Their self-attention mechanism enables precise capture of global semantic relationships, while parallel training fully leverages abundant data and computational resources. When combined with data augmentation techniques to mitigate overfitting, Transformers significantly outperform traditional models in long-sequence processing and multi-task scenarios.

Resource-constrained, real-time edge deployment scenarios: For applications including offline mobile voice assistants, local voice command recognition in smart hardware, and voice interaction on lightweight devices (e.g., Raspberry Pi), adaptable lightweight models are preferred—with LSTM being the optimal choice. These models feature streamlined architectures, low memory footprints, and simple deployment processes. They maintain stable performance in small-sample scenarios without requiring extensive pre-training, deliver fast real-time responses, and allow for convenient hyperparameter tuning—making them ideally suited for resource-constrained environments.

Medium-resource, local-feature-sensitive, noisy scenarios: For applications like in-car speech recognition, industrial voice control, dialect transcription, etc., the TCN model is preferred. It has an expanded convolutional structure, which can precisely capture the local feature details, it can do computations in parallel very effectively, and it is also able to hold up against noise fairly well. Moderate data volume, TCNs reach their best performance-resource tradeoff, proper for complicated real-world situations with lots of noise interference and important local details.

For applications involving both long-form speech and short-interval interactions (e.g., intelligent customer service) or frequent transitions between noisy environments (e.g., voice-enabled navigation), hybrid architectures such as LSTM-Transformer are recommended. This design integrates LSTM's strengths in local temporal modeling and small-sample robustness with the Transformer's advantages in global context capture and parallel efficiency. It can flexibly adapt to cross-scenario and multi-task fusion requirements, demonstrating superior scenario adaptability.

4. Conclusion

This paper systematically reviews the core mechanisms, practical applications, and adaptability characteristics of LSTM, Transformer, and TCN architectures in the auditory domain. Through a

multi-dimensional evaluation framework, it clarifies the differentiated advantages and applicable scenarios of each model. LSTMs, with their lightweight gated structures, are well-suited for low-resource and edge deployment scenarios; Transformers, leveraging self-attention mechanisms, excel in large-scale, high-precision long-time-series tasks; TCNs, through the innovative integration of causal and dilated convolutions, have developed unique competitiveness in scenarios requiring local feature sensitivity and real-time noise mitigation. Collectively, these three model types form a complementary technical ecosystem that provides comprehensive support for the diversified application of speech processing technologies.

References

[1] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation, 2000, 12(10): 2451-2471.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Montreal: Curran Associates Inc, 2017: 5998-6008.

[3] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. 2018.

[4] OpenAI. Whisper: Robust speech recognition via large-scale supervised training. San Francisco: OpenAI, 2022.

[5] Yuan G, Chung Y A, Glass J R. AST: Audio spectrogram transformer. IEEE Transactions on Audio, Speech, and Language Processing, 2022, 30: 3040-3053.

[6] Zhang X, Wang L, Lee H. Speech emotion recognition using LSTM and CNN//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Chengdu: IEEE, 2018: 52-56.

[7] Schuller B, Batliner A, Steidl S. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 5812-5825.

[8] Wang Y, Chen J, Liu W. Noise-robust speech recognition using LSTM with noise adaptive training. Journal of Signal Processing Systems, 2020, 92(7): 893-902.

[9] Hinton G, Deng L, Yu D. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[10] Tsinghua University Speech Laboratory. Zero-resource speech recognition: Cross-lingual transfer learning based on XLSR-53 and HuBERT. Beijing: Tsinghua University, 2021.

[11] Wang S, Chen Y, Zhang H. Acoustic echo cancellation with the dual-signal transformation LSTM network. IEEE Transactions on Audio, Speech, and Language Processing, 2020, 28: 2654-2666.

[12] Lee J, Park S, Kim H. Temporal convolutional networks for environmental sound classification. Applied Acoustics, 2021, 175: 107803.

[13] Li M, Zhang M, Liu G. Speech emotion recognition based on CNN-LSTM with Bayesian optimization. Pattern Recognition and Artificial Intelligence, 2022, 35(4): 321-328.

[14] Pitaojun. A comparison of transformer and LSTM encoder decoder models for ASR. 2024.

[15] Luo Y, Mesgarani N. ConvTasNet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[16]