Fully Convolutional Transformer-Based Speech Emotion Recognition for Automotive Systems

Hanwook Chung and Hyunjin Yoo
Forvia IRYStec Inc., Montreal, QC, Canada

Abstract
We introduce a fully convolutional transformer for speech emotion recognition with application to automotive systems. The proposed architecture is composed of convolutional channel expansion, multi-head attention and feed-forward layers. We employ a trainable emotion query to better capture the characteristics of different emotions. In addition, we consider channel attention to better enable real-time processing. Experiments show that the proposed method provides better performance than the benchmark algorithms.

Author Keywords
Speech emotion recognition, classification, convolutional neural network, transformer, attention mechanism.

1. Introduction
The goal of speech emotion recognition (SER) algorithms is to automatically determine the emotional state of the speaker from the given speech signal [1]. SER has become an important aspect for human-computer interaction (HCI) and finds various applications, e.g., customer support call center, mental depression diagnosis and automotive applications such as driver condition monitoring. A typical SER system consists of acoustic feature extractor and emotion classifier.

In recent years, deep learning (DL)-based techniques have attracted enormous interest and found diverse applications due to its strong nonlinear modeling capabilities that can effectively extract complex features from a given data. Various DL-based algorithms for SER have been introduced. For example, a combination of phoneme sequence and spectrogram-based convolutional neural networks (CNNs) has been proposed in [2]. In [3], the authors proposed to use multiple parallel CNN blocks with different filter sizes to better extract the spectro-temporal information. The authors in [4] introduced an adversarial autoencoder-based classifier by augmenting the data within real data distribution, for a more robust recognition.

Besides, an attention mechanism-based transformer, which was first introduced for machine translation [5], has been widely adopted in various DL-based applications. It is well known to effectively focus on relevant information while disregarding the irrelevant ones. Regarding the SER task, training a recurrent neural network (RNN) with a weighted pooling over time based on local attention has been introduced in in [6]. A convolutional recurrent network with attention layer has been proposed in [7], while the authors in [8] incorporated gender classification as an auxiliary task and introduced multitask learning. In [9], the Mel-spectral coefficients with their derivatives are used for training a 3-D attention-based CNN. In [10], the attention mechanism was introduced into the forgetting gate of long short-term memory (LSTM) cell. The authors in [11] exploited both acoustic and lexical information from speech and proposed a multi-scale CNN. In [12], a CNN-based autoencoder with emotion embedding has been introduced. In [13], the authors proposed a multi-task learning algorithm using emotion attribute information inspired by emotion perceptive process of the human brain. However, most attention-based networks first process the given feature through multiple hidden layers, followed by the multi-layer perceptron (MLP)-based attention layer, which may limit the full capabilities of attention mechanism. Moreover, most approaches are implemented based on the attention score computed across time frames, which may limit a real-time processing for emotion recognition.

In this paper, we introduce a fully convolutional transformer-based architecture, motivated by [14] and [15], for SER with application to automotive systems. The proposed architecture consists of encoder and decoder blocks. The encoder blocks are composed of a channel expansion layer, convolutional multi-head self-attention and convolutional feed-forward layers. The decoder blocks consist of convolutional multi-head self-attention, cross-attention and convolutional feed-forward layers. Specifically, we employ a trainable emotion query to better capture the characteristics of different emotion states. Moreover, instead of computing the attention scores across time frames, we consider frame-wise channel attention to better enable real-time processing, which is especially useful for automotive applications such as driver emotion monitoring for emergency detection. Experimental results showed that the proposed emotion classification method using a fully with the fully convolutional attention-based architecture provided better recognition performance than the selected benchmark algorithms.

2. *Proposed Architecture*
In the proposed framework, we use the log-Mel filterbank (LMFB) coefficients as the input feature, $\mathbf{x} \in \mathbb{R}^{T \times F}$ where T is the number of time frames and F is the feature size. The overall architecture of the proposed fully convolutional transformer is illustrated in Figure 1. The main difference from the original attention mechanism introduced in [5] is that we replace the MLP-based layers with convolutional layers. The underlying motivation is that CNNs have shown great potential on extracting local patterns of the given features efficiently, as well as they require fewer model parameters, in general. Below, we explain the sub-blocks of the proposed transformer architecture in detail.

Proposed Encoder
The proposed encoder is composed of a stack of N_e blocks. In each block, we first perform a channel expansion as illustrated in Figure 1 (a) for a better feature extraction and efficient learning, which is similar to a typical CNN layer. The channel expansion block consists of a convolutional layer, batch normalization, rectified linear unit (ReLU) activation function and average pooling. Specifically, we use $N_e = 5$ with the numbers of channels of $\{64, 128, 128, 256, 256\}$ with $(k, s, p) = (3,1,1)$ for

ISSN 0097-996X/25/5602-0979-$1.00 © 2025

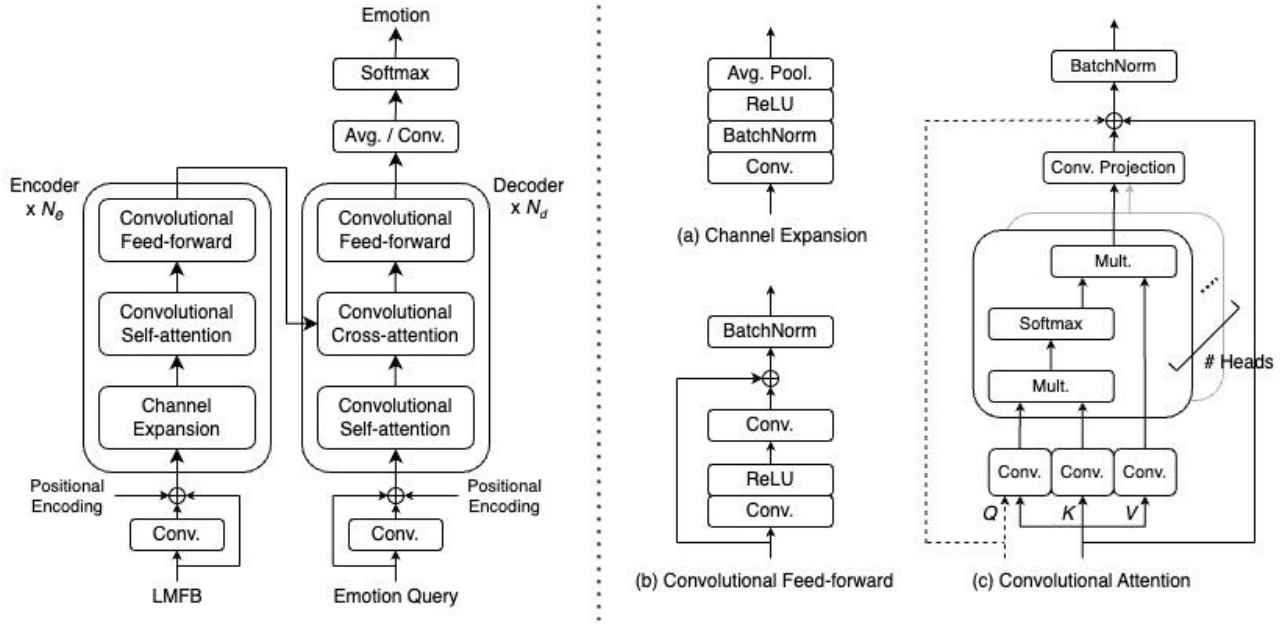

Figure 1. Proposed fully convolutional transformer architecture. Left column: overall block diagram of the proposed transformer. Right column: detailed sub-blocks of the encoder and decoder.

the convolutional layer and $(k, s, p) = (3,2,1)$ for the average pooling layer, resulting in extended channel with reduced feature size. Note that k, s and p respectively denote the kernel size, stride and padding size.

Second, the convolutional self-attention is performed as illustrated in Figure 1 (c). For a given *i*-th encoder block, we first apply convolution operation to the query ($\mathbf{Q}_i$), key ($\mathbf{K}_i$) and value ($\mathbf{V}_i$). The multi-head attention, which enables to better handle various input sequences, is then implemented by dividing them into multiple channels with a total of *H* heads. The output of a single attention head is computed as follows:

$$Attention\left(\widetilde{\mathbf{Q}}_{ih}, \widetilde{\mathbf{K}}_{ih}, \widetilde{\mathbf{V}}_{ih}\right) = Softmax\left(\frac{\widetilde{\mathbf{Q}}_{ih}\widetilde{\mathbf{K}}_{ih}^{T}}{\sqrt{F_i}}\right)\widetilde{\mathbf{V}}_{ih} \quad (1)$$

Where $\widetilde{\mathbf{Q}}_{ih}, \widetilde{\mathbf{K}}_{ih}, \widetilde{\mathbf{V}}_{ih} \in \mathbb{R}^{T_i \times C_{ih} \times F_i}$ respectively denote the query, key and value of the *h*-th attention head after the convolution operation, and C_{ih}, T_i and F_i respectively denote the number of channels, time frames and feature size, and the superscript *T* denotes matrix transformation of the channel and spectral feature domain. Note that we consider the channel attention to better enable real-time processing, i.e., the attention score having the formula of $\widetilde{\mathbf{Q}}_{ih}\widetilde{\mathbf{K}}_{ih}^{T} \in \mathbb{R}^{T_i \times C_{ih} \times C_{ih}}$. In this work, we use the numbers of heads of $\{4, 8, 8, 16, 16\}$ through the encoder blocks. Individual self-attention results are then projected using additional convolutional layer. We use $(k, s, p) = (3,1,1)$ for the initial convolution operation, whereas use $(k, s, p) = (1,1,0)$ for the convolutional projection layer.

Finally, the output of the self-attention layer is passed to the convolutional feed-forward block which is shown in Figure 1 (b). We use $(k, s, p) = (3,1,1)$ for the convolutional layer.

Proposed Decoder

The proposed decoder is composed of a stack of N_d blocks, where we use $N_d = 1$ for computational efficiency. In order to better capture the characteristics of different emotion, we use a trainable emotion query, i.e., $\mathbf{Q} \in \mathbb{R}^{C_Q \times 1 \times F_E}$ where C_Q is the query size and F_E is the feature size of the output of the encoder. In this work, we use $C_Q = 64$. The emotion query is first passed to the self-attention block, where the output is computed as explained in the previous sub-section, followed by the cross-attention block. The latter is also computed using (1) based on the key and value of the encoder output, i.e., $\mathbf{K} \in \mathbb{R}^{T_E \times C_E \times F_E}$ and $\mathbf{V} \in \mathbb{R}^{T_E \times C_E \times F_E}$ where C_E and T_E respectively denote the number of channels and time frames of the encoder output feature, while using the query from the self-attention output of the decoder block i.e., $\mathbf{Q} \in \mathbb{R}^{T_E \times C_Q \times F_E}$. Note that the dotted line in Figure 1 (c) indicates the additive path for the cross-attention. The output of the cross-attention is then passed to the convolutional feed-forward block. As in the encoder, we use $(k, s, p) = (3,1,1)$ for all convolutional and average pooling layers, except the convolutional projection layer where we use $(k, s, p) = (1,1,0)$. Finally, the output of the decoder is passed to the frame-wise average pooling and convolutional layers. The emotion state is then predicted by applying the softmax activation function, resulting in $\mathbf{O} \in \mathbb{R}^{C_O \times T_O \times 1}$, where C_O is the number of emotion states and T_O is the number of final time frames.

Model Training

For both the encoder and decoder, we first apply the positional encoding to the input features, i.e., the LMFB and emotion query, as shown in Figure 1. Regarding the convolutional layer for the feature embedding, we use $(k, s, p) = (3,1,1)$.

The proposed transformer architecture is trained by minimizing the cross-entropy error function. Especially, we adopt a simple, yet efficient masking approach introduced in [16], to handle the various length of the given utterances while training the model. Specifically, the mask values are computed based on the average power spectral coefficients of the given time frames to detect the speech-presence periods and speech-absence periods.

Table 1. Accuracy Comparison (Improvised)

Methods	Attn. Conf.	UA (%)	WA (%)
CNN (baseline)	N/A	65.4	66.3
CNN+MAT	Enc.	66.9	67.2
CNN+MAT	Enc.-Dec.	67.0	68.4
CNN+CAT	Enc.	67.9	67.4
CNN+CAT	Enc.-Dec.	69.1	68.6
(prop.) fully CAT	Enc.	68.9	69.0
(prop.) fully CAT	Enc.-Dec.	**70.2**	**70.4**

Table 2. Accuracy Comparison (Improvised + Scripted)

Methods	Attn. Conf.	UA (%)	WA (%)
CNN (baseline)	N/A	60.7	61.6
CNN+MAT	Enc.	61.0	61.6
CNN+MAT	Enc.-Dec.	61.2	61.9
CNN+CAT	Enc.	61.5	62.9
CNN+CAT	Enc.-Dec.	62.7	63.0
(prop.) fully CAT	Enc.	62.9	63.4
(prop.) fully CAT	Enc.-Dec.	**63.8**	**64.9**

3. *Experiments*

Dataset

We conducted experiments using IEMOCAP [17], where the sampling rate of all signals was 16 kHz. The dataset includes a total of 12 hours of audio-visual data divided into 5 sessions with 10 subjects, i.e., one male and one female speech recording in each session under scripted and improvised scenarios. The scripted part is performed for predetermined emotions, while the improvised part is closer to a natural conversation. The dataset contains various categorical emotion labels. Among them, to compare the results of the proposed method to those of the previous studies, we combined two classes of excited and happy into one class, and used a total of 4 emotion states, i.e., neutral, angry, sad and happy. We performed evaluations using the 5-fold cross-validation method for improvised scenario and all scenario (i.e., improvised and scripted), separately. In each fold, the data from the remaining four sessions are used for training the model.

Methodology

Regarding the implementation of the proposed SER algorithm, we first applied pre-emphasis to the time-domain speech signal. The maximum length of the utterances was set to 8s. That is, we zero-padded with random sizes before and after the given utterances which were shorter than the maximum length, while we randomly cropped the given utterances which were longer than the maximum length. The Hamming window of 512 samples with 50% overlap and 512-point fast Fourier transform were then applied to the pre-processed signal to obtain the short-time Fourier transform (STFT) coefficients. Subsequently, we computed the LMFB coefficients with the feature size of 40. Finally, we applied frame-wise zero-mean normalization to the extracted LMFB feature. The model parameters were updated via error back-propagation and the adaptive moment estimation (Adam) optimizer for 200 epochs, with the batch size of 64. The initial learning rate was set to 10^{-4}, which decreased by 15% for every 20 epochs after 60 epochs of training.

Regarding the objective measures for performance evaluation, we computed the unweighted accuracy (UA) and weighted accuracy (WA). The UA indicates the overall accuracy across all utterances of the testing set, while the WA indicates the average accuracy across all classes.

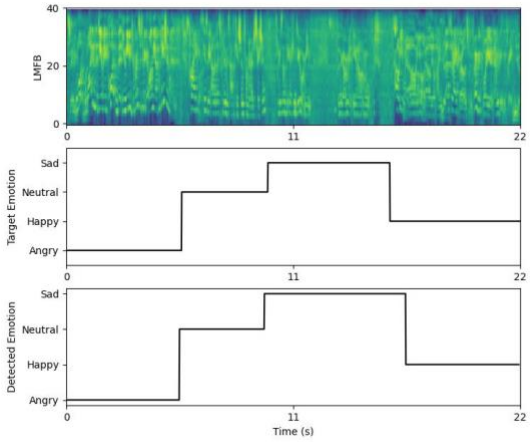

Figure 2. An example of detected emotions over time.

We implemented several benchmark algorithms for performance comparison. Regarding the baseline model, we trained a typical CNN with the layer configurations given by the channel expansion layer in Figure 1 (a). To verify the effectiveness of the proposed fully convolutional transformer, we trained typical attention-based models, which are designed by stacking the baseline CNN and the attention blocks. Regarding the attention blocks, we considered both the MLP-based attention layers (which will be referred to as MAT) and the convolutional attention layers (which will be referred to as CAT) layers. Moreover, we conducted experiments using only the encoder blocks and the encoder-decoder structure for the attention-based architectures. The basic experimental settings, such as the size of the LMFB feature, the emotion query size and training procedure, were kept identical when applicable.

Results

The recognition results of the improvised scenario and all scenarios using different architectures are shown in Tables 1 and 2, respectively. The values in bold indicate the best performance along the corresponding column.

Most of all, we can see that using the proposed fully convolutional transformer-based architecture gave the best recognition performance for both scenarios. When comparing between the baseline and the proposed transformer, we achieved about 4.8% and 4.1% improvements of UA and WA for the improvised scenario and about 3.1% and 3.3% improvements of UA and WA for the improvised and scripted scenarios. Secondly, when comparing between the results of using the MAT and CAT, it is verified through experiments that employing the proposed convolutional attention mechanism improves the recognition performance. Thirdly, when comparing between the results of using only the encoder, i.e., self-attention, and both the encoder and decoder, i.e., with additional emotion query to capture the characteristics of different emotion states, we can see that the latter approach provided better performance.

An example of real-time prediction of the emotion is shown in Figure 2. We generated a test sample speech signal by concatenating female speech utterances from Session 2 (improvised scenario) with the emotion states of angry, neutral, sad and happy consecutively. From top to bottom shows the

LMFB feature, target emotions over time and the predicted emotions over time. Although there was a slight mismatch between the target and predicted emotions at some period, mainly due to the average pooling process across time frames as well as the spectral domain in the channel expansion block, we can see that the proposed model tends to well predict the emotion variations over time.

Finally, we conducted an informal testing using additionally recorded speech signals. The objective was to verify whether the model properly detects the emotions for the recordings obtained from different acoustic environment as well as for the speakers and scripts not included in the training data. We invited a total of 12 participants (9 males and 3 females) internally from Forvia IRYStec *Inc*. We considered 16 scripts from IEMOCAP and prepared 8 sentences regarding a possible real-world driving scenario. Note that we tested using the model trained with the IEMOCAP dataset. The proposed fully convolutional transformer provided reasonable results, i.e., 63.8% of UA and 64.6% of WA.

4. *Conclusion and Future Works*

We introduced a fully convolutional transformer-based neural architecture for SER. The proposed encoder and decoder blocks were mainly composed of convolutional attention and feed-forward layers, where we used trainable emotion queries in the decoder to better capture the characteristics of different emotion states. Moreover, we considered the frame-wise channel attention to better enable real-time processing which is especially useful for automotive applications. Experimental results showed that the proposed emotion classification method provided better recognition performance than the selected benchmark algorithms.

Finally, we comment on some interesting research avenue for our future works. To better handle a real-world noisy environment, we will consider noise reduction and dereverberation algorithms as a pre-processor. In addition, further improve the emotion recognition performance, we plan to conduct research on audio-visual emotion recognition, e.g., [18].

5. References

1. Lieskovská E, Jakubec M, Jarina R, Chmulík M. A review on speech emotion recognition using deep learning and attention mechanism. Electronics. 2021 May 13;10(10):1163.
2. Yenigalla P, Kumar A, Tripathi S, Singh C, Kar S, Vepa J. Speech emotion recognition using spectrogram & phoneme Embedding. In Interspeech 2018 Sep 2 (vol. 2018, pp. 3688-3692).
3. Aftab A, Morsali A, Ghaemmaghami S, Champagne B. LIGHTSERNET: A lightweight fully convolutional neural network for speech emotion recognition. In 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2022 May 23 (pp. 6912-6916). IEEE.
4. Fu C, Shi J, Liu C, Ishi CT, Ishiguro H. AAEC: An adversarial autoencoder-based classifier for audio emotion recognition. In Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop 2020 Oct 16 (pp. 45-51).
5. Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems. 2017.
6. Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) 2017 Mar 5 (pp. 2227-2231). IEEE.
7. Zhang Y, Du J, Wang Z, Zhang J, Tu Y. Attention based fully convolutional network for speech emotion recognition. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2018 Nov 12 (pp. 1771-1775). IEEE.
8. Li Y, Zhao T, Kawahara T. Improved end-to-end speech emotion recognition using self attention mechanism and Multitask Learning. In Interspeech 2019 Sep 15 (pp. 2803-2807).
9. Chen M, He X, Yang J, Zhang H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters. 2018 Jul 26;25(10):1440-4.
10. Xie Y, Liang R, Liang Z, Huang C, Zou C, Schuller B. Speech emotion classification using attention-based LSTM. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019 Jul 1;27(11):1675-85.
11. Peng Z, Lu Y, Pan S, Liu Y. Efficient speech emotion recognition using multi-scale CNN and attention. In 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2021 Jun 6 (pp. 3020-3024). IEEE.
12. Zhang C, Xue L. Autoencoder with emotion embedding for speech emotion recognition. IEEE Access. 2021 Mar 30;9:51231-41.
13. Liu G, Cai S, Wang C. Speech emotion recognition based on emotion perception. EURASIP Journal on Audio, Speech, and Music Processing. 2023 May 12;2023(1):22.
14. Zhu B, Hofstee P, Lee J, Al-Ars Z. An attention module for convolutional neural networks. In International Conference on Artificial Neural Networks 2021 Sep 14 (pp. 167-178). Springer International Publishing.
15. Potlapalli V, Zamir SW, Khan SH, Shahbaz Khan F. PromptIR: Prompting for all-in-one image restoration. Advances in Neural Information Processing Systems. 2024 Feb 13;36.
16. Ma X, Wu Z, Jia J, Xu M, Meng H, Cai L. Emotion recognition from variable-length speech segments using deep learning on spectrograms. In Interspeech 2018 Sep 2 (pp. 3683-3687).
17. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation. 2008 Dec;42:335-59.
18. Zhao S, Ma Y, Gu Y, Yang J, Xing T, Xu P, Hu R, Chai H, Keutzer K. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In Proceedings of the AAAI Conference on Artificial Intelligence 2020 Apr 3 (Vol. 34, No. 01, pp. 303-311).