Speech Emotion Recognition from Acoustic and Text Features

Using Multitask Learning to Improve Prediction of Valence, Arousal, and Dominance in Speech

@btatmaja
October 25th, 2019

Overview

- Background & previous results
- Problems & purpose
- Method & Experiments:
  - MTL vs STL
  - MTL on eGeMAPS & 31 Features
  - Data visualization with t-SNE
- Summary & future works

Backgrounds

- A **categorical speech emotion** recognition are developed with fair accuracy (±70%).
- Instead of predicting emotion in category, **estimating the “degree” of emotion is more important** as it enables deeper analysis in continuous space (2D, 3D, or 4D).
- The previous result on dimensional emotion recognition shows **lower performance score on valence** compared to arousal and dominance score.
- That result shows the **need to improve/balance the prediction** of emotion in dimensional space.

Problems

- How to balance CCC score among three emotional attributes using MTL?
- Which feature set & structure learn better?

Purpose

- Investigate the effectiveness of proposed multitask learning (MTL) method over single task learning (STL) method:
  - In different acoustic feature set: 31 features vs eGemaps.
  - In different network structures.

Datasets

- USC-IEMOCAP
  - As previously used, but now I used dimensional labels (valence, arousal, and dominance)
  - All data is used, i.e. 10,039 utterances.
- **MSP-IMPROV**
  - Improvement of IEMOCAP by the same author
  - Promoting naturalness by improvised speech (beside natural interaction and read sentence)
  - It consists of 8,438 utterances.

[1] C. Busso et al., “MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception,” Trans. Affect. Comput., vol. 8, no. 1, pp. 67–80, Jan. 2017.

Feature Sets

- **31 Features**: 3 time domain features, 5 frequency domain features, 13 Mel-frequency cepstral coefficients (MFCCs), 5 Fundamental frequencies, 5 Harmonics.
- **eGeMAPs**: loudness, alpha ratio, hammarberg index, spectral slope 0-500 Hz, spectral slope 500-1500 Hz, spectral flux, 4 MFCCs, F0, jitter, shimmer, Harmonics-to-Noise Ratio (HNR), Harmonic difference H1-H2, Harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.

[1] F. Eyben et al., “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 190–202, Apr. 2016.

Related Work

Experiment 1: MTL vs STL on IEMOCAP

Loss	Valence	Arousal	Dominance	Average
STL1 (Val)	0.08	-0.002	0.0001	0.028
STL2 (Aro)	0.0007	**0.432**	0.077	0.145
STL3 (Dom)	-0.00006	-0.0039	**0.4027**	0.132
MTL1 [1]	0.07	0.3728	0.0018	0.148
MTL2	0.0818	0.4004	0.3049	0.262
MTL3	**0.1379**	0.4294	0.3952	**0.320**

MTL1: α=0.7, β=0.3, γ=0.0 (MSE)
MTL2: α=1.0, β=1.0, γ=1.0 (CCC)
MTL3: α=0.7, β=0.3, γ=0.6 (CCC)

[1] S. Parthasarathy and C. Busso, “Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning,” in interspeech, 2017, pp. 1103–1107.

Experiment 1: MTL vs STL in MSP-IMPROV

Loss	Valence	Arousal	Dominance	Average
STL1 (Val)	0.249	0.002	-0.0014	0.083
STL2 (Aro)	0.0005	**0.487**	0.0042	0.168
STL3 (Dom)	0.001	0.0074	0.3955	0.135
MTL1 [1]	0.139	0.381	-0.00004	0.173
MTL2	**0.2735**	0.4733	0.39	**0.379**
MTL3	0.2567	0.4611	**0.4046**	0.374

MTL1: α=0.7, β=0.3, γ=0.0 (MSE)
MTL2: α=1.0, β=1.0, γ=1.0 (CCC)
MTL3: α=0.7, β=0.3, γ=0.6 (CCC)

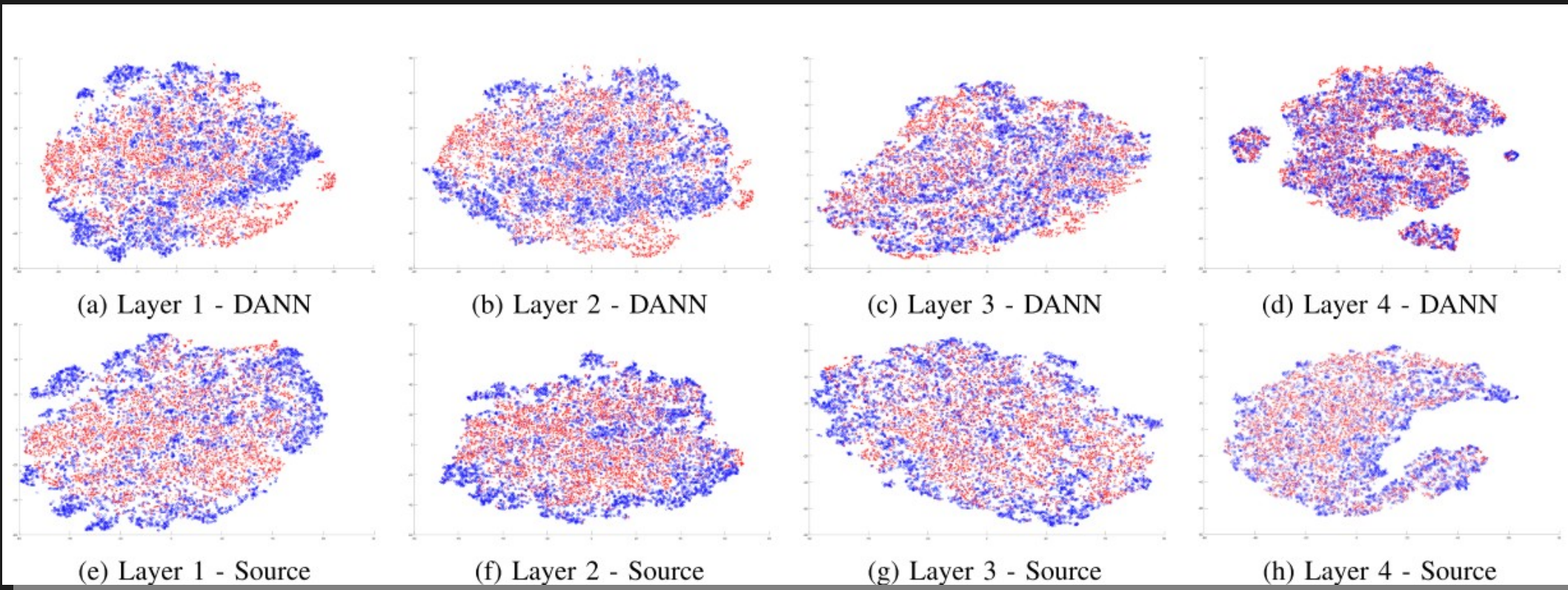
Experiment 2: MTL on eGeMAPs and 31 Features

Method	MSE	MAPE	MAE	CCC
	31 Features			
DNN	1.441	32.372	0.965	0.050
GRU	1.332	30.802	0.925	0.076
LSTM	1.068	28.278	0.823	0.088
	eGeMaps			
DNN	0.955	25.855	0.7	0.198
GRU	0.663	23.488	0.644	0.234
LSTM	0.683	23.814	0.655	0.245

Experiment 3: Data Visualization

- The aim of data/feature representation/visualization in this experiment:
  - Which features learn better.
  - Which (network) structure perform better.
- t-SNE (t-Distributed Stochastic Neighbor Embedding): technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets.
- The more separation among categories, the better process learned.

t-SNE from similar researches [2]

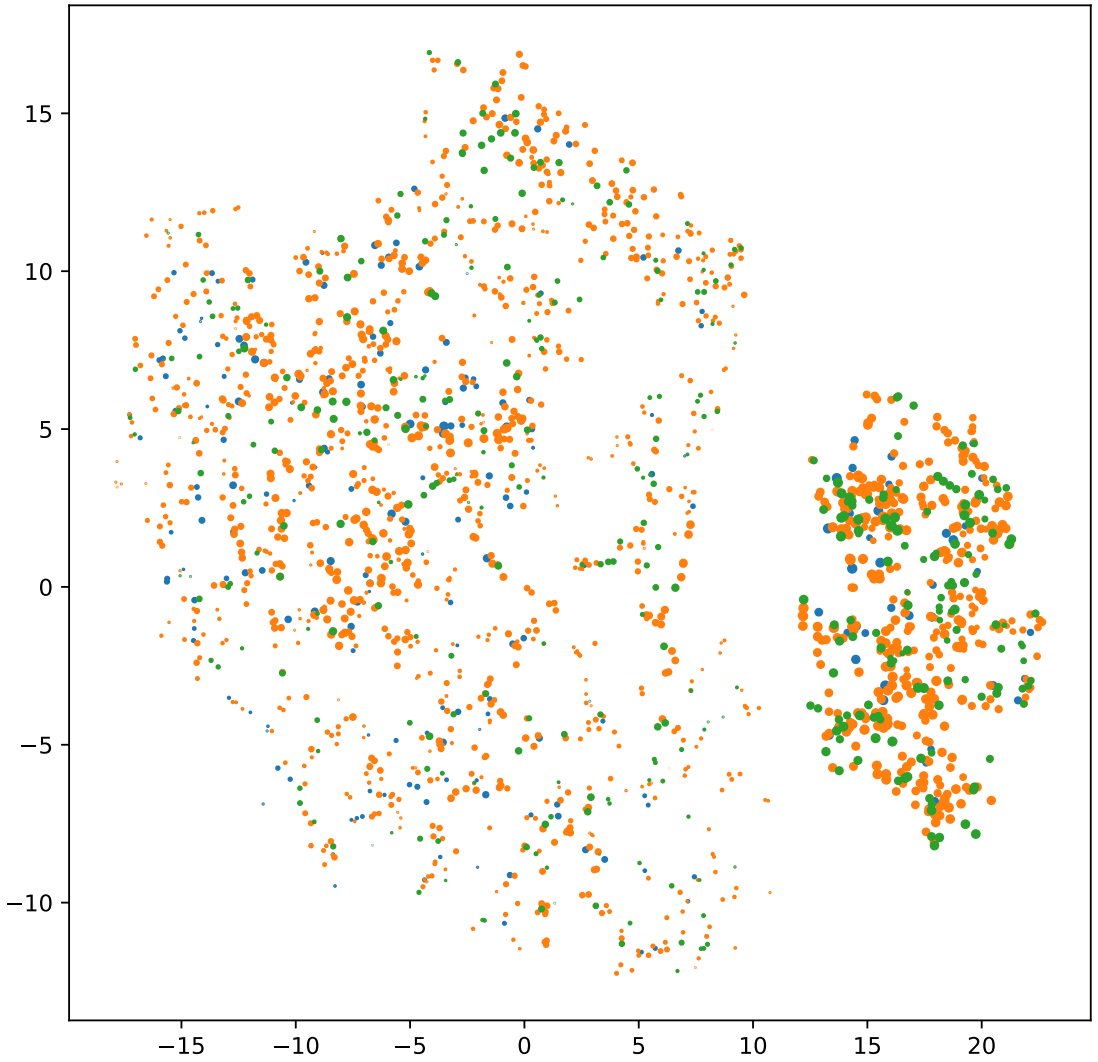


(a) Layer 1 - DANN (b) Layer 2 - DANN (c) Layer 3 - DANN (d) Layer 4 - DANN

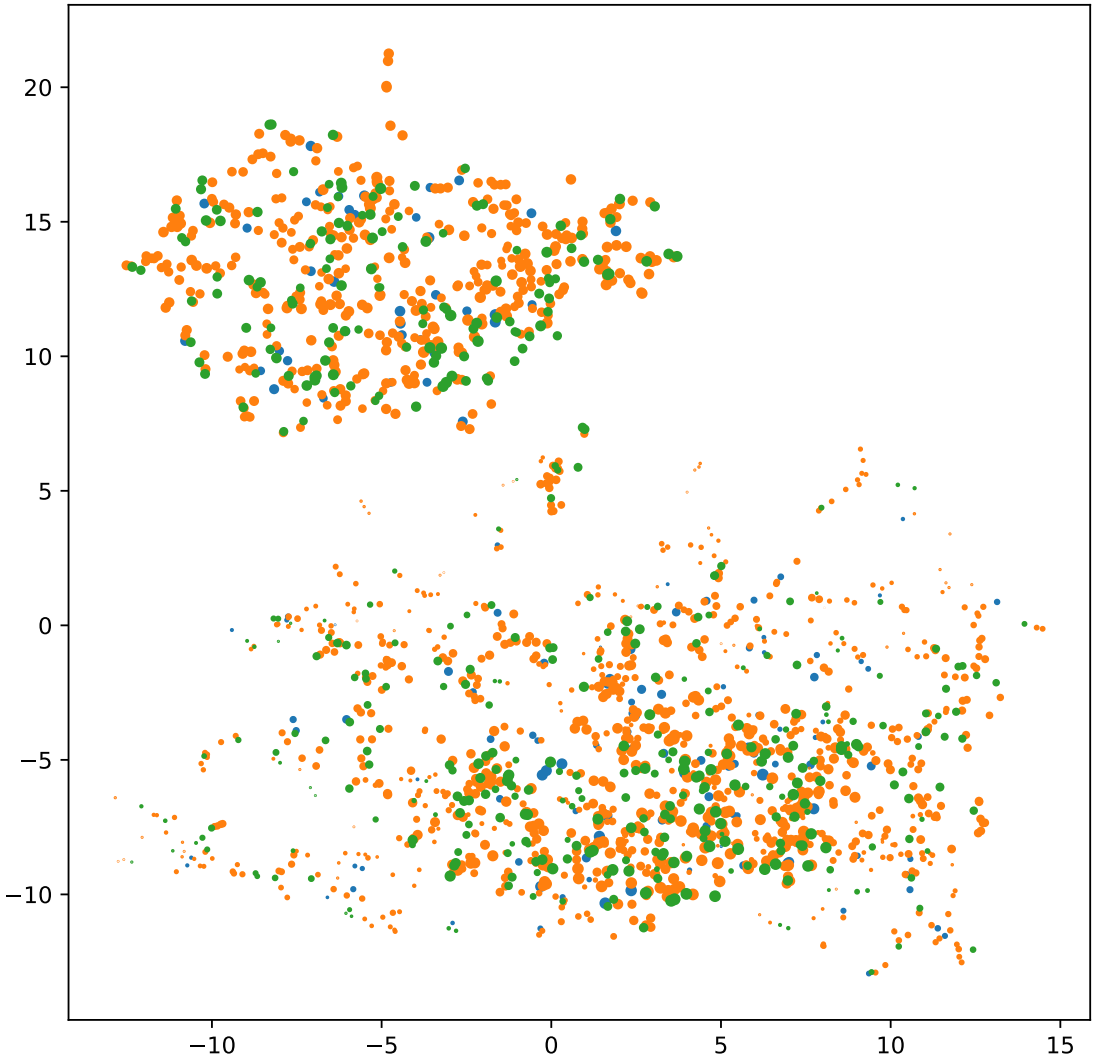
(e) Layer 1 - Source (f) Layer 2 - Source (g) Layer 3 - Source (h) Layer 4 - Source

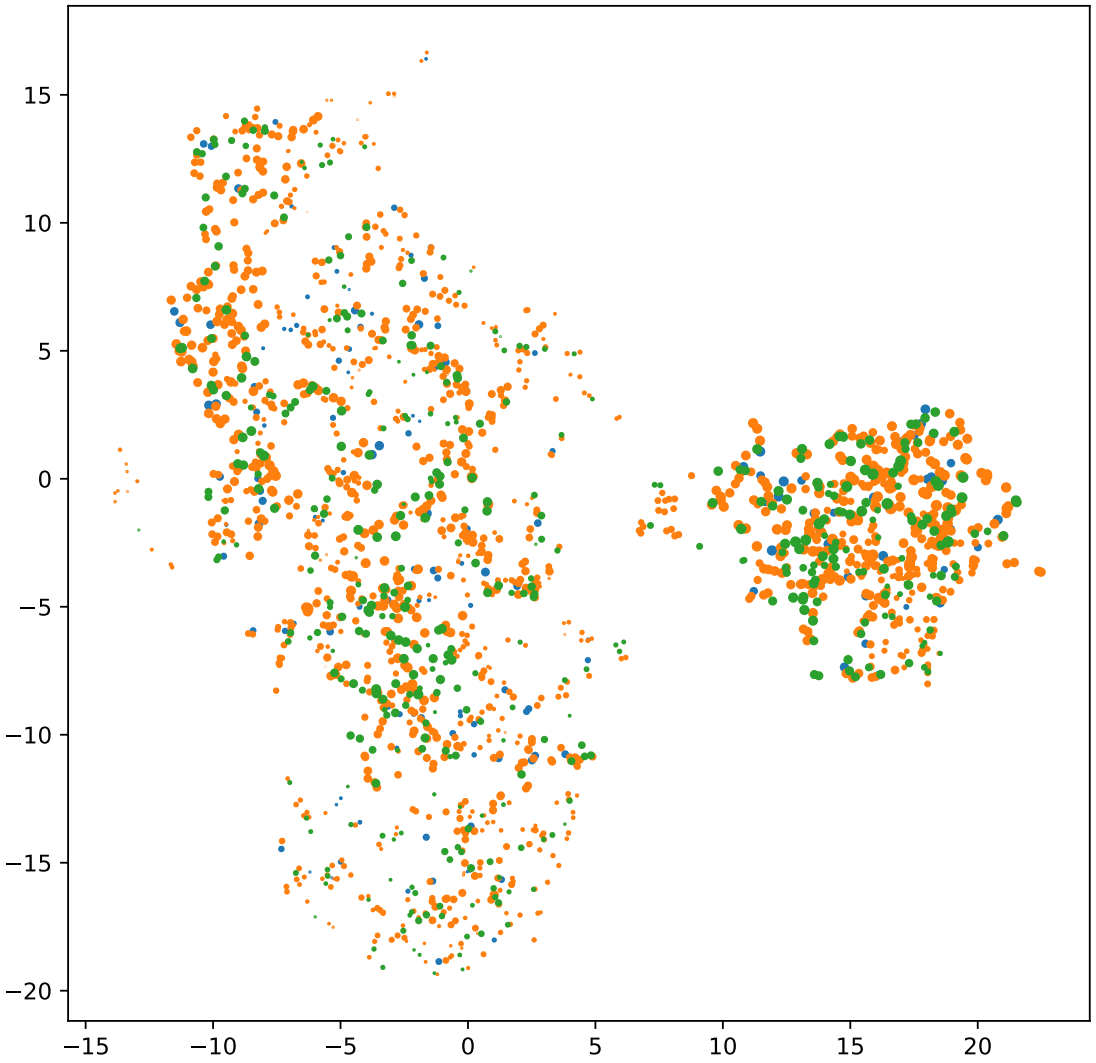
[2] M. AbdelWahab and C. Busso, “Domain Adversarial for Acoustic Emotion Recognition,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 26, no. 12, pp. 2423–2435, 2018.

Obtained t-SNE: **MTL** from eGeMAPs-test data

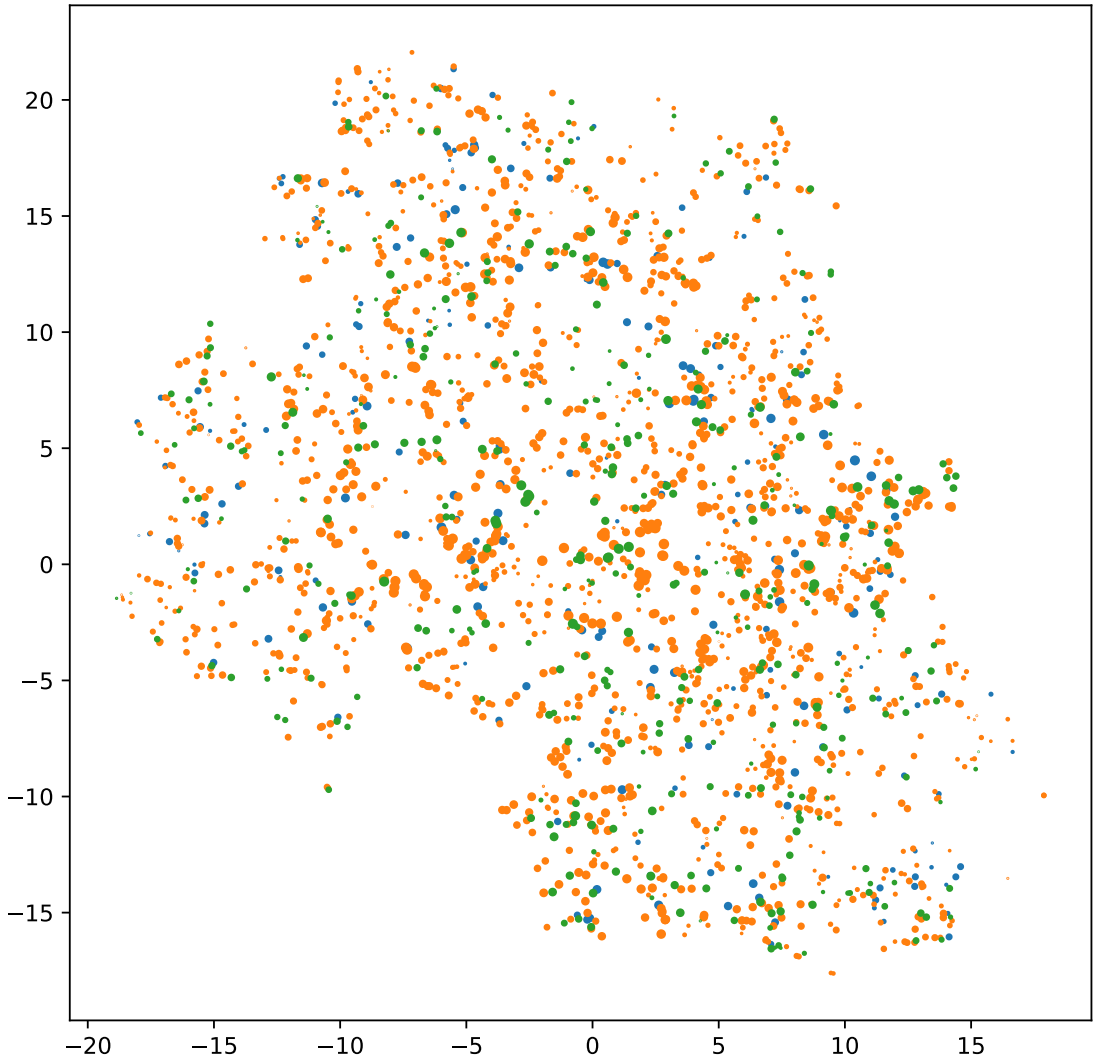


Obtained t-SNE: **STL** from eGeMAPs-test data





Obtained t-SNE: MTL from **31 features**-test data



Obtained t-SNE: MTL from eGeMAPs-**train data**

Summary

- Three experiment scenarios are conducted to investigate the effectiveness of proposed MTL method.
- The result shows improvement on both IEMOCAP and MSP-IMROV dataset with parameter [0.7, 0.3, 0.6] and [1.0, 1.0, 1.0] via random search.
- The result from eGeMAPs feature set shows a better score than 31 acoustic features.

Remaining problems/ Future works

- Parameter optimization via grid search/linear search.
- Correctness and interpretation of t-SNE visualization.