# Submission to DCASE 2023 Task 1: Low-Complexity Acoustic Scene Classification Using Cepstral Analysis

Technical Report

*Yaojun Han, Nengheng Zheng*

College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China
2110436189@email.szu.edu.cn, nhzheng@szu.edu.cn

## ABSTRACT

This technical report describes the submitted system for task 1 of the DCASE 2023 challenge. The goal of this task is to design an acoustic scene classification system for devise-imbalanced datasets under the constraints of low complexity. We applied cepstrum analysis to filter out the channel information contained in the raw audio signals before the feature extraction stage. Moreover, we separate the spectral envelope and the fine structure of the spectrum in the cepstrum domain, and simply analyzed the impact of the two on the classification results. Due to the constraints of low complexity, we use knowledge distillation to allow the simpler student model to learn complex teacher models. In addition, we experimented with different augmentation techniques such as Mixup, random noise, pitch shifting, and time-frequency masking to expand the diversity of the dataset. Through the calculation of NeSsi tool, our model requires 80.845K of memory, with 29.349M MACs. And the accuracy of the model on the development dataset is 51.4%.

***Index Terms*—** Acoustic scene classification, cepstrum analysis, spectral envelope, fine structure, knowledge distillation

## 1. INTRODUCTION

Task 1 aims to identify ten different acoustic scenes from one-second audio recorded by different devices[1]. In order to approach engineering more closely, the official limits model parameters (128K), 30 MMACs (million multiply-accumulate operations) of the model. This year, additional indicators of energy consumption have also been added. These restrictions all increase the difficulty of the competition.

Previous research results have proven that the combination of Log Mel spectrograms and Convolutional Neural Networks (CNNs) is effective for acoustic scene classification tasks[2]. In the past few years, researchers have invested a lot of effort in improving CNNs, but few have focused on audio signals and acoustic features. This report provides a detailed record of our exploration process in the audio source and feature extraction stages. We hope to reduce channel noise caused by different devices before the network input features. Although this may result in more tedious calculations during the feature extraction stage, experiments have shown that it is an effective approach. We believe that this may provide new ideas for others' research.

From the perspective of signal processing, the audio collected by different devices is essentially the result of convolution of environmental audio with different channel filters. In the time domain, the two signals of convolution cannot be separated, but in the cepstrum domain, the linearly indistinguishable convolution signal can be converted to a linearly distinguishable signal. Through this method, we can separate the channel noise caused by different devices from the audio in the cepstrum domain.

And inspired by Phaye et al.[5], we propose that spectral envelope and fine structure may play different roles in acoustic scene classification. Therefore, we separated the envelope and fine structure of the spectrum in the cepstrum domain and validated their impact on acoustic scene classification tasks through experiments. In addition, we also combine the above features to express richer information. Finally, we used knowledge distillation to condense the model into small models while maintaining performance.

The rest of the paper is divided as follows—in section 2, we will provide a detailed explanation of the principles and operational processes of applying cepstrum analysis to datasets, as well as our feature extraction. In section 3, we show the network structure and training parameter settings. Section 4 presents the experimental setup. Section 5 records our experimental results, and discussions of the results. In section 6, we briefly summarized this work.

## 2. CEPSTRUM ANALYSIS AND FEATURE EXTRACTION

In this section, we record the preprocessing process of input data in detail. Firstly, the role of cepstrum analysis in this challenging task was explained from a theoretical perspective. Secondly, we also separate the spectral envelope and fine structure from the audio as additional features. In addition, we also recorded the parameter details of extracting Log Mel features, and finally, we record the data augmentation measures taken.

### 2.1. Cepstrum analysis

Task 1 requires participants to provide a low-complexity network model to cope with audio dataset recorded by different devices. The most popular approach is to enhance the generalization ability of the model by improving convolutional neural networks. We provide another processing approach, which is to use cepstrum analysis technology to weaken the impact of different devices on audio before the feature extraction stage. As described above, the audio signal collected by different devices is essentially the product of convolution. Mathematics expression is as follows:

$$y(n) = x(n) * h(n) \tag{1}$$

Here, $y(n)$ is the recorded audio, $x(n)$ is the ideal original environment audio, and $h(n)$ represents the channel filters of different devices. The process of collecting environmental audio with different devices inevitably leads to mixes different channel noise in the environmental audio. And our purpose is to restore $x(n)$ as much as possible, which can weaken the channel noise caused by different devices. Take Z-transform as an example, we convert it to the frequency domain:

$$Z[y(n)] = X(z) \cdot H(z) \tag{2}$$

Take logarithms on both sides at the same time:

$$ln\big(Y(z)\big) = \ln\big(X(z)\big) + \ln\big(H(z)\big) \tag{3}$$

Then, the two can be separated by inverse transformation, and the signal is in the cepstrum domain at this time:

$$\hat{y}(n) = \hat{x}(n) + \hat{h}(n) \tag{4}$$

According to the central limit theorem, when we analyze a sufficient number of audio samples, the short-time spectral features of the audio signal obey the Gaussian distribution when we analyze a sufficient number of audio samples. In the assumptions of the Gaussian model, it can be considered that the Fourier expansion coefficient is an independent Gaussian random variable, with the average value of zero. Therefore, the mean of $\hat{x}(n)$ can be approximately zero. Meanwhile, the device filter can be regarded as a time invariant system, and its mean is itself.

Based on the above, we can subtract the mean of in the cepstrum domain. As follows:

$$\begin{aligned} \hat{y}(n) - mean\big(\hat{y}(n)\big) &= \hat{x}(n) + \hat{h}(n) - mean\left(\hat{x}(n)+\hat{h}(n)\right) \\ &\approx \hat{x}(n) + \hat{h}(n) - 0 - \hat{h}(n) \\ &= \hat{x}(n) \end{aligned} \tag{5}$$

Here, we have separated the influence of devices theoretically. We restored it to the frequency domain and extract the Log Mel energy of the processed signal

## 2.2. Spectral envelope and fine structure

In automatic speech recognition (ASR), spectral envelope plays a crucial role. And Phaye et al. proposed that there are significant differences in the spectral envelope of different acoustic scenes through experiments[5]. Therefore, spectral envelope may be an important feature in acoustic scene classification tasks. On the other hand, the reason why the spectrum envelope works well in ASR is because it portrayed sound tract information well. However, there is no concept similar to sound tract in environmental audio, and the fine structure represents the high-frequency information of the spectrum. So we cannot ignore the fine structure also.

The extraction process of envelope and fine structure is roughly described as follows. The low-frequency part of the cepstrum represents a slow changing trend in the spectrum, which is the spectrum envelope. Conversely, its high -frequency part means the fine structure of the spectrum. We can take the first N values in the cepstrum domain to restore the spectral envelope. Similarly, we can easily obtain the fine structure of the spectrum. This method is similar to the process of extracting MFCC.

## 2.3. Feature extraction

The input audio maintains a sampling rate of 44.1 kHz. We used a Short Time Fourier Transform (STFT) with a window size of 2048 and 50% overlap to extract the input features. We chose 256 Mel frequency bins instead of 128, because we found in the experiment that 256 frequency bins contain richer information, which can help us better classify. According to the previous description, we can perform the operation of subtracting the mean and extract the spectral envelope and fine structure in the cepstrum domain. In this way, we obtained three feature maps of the same size, all of which are 128×44. For ease of expression, we record them in order as feat_m, feat_e and feat_s.

## 2.4. Data augmentation

In order to improve the generalization ability of the network, we applied we applied Mixup, time -frequency masking, random noise, random pitch-shifting.

# 3. ARCHITECTURES

Our teacher-student network is based on TensorFlow and Keras. Their versions are tensorflow-gpu 2.6.0 and keras 2.6.0 respectively.

## 3.1. teacher model

The teacher model structure refers to the structure submitted by Tobias in 2022[6]. Our differences include the following points:

- Divided frequency band training: SubSpectralNet, proposed by Phaye et al. in 2018, provided a great help for the later studies[5]. Their work shows that depending on the scene class, there is a specific frequency band showing most activity, hence providing discriminative features for that class. In the system we submitted, we only divided two frequency bands evenly along the frequency axis, and the experimental results showed that this is still an effective method.

- Channel Attention: Channel Attention was initially proposed by Jie Hu et al.[7], and improved by Sanghyun Woo et al. later[8]. In simple terms, it can improve accuracy by modeling the correlation between feature channels and increasing the weight of important features. We add channel attention after the final convolutional layer. It will calculate the maximum pooling and average pooling of the input features of this layer separately, and then model the dependency relationship between channels through a two-layer fully connected network. Finally, channel weights are obtained by adding.

- Simpler convolutional layers: We must reduce the learning ability of the network, because our features are relatively complex. The experimental results show that too large convolution kernel or too deep network layer will lead to serious overfitting. We set the number of convolution kernels in three stages to 64, 128 and 256. And we used conventional convolutional layers instead of the damped version.

Table 1 The structure of the student model

| Student-network | | | | |
|---|---|---|---|---|
| Stage | Blocks/Layers | C | K | O |
| Stage 1 | Inputs | 1 | / | 256×44 |
| | [Conv2d/BN/ELU]×2 | 18 | 5/3 | 128×22 |
| Stage 2 | MaxPooling | / | 2 | 64×11 |
| | [Conv2d/BN/ELU]×3 | 32 | 3/3/1 | 64×11 |
| Stage 3 | MaxPooling | / | [3, 2] | 21×5 |
| | [Conv2d/BN/ELU]×3 | 64 | 3/3/1 | 21×5 |
| | Dropout | / | 0.2 | 21×5 |
| Stage 4 | Conv2d/BN | 10 | 3 | 21×5 |
| | Channel Attention | / | / | 10 |
| | GlobalAveragePooling | / | / | 10 |
| | Softmax | / | / | 10 |

### 3.2. student model

We propose a new model structure for our complex features, which is also the system we ultimately submitted. The parameter quantity of this model is 80.845K, with 29.349M MACs. Both indexes meet the requirements of the challenge. The structure of the student model is shown in Table 1.

In the process of student model training, there are two key parameters in knowledge distillation, namely the distillation temperature T and a ratio called alpha.

The temperature T can control the negative label ratio given by the teacher model. When T=1, the output of the teacher model is normal. As T increases, the output of the teacher model will increase the proportion of negative labels, and the student model can learn more information about negative labels. But the higher the temperature T is not the better. The higher the temperature T, the more negative label information the teacher model outputs. This requires a more complex student model to learn.

Alpha is a ratio that represents the ratio between soft loss and hard loss. The larger alpha, the greater the proportion of soft loss. According to the previous research, student model tends to achieve good results when the soft loss is relatively high[9].

We tried many different combinations of T and alpha, and eventually in the system we submitted, T is set to 2, alpha is set to 0.8.

## 4. EXPERIMENTAL SETUP

We conducted all experiments on a NVIDIA RTX 3090 GPU with 24GB memory.

In the training of the teacher model, we used several different data augmentation methods. It should be noted that the parameter alpha of Mixup is set to 0.4. In the feature map, we randomly mask 16 frequency bins along the frequency axis and repeat twice. We use Stochastic Gradient Descent and introduce momentum. The maximum learning rate is 0.1, the minimum learning rate is 1e-6, and the momentum parameter is set to 0.9. We use the decay method of CosineAnnealing and let it restart the learning rate at 3/7/15/31/63/127 epochs. The training lasts for a total of 255 epochs.

In the training of the student model, we set the Mixup parameter to 0.1. We use the Adam optimizer with a learning rate of 0.001 and CosineAnnealing strategy. In addition, we also set an early stop mechanism to prevent overfitting.

## 5. RESULTS

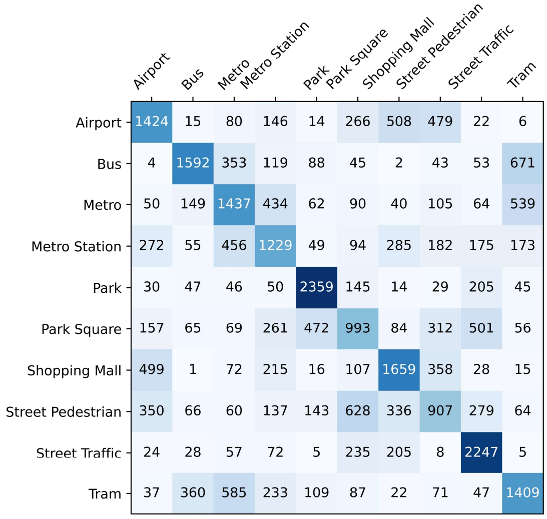
In this section, we first show the performance of the teacher model and the student model, and show the output details of the student model in detail by calculating the confusion matrix. Secondly, we verify the effectiveness of the three channel combination features through simple ablation experiments. In addition, we also recorded the exploration process of parameter selection during knowledge distillation.

We validated the performance of the teacher-student system on the development dataset[10] through experiments, as shown in Table 2. We set the distillation temperature to 2 and alpha to 0.8, and under the influence of the early stop mechanism, we can obtain a small student model within 100 epochs of training. Although the performance of the student model has slightly decreased, it still significantly outperforms the baseline system on S1~S6 devices. The specific output details are shown in Figure 1.

We validated the role of three channel features through ablation experiments. The experimental results are shown in Table 3. In our model structure, the performance of Log Mel was lower than expected. This is because we reduced the scale of the teacher model on the basis of the work of Tobias et al., which will lead to a decrease in the learning ability of the model[6]. The feat_m obtained in the cepstrum domain can significantly improve the performance of the model on different devices, which also proves the effectiveness of the previous derivation. When we introduce additional features from the other two channels, the model can learn more information from them. But when they work alone, their performance ability is not good. The envelope lacks the high-frequency information of the spectrum, while the fine structure lacks the low-frequency information, and both have an equally important impact on the classification results.

Table 2 The performance of the teacher-student model

| | Teacher-model | | Student-model | |
|---|---|---|---|---|
| Devices | Log loss | Accuracy | Log loss | Accuracy |
| A | 0.911 | 66.3% | 1.006 | 63.7% |
| B | 1.132 | 58.9% | 1.390 | 51.7% |
| C | 1.091 | 63.4% | 1.161 | 58.4% |
| S1 | 1.242 | 55.4% | 1.423 | 48.2% |
| S2 | 1.281 | 55.9% | 1.454 | 48.4% |
| S3 | 1.175 | 56.8% | 1.285 | 43.2% |
| S4 | 1.423 | 50.5% | 1.505 | 47.4% |
| S5 | 1.333 | 52.0% | 1.572 | 44.8% |
| S6 | 1.538 | 48.3% | 1.609 | 46.9% |
| Average | 1.228 | 56.4% | 1.378 | 51.4% |



| | Airport | Bus | Metro | Metro Station | Park | Park Square | Shopping Mall | Street Pedestrian | Street Traffic | Tram |
|---|---|---|---|---|---|---|---|---|---|---|
| Airport | 1424 | 15 | 80 | 146 | 14 | 266 | 508 | 479 | 22 | 6 |
| Bus | 4 | 1592 | 353 | 119 | 88 | 45 | 2 | 43 | 53 | 671 |
| Metro | 50 | 149 | 1437 | 434 | 62 | 90 | 40 | 105 | 64 | 539 |
| Metro Station | 272 | 55 | 456 | 1229 | 49 | 94 | 285 | 182 | 175 | 173 |
| Park | 30 | 47 | 46 | 50 | 2359 | 145 | 14 | 29 | 205 | 45 |
| Park Square | 157 | 65 | 69 | 261 | 472 | 993 | 84 | 312 | 501 | 56 |
| Shopping Mall | 499 | 1 | 72 | 215 | 16 | 107 | 1659 | 358 | 28 | 15 |
| Street Pedestrian | 350 | 66 | 60 | 137 | 143 | 628 | 336 | 907 | 279 | 64 |
| Street Traffic | 24 | 28 | 57 | 72 | 5 | 235 | 205 | 8 | 2247 | 5 |
| Tram | 37 | 360 | 585 | 233 | 109 | 87 | 22 | 71 | 47 | 1409 |

Figure 1 Confusion matrix based on student model

Table 3 Comparison of performance of different features

| Features | Log loss | Accuracy |
|---|---|---|
| Log Mel | 1.638 | 43.3% |
| feat_m | 1.519 | 46.4% |
| feat_e | 1.731 | 41.7% |
| feat_s | 1.848 | 38.2% |
| feat_m+ feat_e | 1.437 | 48.9% |
| feat_m+ feat_s | 1.501 | 46.6% |
| feat_m +<br>feat_e+ feat_s | **1.378** | **51.4%** |

Regarding the parameter selection of knowledge distillation, the relevant experimental results are shown in Table 4. All experiments were based on three channel features.

## 6. CONCLUSION

In this technical report, we described our submission to Task 1 of the DCASE 23 challenge. We add cepstrum analysis before traditional feature extraction. This signal processing method can weaken the device channel noise in advance. In addition, we also separated spectral envelope and fine structure in the cepstrum domain, and used these two as supplementary features. Experiments have shown that this is an effective combination feature. Based on the above feature, we trained a large-scale teacher model and then trained a student model that meets the challenge requirements through knowledge distillation. Through the calculation of NeSsi tool, our model requires 80.845K of memory, with a MACs value of 29.349M. The student model demonstrated good performance on the development dataset, with an accuracy rate 8.5% higher than the baseline system.

Table 4 Parameter selection of knowledge distillation

| T \ alpha | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|
| 2 | 49.5 | 50.0 | **51.4** | 50.4 |
| 3 | 48.5 | 48.2 | 50.6 | 49.1 |
| 5 | 49.6 | 49.8 | 51.1 | 50.3 |
| 10 | 47.9 | 49. 6 | 49.6 | 48.8 |

## 7. REFERENCES

[1] Irene Martín-Morató, Francesco Paissan, Alberto Ancilotto, Toni Heittola, Annamaria Mesaros, Elisabetta Farella, Alessio Brutti, and Tuomas Virtanen. Low-complexity acoustic scene classification in dcase 2022 challenge. 2022.

[2] ] I. Mart´ın-Morato, T. Heittola, A. Mesaros, and T. Virtanen, “Low-complexity acoustic scene classification for multidevice audio: analysis of DCASE 2021 Challenge systems,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), 2021.

[3] B. Kim, S. Yang, J. Kim, and S. Chang, “QTI Submission to DCASE 2021: Residual Normalization for Device-Imbalanced Acoustic Scene Classification with Efficient Design,” DCASE2021 Challenge, Tech. Rep., June 2021.

[4] C.-H. H. Yang, H. Hu, S. M. Siniscalchi, Q. Wang, W. Yuyang, X. Xia, Y. Zhao, Y. Wu, Y. Wang, J. Du, and C.-H. Lee, “A lottery ticket hypothesis framework for low-complexity device-robust neural acoustic scene classification,” DCASE2021 Challenge, Tech. Rep., June 2021.

[5] Phaye S, Benetos E, Wang Y. SubSpectralNet - Using Sub-Spectrogram based Convolutional Neural Networks for Acoustic Scene Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 825–829.

[6] T. Morocutti, D. Shalaby, “Receptive Field Regularized CNNs with Traditional Audio Augmentations,” DCASE2022 Challenge, Tech. Rep., June 2022.

[7] Jie H , Li S , Gang S . Squeeze-and-Excitation Networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[8] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19).

[9] Hinton G , Vinyals O , Dean J . Distilling the Knowledge in a Neural Network[J]. Computer Science, 2015, 14(7):38-39.

[10] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), 56–60. 2020. C. D. Jones, A. B. Smith, and E. F. Roberts, “A sample paper in conference proceedings,” in *Proc. IEEE ICASSP*, 2003, vol. II, pp. 803-806.