High-Resolution Attention Network with Acoustic Segment Model for Acoustic Scene Classification

Xue Bai[1], Jun Du[1], Jia Pan[1], Heng-shun Zhou[1], Yan-Hui Tu[1];Chin-Hui Lee[2]

[1]University of Science and Technology of China, Hefei, China

[2]Georgia Institute of Technology, Atlanta, Georgia, USA

ICASSP 2020

CONTENTS

CONTENTS

Background & Motivation

- **The goal of Acoustic Scene Classification (ASC) task is to classify the audio to specific scenes, like park, airport, etc.**

- **For ASC, there are several difficulties in developing high-performance systems.**
  - Existence of overlapping sound events
  - Lack of distinguishing audio segments
  - Commonalities between different scene categories.

Background & Motivation

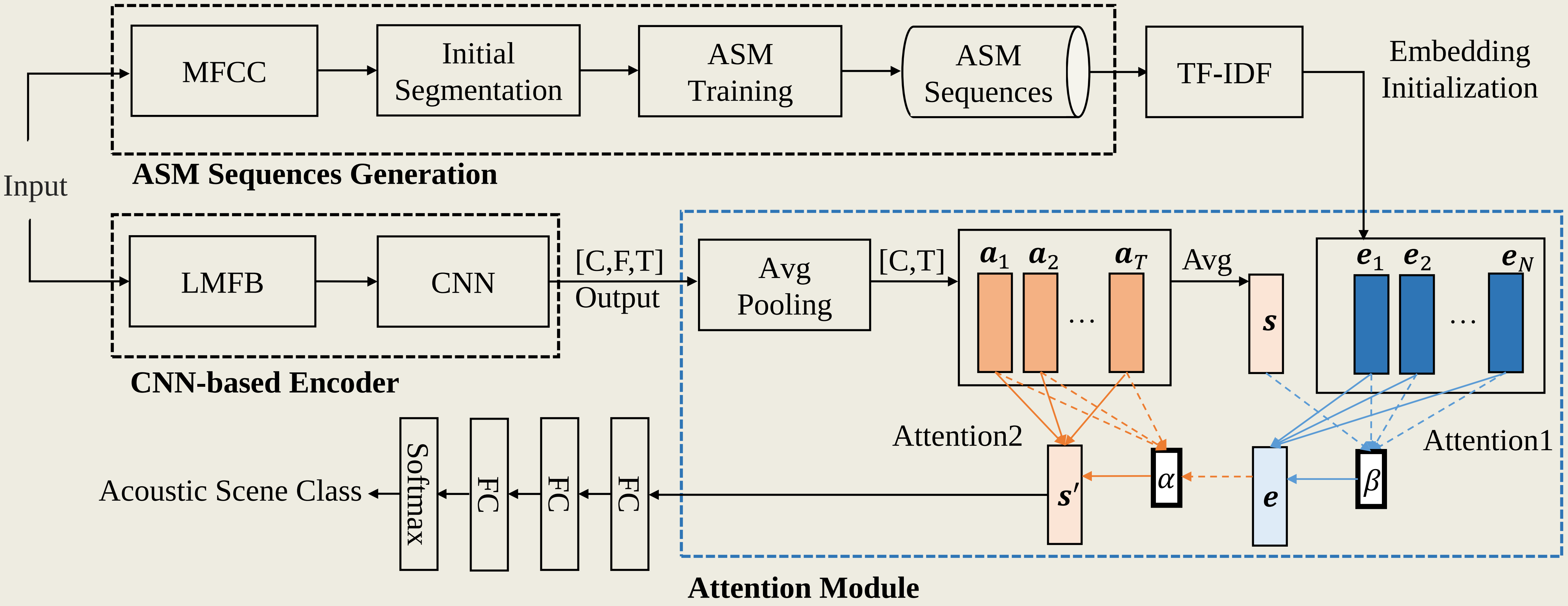
➢ **In this paper, we propose a novel strategy for acoustic scene classification, namely high-resolution attention network with acoustic segment model (HRAN-ASM) to improve the classification performance.**

- Use fully CNN to obtain high-level semantic information.
- The acoustic segment model (ASM) proposed in our recent work provides embedding vectors for our attention mechanism.
- Adopt two-stage attention strategy to select the relevant acoustic scene segments.

CONTENTS

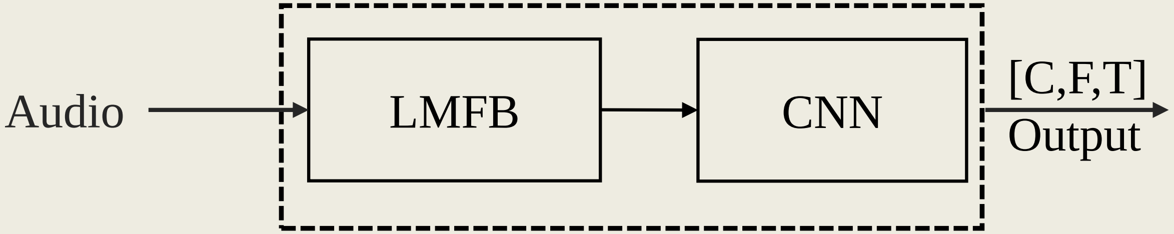
The proposed HRAN-ASM

➢ Overall framework



The proposed HRAN-ASM

➢ CNN-based Encoder



- Log mel-filterbank (LMFB) is our input feature with the size of $c \times f \times t$.
- VGGNet-16 is converted into a fully convolutional network (FCN) by simply removing its fully connected layers and used as our CNN-based encoder.
- The output is a 3-dimensional array of size $C \times F \times T$.

The proposed HRAN-ASM

➢ ASM Sequences Generation

- Use acoustic scene model to generate ASM sequences for each audio.
- Together the ASM sequences belonging to the same category.
- Term frequency (TF) and inverse document frequency (IDF) (TF-IDF) are used to obtain the ASM unit counts in each scene.

The proposed HRAN-ASM

ASM Sequences Generation

- The TF of ASM unit m in the nth scene is given by (1), where $c_{m,n}$ is the count of m in the nth scene.

$$TF_{m,n} = \frac{c_{m,n}}{\sum_{k=1}^{K} c_{k,n}} \quad (1)$$

- The IDF is given by (2), where L is the number of all scene types and $L(m)$ is the total number of times that ASM unit m appears in all scenes.

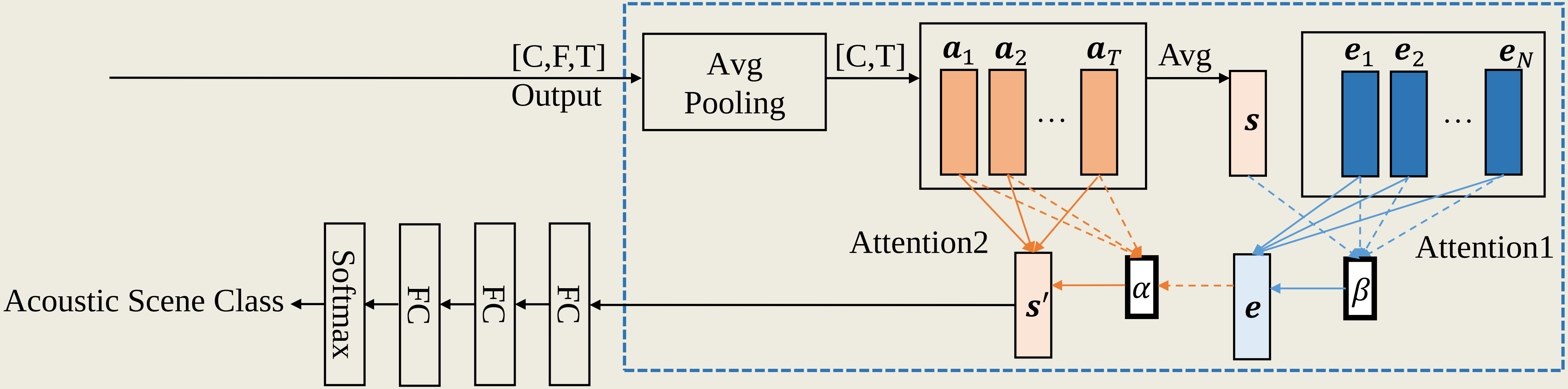
$$IDF_m = \log \frac{L+1}{L(m)+1} \quad (2)$$

- Each element in the embedding e_n is given by

$$e_{m,n} = TF_{m,n} \times IDF_m \quad (3)$$

The proposed HRAN-ASM

➢ Attention Module



- Get a vector representation $\boldsymbol{s}$ of the scene.

$$A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_T\} \tag{5}$$

$$\boldsymbol{s} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{a}_t \tag{6}$$

The proposed HRAN-ASM

➢ Attention Module

- The first attention of HRAN-ASM approach

  To explore the intrinsic connection between the current utterance and different scenes

$$\beta_i = \frac{\exp(\boldsymbol{e}_i^\top \cdot \boldsymbol{s})}{\sum_{n=1}^{N} \exp(\boldsymbol{e}_n^\top \cdot \boldsymbol{s})}, i \in (1, N) \tag{7}$$

$$\boldsymbol{e} = \sum_{i=1}^{N} \beta_i \boldsymbol{e}_i \tag{8}$$

- The second attention of HRAN-ASM approach

  To focus on effective time regions of the current utterance

$$\alpha_j = \frac{\exp(\boldsymbol{e}^\top \cdot \boldsymbol{a}_j)}{\sum_{t=1}^{T} \exp(\boldsymbol{e}^\top \cdot \boldsymbol{a}_t)}, j \in (1, T) \tag{9}$$

$$\boldsymbol{s}' = \sum_{j=1}^{T} \alpha_j \boldsymbol{a}_j \tag{10}$$

CONTENTS

Results and Analysis

- Experimental setup
  - Data set: DCASE2018 Task1a
  - CNN-based encoder: VGGNet-16 without fully connected layer
  - ASM sequences: 20 ASM units, 405-dimensional embedding vectors
  - Model Training:
    - Stochastic gradient descent (SGD)
    - Learning rate is 0.005
    - The number of iterations is 60

Results and Analysis

- The performance of different approaches on test set.

System	VGGNet-16[24]	ASM[12]	Self-Attention
Accuracy	67.4%	66.1%	68.9%

- The performance comparison of our HRAN approach with different initialization methods for the embedding vectors on test set.

System	Self-Attention	HRAN-Orth	HRAN-ASM
Accuracy	68.9%	68.3%	70.5%

Results and Analysis

- Fig. 2 (a) shows the LMFB spectrogram of a Bus scene.
- From Fig. 2 (b), the LMFB features at different time points are assigned with different weights and HRAN-ASM approach find critical segments.
- Our approach can generate different weights (in Fig. 2 (c)) to the set of embedding vectors initialized by ASM while only a global embedding vector is adopted in the self-attention case.

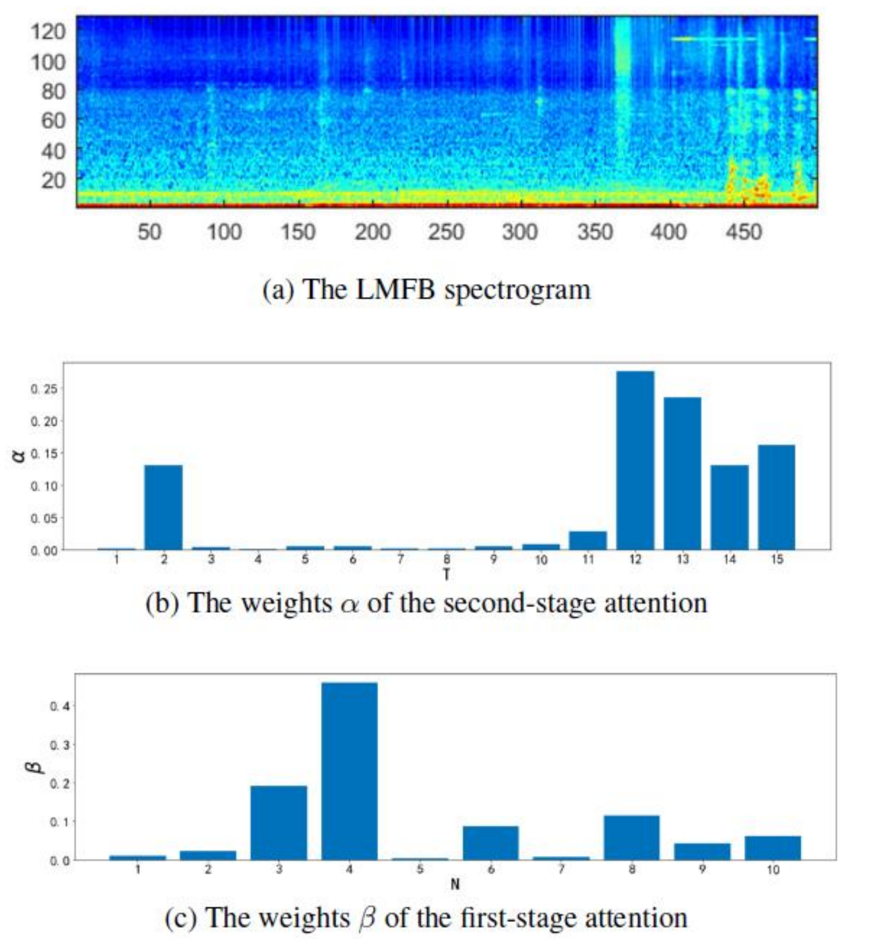


Fig. 2. Visualization of attention for one example in Bus scene.

CONTENTS

Conclusion and Future Work

➢ Conclusion

- The acoustic segment model is used to generate representative embedding for each scene as a guided information.
- A two-stage attention mechanism is utilized to get salient frames of each scene and improve recognition.
- Our approach can achieve highly competitive performance **under single system and no data expansion.**

➢ Future Work

- More fusion methods will be tried to improve the recognition rate of ASC.

CONTENTS

Thanks for listening!

If you have any questions about this paper, Please contact me and I will answer it.