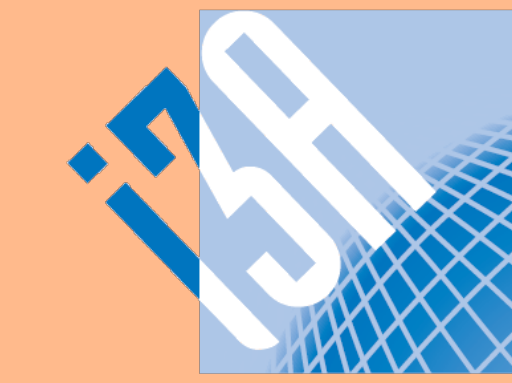


Generalizing AUC Optimization to Multiclass Classification for Audio Segmentation With Limited Training Data



Pablo Gimeno Victoria Mingote Alfonso Ortega Antonio Miguel Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain

Introduction

Audio segmentation aims to obtain a set of labels so that an audio signal can be classified into a predefined set of classes, e.g., speech, music or noise, and thus be separated into homogeneous regions.

- *Music-related audio segmentation*: speech and music separation, music detection, **relative music loudness estimation**
- Relevance in broadcast content:
 - monitor copyright infringements
 - document information retrieval

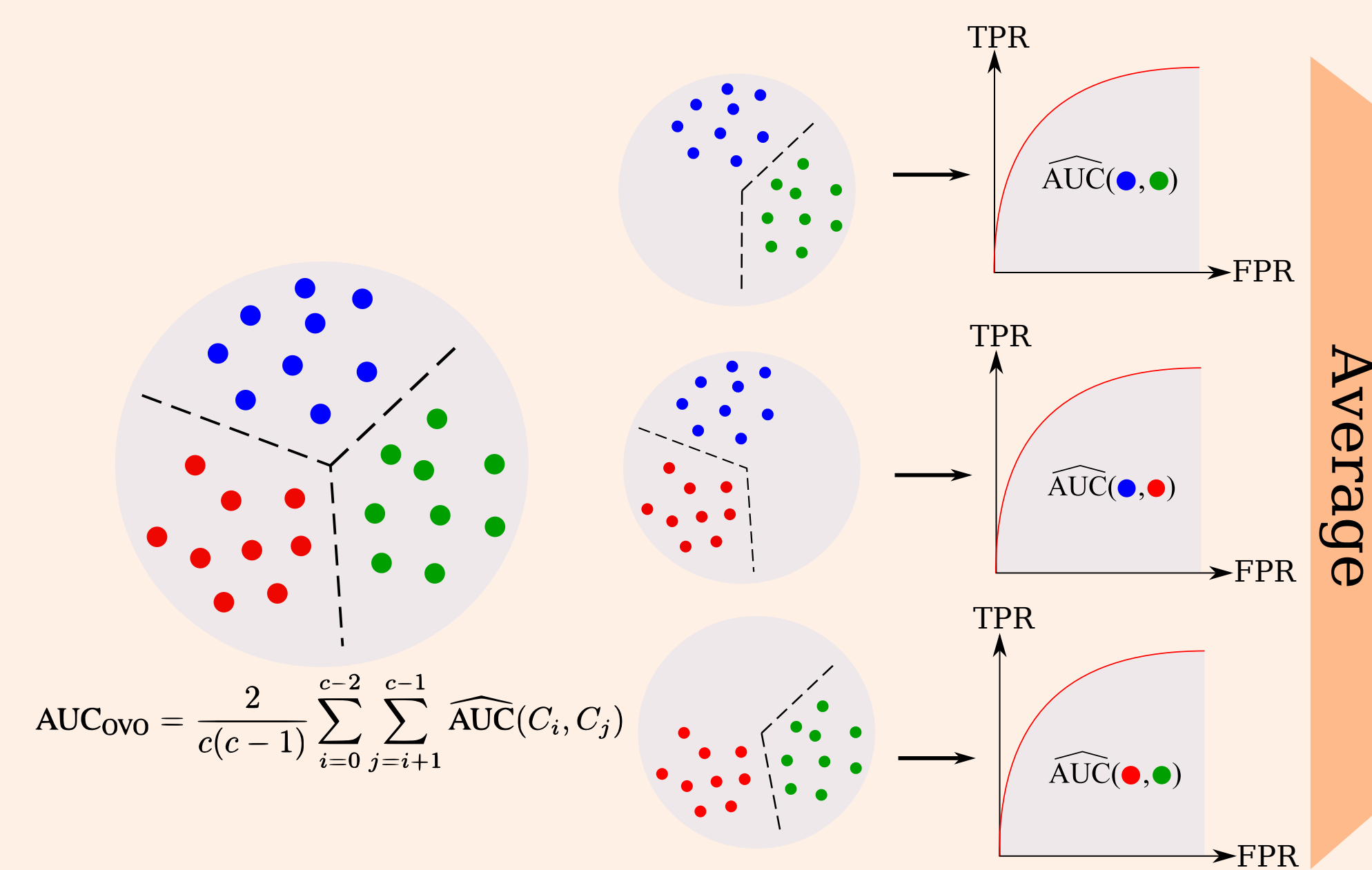
From binary to multiclass AUC

Multiclass AUCs can be computed by averaging binary AUCs

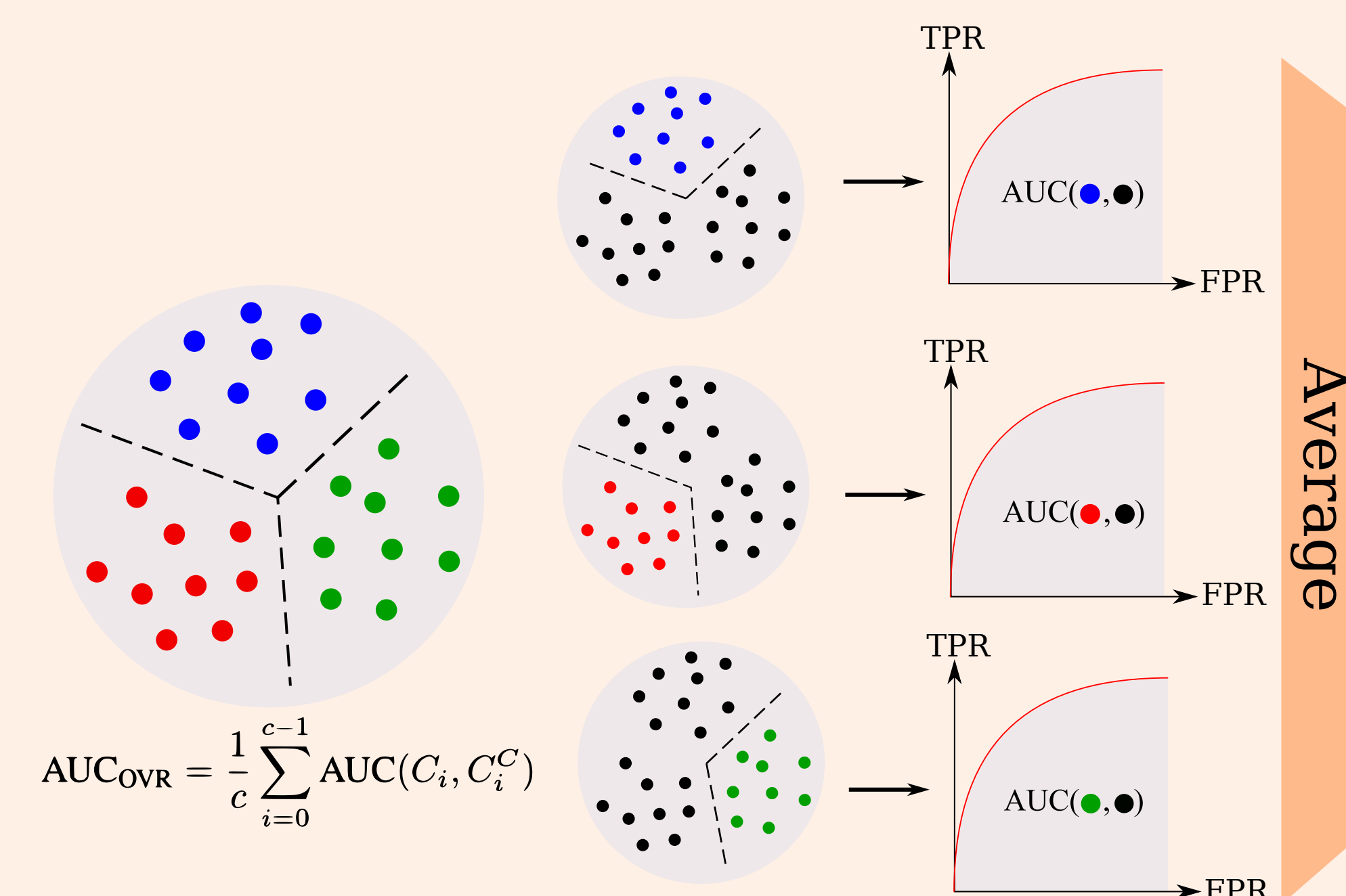
AUC one versus one (OVO)

In a multiclass setup, $AUC(\bullet, \bullet) \neq AUC(\bullet, \bullet)$

use a modified version, $\widehat{AUC}(\bullet, \bullet) = \frac{1}{2} AUC(\bullet, \bullet) + \frac{1}{2} AUC(\bullet, \bullet)$



AUC one versus rest (OVR)



Multiclass AUC optimization

Several works have already tried to optimise AUC metric from its binary expression.

$$AUC = \frac{1}{N^+N^-} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} u(s_i^+ - s_j^-). \quad (1)$$

Using sigmoid approximation to overcome differentiability issues:

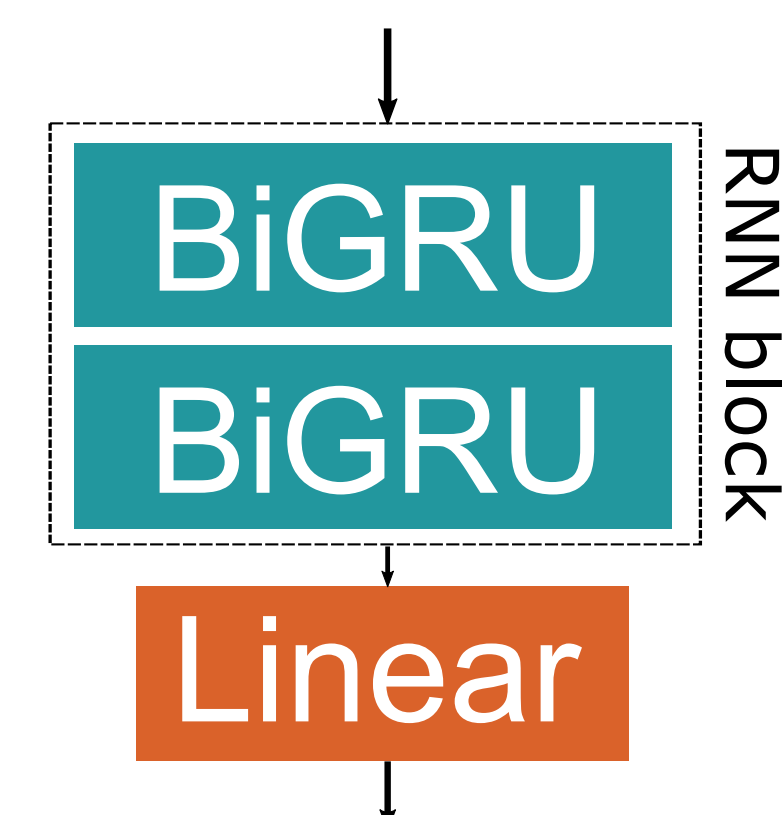
$$aAUC = \frac{1}{N^+N^-} \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \sigma(\delta(s_i^+ - s_j^-)), \quad (2)$$

Multiclass AUCs are computed averaging N binary AUCs

Apply N times sigmoid approximation and average to obtain a differentiable expression

Experimental setup

Neural network



- 2 stacked BiGRU with 128 neurons each & linear layer for final classification
- **Feature extraction**: 128 Mel filter bank + chroma features
- Fixed setup in all our experiments

Data description

OpenBMAT dataset: Broadcast domain data

No music	3 class audio segmentation task aiming to separate foreground and background music
Background music	
Foreground music	

- **Train**: Splits 0 to 7, 22 hours of audio
- **Validation**: Split 8, 3 hours of audio
- **Test**: Split 9, 3 hours of audio

Results

Training objective	AUC _{OVO} (%)	AUC _{OVR} (%)	Avg. AUC(%) prec vs recall
Softmax CE	81.69±0.84	79.42±0.69	66.05±0.97
Angular softmax	80.95±0.71	79.23±0.65	65.89±0.91
aAUC _{OVO}	83.67±0.54	81.28±0.61	69.55±0.80
aAUC _{OVR}	82.46±0.81	80.33±0.71	68.51±0.90

Table 1. AUC_{OVO}, AUC_{OVR} and average area under the precision versus recall curve on test data for the audio segmentation systems trained using the proposed multiclass AUC training objectives compared to two variants of cross entropy based training. (Mean ± standard deviation over 10 different experiments)

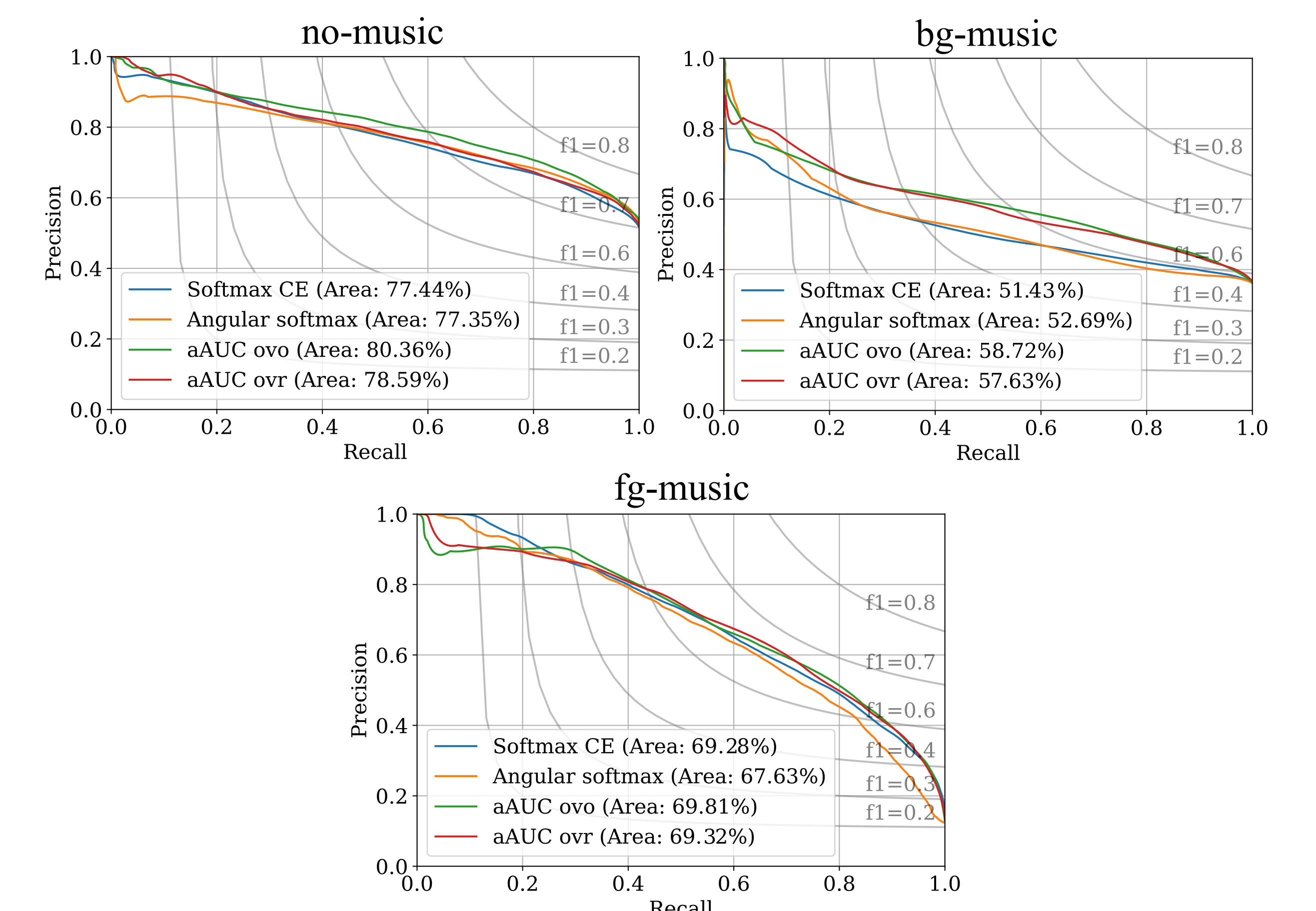


Figure 1. Precision versus recall curves, f1 isocurves, and area under the precision versus recall curve per class on the test data for the proposed multiclass AUC training objectives compared to two variants of cross entropy based training. (Average curve obtained over 10 different experiments)

Conclusions

- Introduced a generalization of the **AUC optimization** framework that can be applied to an **arbitrary number of classes**
- **Multiclass AUC optimisation** techniques show **better performance than traditional training objectives** in a limited training data scenario
 - 14% relative improvement in overall accuracy using aAUC_{OVO}
- Results show that **OVO approach**, using combinations of pairs of classes, is a **more robust training criterion** than the use of one-versus-rest binarisation solutions