POLYPHONIC SOUND EVENT DETECTION AND LOCALIZATION USING A TWO-STAGE STRATEGY

Technical Report

Lihong Pi and Xue Zheng

ABSTRACT

The joint training of SED and DOAE affects the performance of both. We adopt a two-stage polyphonic sound event detection and localization method. The method learns SED first, after which the learned feature layers are transferred for DOAE. It then uses the SED ground truth as a mask to train DOAE.
We select Log mel spectrograms and GCCPHAT as the input features, and the GCCPHAT feature which contains phase difference information between any of the two microphones improves the performance of DOAE.

Index Terms**—**Sound event detection, Source Localization, Direction Of Arrival, Log Mel, GCCPATH, Convolutional Recurrent Neural Networks

1. PROPOSED FRAMEWORK

1.1. Preprocessing and Feature Extracting

1.1.1. Logmel Feature

Ambisonics was developed as a spatial sound encoding approach several decades ago. It is based on the spherical harmonic (SH) decomposition of the sound field. Ambisonics encoding for planewave sound fields can be expressed as

$$b(t)=\sum_{n=0}^{N} y_n s_n(t) \tag{1}$$

where $s_n(t)$ is the n-th plane-wave source signal, N is the total number of sources, and yn is the vector of the spherical harmonic function values for direction (θ_n, φ_n), and can be expressed as

$$\begin{aligned} y_n &= [Y_0^0(\theta_n,\phi_n), Y_1^{-1}(\theta_n,\phi_n), Y_1^0(\theta_n,\phi_n),\\ &Y_1^1(\theta_n,\phi_n),\cdots,Y_L^{-L}(\theta_n,\phi_n),\cdots,Y_L^0(\theta_n,\phi_n),\\ &\cdots,Y_L^L(\theta_n,\phi_n)]^T \end{aligned} \tag{2}$$

where L indicates the order of Ambisonics. It can be seen that Ambisonics contains the information of the source DOA. In addition, a higher directional resolution relates to a higher order of Ambisonics. Order-L of Ambisonics needs at least $(L+1)^2$ microphones to encode. In real applications, the sound field is recorded using a spherical microphone array and converted into Ambisonics.

1.1.2. Generalized Cross-Correlation

GCC is widely used in TDOA estimation by means of maximizing the cross-correlation function to obtain the lag time between two microphones. The cross-correlation function is usually calculated through the inverse-FFT of the cross power spectrum. GCC-PHAT is the phase-transformed version of GCC, which whitens the cross power spectrum to eliminate the influence of the amplitude, leaving only the phase. GCC-PHAT can be expressed as

$$GCC_{i,j}(t,\tau)=F_{f\to\tau}^{-1}\frac{X_i(f,t)X_j^*(f,t)}{\left|X_i(f,t)\right|\left|X_j^*(f,t)\right|} \tag{3}$$

Where $F_{f\to\tau}^{-1}$ is the inverse-FFT from f to τ ,X_i(f, t) is the ShortTime Fourier Transform (STFT) of the i-th microphone signal, and *denotes the conjugate. TDOA, which is the lag time Δτ betweentwo microphones, can then be estimated by maximizing GCC withrespect to τ . Nevertheless, this estimation is usually not stable, especially in high reverberation and low SNR environments, and doesnot directly work for multiple sources. However, GCCij (t, τ) contains all of the time delay information and is generally short-timestationary. GCC_{ij}(t, τ) can also be considered as a GCC spectrogram with τ corresponding to the number of mel-band filters. That is, GCC-PHAT can be stacked with a log mel spectrogram as the input features. In order to determine the size of GCC-PHAT, the largest distance between two microphones dmax needs to be used.

The maximum delayed samples corresponding to Δτmax can be estimated by dmax/c · fs, where c is the sound speed and fs is the sample rate. In this paper, log mel and GCC-PHAT will be stacked as the input features, considering the possibility of the advance and the delay of GCC. The number of mel-bands, therefore, should be no smaller than the doubled number of delayed samples plus one.

1.2. Model architecture

1.2.1. The model

A logarithm of mel-scale spectrogram (logmel) is widely used pre- processing step in audio signal analysis. In this work, we applied Gcc-path and Log-mel as our input features and they are concatenated to data of 10 channels. The full architecture for our model is presented in Fig. 1.

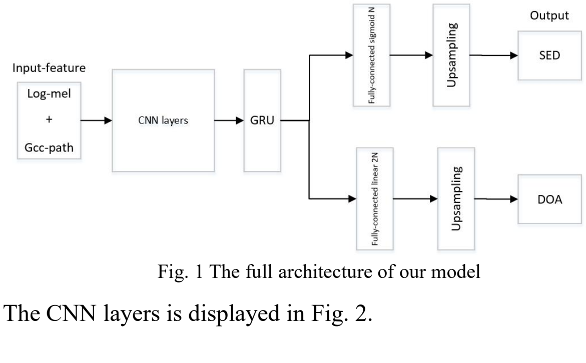

Fig. 1 The full architecture of our model

The CNN layers is displayed in Fig. 2.

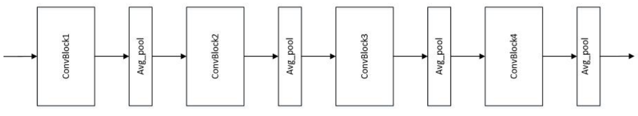

Fig. 2 The CNN layers

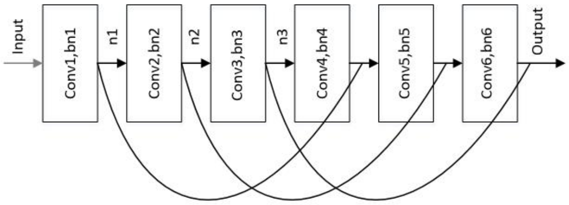

Fig. 3 The ConvBlock of CNN layers

The ConvBlock of CNN layers is illustrated in Fig. 3.

2. REFERENCES

[1]Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, Mark D. Plumbley. Polyphonic Sound Event Detection and Localization Using Two-Stage Strategy. arXiv preprint arXiv: 1905.00268v2 Paper URL:

[2]S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[3]R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," arXiv preprint arXiv:1505.00387, 2015.

[4]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770– 778.

[5]G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in CVPR, vol. 1, no. 2, 2017, p. 3.

[6]J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation net- works," arXiv preprint arXiv:170