MULTI-MODAL EMOTION DETECTION WITH TRANSFER LEARNING

Amith Ananthram[1] **Kailash Karthik Saravanakumar**[1] **Jessica Huynh**[1] **Homayoon Beigi**[1,2]

[1]Columbia University
New York, NY 10027

[2]Recognition Technologies, Inc.
South Salem, NY 10590

ABSTRACT

Automated emotion detection in speech is a challenging task due to the complex interdependence between words and the manner in which they are spoken. It is made more difficult by the available datasets; their small size and incompatible labeling idiosyncrasies make it hard to build generalizable emotion detection systems. To address these two challenges, 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 and domains. 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: the TDNN alone, speech embeddings from each of its layers alone, text embeddings alone and every combination thereof. Our best variant, trained on only VoxCeleb and Crema-D and evaluated on IEMOCAP, achieves an EER of 38.05%. Including a portion of IEMOCAP during training produces a 5-fold averaged EER of 25.72% (For comparison, 44.71% of the gold-label annotations include at least one annotator who disagrees).

Keywords Emotion Detection, Multimodal Embeddings, Transfer Learning, Domain Adaptation, pLDA

1 Introduction

Due to the growing presence of AI-powered systems in our lives, affective computing has become an important part of human-computer interaction. Emotion plays a role in our thoughts and actions and is an integral part of the way we communicate (Choi et al., 2018). The ability to leverage context to understand emotions communicated both verbally and non-verbally is trivial for humans but remains difficult for machines (Chen et al., 2019). Emotional responses depend on both our psyche and physiology and are governed by our perception of situations, people and objects. They also depend on our mental state (mood, motivation, temperament) (Tripathi et al., 2018a). The way we exhibit and perceive emotion may also differ based on our age, gender, race, culture and accent (Latif et al., 2019). In addition to all of this, unlike targets in other classification tasks, the emotions we experience are rarely distinct: they often coexist without clear temporal boundaries, adding considerable complexity to the task (Tzirakis et al., 2017).

Despite these difficulties, automated emotion recognition has social and commercial applications that make it worth pursuing. In the medical domain, it has exciting potential: to identify and diagnose depression and stress in individuals (Zhu et al., 2017; Rana et al., 2019), to monitor and help people with bipolar disorder (Rana, 2016) and to assist the general public in maintaining mental health. Commercial applications include call center customer management, advertising through neuro-marketing and social media engagement (Tzirakis et al., 2017; Choi et al., 2018; Chen et al., 2019). As intelligent chatbots and virtual assistants have become more widely used, emotion detection has become a vital component in the design, development and deployment of these conversational agents (Yoon et al., 2018).

Early research in emotion detection focused on binary classification in a single modality, whether in text, speech (Chernykh and Prikhodko, 2017; Neumann and Vu, 2017), or images (Dhall et al., 2015). Text-based classifiers used the n-gram vocabulary of sentences to predict their polarity and speech models modeled the vocal dynamics that characterize these emotions. These approaches are inherently limited: a binary granularity and cues from a single modality are far removed from the actual human process they're meant to model. As a result, joint approaches which leverage all available modalities (e.g., both speech and text in applications like home assistants) are promising.

While existing multi-modal emotion corpora like IEMOCAP (Busso et al., 2008) and Crem-D (Cao et al., 2014) have been critical for the progress in affective computing to date, they suffer from three issues that are the focus of our work. First, these corpora tend to be small due to the high costs of annotating for emotion. This precludes the use of deep neural models with high model complexity as they require many training samples to generalize well. This also compounds the second difficulty inherent to many emotion datasets: while there are usually many neutral, happy and sad training examples, there are often very few examples of rarer emotions like disgust making them difficult to classify. This issue is not easily solved by combining different corpora due to the third issue, their lack of mutual compatibility – they differ in the emotions identified, the types of dialogue and number of speakers represented and the naturalness of the recordings (see Figure 1). This severely restricts the generalizability of models trained on a single corpus.

Contemporary literature has dealt with these problems by dropping labels (Pappagari et al. (2020); Chen and Zhao (2020); Yoon et al. (2020)). Hard and scarce emotions like disgust are dropped from the corpus and the models are trained and evaluated on the trimmed corpus. This allows evaluating models on different corpora by using utterances exhibiting only the most common emotions. While this is a reasonable, the resulting performance is not a complete reflection of how these models perform once deployed to production. When emotion models are used in real-world applications, we can expect them to encounter utterances corresponding to dropped labels. For such cases, these models are likely to exhibit degraded performance by predicting one of the known, but incorrect labels.

In this work, we address the problem of data sparsity by transfer learning via the pretrain-then-finetune paradigm. Deep complex models can be trained on large datasets for an auxiliary but related task to learn network parameters that reflect abstract notions related to the target task. As the expression of emotions is highly dependent on the individual, we train a multilayer TDNN (Waibel et al. (1989)) on the task of speaker identification using the VoxCeleb corpus (Chung et al. (2018)) and then fine-tune its final few layers on the task of emotion identification using the Crema-D corpus (Cao et al. (2014)). 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 (Devlin et al. (2018)) and then train an LDA - pLDA (Fisher (1936); Ioffe (2006)) model on the resulting dense representations. pLDA allows our model to more easily adapt to previously unseen classes and domains, a requirement for both evaluating against a different emotion corpus with an incompatible label set and performing well in the wild.

To understand the merits of each component, we exhaustively evaluate the predictive power of every permutation: the TDNN alone, speech embeddings from each of its layers alone, text embeddings alone and every combination thereof. Our best variant, trained on only VoxCeleb and Crema-D and evaluated on IEMOCAP, achieves an Equal Error Rate (EER) of 38.05%. Including a portion of IEMOCAP during training produces a 5-fold averaged EER of 25.72%.

2 Related Work

2.1 Problem Formulation

In this work, we focus on two tasks related to emotions. The first is *emotion identification*. Given the audio and accompanying text for an utterance $\mathbf{u}$, *emotion identification* is the task of identifying the emotion $\mathbf{e}$ expressed in $\mathbf{u}$ from a fixed set of emotions $\mathbf{E}$. This is the standard classification task found in the literature.

The second is *emotion confirmation*. Given the audio and accompanying texts for two utterances $\mathbf{u_1}$ and $\mathbf{u_2}$, *emotion confirmation* is the task of identifying whether the two utterances express the same emotion. This task can be thought of as analogous to either hypothesis testing in speaker recognition or a one-shot classification task. It is motivated by the labeling mismatches among the various emotion corpora and is meant to better reflect the requirement that emotion detection systems be able to adapt to emotions unseen during training once deployed to production.

2.2 Unimodal Speech Emotion Detection

Early work on emotion detection in speech focused on the extraction of hand-crafted features for classification. Liscombe et al. (2003) extracted a set of continuous features based on the fundamental frequency, amplitude and spectral tilt of speech and analyzed its correlation with different emotions. Contemporary literature has focused on deep neural networks with particular successes in transfer leraning. Pappagari et al. (2020) studied the transfer of embeddings from

Table 1: Contemporary emotion detection models, with their modalities, emotion sets and average IEMOCAP (Busso et al., 2008) performance. These models were trained on $4/5$ of IEMOCAP and evaluated on the held out fifth.

Model	Emotion Labels	Metrics	Performance
	Unimodal - Speech		
Pappagari et al. (2020)	happy+exc, angry, sad, neutral	Weighted F1	0.70
Goel and Beigi (2020)	happy, angry, sad, neutral and fear	Accuracy	0.51
Zhou and Beigi (2020)	happy+exc, angry, sad, neutral	Accuracy	0.72
	Bimodal - Speech and Text		
Chen and Zhao (2020)	happy, angry, sad, neutral	Weighted Accuracy Unweighted Accuracy	0.71 0.72
Heusser et al. (2019)	happy, angry, sad, neutral	Weighted Accuracy Unweighted Accuracy	0.70 0.67
Lian et al. (2019)	happy+exc, angry, sad, neutral	Weighted Accuracy	0.83
Tripathi et al. (2019)	happy, angry, sad, neutral	Accuracy Class Accuracy	0.76 0.69
	Trimodal - Speech, Text and Vision		
Yoon et al. (2020)	happy, angry, sad, neutral excited, frustrated, surprised	Weighted Accuracy Unweighted Accuracy	0.62 0.60
Mittal et al. (2019)	happy, angry, sad, neutral	Mean Accuracy F1	0.82 0.82
Tripathi et al. (2018b)	happy+exc, angry, sad, neutral	Accuracy	0.71

Table 2: Mapping of Emotions from Corpus Labels to Canonical Classes

Canonical Emotion	IEMOCAP	Crema-D	DailyDialog
Happiness	Happiness, Excitement	Happiness, Excitement	Happiness
Sadness	Sadness	Sadness	Sadness
Fear/Surprise	Fear, Surprise	Fear	Fear, Surprise
Anger/Disgust	Anger, Disgust, Frustration	Anger, Disgust	Anger, Disgust
Neutral	Neutral	Neutral	Other

a ResNet-based speaker verification model using linear methods. Transfer learning with TDNNs has also been shown to be effective. Zhou and Beigi (2020) fine-tune a pre-trained ASR model and achieve strong results. An alternate approach seen in the literature is to train in a multi-task context on a useful auxiliary task (Goel and Beigi, 2020).

2.3 Multi-modal Emotion Detection

Deep neural techniques have often been applied to integrate information from both the speech and text modalities. Heusser et al. (2019) trained separate speech and text emotion classifiers and then jointly optimized them for multimodal emotion detection. An alternate method for creating multimodal classifiers is to use the embeddings from a hidden layer of the unimodal models for multimodal analysis. While Chen and Zhao (2020) created an ensemble of classifiers using the unimodal embeddings individually and their multimodal concatenations, Tripathi et al. (2019) fed the concatenation to a fully-connected neural layer and backpropagated the classification loss. Fusion using attention is yet another method for combining embeddings from different modalities (Lian et al., 2019). Even in the trimodal setting (speech, text and vision), the primary approaches for integrating modalities are concatenation and attentive fusion (Yoon et al., 2020; Mittal et al., 2019; Tripathi et al., 2018b).

Contemporary results in emotion detection on IEMOCAP are shown in Table 1 with their accompanying modalities, supported emotion sets and performances.

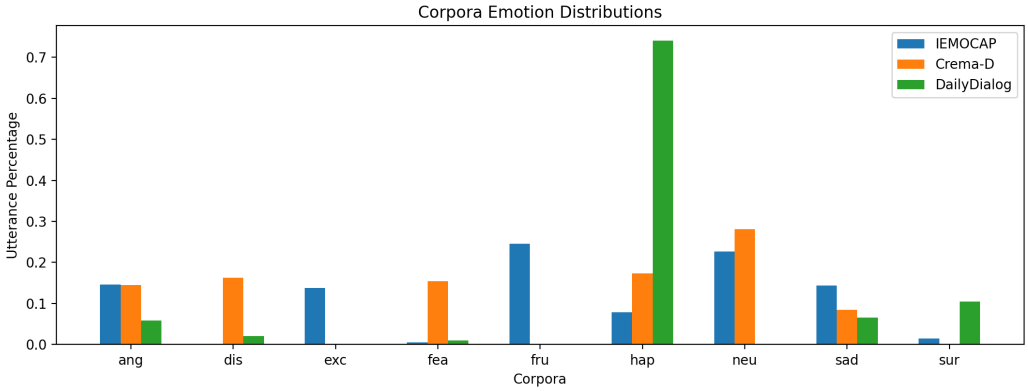

Figure 1: Emotion label distributions in our training and evaluation corpora.

3 Data

3.1 Emotion Classes

As previously discussed, the set of emotions labeled in different corpora is not uniform. A classic model of emotions from psychology comes from Ekman (Scherer and Ekman, 2014) who identifies six fundamental emotions - happiness, sadness, anger, disgust, surprise and fear. Computational models of emotions have attempted classification in this six emotion setting. These approaches have struggled with anger, disgust and fear as they are usually underrepresented in training corpora (Poria et al., 2019). As a result, literature has typically resorted to label dropping, reporting substantially improved performance on the resulting trimmed corpus.

Contemporary research in psychology, informed by facial dynamics, has proposed further clustering into four basic emotions (Jack et al., 2014), arguing against the six-emotion theory. The authors suggest that two pairs of emotions - (*anger, disgust*) and (*fear, surprise*) are psychologically irreducible, a possible explanation for the performance drop seen in systems attempting to disambiguate between these pairs. In light of this research, we adopt their suggested grouping to better balance the classes in our training corpus, Crema-D. When evaluating our fine-tuned neural model on the task of *emotion identification* in IEMOCAP, we rely on this grouping again. While we do the same when evaluating our pLDA classifier on the task of *emotion confirmation*, we also present its performance on IEMOCAP's ungrouped emotions, testing its ability to generalize to previously unseen classes, a key focus of this work. When grouping utterances, we combine the indistinguishable classes into the canonical classes shown in Table 2.

3.2 Pre-Training Corpora for Auxiliary Tasks

3.2.1 VoxCeleb1 & VoxCeleb2, for Speaker Recognition

Our speech model is pre-trained on the auxiliary task of speaker recognition (Beigi (2011)) to learn embeddings that model basic vocal characteristics. VoxCeleb1 and VoxCeleb2 (Chung et al., 2018) are natural speech datasets with audio and video clips of nearly 8,000 celebrities uploaded to YouTube. In total, the two datasets contain more than 1 million utterances. They are fairly well-balanced and the speakers included are reasonably diverse (66% male with many different ethnic groups and nations represented). These recordings are not acted and include background noise.

3.3 Emotion Fine-Tuning and Evaluation Corpora

3.3.1 Crema-D

Crema-D is a multi-modal emotion dataset in the audio and visual modalities (Cao et al. (2014)). Actors perform a limited set of 12 utterances in every target emotion (masking the relationship between their semantic content and their emotional thrust). The emotions labeled are anger, disgust, excitement, fear, happiness, neutral and sadness and include an associated intensity rating (low, medium, high or unspecified). The dataset contains 7,442 utterances from 91 actors who are chosen across different age and ethnic groups. We use Crema-D as our emotion fine-tuning corpus.

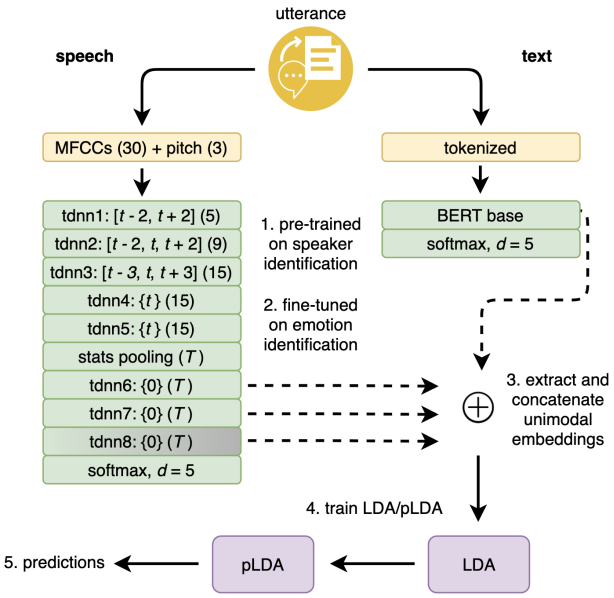

Figure 2: Architecture of our multi-modal emotion detection pipeline

3.4 DailyDialog

While Crema-D is a multi-modal emotion dataset, the lack of diversity in the content of the utterances makes it ineffective to train a text-based emotion recognition model. To address this, we use DailyDialog (Li et al. (2017)), a multi-turn dialog dataset with emotion annotations. The dataset contains more than 13,000 dialogues corresponding to over 100,000 utterances. The emotions labeled are anger, disgust, fear, happiness, sadness, surprise and "other".

3.4.1 IEMOCAP

Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2008) consists of 12 hours of audio and video clips performed by 5 male and 5 female actors. These recordings include improvisations and scripted conversations, both designed to elicit certain emotions. They are then split into individual utterances and annotated for both dimensional attributes and categorical attributes: anger, disgust, excitement, fear, frustration, happiness, neutral, sadness and surprise. Utterances for which human annotator agreement could not be reached are labelled *xxx* and the fraction of the dataset with this label is around 25%. As most emotion recognition systems evaluate against IEMOCAP, we do as well in two contexts: without including any IEMOCAP during training and with 4 out of 5 sessions included during training with the fifth held out for evaluation as part of a five-fold cross validation.

4 Methodology

Our emotion detection pipeline is comprised of the following three stages: (1) extraction of unimodal embeddings; (2) concatenation of unimodal embeddings to create a single multimodal embedding; and (3) inference on the multimodal embedding vector with an LDA/pLDA classifier. This architecture is depicted in Figure 2.

4.1 Unimodal Neural Networks

Our embedding networks follow the pretrain-then-finetune training paradigm.

Our speech embedding model is a time-delay neural network (TDNN, Waibel et al. (1989)). TDNNs are designed to capture long term temporal dependencies efficiently; lower layers aggregate information within narrow contexts while higher layers learn more abstract representations. In Snyder et al. (2018), the authors show that training a pLDA classifier on fixed-length embeddings extracted from the higher layers of a speaker recognition TDNN (which they refer to as "x-vectors") achieves superior performance on out-of-class speaker recognition. Inspired by their result, we choose the same auxiliary task as we hypothesize that such a network learns dense representations of speech segments in its upper layers that contain signals relevant to emotion detection too. We extract hand-crafted speech features and

Table 3: Speaker recognition model architecture, where N is the number of speakers in the training corpus

Layer (Context)	Layer Context	Input x Output
tdnn1	$[t-2, t+2]$ (5)	165 x 512
tdnn2	$\{t-2, t, t+2\}$ (9)	1536 x 512
tdnn3	$\{t-3, t, t+3\}$ (15)	1536 x 512
tdnn4	$\{t\}$ (15)	512 x 512
tdnn5	$\{t\}$ (15)	512 x 1500
stats pooling	$[0, T)$ (T)	$1500T$ x 3000
tdnn6	$\{0\}$ (T)	3000 x 512
tdnn7	$\{0\}$ (T)	512 x 512
softmax	$\{0\}$ (T)	512 x N

pre-train a TDNN on speaker identification (Beigi (2011)). We adopt their architecture (described in Table 3) and their training methodology, building on its accompanying training script published as part of the Kaldi toolkit (Povey et al. (2011)) with a slight modification. We include 3 pitch features, hypothesizing that these contain signal useful for our eventual fine-tuning on emotion identification.

After pre-training on speaker identification, we fine-tune this model on the task of emotion identification using the canonical labelings identified in 3.1. We extract the same set of features and fine-tune several variants, experimenting with learning rates for the first six layers, adding an eighth layer and augmenting our emotion corpus with noise (which has proven to produce more robust embeddings for speaker recognition, Snyder et al. (2018)) .

For embedding the audio transcript text, we fine-tune BERT on the task of emotion identification. The resulting speech and text embeddings are then concatenated to produce a multimodal embedding vector.

4.2 LDA/pLDA

The final stage in our pipeline is an LDA/pLDA (Ioffe, 2006; Beigi, 2011) classifier. Linear discriminant analysis, or LDA, is a dimensionality reduction technique designed to find a set of linear features that maximize the between-class separation of data while minimizing the within-class scatter. Probabilistic linear discriminant analysis, or pLDA, is a generative model that associates probability distributions with those linear features. This makes it adaptable to classes unseen during training, enabling both *emotion identification* and *emotion confirmation*. We train this classifier on our multimodal embeddings using the canonical labelings identified in 3.1.

Experimental details for training each component in this pipeline can be found in Section 5.

5 Experiments

5.1 Experimental Setup

5.1.1 Data Preprocessing

We extract hand-engineered speech features for training our speaker recognition model in the Kaldi toolkit (Povey et al. (2011)): the top 30 MFCCs (Beigi (2011)) and 3 additional pitch features (probability of voicing, mean-subtracted log, and raw pitch deltas) over a frame-length of 25ms. These features are then normalized with sliding-window cepstral mean normalization with a window of 300 frames.

5.1.2 Pre-Training on Speaker Recognition

Using the methodology presented in 4.1, we pre-train our TDNN (Table 3) on the task of speaker identification using the VoxCeleb1 and VoxCeleb2 corpora. We adopt the hyperparameters used by Snyder et al. (2018) in their entirety.

5.1.3 Fine-Tuning Speech on Emotion Detection

We then fine-tune our speaker recognition model on the task of emotion detection using the Crema-D corpus with the canonical label clustering described in Table 3.1. Employing the methodology described in 4.1, we optimize a cross

entropy loss using stochastic gradient descent with momentum. We use an initial learning rate of $1e^{-3}$, a final learning rate of $1e^{-4}$, a batch size of 64, a dropout rate of 50% and we train for 3 epochs. We fine-tune several variants by varying the learning rates on the first six layers, adding an eighth layer and augmenting our emotion corpus with noise.

5.1.4 Fine-Tuning Text on Emotion Detection

We use the uncased version of BERT base (twelve 768 dimensional layers) as our pre-trained text embedding model[1]. We fine-tune this model on emotion identification with the DailyDialog corpus using our canonical emotion classes. The model is trained using cross entropy loss for 4 epochs. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $2e^{-5}$ and epsilon $1e^{-8}$. We employ gradient clipping to bound the norm of gradients to 1.0. Once the BERT model is fine-tuned, we use the embedding of the [CLS][2] token from the final layer as our text embedding.

5.1.5 Training LDA-pLDA

To train our LDA/pLDA classifier, we extract speech embeddings for all utterances in Crema-D and IEMOCAP from layers six and above of our fine-tuned speech models. We concatenate each utterance's corresponding text embedding – for IEMOCAP, this is straightforward. We use the embedding for its accompanying transcript. For Crema-D, as the spoken text is repeated across all emotions, we instead randomly sample text from DailyDialog with the same emotion label and use its embedding instead.

We train an LDA classifier to reduce the dimensionality (Beigi, 2011) of this concatenated representation to 200 and then train a pLDA classifier (Ioffe, 2006) on the resulting dense representations as described in 4.2.

6 Results

To understand the strengths of each component in our pipeline, we present the results from an exhaustive evaluation of their individual and joint predictive powers.

6.1 Emotion Identification

6.1.1 Speech TDNN

We begin with our fine-tuned speech TDNNs whose results can be found in Table 4. We produce six different variants that vary how much learning we allow for the weights in the first six layers from our pre-trained speaker identification model, the number of layers in the fine-tuned model and the inclusion of noise augmented emotion data. We evaluate against all the utterances in IEMOCAP with good inter-annotator agreement using the labeling presented in Table 2.

A few things are immediately clear from these results. While noise augmentation has been shown to improve generalizability in tasks like speaker recognition by encouraging models to learn representations that are more robust, augmenting our emotion corpus with music, noise and babble degraded the performance of our TDNN. We hypothesize that this kind of noise augmentation masks signals relevant to emotion identification. We also find that allowing some fine-tuning of the pre-trained weights in the first six layers generally improved performance but adding an extra randomly initialized layer at the top of the network did not. It's likely that this additional layer requires more data to avoid overfitting than is typically available in an emotion corpus.

A quick comparison of our $4/5$ IEMOCAP results to the unimodal speech baselines presented in Table 1 makes clear that the performance of this fine-tuned TDNN by itself is not competitive with the current state-of-the-art. There are two likely reasons for this: 1) this is partly due to our decision to evaluate against all the utterances in IEMOCAP with the label grouping presented in Table 2 – many of the baselines drop all the rare emotions that are difficult to classify; 2) the baselines transfer from pre-trained models that are much deeper and trained on much more data. Moreover, our TDNN is just the first component in our emotion detection pipeline, meant to generate dense representations useful for subsequent in-class and out-of-class classification by pLDA. As such, we evaluate the predictive power of pLDA classifiers trained on dense representations extracted from all six of our of TDNN variants.

[1] We use the BertForSequenceClassification provided by the HuggingFace Transformers library (Wolf et al., 2020)

[2] [CLS] is a special token introduced by Devlin et al. (2018) that is used to embed an entire text sequence into a single vector.

Table 4: Emotion identification results from the different variants of our fine-tuned speech TDNN. Reported numbers are 1) "No IEMOCAP", the accuracy/F1 on IEMOCAP without training on any IEMOCAP and 2) "4/5 IEMOCAP", the average accuracy/F1 across the five subsets of IEMOCAP while training on the remaining four.

TDNN: *Speech Only*					
Variant	**First Six LR**	**# Layers**	**Noise Aug**	**No IEMOCAP**	**4/5 IEMOCAP**
1	0	7	*no*	0.408/0.269	**0.500/0.418**
2	0	7	*yes*	0.370/0.215	0.488/0.398
3	0.0001	7	*no*	**0.426/0.278**	0.485/0.369
4	0.0001	7	*yes*	0.322/0.186	0.496/0.391
5	0.0001	8	*no*	0.303/0.208	0.484/0.364
6	0.0001	8	*yes*	0.323/0.215	0.482/0.337

6.1.2 Speech LDA/pLDA

In Table 6, we present the emotion identification results of several LDA/pLDA classifiers trained on speech embeddings extracted from every layer of our six TDNN variants. We evaluate against all the utterances in IEMOCAP with the label grouping presented in 2.

Interestingly, the strongest speech-only pLDA results come from embeddings extracted from our second TDNN variant (no learning on the first size layers but with noise augmentation). We hypothesize that, though the noise augmentation degrades the actual predictive performance of our TDNN, it produces robust embeddings in its layers that are more easily separable when classified by our pLDA model.

Including a subset of IEMOCAP in training improves our speech-only pLDA accuracy from 0.36 to 0.46. We note that while this topline performance is slightly degraded from the TDNN, the pLDA classifier is more easily adaptable to emotion classes not present in the training set.

6.1.3 Text LDA/pLDA

In Table 5, we present the emotion identification results of an LDA/pLDA classifier trained on text embeddings extracted from our fine-tuned BERT model. Both tasks are evaluated in the same manner as the speech pLDA classifiers.

When fine-tuned on no IEMOCAP, the performance of our text-only pLDA classifier is fairly similar to our speech-only classifier. But, the inclusion of IEMOCAP during training dramatically boosts performance well past our speech only results (an accuracy of 0.64). As IEMOCAP is a scripted / acted corpus, the faithfulness of its transcripts to the emotions expressed by its utterances is high – as a result, the text of the utterances contains a lot of signal. Our text-based pLDA classifier is able to use this signal to make good predictions while still being easily adaptable to unseen emotion classes.

6.1.4 Multi-modal LDA/pLDA

In Table 6, we present the emotion identification results of LDA/pLDA classifiers trained on multi-modal concatenations of speech embeddings extracted from some of our TDNN variants and text embeddings extracted from our fine-tuned BERT model. Both tasks are evaluated in the same manner as the speech pLDA classifiers.

We first note than in both cases, without IEMOCAP and with IEMOCAP, the performance of our multi-modal pLDA classifier is improved. This suggests that the joint representation contains some signal that wasn't present in the unimodal representations (i.e., the unimodal representations are not redundant). While it is the case that the improvement we get with the multi-modal representation when trained on IEMOCAP is not large, the improvement we get without training on IEMOCAP (that is, evaluated against a corpus unseen during training) is meaningful.

6.2 Emotion Confirmation

Finally, we present *emotion confirmation* results on the full IEMOCAP corpus where each pairwise test case uses the original labels in IEMOCAP (i.e., without any re-grouping) to determine whether or not two utterances are in the same class. This emotion hypothesis testing provides a way to adapt our model to emotion classes unseen during testing, provided we have a single domain-specific example of the emotion of interest.

Table 5: Emotion identification results from LDA/pLDA classifiers trained on only text embeddings extracted from our fine-tuned BERT model. Reported numbers are 1) "No IEMOCAP", the accuracy/F1/EER on IEMOCAP without training on any IEMOCAP and 2) "4/5 IEMOCAP", the average accuracy/F1/EER across the five subsets of IEMOCAP while training on the remaining four.

pLDA: *Text Only*	
No IEMOCAP	4/5 **IEMOCAP**
0.37/0.30/0.42	0.64/0.59/0.29

Table 6: Emotion identification results from the different variants of our LDA/pLDA model trained on only speech embeddings extracted from different layers of our fine-tuned TDNN variants or on multi-modal concatenations of those embeddings and text embeddings extracted from our fine-tuned BERT model. Reported numbers are 1) "No IEMOCAP", the accuracy/F1/EER on IEMOCAP without training on any IEMOCAP and 2) "4/5 IEMOCAP", the average accuracy/F1/EER across the five subsets of IEMOCAP while training on the remaining four.

	pLDA: *Speech Only*		pLDA: *Speech + Text*	
Speech Embeddings	**No IEMOCAP**	4/5 **IEMOCAP**	**No IEMOCAP**	4/5 **IEMOCAP**
variant 2, layer 6	**0.36/0.28/0.41**	0.45/0.36/0.39	0.34/0.26/0.41	-
variant 2, layer 7	0.36/0.23/0.42	**0.46/0.36/0.39**	**0.48/0.37/0.38**	**0.65/0.59/0.25**
variant 3, layer 6	0.31/0.28/0.44	0.45/0.35/0.41	-	-
variant 3, layer 7	0.33/0.28/0.43	0.42/0.33/0.41	-	-
variant 6, layer 6	0.34/0.29/0.43	0.44/0.35/0.40	0.34/0.25/0.41	-
variant 6, layer 7	0.35/0.30/0.43	0.46/0.36/0.39	0.41/0.34/0.38	0.65/0.58/0.28
variant 6, layer 8	0.32/0.28/0.45	0.44/0.35/0.40	0.32/0.21/0.42	-

As is clear from the results in Table 7, without fine-tuning on any IEMOCAP, the best performing variant of our pLDA classifier is one that is trained on the concatenatation of speech and text embeddings. Including 4/5 IEMOCAP reduces this error rate from 0.457 to 0.342. While these rates are admittedly high, the test scenario is a difficult one – we include utterances whose labels were entirely unseen during training.

Table 7: Emotion confirmation results from LDA/pLDA classifiers trained on on only speech embeddings extracted from different layers of our fine-tuned TDNN variants, only text embeddings extracted from our fine-tuned BERT model or on multi-modal concatenations of both. Reported numbers are 1) "No IEMOCAP", the EER on IEMOCAP without training on any IEMOCAP and 2) "4/5 IEMOCAP", the average EER across the five subsets of IEMOCAP while training on the remaining four.

Modality	**No IEMOCAP**	4/5 **IEMOCAP**
Speech Only	0.457	0.433
Text Only	0.466	0.360
Speech + Text	0.457	0.342

7 Conclusion

In this work, we present a multi-modal approach to emotion detection 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 and domains. We show that:

1. when fine-tuning on no IEMOCAP, our multi-modal pLDA classifier performs reasonably well when evaluated on the entirety of IEMOCAP (without dropping any utterances)
2. this pLDA classifier is also able to adapt to emotions unseen during training in a one-shot classification context