Human Sound Classification based on Feature Fusion Method with Air and Bone Conducted Signal

Liang Xu\textsuperscript{1}, Jing Wang\textsuperscript{1}, Lizhong Wang\textsuperscript{2}, Sijun Bi\textsuperscript{1}, Jianqian Zhang\textsuperscript{1}, Qiuyue Ma\textsuperscript{2}

\textsuperscript{1}Beijing Institute of Technology, Beijing, China
\textsuperscript{2}Samsung Research China-Beijing (SRC-B)

\texttt{\{xliang, wangleing\}@bit.edu.cn, \{liz.wang, qiuyue.ma\}@samsung.com}

Abstract

The human sound classification task aims at distinguishing different sounds made by human, which can be widely used in medical and health detection area. Different from other sounds in acoustic scene classification task, human sounds can be transmitted either through air or bone conduction. The bone conducted (BC) signal generated by a speaker has strong anti-noise properties and can assist the air conducted (AC) signal to extract additional acoustic features. In this paper, we explore the effect of the BC signal on human sound classification task. Two stream audios combining BC and AC signals are input to a CNN-based model. An attentional feature fusion method suitable for BC and AC signal features is proposed to improve the performance according to the complementarity between the two signal features. Further improvement can be obtained by using a BC signal feature enhancement method. Experiments on an open access and a self-built dataset show that fusing bone conducted signal can achieve 6.2%/17.4% performance improvement over the baseline with only AC signal as input. The results demonstrate the application value of bone conducted signals and the superior performance of the proposed methods.

Index Terms: human sound classification, bone conducted signal, attentional feature fusion, feature enhancement

1. Introduction

Acoustic scene classification (ASC) is the task of predicting sound events and acoustic information [1]. In recent years, the methods with deep learning have greatly improved the accuracy of sound scene classification and several networks have been proposed in classification task.

Human sound classification is similar to ASC and aims to distinguish the kinds of sounds that the human body makes. This technology can be used in scenes such as human health detection [2, 3, 4], but there are few related works on human sound classification. Different from others ASC, human sounds can be captured by human wearable devices to obtain signals of different modalities, such as brain electricity, electrocardiogram, and bone conduction signals. These signals can play an auxiliary role in human sound classification tasks.

For ASC, the convolutional neural network (CNN) is widely used in extracting the characteristics of the time-frequency domain signal, and is used to learn the correlation between the time-frequency bins [5, 6]. In addition, the emergence of residual network improve the model convergence effect without increasing parameters. The use of residual network also greatly improves the accuracy of the ASC task [5, 7].

The CRNN structure [8, 9] has been proved to be effective in sound event detection and classification tasks by extracting features in both time and feature dimensions. A gated convolutional neural network had been proposed in CRNN structure to select the useful features [10]. In [11], Soonshin et al proposed a shallow Conformer-DNC network by combining Conformer [12, 13] and differentiable neural computer (DNC), this network can simultaneously learn the short-term and long-term features of variable-length acoustic scenes. For small datasets, deep CNN networks are difficult to train. In order to increase the robustness of the dataset, data augment methods such as SpecAugment and SpecAugment++ [14, 15] are proposed based on the input feature of neural network.

The feature fusion method helps to improve the classification ability of the network. In [16], the feature fusion method improves the classification accuracy by merging the features obtained from CNN and LSTM parallelly. Since the limitation about features extracted by a single convolution kernel, fusion of multi-scale features can improve the network classification ability [17, 18]. Tatsuya et al. proposed a classification chain method based on the correlation between classes [19]. With the emergence of attention, it is applied to better fuse features of inconsistent semantics and scales [20].

Since human sound is rarely involved in the current ASC task, its related research mostly focuses on one stream signal which is mainly collected from the air-conducted (AC) microphone, and the event itself is extremely susceptible to external noise interference. For human sound classification, bone-conducted (BC) signal could be used to reduce the noise interference. BC signals have strong anti-noise ability, but due to different propagation media from AC signal, the velocity of sound would be lost [21, 22]. When a person makes a sound (such as speaking, coughing, exercising, etc.), it can be transmitted not only through the air, but also through the bone. The bone conduction microphone can collect such weak signals from the skull and the larynx to obtain BC signal [23]. Li et al. used bone conduction microphone to monitor the human oral sound activities [24].

Thus, the AC and BC signal characteristics are complementary. With the development of neural networks, BC signals have been used in speech enhancement [23] and air-bone conducted conversion tasks [25]. Therefore, a network combining both BC signals and AC signals can possibly improve the classification ability.

In this paper, we combine both BC and AC signals as input to train a CNN based model, and demonstrate the improvement of using BC signals on the human sound classification task. An attentional feature fusion method is proposed based on the complementarity between two signals. Moreover, a feature enhancement method is used for BC signal to improve the high frequency band features. Experimental results show feature fusion method and BC signals feature enhancement method improve the effect of the model in the human sound classification task.
2. Proposed method

2.1. Overview

As shown in Figure 1, the proposed methods are implemented on a CNN and fully connected (FC) based model. There are four continuous convolutional layers to extract features from the AC and BC signal respectively, and each layer follows a Batch Normalization (BN) and an Average Pooling (AP). The extracted BC and AC features are fused according to the proposed method in Section 2.2. The proposed BC,E module is applied to enhance the feature of BC signal. Finally, the FC layer is to estimate the probabilities of each classes.

The drop out is applied between the second and third CNN block to prevent overfitting and we extract acoustic features of two streams from the waveform. Data augmentation method SpecAugment is applied to the input layer space.

The harmonic position of the enhanced BC signal in the spectrum is more obvious than that of the AC signal. However, the energy distribution in the frequency band of the BC signal is limited by the enhancement process, resulting in a more uniform energy distribution than that of the AC signal, which is not conducive to judge sound events from an energy perspective. We also believe that the complementarities still exist in the two stream signal feature layer.

We propose an attentional feature fusion method motivated by selective kernel networks (SKnet) [18] to select more useful features between AC and BC signals. The main difference from SKnet is that we use the network as a two-signal fusion method instead of a multi-scale fusion method. The structural overview of the BC-AC fusion method is shown in Figure 3, which is applied after signals feature extraction.

Given two stream features, we firstly fuse them via an element-wise summation $U = U_{BC} + U_{AC}$ to concentrate both features, where $U, U_{BC}, U_{AC} \in R^{C \times T \times F}$. Then, Global Average Pooling (GAP) is applied to squeeze the fused output $U$ into one channel representation $S$. A simple FC layer is used to reduce dimensionality for better efficiency, followed by an activation function ReLU. The degree of compression between the input $S$ and the output $Z$ is determined by the reduction ratio $r$.

A soft attention across channels contains FC layers and a Softmax function is used to adaptively select different spatial scales of information, which is guided by the compact feature descriptor $Z$. $A, B \in R^{C \times 1}$ denote the output of FC with $Z$ as input. Specifically, the Softmax is applied on the channel-wise digits.

$$Z = ReLU(WS), W \in R^{C \times C}$$ (1)

$$a(c) = \frac{e^{A(c)}}{e^{A(c)} + e^{B(c)}}, c \in 1, 2, ..., C$$ (2)
\[ b(c) = \frac{e^{B(c)}}{e^{A(c)} + e^{B(c)}}, c \in 1, 2, ..., C \quad (3) \]

\[ a, b \in \mathbb{R}^{(C \times 1)} \] are the results of performing soft attention and are used to guide the weight of the BC and AC signal. Here, \( a(c) + b(c) = 1 \). Finally, the weighted BC and AC signals are added as follows.

\[ U_{\text{fuse}} = b \otimes U_{\text{BC}} + a \otimes U_{\text{AC}} \quad (4) \]

Where \( \otimes \) denotes element-wise multiplication.

2.3. BC signal feature enhancement method

In the audio classification task, the degree to which the event features are disturbed by noise will affect the final effect of the classification. From Figure 2(b), we have known that the BC signal is severely attenuated at high frequencies. However, the remaining harmonics are very clean and there is little noise between harmonics. So it is hard to use the BC signal directly.

Inspired by frequency band extension (BWE), we propose a method for BC signals enhancement without neural network. The full-band enhancement of the BC signal can be performed with only low frequency band harmonics and the positions of all harmonics can be recovered. Therefore, the enhanced BC signal has an auxiliary effect on the AC signal. The method consists of three main steps in total, as shown in Figure 4.

**Crop**: The first step in processing the BC signal is to crop. In order to make the energy of harmonics account for a larger proportion of the overall signal energy, low pass filtering (LPF) is performed on the full-band signal, and only the signal below the cut-off frequency (CoF) is retained. Given an input \( X \in \mathbb{R}^{(1 \times L)} \), \( L \) is the length of sequence. Here we use Butterworth filter, and the processing is shown in Eq.5.

\[ X_L = \text{Butterworth}(X) \quad (5) \]

**Enhance**: The enhance operation is applied to recover the signal harmonics above CoF through the signal \( X_L \). Then, the signal \( X_e \) pass through a non-linear device (NLD) which shifts the frequencies to higher region [27]. Here we choose full wave rectifier as the NLD to generate harmonics above existing frequency bands [28]. There are still other NLDs such as half-wave rectification, square and cube law that perform well in BC_E module. Then a high pass filter (HPF) is used to select extensions frequency (>1KHz).

**Combine**: After obtaining the low-frequency band and the covered high-frequency band information separately, we combine the two parts into one enhanced BC signal \( X_{en} \). The recovered signal \( X_{en} \) has almost same energy as the signal \( X_L \), thus we adopt a gain factor \( \alpha \) to balance the energy distribution of the whole band.

\[ X_{en} = \alpha \cdot X_{ex} + X_L \quad (6) \]

The enhancement method BC_E not only retains the low frequency part in the original bone conducted signal, but also enhances the features in the high frequency band.

3. Experiments

3.1. Datasets

We conducted the experiments using two datasets: Audioset and a dataset recored by real BC and AC signals microphones. We choose the ‘human sound’ part in this large speech dataset Audioset released by Google with a 32KHz sampling rate and 16-bit WAV format [29]. Here, each selected sample in both datasets contains only one target event.

The human sound in Audioset includes eight classes of human activity data including human voice, whistling, respiratory sounds, etc. Based on stratified sampling, we split each type of data into training set, validation set and test set according to the ratio of 8:1:1. Since the Audioset contains only one AC signal, we use a 1KHz low-pass filter and an attenuation curve to simulate its BC signal.

We recorded a real dataset using integrated air conduction and bone conduction microphones to collect human sounds. The recorded human sound dataset contains six classes: speaking sounds with different loudness, cough sound, laugh sound, breath sound, beat sound, and humming sound. A total of 24 people participated in the recording, with a total of 7000 samples, and each sample has a sampling rate of 16KHz.

3.2. Experimental conditions

The input features used in the proposed system are MFCC extracted by 64 Mel-scale filters and a window size of 32ms with 30% overlap between windows. Here we set the gain factor of the BC_E module to 5.

The kernel size of each CNN layer in Figure 1 is (3, 3), and the number of channels for each branch is \{8, 16, 32, 32\}, respectively. The stride and pooling size of the AP layer are set to (2, 2) to halve the time dimension and frequency dimension. The model is trained using the Adam optimizer with the initial learning rate 0.001. A standard cross-entropy loss is applied as the loss function with a batch size of 128 and epoch 40.

3.3. Experimental results

In order to verify the effect of the proposed BC signal enhancement method and the fusion method, we conduct ablation experiments on two datasets. As an evaluation of the proposed methods, we trained five models: our proposed model in section 2.1, the proposed model without BC signal stream (and BC_E module), the proposed model without BC_E module, the proposed model using addition operation instead BC-AC fusion method to combine signals, and the baseline of the DCASE2021 task1a [30] which uses AC signal as input. In this part, the CoF of BC_E module is 600Hz, and the NLD is a full wave rectifier.

Table 1 shows the comparison results of the five models on the human sound part in Audioset and the real recored dataset.
Table 1: The F1-score on the two datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Audioset</th>
<th>real dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>model in section 2.1</td>
<td>0.491</td>
<td>0.714</td>
</tr>
<tr>
<td>w/o BC signal</td>
<td>0.462</td>
<td>0.608</td>
</tr>
<tr>
<td>w/o BC_E</td>
<td>0.481</td>
<td>0.633</td>
</tr>
<tr>
<td>adding operation</td>
<td>0.468</td>
<td>0.622</td>
</tr>
<tr>
<td>DCASE2021 task 1a baseline</td>
<td>0.458</td>
<td>0.618</td>
</tr>
</tbody>
</table>

Table 2: The F1-score of each class on the Audioset.

<table>
<thead>
<tr>
<th>class</th>
<th>w/ BC signal</th>
<th>w/o BC signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>human voice</td>
<td>0.645</td>
<td>0.563</td>
</tr>
<tr>
<td>whistling</td>
<td>0.389</td>
<td>0.352</td>
</tr>
<tr>
<td>respiratory sound</td>
<td>0.560</td>
<td>0.533</td>
</tr>
<tr>
<td>human locomotion</td>
<td>0.524</td>
<td>0.514</td>
</tr>
<tr>
<td>digestive</td>
<td>0.397</td>
<td>0.401</td>
</tr>
<tr>
<td>hands</td>
<td>0.428</td>
<td>0.407</td>
</tr>
<tr>
<td>heart sounds, heartbeat</td>
<td>0.410</td>
<td>0.393</td>
</tr>
<tr>
<td>human group actions</td>
<td>0.577</td>
<td>0.536</td>
</tr>
</tbody>
</table>

average F1-score 0.491 0.462

We count the F1-score on the two datasets, and in the comparison results, we can conclude that whether the model uses the simulated BC signal or the real BC signal, its classification effect is better than the model only uses AC signal. Compared to models using only AC signals, our proposed methods achieve 6.2% and 17.4% improvement in performance on the two datasets, respectively.

Comparing the results of model with or without BC_E module, the BC signal feature enhancement method proposed in this paper can effectively improve the characteristics of BC signals. However, in the Audioset, the improvement effect achieved by using the BC signal feature enhancement method is limited. We consider that this is due to the influence of the simulation for the BC signal: the noise other than human sound cannot be completely removed during the simulation process. It results in enhancing the characteristics of the human sound while also enhancing the characteristics of the noise.

Furthermore, we can see that the proposed attentional feature fusion method effectively selects the importance of co-located channels between two signal features. Compared with the proposed model that uses adding operation to fuse features, the experimental results are significantly improved on both datasets (4.9% and 14.7%). This proves that there is complementarity between the features of AC and BC signals, and the BC signal enables the model to acquire more features.

Table 2 shows the performance improvement of each human sound category on the Audioset. We separately count the F1-score of the model in section 2.1 with or without BC signal (and BC_E module). From the results, the model with BC signal and proposed methods has a significant improvement in the human voice and human group actions, which proves that the BC signal can effectively capture and expand features for sounds with obvious harmonics. Besides, the discrimination effect of almost all categories is still improved after adding the BC signal, which means the features of the BC signal cannot be obtained by the AC signal.

3.4. Impact of the BC_E module

To further explore the performance of BC_E module, we then evaluate the performance of the BC_E module with different CoF and NLD [28]. In Table 3, we calculated the F1-score on the two datasets to verify the effect of different operations in the BC_E module on the simulated and real BC signals. Considering that the characteristics of the BC signal are mainly concentrated below 1000Hz, we set the upper limit of the cut-off frequency as 1000Hz. From the result in Table 3, the 400Hz as the cut-off frequency and cube law as the NLD, provides the best classification performance in Audioset with simulated BC signal and the real dataset get the best performance at the 800Hz with the square law.

Table 3: Evaluation of the BC_E module.

<table>
<thead>
<tr>
<th>CoF (Hz)</th>
<th>NLD</th>
<th>Audioset</th>
<th>real dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>half wave rectifier</td>
<td>0.479</td>
<td>0.670</td>
</tr>
<tr>
<td></td>
<td>square law</td>
<td>0.466</td>
<td>0.641</td>
</tr>
<tr>
<td></td>
<td>cube</td>
<td>0.511</td>
<td>0.696</td>
</tr>
<tr>
<td>500</td>
<td>half wave rectifier</td>
<td>0.487</td>
<td>0.692</td>
</tr>
<tr>
<td></td>
<td>square law</td>
<td>0.494</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>cube</td>
<td>0.502</td>
<td>0.705</td>
</tr>
<tr>
<td>600</td>
<td>half wave rectifier</td>
<td>0.478</td>
<td>0.683</td>
</tr>
<tr>
<td></td>
<td>square law</td>
<td>0.475</td>
<td>0.713</td>
</tr>
<tr>
<td></td>
<td>cube</td>
<td>0.459</td>
<td>0.641</td>
</tr>
<tr>
<td>800</td>
<td>half wave rectifier</td>
<td>0.468</td>
<td>0.727</td>
</tr>
<tr>
<td></td>
<td>square law</td>
<td>0.465</td>
<td>0.662</td>
</tr>
<tr>
<td></td>
<td>cube</td>
<td>0.509</td>
<td>0.719</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.478</td>
<td>0.611</td>
</tr>
<tr>
<td>1000</td>
<td>half wave rectifier</td>
<td>0.465</td>
<td>0.700</td>
</tr>
<tr>
<td></td>
<td>square law</td>
<td>0.493</td>
<td>0.656</td>
</tr>
<tr>
<td></td>
<td>cube</td>
<td>0.484</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.470</td>
<td>0.656</td>
</tr>
</tbody>
</table>

From Table 3 we can obtain the following three conclusions about BC_E module. First of all, it is not that the wider the frequency band intercepted in the crop stage, the more information will be obtained after expansion, and the classification performance will decrease after a certain frequency band reaches the highest level. Secondly, the optimal gain effect brought by different NLD modules occurs at different cut-off frequencies. For example, full-wave rectifier as an NLD module will obtain the best performance at the cut-off frequency of 600Hz, while the square law will obtain the highest performance at 800Hz. Finally, on the same CoF, both real and simulated BC signals basically obtain the best results on the same NLD. However, square law performs better on real BC signals, while cube law is more suitable for simulated BC signals. Compared with the proposed model without BC_E module in section 3.3, the BC_E module improves the performance by 6.2% and 20.4%, respectively.

4. Conclusion

In this paper, we combine both BC and AC signals in the human sound classification task. An attentional feature fusion method for AC and BC signals is proposed and a feature enhancement method based on frequency band extension is used on BC signal. This paper proves the complementarity between BC signal and AC signal, and the experimental results show the BC signal and proposed methods can effectively improve the model classification ability. In the future, we aim to explore other techniques to make use of the BC signal features in human sound classification.
5. References


