SpeechFormer: A Hierarchical Efficient Framework Incorporating the Characteristics of Speech

Weidong Chen¹, Xiaofen Xing¹, Xiangmin Xu¹*, Jianxin Pang², Lan Du³

¹School of Electronic and Information Engineering, South China University of Technology, China
²UBTECH Robotics Corp, China
³iFLYTEK Research, China

eewdchen@mail.scut.edu.cn, {xfxing, xmmxu}@scut.edu.cn, walton@ubtrobot.com, landu@iflytek.com

Abstract

Transformer has obtained promising results on cognitive speech signal processing field, which is of interest in various applications ranging from emotion to neurocognitive disorder analysis. However, most works treat speech signal as a whole, leading to the neglect of the pronunciation structure that is unique to speech and reflects the cognitive process. Meanwhile, Transformer has heavy computational burden due to its full attention operation. In this paper, a hierarchical efficient framework, called SpeechFormer, which considers the structural characteristics of speech, is proposed and can be served as a general-purpose backbone for cognitive speech signal processing. The proposed SpeechFormer consists of frame, phoneme, word and utterance stages in succession, each performing a neighboring attention according to the structural pattern of speech with high computational efficiency. SpeechFormer is evaluated on speech emotion recognition (IEMOCAP & MELD) and neurocognitive disorder detection (Pitt & DAIC-WOZ) tasks, and the results show that SpeechFormer outperforms the standard Transformer-based framework while greatly reducing the computational cost. Furthermore, our SpeechFormer achieves comparable results to the state-of-the-art approaches.

Index Terms: hierarchical framework, speech signal processing, speech emotion recognition, cognitive disorder detection

1. Introduction

Speech signal is able to express the most information in the simplest way [1]. Speech based emotion and neurocognitive disorder analysis, which is collectively referred to as cognitive speech signal processing (CoSSP), has covered a wide area of applications, including speech emotion recognition (SER), depression classification, Alzheimer’s disease (AD) detection and so on. Because of its broad application value, CoSSP is becoming an increasing interest in speech signal processing field.

In the last century, Hidden Markov model, which is a statistical Markov model and assumes the system being modeled to be a Markov process, is proposed to model speech signal [2, 3]. After that, more and more machine learning methods, such as decision tree [4, 5] and restricted Boltzmann machine [6, 7] and so on, are applied to CoSSP. Recently, with the development of deep learning, Convolutional Neural Network [8, 9, 10, 11], Recurrent Neural Network and its variants [12, 13, 14, 15] are proposed and achieve promising results. After that, deep learning methods deliver superior performances in CoSSP field.

Inspired by the global attention mechanism, Transformer [16], which is outstanding in modeling long-range dependencies in the sequence, has achieved great success in natural language processing (NLP). Although the original Transformer is designed for machine translation task in NLP, researchers are active in investigating its adaptation to many other fields, specifically computer vision [17, 18]. Certainly, several attempts also have been made in CoSSP field [19, 20, 21, 22].

However, most studies omit the natural characteristics of speech while using the attention with Transformer, leading to the neglect of the pronunciation structure that is unique to speech and conveys lots of messages. For example, even if I don’t understand Greek, I can still determine the emotion in a Greek recording by utilizing the characteristics in speech, such as articulation, prolongation and the dynamic change of speech sound. Meanwhile, Transformer has heavy computational burden, as the computational complexity of its full attention is quadratic to input length. In other words, the standard Transformer needs to incorporate the characteristics of speech before all its performance can be exploited in CoSSP field.

To solve the above problems, we should rethink the structure of speech signal first. As shown in Figure 1, we can observe that an utterance consists of several words, a word consists of several phonemes (e.g. word ‘WANT’ includes phonemes ‘W’, ‘AA1’, ‘N’ and ‘T’) and a phoneme consists of several frames (e.g. phoneme ‘AA1’ includes four frames). This progressive structure reveals the importance of the interaction between adjacent elements and indicates that we can model the speech signal hierarchically based on the nature of pronunciation. Therefore, a hierarchical framework, called SpeechFormer, which consists of frame, phoneme, word and utterance stages, is proposed to model the signal step-by-step. Firstly, we capture the frame, phoneme and word-level features through the first three stages and merge the features between two successive stages, both of which are performed under the instruction of the characteristics of speech with high computational efficiency. At last, an utterance stage is applied to gather all the word-level features and generate an utterance-level representation for classification.

The contributions of this paper can be summarized as follows:
The proposed SpeechFormer, as shown in Figure 2, mainly consists of four stages and three merging blocks (M-Blocks). In which, frame stage (F-Stage), phoneme stage (P-Stage) and word stage (W-Stage) are used to learn features of different levels, and utterance stage (U-Stage) aims to generate a global representation for classification. Three M-Blocks refine the redundant features between two consecutive stages by reducing the number of tokens. Moreover, an additional branch is employed to provide the statistical natures of speech signal.

2. Methodology

2.1. Vanilla Transformer

Transformer (refers only to its encoder part in this paper), as shown in the left part of Figure 3, consists of two sub-layers of Multi-Head Self-Attention (MSA) and Feed-Forward Network (FFN). MSA is at the core of Transformer and we will give a brief introduction to it. More details can be found in [16].

2.1.1. MSA in vanilla Transformer

For a sequential input \( x \in \mathbb{R}^{T \times d_{m}} \), where \( T \) and \( d_{m} \) are the length and dimension of input, respectively. Transformer first obtains query \( Q \), key \( K \) and value \( V \) by applying three projections to \( x \), \( QK/V \) are further divided into \( h \) parts, producing \( d_{k} \) dimensional features, where \( d_{k} = d_{m}/h \) and \( h \) is the number of heads. Each head performs Single-Head Self-Attention (SSA) and the output value of each head is concatenated to form the final output of MSA. SSA is depicted as fellows:

\[
\text{SSA}(Q, K, V) = \text{Softmax}\left( \frac{QK^T}{\sqrt{d_k}} \right)V
\]  

2.2. SpeechFormer framework

The vanilla Transformer-based framework, illustrated in the right part of Figure 3, neglects the implicit relations in speech. Conversely, our SpeechFormer, shown in Figure 2, makes use of these relations to model speech signal in a hierarchical manner. To be specific, SpeechFormer block employs Speech-based Multi-Head Self-Attention (Speech-MSA) to capture the relations between adjacent elements. Merging block is applied to refine the redundant feature. Both of them are performed under the instruction of the statistical characteristics of speech.

2.2.1. Statistical characteristics of speech

Elements of speech signal. Phoneme is the minimal sound unit in language, recycled to form all our spoken words. Multiple words are arranged together to form an utterance that is recorded in a wave file. In data format terms, the digital speech signal is divided into numerous frames. Each frame contains information at that particular point in time. Therefore, frame is the basic processing unit in digital system, which then gradually forms a phoneme, a word, and finally an utterance.

Time duration. The frame length is literally the size of the window during, which can be set manually. Phonemes are of different lengths, varying from 50 ms to 200 ms. To analyse the duration of word, we use P2FA [28] to extract the phonemes from corpora used and find that more than 90% of words contain less than 5 phonemes. Thus, the duration of word is considered as 250 ms to 1000 ms (5x the duration of phoneme). Time duration divided by \( hop \) is the approximate number of tokens in the feature, where \( hop \) indicates the feature’s hop length.
2.2.2. Speech-MSA in SpeechFormer block

The only difference between SpeechFormer block and standard Transformer is the replacement of MSA with Speech-MSA. As shown in Figure 4, Speech-MSA applies a window \(T_w\) to limit the full attention computation to a small scope of adjacent tokens, which can greatly relieve the computational burden. Furthermore, as displayed in Table 1, the value of \(T_w\) ensures that the Speech-MSA in the first three stages can learn the interactions between neighboring frames, phonemes and words, respectively. And in U-Stage, \(T_w\) is set to the length of its input such that a global representation is learnt. The complete output of Speech-MSA is calculated by following steps:

(i) Apply for a window \(T_w\) according to the current stage-level. The value and explanation are shown in Table 1.

(ii) Select the \( (t - \frac{T_w}{2}) \)-th to \( (t + \frac{T_w}{2}) \)-th tokens in \( K \), each performing a scaled dot-product with the \( t \)-th token in \( Q \) to produce a score. All the scores are concatenated and scaled by Softmax to generate the attention weights.

(iii) Fetch the \( (t - \frac{T_w}{2}) \)-th to \( (t + \frac{T_w}{2}) \)-th tokens in \( V \) and calculate the sum of each token multiplied by its weight. The result is the output of the \( t \)-th token in Speech-MSA.

(iv) Return to (ii) until the value of \( t \) varies from 1 to \( T \).

2.2.3. Merging block in SpeechFormer framework

The merging blocks between each two successive stages refine features under the instruction of the characteristics of speech, making a hierarchical framework which follows the natural structure of speech signal. Each merging block consists of an average pooling layer and a linear layer in succession. The first layer merges \( M \) consecutive tokens in the input, where \( M \) is the merging scale, and the second layer expands the embedding dimension of the input by a factor of \( r \). Each merging block aims to prepare the appropriate feature for its following stage as input. Specifically, each token should represent a sub-phoneme in P-Stage such that P-Stage can model the interaction between neighboring phonemes. Therefore, the \( M_1 \) before P-Stage should be no less than the minimum length of phoneme. Similarly, the \( M_2 \) before W-Stage should be no less than the minimum length of word. More details are displayed in Table 1.

### Table 1: The statistical values of window \( T_w \) and merging scale \( M \). Statistical duration is used instead of the actual duration.

<table>
<thead>
<tr>
<th>Module</th>
<th>( T_w ) or ( M )</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-Stage</td>
<td>50 ms / hop_1</td>
<td>Min. length of phoneme</td>
</tr>
<tr>
<td>M-Block</td>
<td>50 ms / hop_1</td>
<td>Min. length of phoneme</td>
</tr>
<tr>
<td>P-Stage</td>
<td>400 ms / hop_2</td>
<td>2×Max. length of phoneme</td>
</tr>
<tr>
<td>M-Block</td>
<td>250 ms / hop_2</td>
<td>Min. length of word</td>
</tr>
<tr>
<td>W-Stage</td>
<td>2000 ms / hop_3</td>
<td>2×Max. length of word</td>
</tr>
<tr>
<td>M-Block</td>
<td>1000 ms / hop_3</td>
<td>Max. length of word</td>
</tr>
<tr>
<td>U-Stage</td>
<td>( T )</td>
<td>Input length</td>
</tr>
</tbody>
</table>

Notes: \( hop_2 = M_1 \times hop_1, hop_3 = M_2 \times hop_2 \)

### Table 2: The training settings on four corpora. The learning rate drops to \( 1\% \) of the original gradually by cosine annealing. WA: weighted accuracy, UA: unweighted accuracy, WF1: weighted average F1, MF1: macro average F1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Epoch</th>
<th>Batch</th>
<th>LR</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEMOCAP</td>
<td>120</td>
<td>32</td>
<td>0.0005</td>
<td>WA &amp; UA</td>
</tr>
<tr>
<td>MELD</td>
<td>120</td>
<td>32</td>
<td>0.001</td>
<td>WF1</td>
</tr>
<tr>
<td>Pitt⁴</td>
<td>80</td>
<td>32</td>
<td>0.0005</td>
<td>WA &amp; UA</td>
</tr>
<tr>
<td>DAIC-WOZ⁵</td>
<td>60</td>
<td>16</td>
<td>0.0005</td>
<td>MF1</td>
</tr>
</tbody>
</table>

⁴ Segment-level samples are used for training, and the sublevel result obtained by majority vote is used for evaluation.

### 3. Experiments

#### 3.1. Datasets

IEMOCAP is the most widely used dataset in SER field. We select 5,531 utterances from four emotion categories: angry, neutral, happy (& excited) and sad. Experiments are conducted in leave-one-session-out cross-validation strategy.

MELD is another dataset used in SER. It consists of 13,708 utterances with seven emotions. The dataset is split into train, validation and test sets, and the scores on test set are reported.

Pitt corpus is used in AD detection field. The AD patients and healthy controls are asked to take the “Cookie Theft” picture description task [29] to produce recordings. Experiments are conducted in speaker-independent 10-fold cross-validation strategy.

DAIC-WOZ corpus, used in AVEC 2016 [30], contains recordings labeled depressed / not depressed. Since the labels of test data are not provided, we report the results on validation set.

#### 3.2. Experimental setup

**Acoustic features.** Three types of acoustic features, named spectrogram (Spec), Log-Mel spectrogram (Logmel) and pre-trained Wav2vec [31], are extracted in this paper. The window sizes of Spec and Logmel are set to 20 and 25 ms, respectively.

**Hyper-parameters.** The hyper-parameters used in different datasets for training are listed in Table 2. SGD [32] is employed to optimize the model. The number of Transformers \( N \) used in baseline framework is 12. The number of heads used in multi-head attention is 8. As for the SpeechFormer framework, we introduce two versions with different model sizes and computational complexity. Their hyper-parameters are:

- SpeechFormer-S: \( N_1 \sim N_4 = \{2, 2, 4, 4\}, r = \{1, 1, 1\} \)
- SpeechFormer-B: \( N_1 \sim N_4 = \{2, 2, 4, 4\}, r = \{1, 1, 2\} \)
### 3.3. Experimental results and analysis

#### 3.3.1. Comparison to the baseline framework

The comparison results of the proposed SpeechFormer and the baseline framework on IEMOCAP, MELD, Pitt and DAIC-WOZ are summarized in Table 3. Speech emotion recognition, Alzheimer’s disease detection and depression classification tasks are involved. From Table 3, we can observe that the model sizes of SpeechFormer-S and baseline are very close, where the former is slightly larger than the later. However, the theoretical computational complexity\(^1\) (FLOPs) of SpeechFormer-S on four corpora are about 10.2%, 15.1%, 13.2% and 7.8% of the baseline, respectively. SpeechFormer-B expands the model size of SpeechFormer-S by approximately two times, while keeping the FLOPs comparable. While conducting speech emotion recognition, the baseline results are lower than those of SpeechFormer, regardless of the feature used. When performing Alzheimer’s disease detection, the baseline achieves better WA using Logmel as input. In other cases, SpeechFormer obtains higher performances. As for depression classification, SpeechFormer framework substantially outperforms the baseline in all cases. In summary, our SpeechFormer framework generally outperforms the vanilla Transformer-based framework in terms of performance and computational efficiency.

#### 3.3.2. Comparison to previous state-of-the-art

Table 4 gives the comparison among SpeechFormer with some known systems on four corpora, in which, all systems utilize only speech feature as input to allow a fair comparison. On IEMOCAP, SpeechFormer-S achieves comparable performances to [33]: +0.6% WA. When evaluated on MELD, Pitt and DAIC-WOZ, SpeechFormer-B outperforms other comparisons with promising gains: +1.7% WF1 over [20], +1.8% (+1.1%) WA (UA) over [34] and +3.4% MF1 over [15], respectively.

---

1. We omit Softmax computation in determining complexity.

### 4. Conclusion

A hierarchical efficient framework, named SpeechFormer, is proposed for cognitive speech signal processing. The Speech-MSA and merging block in SpeechFormer model speech signal by incorporating the features of speech. Experimental results on four corpora demonstrate the superiority of our SpeechFormer.

### 5. Acknowledgements

This work is funded in part by the National Natural Science Foundation of China U1801262, Key-Area Research and Development Program of Guangdong 2019B010154003, Science and Technology Project of Guangzhou 202103010002, Guangdong Provincial Key Laboratory of Human Digital Twin 2022B1212010004 and Natural Science Foundation of Guangdong Province 2022A1515011588.
6. References


