XIAOFEI WANG, Hefei University of Technology
JIEHUI TANG, Hefei University of Technology
ZHEN XIAO, Hefei University of Technology
XUELIANG LIU*, Hefei University of Technology

1 Dual-branch Model Structure

In our research, we adopted a dual-branch structure consisting of the pre-trained $\mathbf{FOP}_{frozen}$ and the $\mathbf{FOP}_{update}$ that requires updates, fusing the output results of the two feature output layers (FOPs) through a learnable weight W (see Figure 1). Specifically, we froze the parameters of the $\mathbf{FOP}_{frozen}$ during training, while the $\mathbf{FOP}_{update}$ was trained from scratch. This design allows the $\mathbf{FOP}_{frozen}$ model to serve as a fixed feature extractor, while the $\mathbf{FOP}_{update}$ dynamically complements the output results of the $\mathbf{FOP}_{frozen}$. We trained the regularly $\mathbf{FOP}_{update}$ using augmented face-audio sample pairs to enhance its generalization capability and robustness. Furthermore, to modulate the supplementary information, we introduced a strategy for dynamically adjusting the weight W, automatically adjusting the weight based on loss changes during the training process to optimize the fusion of the two model outputs. This strategy not only enhances the model's ability to handle multimodal data but also improves performance stability in complex environments.

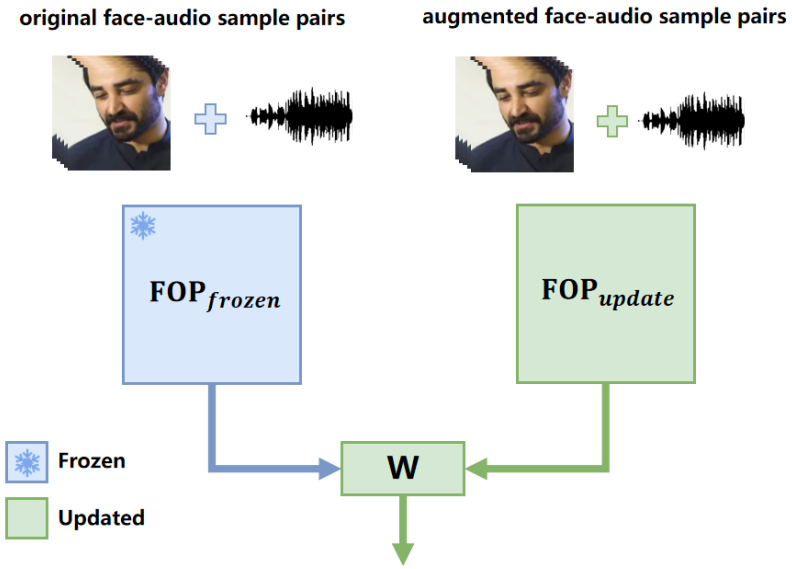

Fig. 1. Model structure diagram

*Corresponding author

Authors' Contact Information: Xiaofei Wang, Hefei University of Technology, xiaofei2027@gmail.com; Jiehui Tang, Hefei University of Technology, hfut2021214635@gmail.com; Zhen Xiao, Hefei University of Technology, xiaozhen0305@hotmail.com; Xueliang Liu*, Hefei University of Technology, liuxueliang1982@gmail.com.

2 Positive and Negative Sample Pair Weight Configuration with Data Augmentation

2.1 Positive and Negative Sample Pair Weight Configuration

In our research, to more effectively handle difficult classification scenarios (i.e., misclassifying positive sample pairs as different classes and negative sample pairs as the same class), we adjusted the weight parameters of positive and negative sample pairs in the loss function. Specifically, we modified the weights of positive and negative sample pairs based on their similarity. For positive sample pairs, we assigned smaller weights to pairs with higher similarity and larger weights to pairs with lower similarity. For negative sample pairs, we assigned larger weights to pairs with higher similarity and smaller weights (or even zero) to pairs with lower similarity.

As shown in Figure 2, the same color anchors represent facial-audio samples under the same label, with hollow anchors indicating audio data and solid anchors indicating facial data. Different colors represent different identity labels. By constructing inter-class sample pairs and intra-class sample pairs, we create positive and negative sample pairs and then apply weighting to these sample pairs. The purpose of this strategy is to enhance the model's ability to focus on the misclassification of positive samples (i.e., samples that should be similar but are predicted to be dissimilar) and the correct classification of negative samples (i.e., samples that should be dissimilar but are predicted to be similar). Through this weight adjustment, we successfully improved the model's robustness and accuracy in handling challenging classification tasks. Particularly, the model demonstrated superior performance in recognizing samples with complex features.

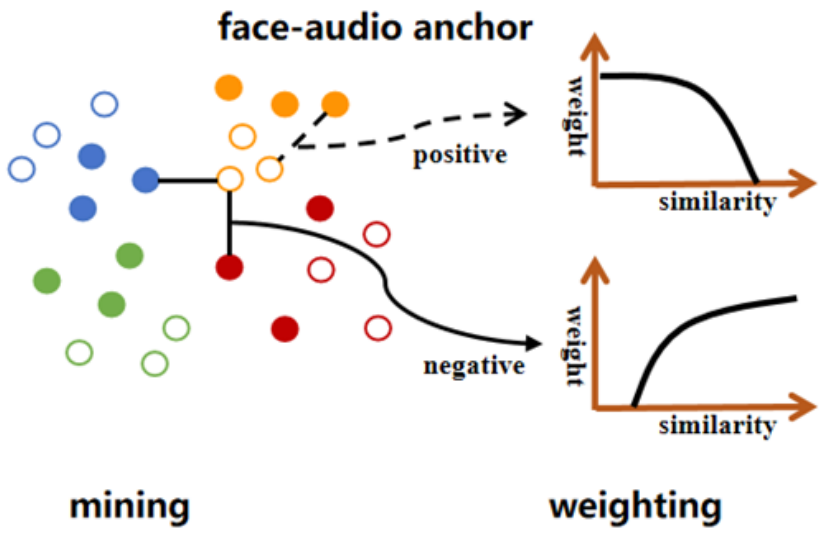

Fig. 2. Sample Pair Construction and Weighting Strategy

2.2 Data Augmentation

In our research, we used a dataset containing pairs of audio and facial features, where each face corresponds to a specific audio clip and an identity label.We found that in the original dataset, under the same label (i.e., the same speaker), data

often include multiple different scenes. In the original features, facial data from one scene only corresponds to audio data from the same scene, existing as one-to-one pairs. We believe that even in different scenes within the same label, audio and facial data can still form valid positive sample pairs.Therefore, to enhance the robustness and diversity of the training set, we employed a data augmentation strategy that disrupts the original pairing relationships within the same label category of facial and audio samples (see Figure 3). This method randomly generated additional training sample pairs. By doing so, we aim to simulate a broader range of real-world scenarios the model might encounter, thereby improving the system's generalization ability. Additionally, this approach allows the model to accurately recognize diverse scenarios without the need for new data collection, effectively maximizing the potential of the existing dataset.

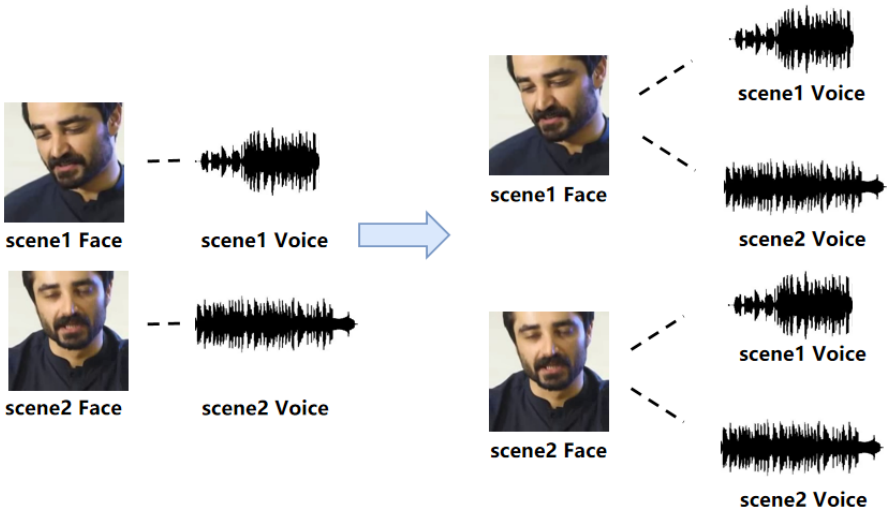

Fig. 3. Data augmentation strategy diagram.

3 Score Polarization Processing

By observing multiple sets of generated L2 results, we adopted a threshold technique to achieve score polarization, resulting in outcomes with greater distinction. The core of this method lies in amplifying the differences between high and low scores, making high scores more pronounced and low scores more evident. This approach can significantly enhance the discrimination ability of the model's predictions. In the specific operation process, we fine-tune the model's output scores by setting appropriate thresholds, making scores above the threshold higher and scores below the threshold lower. Through this method of polarizing scores, we aim for the model to provide clearer and more reliable prediction results.

4 Other Attempts

Additionally, we explored several other methods to further enhance the extraction capabilities of facial and audio features, as well as to improve the model's performance and prediction accuracy. One of the approaches considered was the removal of hard-to-learn samples. These samples may contain noise, outliers, or other interfering factors that make it difficult for the model to learn and generalize effectively. By eliminating these challenging samples, we aim to reduce interference during the training process, thereby improving the overall performance of the model.

We employed various techniques for facial and audio feature extraction, such as DeepSpeech, InsightFace, and metric learning, to enhance the model's performance and prediction accuracy. Using pre-trained DeepSpeech models, we extracted high-quality per-frame speech features from audio data. The raw audio signal was reorganized into overlapped windows of size 16 time intervals, with each window centered on the corresponding video frame. With InsightFace, we utilized pre-trained models trained on large-scale facial datasets to extract precise facial features from images. Metric learning, by constructing loss functions based on positive and negative sample pairs, is particularly suitable for classification tasks. It introduces margin constraints, enhancing the model's ability to distinguish features of different categories.

We explored several different attention mechanism alternatives and aimed to improve the model's ability to understand complex data and enhance accuracy and robustness in classification tasks through these modifications. Specifically, We processed input features using convolutional layers and attention mechanisms. First, meaningful features were extracted through convolution, and important segments were emphasized by weighting them with an attention network. Additionally, we implemented a cross-modal attention mechanism to align and fuse features from facial and audio inputs, thereby enhancing the combined feature representation. Through these replacements and optimizations, we aim for the model to better understand and handle complex data, and to exhibit higher accuracy and robustness in classification tasks.