Multi-modal Emotion Detection with Transfer Learning

Amith Ananthram[1], Kailash Karthik Saravanakumar[1], Jessica Huynh[1], Homayoon Beigi[12]

Columbia University[1], NY, Recognition Technologies, Inc.[2], NY | [amith.ananthram,kailashkarthik.s,jyh2127]@columbia.edu, beigi@recotechnologies.com

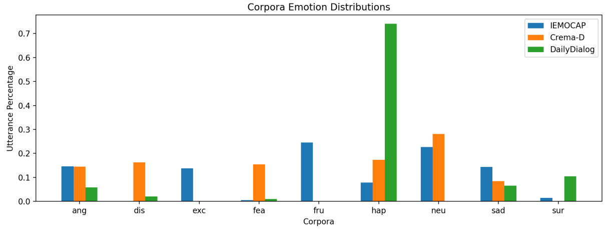
1. Overview: Automated emotion detection in speech is a challenging task due to 1) the complex interdependence between words and the manner in which they are spoken and 2) the small size and incompatible labeling idiosyncrasies of available training corpora. To address these challenges, in this work, we present a multi-modal approach that first transfers learning from related tasks in speech and text to produce robust neural embeddings and then uses these embeddings to train a pLDA classifier that is able to adapt to previously unseen emotions. We begin by training a multilayer TDNN on the task of speaker identification with the VoxCeleb corpora and then fine-tune it on the task of emotion identification with the Crema-D corpus. Using this network, we extract speech embeddings for Crema-D from each of its layers, generate and concatenate text embeddings for the accompanying transcripts using a fine-tuned BERT model and then train an LDA-pLDA classifier on the resulting dense representations. We exhaustively evaluate the predictive power of every component. Our best variant, trained on only VoxCeleb and Crema-D and evaluated on IEMOCAP, achieves an EER of 38.05%. Including some IEMOCAP during training produces a 5-fold averaged EER of 25.72% (For comparison, 44.71% of the gold-label annotations include an annotator who disagrees). Full paper: https://rebrand.ly/emotion_detection

2. Background and Motivation:

- emotion is **hard:** multimodal; emotion corpora are 1) small, 2) have non-uniform, mutually incompatible labeling distributions
- **key idea:** pre-train on auxiliary task (speaker identification) to address data sparsity, train a pLDA classifier on multi-modal embeddings to allow adapting to new emotion classes
- **prior work:** transfer from ASR [1], pLDA [2] on neural embeddings for speaker recognition [3]

3. Datasets

Speaker ID: VoxCeleb1/VoxCeleb2 (100k+/1m+ utts)
Emotion ID: Crema-D/DailyDialog/IEMOCAP (7k/100k/7k utts)



4. Methods

- implemented in Kaldi [4]
- collapse emotion labels into larger canonical emotions
- **speech:** extract 30 MFCC, 3 pitch features; pre-train TDNN on speaker ID (VoxCeleb1|2), fine-tune on emotion ID (Crema-D)
- **text:** fine-tune BERT on emotion identification (DailyDialog)
- extract and concatenate neural embeddings, train pLDA [2]

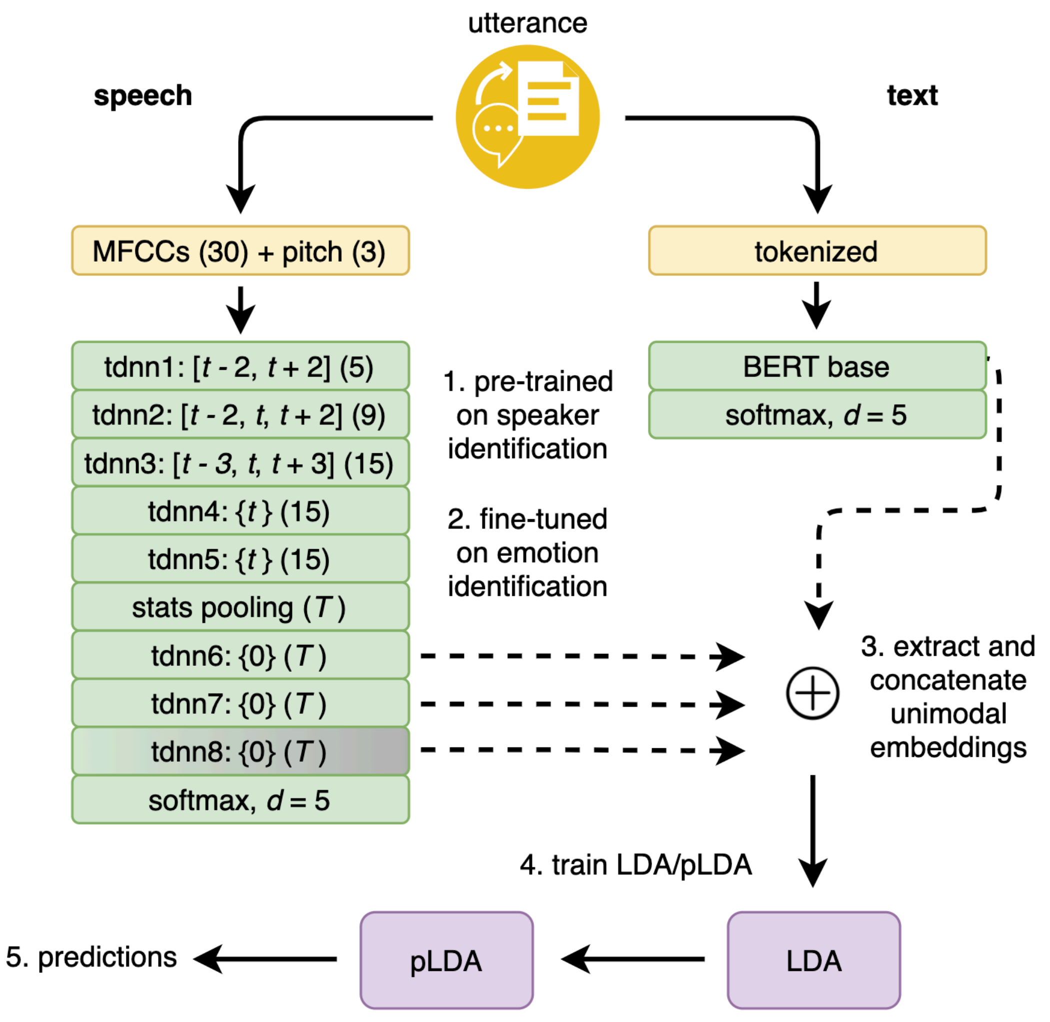
Selected References

[1] S. Zhou and H. Beigi. A transfer learning method for speech emotion recognition from automatic speech recognition, 2020.
[2] S. Ioffe, Probabilistic linear discriminant analysis, 2006.
[3] D. Snyder et al., X-vectors: Robust dnn embeddings for speaker recognition, 2018.
[4] D. Povey et al., The kaldi speech recognition toolkit, 2011.

5. Experiments

- TDNN: w/ and w/o learning on first six layers, w/ additional 8th layer, w/ noise augmentation, w/ and w/o IEMOCAP
- pLDA: speech only, text only, speech + text, w/ and w/o IEMOCAP

6. Architecture Diagram



utterance

speech

text

MFCCs (30) + pitch (3)

tdnn1: [t - 2, t + 2] (5)
tdnn2: [t - 2, t, t + 2] (9)
tdnn3: [t - 3, t, t + 3] (15)
tdnn4: {t} (15)
tdnn5: {t} (15)
stats pooling (T)
tdnn6: {0} (T)
tdnn7: {0} (T)
tdnn8: {0} (T)
softmax, d = 5

tokenized

BERT base
softmax, d = 5

1. pre-trained on speaker identification

2. fine-tuned on emotion identification

3. extract and concatenate unimodal embeddings

4. train LDA/pLDA

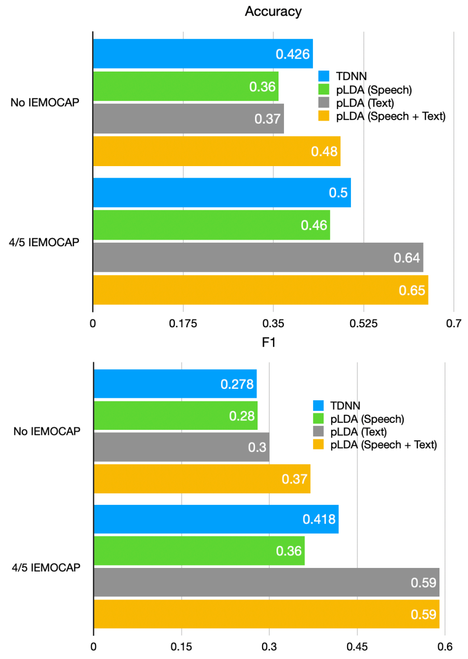
LDA

pLDA

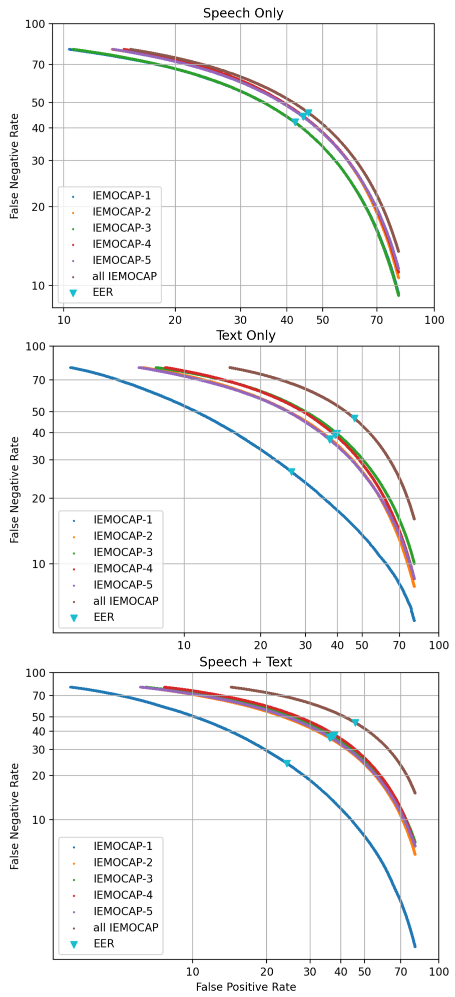
5. predictions

7. Results

Emotion Identification



Emotion Confirmation



- wo IEMOCAP, speaker ID based speech + text pLDA improves upon unimodal approaches, adapts well to unseen emotions
- w IEMOCAP, text alone suffices, perhaps a corpus-specific artifact

8. Future Work

1. explore deeper networks / larger pre-training corpora
2. explore transferring from other tasks like ASR
3. explore joint optimization techniques across modalities