M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues

Trisha Mittal,[1] **Uttaran Bhattacharya,**[1] **Rohan Chandra,**[1] **Aniket Bera,**[1] **Dinesh Manocha**[1]
[1]Department of Computer Science, University of Maryland, College Park, USA
{trisha, uttaranb, rohan, ab, dm}@cs.umd.edu
Project URL: https://gamma.umd.edu/m3er

Abstract

We present M3ER, a learning-based method for emotion recognition from multiple input modalities. Our approach combines cues from multiple co-occurring modalities (such as face, text, and speech) and also is more robust than other methods to sensor noise in any of the individual modalities. M3ER models a novel, data-driven multiplicative fusion method to combine the modalities, which learn to emphasize the more reliable cues and suppress others on a per-sample basis. By introducing a check step which uses Canonical Correlational Analysis to differentiate between ineffective and effective modalities, M3ER is robust to sensor noise. M3ER also generates proxy features in place of the ineffectual modalities. We demonstrate the efficiency of our network through experimentation on two benchmark datasets, IEMOCAP and CMU-MOSEI. We report a mean accuracy of 82.7% on IEMOCAP and 89.0% on CMU-MOSEI, which, collectively, is an improvement of about 5% over prior work.

1 Introduction

The perception of human emotions plays a vital role in our everyday lives. People modify their responses and behaviors based on their perception of the emotions of those around them. For example, one might cautiously approach a person they perceive to be angry, whereas they might be more forthcoming when approaching a person they perceive to be happy and calm. Given the importance of emotion perception, emotion recognition from sensor data is important for various applications, including human-computer interaction (Cowie et al. 2001), surveillance (Clavel et al. 2008), robotics, games and entertainment, and more. In this work, we address the problem of perceived emotion recognition rather than recognition of the actual emotional state.

One of the primary tasks in developing efficient AI systems for perceiving emotions is to combine and collate information from the various modalities by which humans express emotion. These modalities include, but are not limited to, facial expressions, speech and voice modulations, written text, body postures, gestures, and walking styles. Many researchers have advocated combining more than one modality to infer perceived emotion for various reasons, including:

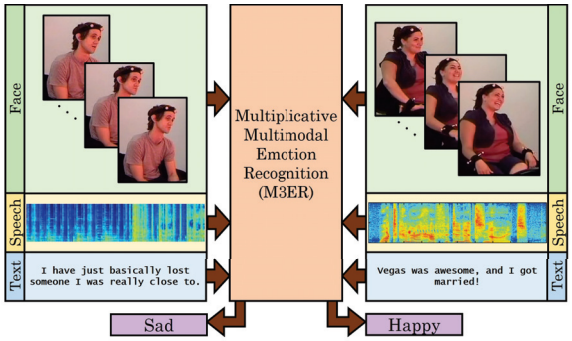

Figure 1: Multimodal Perceived Emotion Recognition: We use multiple modalities to perform perceived emotion prediction. Our approach uses a deep learning model along with a multiplicative fusion method for emotion recognition. We show results on two datasets, IEMOCAP and CMU-MOSEI both of which have face, speech and text as the three input modalities. Above is one sample point extracted from the IEMOCAP dataset.

(a) *Richer information:* Cues from different modalities can augment or complement each other, and hence lead to more sophisticated inference algorithms.
(b) *Robustness to Sensor Noise:* Information on different modalities captured through sensors can often be corrupted due to signal noise, or be missing altogether when the particular modality is not expressed, or cannot be captured due to occlusion, sensor artifacts, etc. We call such modalities *ineffectual*. Ineffectual modalities are especially prevalent in in-the-wild datasets.

However, multimodal emotion recognition comes with its own challenges. At the outset, it is important to decide which modalities should be combined and how. Some modalities are more likely to co-occur than others, and therefore are easier to collect and utilize together. For example, some of the most popular benchmark datasets on multiple modalities, such as IEMOCAP (Busso et al. 2008) and CMU-MOSEI (Zadeh et al. 2018b), contain commonly co-occurring modalities of facial expressions with associated speech and transcribed text. With the growing number of

social media sites and data on internet (e.g., YouTube), often equipped with automatic caption generation, it is easier to get data for these three modalities. Many of the other existing multimodal datasets (Ringeval et al. 2013; Kossaifi et al. 2017) are also a subset of these three modalities. Consequently, these are the modalities we have used in our work.

Another challenge is the current lack of agreement on the most efficient mechanism for combining (also called "fusing") multiple modalities (Baltrusaitis, Ahuja, and Morency 2017). The most commonly used techniques are early fusion (also "feature-level" fusion) and late fusion (also "decision-level" fusion). Early fusion combines the input modalities into a single feature vector on which a prediction is made. In late fusion methods, each of the input modalities is used to make an individual prediction, which is then combined for the final classification. Most prior works on emotion recognition works have explored early fusion (Sikka et al. 2013) and late fusion (Gunes and Piccardi 2007) techniques in additive combinations. Additive combinations assume that every modality is always potentially useful and hence should be used in the joint representation. This assumption makes the additive combination not ideal for in-the-wild datasets which are prone to sensor noise. Hence, in our work, we use multiplicative combination, which does not make such an assumption. Multiplicative methods explicitly model the relative reliability of each modality on a per-sample basis, such that reliable modalities are given higher weight in the joint prediction.

Main Contributions: We make the following contributions:

1. We present a multimodal emotion recognition algorithm called M3ER, which uses a data-driven multiplicative fusion technique with deep neural networks. Our input consists of the feature vectors for three modalities — face, speech, and text.
2. To make M3ER robust to noise, we propose a novel preprocessing step where we use Canonical Correlational Analysis (CCA) (Hotelling 1936) to differentiate between an ineffectual and effectual input modality signal.
3. We also present a feature transformation method to generate proxy feature vectors for ineffectual modalities given the true feature vectors for the effective modalities. This enables our network to work even when some modalities are corrupted or missing.

We compare our work with prior methods by testing our performance on two benchmark datasets IEMOCAP and CMU-MOSEI. We report an accuracy of 82.7% on the IEMOCAP dataset and 89.0% on the CMU-MOSEI dataset, which is a collective 5% accuracy improvement on the absolute over prior methods. We show ablation experiment results on both datasets, where almost 75% of the data has at least one modality corrupted or missing, to demonstrate the importance of our contributions. As per the annotations in the datasets, we classify IEMOCAP into 4 discrete emotions (angry, happy, neutral, sad) and CMU-MOSEI into 6 discrete emotions (anger, disgust, fear, happy, sad, surprise). According to the continuous space representations, emotions are seen as points on a 3D space of arousal, valence, and dominance (Ekman and Friesen 1967). The discrete emotions are related to the continuous space through an eigen-transform; therefore we can switch between the representations without adding any noise.

2 Related Work

In this section, we give a brief overview of previous works on unimodal and multimodal emotion recognition, as well as modality combination techniques that have been used in the broader field of multimodal machine learning.

Emotion Recognition in Psychology Research: Understanding and interpreting human emotion is of great interest in psychology. The initial attempts (Russell, Bachorowski, and Fernández-Dols 2003) at predicting emotion only from facial expressions were not considered very reflective of the human sensory system and were questioned. There is also unreliability in using facial expressions, because of the ease of displaying "mocking" expressions (Ekman 1993), especially in the presence of an audience (Fernández-Dols and Ruiz-Belda 1995). Psychology research also points to the importance of considering cues other than facial expressions to make more accurate predictions. Sebe et al. (2011), Aviezer et al. (2012) and Pantic et al. (2005) highlight the fact that an ideal system for automatic human emotion recognition should be multimodal, because this is more close to the human sensory system. Meeran et al. (2005) suggest that the integration of modalities is an inevitable step learned very early-on in the human sensory system.

Unimodal Emotion Recognition: The initial attempts in human emotion recognition have been mostly unimodal. Even in that domain, the most predominantly explored modality has been facial expressions (Saragih, Lucey, and Cohn 2009; Akputu, Seng, and Lee 2013), owing to the availability of face datasets and advances in computer vision methods. Other modalities that have been explored include speech or voice expressions (Scherer, Johnstone, and Klasmeyer 2003), body gestures (Navarretta 2012), and physiological signals such as respiratory and heart signals (Knapp, Kim, and André 2011).

Multimodal Emotion Recognition: Multimodal emotion recognition was initially explored using classifiers like Support Vector Machines, and linear and logistic regression (Sikka et al. 2013; Gunes and Piccardi 2007; Castellano, Kessous, and Caridakis 2008), when the size of the datasets was less than 500. As bigger datasets were developed, deep learning architectures (Yoon et al. 2019; Kim, Lee, and Provost 2013; Majumder et al. 2018; Zadeh et al. 2018c; Lee et al. 2018; Sahay et al. 2018) were explored. All multimodal methods also perform feature extraction steps on each of the input modalities, using either hand-crafted formulations or deep learning architectures. Some of the architectures that have been explored are Bi-Directional Long Short Term Memory (BLSTM) networks (Yoon et al. 2019), Deep Belief

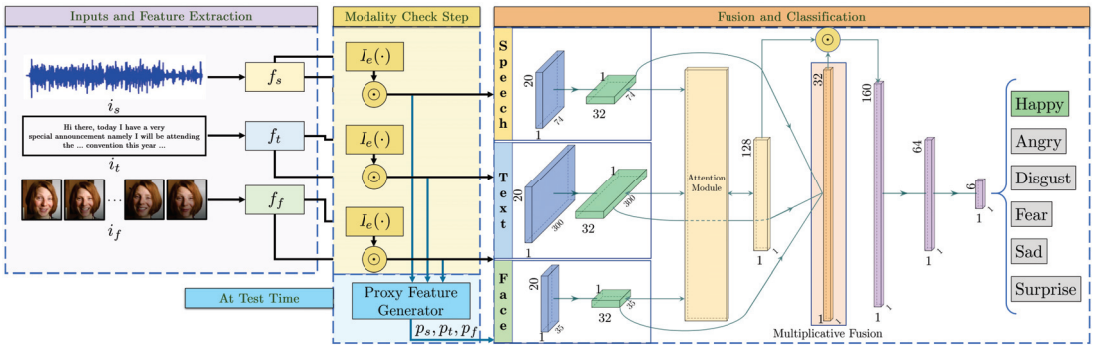

Figure 2: M3ER: We use three modalities, speech, text and the facial features. We first extract features to obtain f_s, f_t, f_f from the raw inputs, i_s, i_t and i_f (purple box). The feature vectors then are checked if they are effective. We use a indicator function I_e (Equation 1) to process the feature vectors (yellow box). These vectors are then passed into the classification and fusion network of M3ER to get a prediction of the emotion (orange box). At the inference time, if we encounter a noisy modality, we regenerate a proxy feature vector (p_s, p_t or p_f) for that particular modality (blue box).

Networks (DBNs) (Kim, Lee, and Provost 2013), and Convolutional Neural Networks (Lee et al. 2018). Other methods are based on hierarchical networks (Majumder et al. 2018) and Relational Tensor Networks (Sahay et al. 2018).

Modality Combination: Prior works in emotion recognition (Sikka et al. 2013; Gunes and Piccardi 2007; Castellano, Kessous, and Caridakis 2008; Yoon et al. 2019; Kim, Lee, and Provost 2013) using either late or early fusion have relied on additive combinations. The performance of these additive approaches relies on figuring out the relative emphasis to be placed on different modalities. However, in the real-world, not every modality is equally reliable for every data point due to sensor noise, occlusions, etc. Recent works have also looked at variations on more sophisticated data-driven (Lee et al. 2018), hierarchical (Majumder et al. 2018), and attention-mechanism based (Yoon et al. 2019; Lee et al. 2018) fusion techniques. Multiplicative combination methods (Liu et al. 2018) explicitly models the relative reliability of each modality, such that more reliable modalities are given more weight in the joint prediction. Reliable modalities can also change from sample to sample, so it is also important to learn which modalities are more reliable on a per sample basis. This method has previously been shown to be successful on tasks like user profiling and physical process recognition (Liu et al. 2018).

Canonical Correlational Analysis (CCA): The objective of CCA (Hotelling 1936) is to project the input vectors into a common space by maximizing their component-wise correlation. There have been extensions to CCA, namely Deep CCA (Andrew et al. 2013), Generalized CCA (Kettenring 1971), and Kernel CCA (Welling 2005), which learn parametric non-linear transformations of two random vectors, such that their correlation is maximized. CCA approaches have also been explored for the task of multimodal emotion recognition (Shan, Gong, and McOwan 2007), to get maximally correlated feature vectors from each input modality before combining them. In our work, we use CCA to check for correlation among the input modalities and to check for effective and ineffective modalities.

3 M3ER: Our Approach

3.1 Notation

We denote the set of modalities as $\mathcal{M} = \{\text{face}, \text{text}, \text{speech}\}$. The feature vectors for each modality are denoted as f_f, f_t, and f_s, respectively. We denote the set of predicted emotions as $\mathcal{E} = \{\text{happy}, \text{sad}, \text{angry}, \text{neutral}\}$. The proxy feature vectors generated for speech, text, and face vectors are represented by p_s, p_t, p_f, respectively. Finally, we define an indicator function, $I_e(f)$ that outputs either a vector of zero or one of the same dimension as f, depending on the conditions of the function definition.

3.2 Overview

We present an overview of our multimodal perceived emotion recognition model in Figure 2. During training, we first extract feature vectors (f_s, f_t, f_f) from raw inputs (i_s, i_t, i_f) (purple box in the Figure 2). These are then passed through the modality check step (yellow box in the Figure 2) to distinguish between effective and ineffectual signals, and discarding the latter if any (See Section 3.3). The feature vectors as returned by the modality check step go through three deep-layered feed-forward neural network channels (orange box in Figure 2). Finally, we add our multiplicative fusion layer to combine the three modalities. At test time, the data point once again goes through the modality check step. If a modality is deemed ineffectual, we regenerate a proxy feature vector (blue box in Figure 2) which is passed to the network for the emotion classification. In the following subsections, we explain each of the three novel components of our network in detail.

3.3 Modality Check Step

To enable perceived emotion recognition in real world scenarios, where sensor noise is inevitable, we introduce the

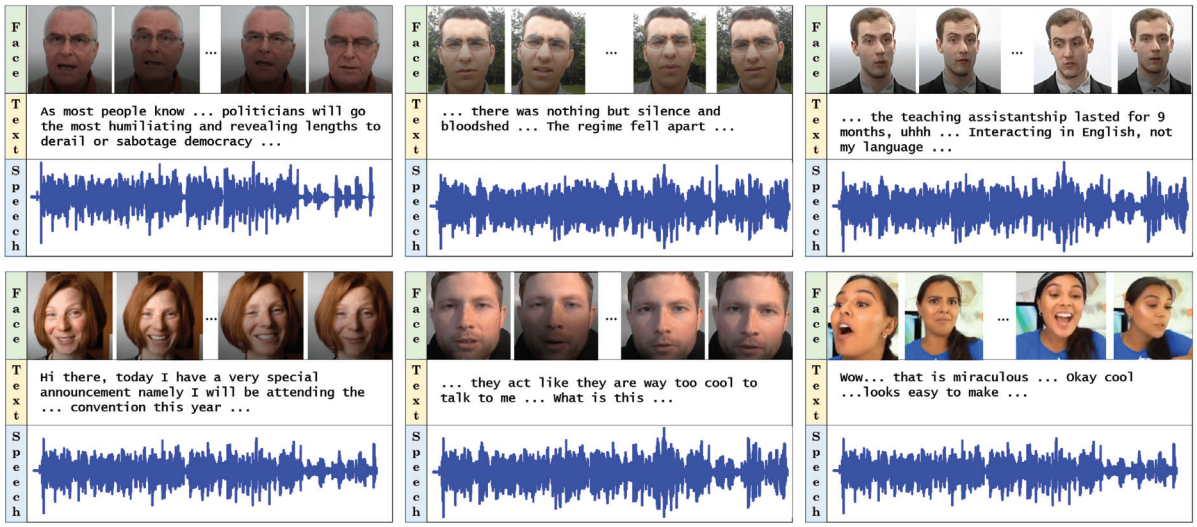

Figure 3: Qualitative Results on CMU-MOSEI: We qualitatively show data points correctly classified by M3ER from all the 6 class labels of CMU-MOSEI. The labels as classified by M3ER in row order from top left, are *Anger, Disgust, Fear, Happy, Sad, Surprise*.

Figure 4: Qualitative Results on IEMOCAP: We qualitatively show data points correctly classified by M3ER from all the 4 class labels of IEMOCAP. The labels as classified by M3ER in row order from top left, are *Angry, Happy, Neutral, Sad*.

Modality Check step which filters ineffectual data. It has been observed in emotion prediction studies (Shan, Gong, and McOwan 2007), that for participants whose emotions were predicted correctly, each of their corresponding modality signals correlated with at least one other modality signal. We directly exploit this notion of correlation to distinguish between features that could be effective for emotion classification (effective features) and features that are noisy (ineffectual features).

More concretely, we use Canonical Correlation Analysis (CCA) to compute the correlation score, ρ, of every pair of input modalities. Given a pair of feature vectors, f_i, f_j, with $i, j \in \mathcal{M}$, we first compute the the projective transformations, $H^i_{i,j}$ and $H^j_{i,j}$, for both feature vectors, respectively. Also note that these feature vectors f_i, f_j are reduced to the same lower dimensions (100, here). We obtain the projected vector by applying the projective transformation. Thus, in our example above,

$$f_i^{'} = H^i_{i,j} f_i,$$

and,

$$f_j^{'} = H^j_{i,j} f_j,$$

Finally, we can compute the correlation score for the pair $\{f_i, f_j\}$ using the formula:

$$\rho(f_i^{'}, f_j^{'}) = \frac{\text{cov}(f_i^{'}, f_j^{'})}{\sigma_{f_i^{'}} \sigma_{f_j^{'}}}$$

and check them against an empirically chosen threshold (τ). $\forall i \in m$, we check

$$\rho(f_i^{'}, f_j^{'}) < \tau,$$

where $\forall\,(i,j) \in \mathcal{M}, i \neq j$.

For implementation purposes, we keep the $H^j_{i,j}$ for all pairs of modalities precomputed based on the training set. At inference time, we simply compute the projected vectors $f_i^{'}, f_j^{'}$ and $\rho(f_i^{'}, f_j^{'})$.

We compare the correlation against a heuristically chosen threshold, τ and introduce the following indicator function,

$$I_e(f_i) = \begin{cases} \mathbb{0} & \rho(f_i, f_j) < \tau, (i,j) \in \mathcal{M}, i \neq j, \\ \mathbb{1} & else. \end{cases} \quad (1)$$

For all features, we apply the following operation, $I_e(f) \odot f$, which discards ineffectual features and retains the effective ones. Here, $\odot$ denotes element-wise multiplication.

3.4 Regenerating Proxy Feature Vectors

When one or more modalities have been deemed ineffectual at test time in the modality check step, we generate proxy feature vectors for the ineffectual modalities using the following equation, $p_i = \mathcal{T} f_i$, where $i \in \mathcal{M}$ and $\mathcal{T}$ is any linear transformation. We illustrate the details below.

Generating exact feature vectors for missing modalities is challenging due to the non-linear relationship between the modalities. However, we empirically show that by relaxing the non-linear constraint, there exists a linear algorithm that approximates the feature vectors for the missing modalities with high classification accuracy. We call these resulting vectors: proxy feature vectors.

Suppose that during test time, the feature vector for the speech modality is corrupt and identified as ineffectual, while f_f is identified as effective during the Modality Check Step. Our aim is then to regenerate a proxy feature vector, p_s, for the speech modality. More formally, we are given, say, a new, unseen face modality feature vector, f_f, the set of observed face modality vectors, $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$, and the set of corresponding observed speech modality vectors, $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$. Our goal is to generate a proxy speech vector, p_s, corresponding to f_f. We begin by preprocessing the inputs to construct bases, $\mathcal{F}_b = \{v_1, v_2, \ldots, v_p\}$ and $\mathcal{S}_b = \{w_1, w_2, \ldots, w_q\}$ from the column spaces of $\mathcal{F}$ and $\mathcal{S}$. Under the relaxed constraint, we assume there exists a linear transformation, $\mathcal{T} : \mathcal{F}_b \rightarrow \mathcal{S}_b$. Our algorithm proceeds without assuming knowledge of $\mathcal{T}$:

1. The first step is to find $v_j = \text{argmin}_j d(v_j, f_f)$, where d is any distance metric. We chose the L_2 norm in our experiment s. We can solve this optimization problem using any distance metric minimization algorithm such as the K-nearest neighbors algorithm.
2. Compute constants $a_i \in \mathbb{R}$ by solving the following linear system, $f_f = \sum_{i=1}^{p} a_i v_i$. Then,

$$p_s = \mathcal{T} f_f = \sum_{i=1}^{p} a_i \mathcal{T} v_i = \sum_{i=1}^{p} a_i w_i.$$

Our algorithm can be extended to generate proxy vectors from effective feature vectors corresponding to multiple modalities. In this case, we would apply the steps above to each of the effective feature vectors and take the mean of both the resulting proxy vectors.

3.5 Multiplicative Modality Fusion

The key idea in the original work (Liu et al. 2018) for multiplicative combination is to explicitly suppress the weaker (not so expressive) modalities, which indirectly boost the stronger (expressive) modalities. They define the loss for the i^{th} modality as follows.

$$c^{(y)} = -\sum_{i=1}^{M} \prod_{j \neq i} \left(1 - p_j^{(y)}\right)^{\beta/(M-1)} \log p_i^{(y)} \qquad (2)$$

where y is the true class label, M is the number of modalities, β is the hyperparameter that down-weights the unreliable modalities and $p_i^{(y)}$ is the prediction for class y given by the network for the i^{th} modality. This indirectly boosts the stronger modalities. In our approach, we reverse this concept and propose a modified loss. We explicitly boost the stronger modalities in the combination network. The difference is subtle but has key significance on the results. In the original formulation, the modified loss was given by Equation 2. We empirically show that the modified loss gives better classification accuracies than the originally proposed loss function in Section 5. The original loss function tries to ignore or tolerate the mistakes of the modalities making wrong predictions by explicitly suppressing them, whereas in our modified version, we ignore the wrong predictions by simply not addressing them and rather focusing on modalities giving the right prediction. In the original loss, calculating the loss for each modality depends on the probability given by all the other modalities. This has a higher computation cost due to the product term. Furthermore, if either of the input modalities produces an outlier prediction due to noise in the signal, it affects the prediction of all other modalities. Our proposed modified loss is as follows:

$$c^{(y)} = -\sum_{i=1}^{M} \left(p_i^{(y)}\right)^{\beta/(M-1)} \log p_i^{(y)} \qquad (3)$$

This fusion layer is applied to combine the three input modalities.

M3ER is a modular algorithm that can work on top of existing networks for multimodal classification. Given a network for multiple modalities, we can replace the fusion step and incorporate the modality check and proxy vector regeneration of the M3ER and improve classification accuracies. In the next Section, we demonstrate this point by incorporating M3ER in SOTA networks for two datasets, IEMOCAP and CMU-MOSEI.

4 Implementation Details

We state the implementation and training details for training with M3ER on the CMU-MOSEI dataset in this section. Details on the network, implementation, and training on the IEMOCAP dataset can be found here [1]).

4.1 Feature Extraction

To extract f_t from the CMU-MOSEI dataset, we use the 300-dimensional pre-trained GloVe word embeddings (Pennington, Socher, and Manning 2014). To compute f_s from the CMU-MOSEI dataset, we follow the approach of Zadeh et al. (2018c) and obtain the 12 Mel-frequency cepstral coefficients, pitch, voiced/unvoiced segmenting features, glottal source parameters among others. Lastly, to obtain f_f, we use the combination of face embeddings obtained from state-of-the-art facial recognition models, facial action units, and facial landmarks for CMU-MOSEI.

4.2 Classification Network Architecture

For training on the CU-MOSEI dataset, we integrate our multiplicative fusion layer into Zadeh et al.'s (2018a) memory fusion network (MFN). Each of the input modalities

[1] https://github.com/TrishaMittal/M3ER

Dataset	Method	F1	MA
IEMOCAP	Kim et al. (2013)	-	72.8%
	Majumdar et al. (2018)	-	76.5%
	Yoon et al. (2019)	-	77.6%
	M3ER	**0.824**	**82.7%**
CMU-MOSEI	Sahay et al. (2018)	0.668	-
	Zadeh et al. (2018c)	0.763	-
	Choi et al. (2018)	0.895	-
	M3ER	**0.902**	**89.0%**

Table 1: M3ER for Emotion Recognition: We compare the F1 scores and the mean classification accuracies (MA) of M3ER on the two datasets, IEMOCAP and CMU-MOSEI, with three prior SOTA methods. Numbers not reported by prior methods are marked with '-'. We observe around 5-10% increase in MA and 1-23% increase in F1 score.

is first passed through single-hidden-layer LSTMs, each of output dimension 32. The outputs of the LSTMs, along with a 128-dimensional memory variable initialized to all zeros (yellow box in the network Figure 2), are then passed into an *attention module* as described by the authors of MFN. The operations inside the attention module are repeated for a fixed number of iterations t, determined by the maximum sequence length among the input modalities ($t = 20$ in our case). The outputs at the end of every iteration in the attention module are used to update the memory variable as well as the inputs to the LSTMs. After the end of t iterations, the outputs of the 3 LSTMs are combined using multiplicative fusion to a 32 dimensional feature vector. This feature vector is concatenated with the final value of the memory variable, and the resultant 160 dimensional feature vector is passed through a 64 dimensional fully connected layer followed by a 6 dimensional fully connected to generate the network outputs.

4.3 Training Details

For training with M3ER on the CMU-MOSEI dataset, we split the CMU-MOSEI dataset into training (70%), validation (10%), and testing (20%) sets. We use a batch size of 256 and train it for 500 epochs. We use the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.01. All our results were generated on an NVIDIA GeForce GTX 1080 Ti GPU.

5 Experiments and Results

We perform experiments on the two large-scale benchmark datasets, IEMOCAP and CMU-MOSEI, described in Section 5.1. In Section 5.2, we list the SOTA algorithms with which we compare M3ER using standard classification evaluation metrics. We report our findings and analysis in Section 5.3. We perform exhaustive ablation experiments to motivate the benefits of our contributions in Section 5.4. Finally, we provide details of all hyperparameters and the hardware used for training M3ER in Section 4.3.

5.1 Datasets

The *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) dataset (Busso et al. 2008) consists of text, speech, and face modalities of 10 actors recorded in the form of conversations using a Motion Capture camera. The conversations include both scripted and spontaneous sessions. The labeled annotations consists of four emotions — angry, happy, neutral, and sad. The *CMU Multimodal Opinion Sentiment and Emotion Intensity* (CMU-MOSEI) (Zadeh et al. 2018b) contains 23,453 annotated video segments from 1,000 distinct speakers and 250 topics acquired from social media channels. The labels in this dataset comprise six emotions — angry, disgust, fear, happy, sad and surprise.

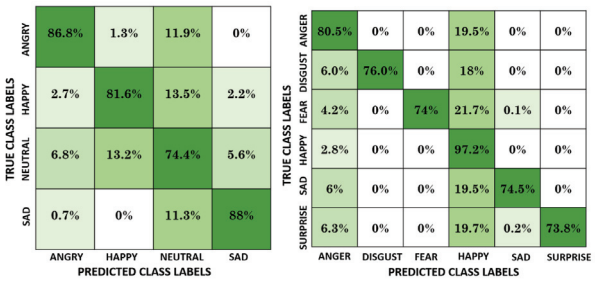

Figure 5: Confusion Matrix: For each emotion class, we show the percentage of inputs belonging to that class that were correctly classified by M3ER (dark green cells) and the percentage of inputs that were misclassified into other classes (pale green and white cells) for both the datasets. *Left:* Confusion matrix for classification on IEMOCAP dataset. *Right:* Confusion matrix for classification on CMU-MOSEI dataset.

5.2 Evaluation Metrics and Methods

We use two standard metrics, F1 scores and mean classification accuracies (MAs), to evaluate all the methods. However, some prior methods have not reported MA, while others have not reported F1 scores. We, therefore, leave out the corresponding numbers in our evaluation as well and compare the methods with only the available numbers. For the IEMOCAP dataset, we compare our accuracies with the following SOTA methods.

1. **Yoon et al. (2019)** use only two modalities of the IEMOCAP dataset, text and speech, using an attention mechanism that learns to aligns the relevant text with the audio signal instead of explicitly combining outputs from the two modalities separately. The framework uses two Bilinear LSTM networks.
2. **Kim et al. (2013)** focus on feature selection parts and hence use DBNs which they claim are better equipped at learning high-order non-linear relationships. They empirically show that non-linear relationships help in emotion recognition.
3. **Majumdar et al. (2018)** recognize the need of a more explainable and intuitive method for fusing different modalities. They propose a hierarchical fusion that learns bimodal and trimodal correlations for data fusion using deep neural networks.

For the CMU-MOSEI dataset, we compare our F1 scores with the following SOTA methods.

(a) Ablation Experiments performed on IEMOCAP Dataset.

Ineffectual modalities?	Experiments	Angry		Happy		Neutral		Sad		Overall	
		F1	MA	F1	MA	F1	MA	F1	MA	F1	MA
No	Original Multiplicative Fusion (Liu et al. 2018)	0.794	80.6%	0.750	76.9%	0.695	68.0%	0.762	80.8%	0.751	76.6%
	M3ER	**0.862**	**86.8%**	**0.862**	**81.6%**	**0.745**	**74.4%**	**0.828**	**88.1%**	**0.824**	**82.7%**
Yes	M3ER – Modality Check Step – Proxy Feature Vector	0.704	71.6%	0.712	70.4%	0.673	64.7%	0.736	79.8%	0.706	71.6%
	M3ER – Proxy Feature Vector	0.742	75.7%	0.745	73.7%	0.697	66.9%	0.778	84.0%	0.741	75.1%
	M3ER	**0.799**	**82.2%**	**0.743**	**76.7%**	**0.727**	**67.5%**	**0.775**	**86.3%**	**0.761**	**78.2%**

(b) Ablation Experiments performed on CMU-MOSEI Dataset.

Ineffectual modalities?	Experiments	Angry		Disgust		Fear		Happy		Sad		Surprise		Overall	
		F1	MA	F1	MA	F1	MA	F1	MA	F1	MA	F1	MA	F1	MA
No	Original Multiplicative Fusion (Liu et al. 2018)	0.889	79.9%	0.945	89.6%	0.963	93.1%	0.587	55.8%	0.926	85.3%	0.949	90.0%	0.878	82.3%
	M3ER	**0.919**	**86.3%**	**0.927**	**92.1%**	**0.904**	**88.9%**	**0.836**	**82.1%**	**0.899**	**89.8%**	**0.952**	**95.0%**	**0.902**	**89.0%**
Yes	M3ER – Modality Check Step – Proxy Feature Vector	0.788	73.3%	0.794	80.0%	0.843	85.0%	0.546	55.7%	0.832	79.5%	0.795	80.1%	0.764	75.6%
	M3ER – Proxy Feature Vector	0.785	77.8%	0.799	83.2%	0.734	77.5%	0.740	77.1%	0.840	86.0%	0.781	83.5%	0.783	80.9%
	M3ER	**0.816**	**81.3%**	**0.844**	**86.8%**	**0.918**	**89.4%**	**0.780**	**75.7%**	**0.873**	**86.1%**	**0.932**	**91.3%**	**0.856**	**85.0%**

Table 2: Ablation Experiments: We remove one component of M3ER at a time, and report the F1 and MA scores on the IEMOCAP and the CMU-MOSEI datasets, to showcase the effect of each of these components. Modifying the loss function leads to an increase of 6-7% in both F1 and MA. Adding the modality check step on datasets with ineffectual modalities leads to an increase of 2-5% in F1 and 4-5% in MA, and adding the proxy feature regeneration step on the same datasets leads to a further increase of 2-7% in F1 and 5-7% in MA.

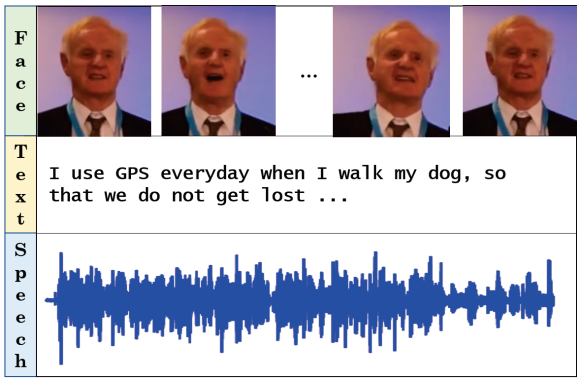

Figure 6: Misclassification by M3ER: This is the text and face input of a 'happy' data point from CMU-MOSEI dataset that our model, M3ER misclassifies as 'angry'. Here, the man is giving a funny speech with animated and exaggerated facial looks which appear informative but lead us to a wrong class label.

1. **Zadeh et al. (2018c)** propose a Dynamic Fusion Graph (DFG) for fusing the modalities. The DFG can model n-modal interactions with an efficient number of parameters. It can also dynamically alter its structure and choose a fusion graph based on the importance of each n-modal dynamics. They claim that this is more interpretable fusion as opposed to the naive late fusion techniques.
2. **Choi et al. (2018)** use the text and speech modality of the CMU-MOSEI dataset. They extract feature vectors for text and speech spectrograms using Convolutional Neural Networks (CNNs) architectures. They then use a trainable attention mechanism to leaner non-linear dependence between the two modalities.
3. **Sahay et al. (2018)** propose a tensor fusion network that explicitly models n-modal inter-modal interactions using an n-fold Cartesian product from modality embeddings.

5.3 Analysis

Comparison with SOTA: Evaluation of F1 scores and MAs of all the methods is summarized in Table 1. We observe an improvement of 1-23% in F1 scores and 5-10% in MAs when using our method.

Confusion Matrix: We also show the confusion matrix (Figure 5) to analyze the per-class performance of M3ER on IEMOCAP and CMU-MOSEI. We observe that more than 73% of the samples per class were correctly classified by M3ER. We see no confusions (0%) between some emotion labels in the two confusion matrices, for instance 'sad' and 'happy' in IEMOCAP and 'fear' and 'surprise' in CMU-MOSEI. Interestingly, we see a small set of data points getting confused between 'happy' and 'angry' labels for both datasets. We reason that this is because, in both situations, people often tend to exaggerate their cues.

Interpretability of Multiplicative Layer: We ran experiments to see the change of weights per sample point for each modality at the time of fusion to validate the importance of multiplicative fusion. Averaged over all the data points in the test set, when we corrupted the face modality, the average weight for the face modality dropped by 12%, which was distributed tot he other modalities, text and speech. This is expected of the multiplicative layer, i.e. to adjust weights for each modality depending on the quality of the inputs.

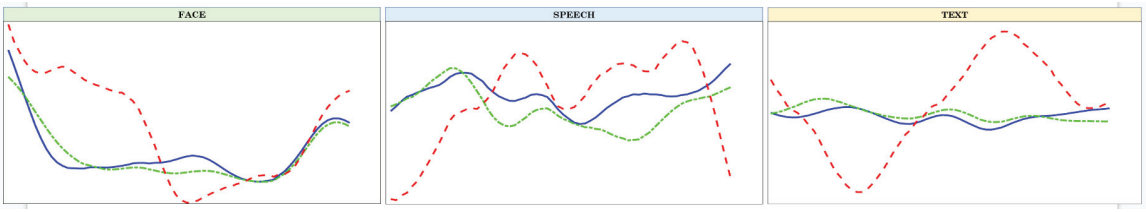

Figure 7: Regenerated Proxy Feature Vector: We show the quality of the regenerated proxy feature vectors for each of the three modalities. For the three graphs, we demonstrate the original feature vector (blue), the ineffectual version of the modality because of added white Gaussian noise (red) and the regenerated feature vector (green). The mean L_2 norm distance between the original and the regenerated vector for the speech, text and face modality are all around 0.01% of the L_2 norm of the respective data.

Qualitative Results: Additionally, we show one sample per class from the CMU-MOSEI and IEMOCAP dataset that were correctly classified by M3ER in Figure 3 and Figure 4.

Failure Case: We also qualitatively show a data point in Figure 6 where M3ER fails to classify correctly. We observe that exaggerations of facial expressions and speech have led to a 'happy' sample being classified by our model as 'angry', a pattern also observed from the confusion matrices.

5.4 Ablation Experiments

Original vs M3ER Multiplicative Fusion Loss. We first compare the original multiplicative fusion loss (Liu et al. 2018) (Equation 2) with our modified loss (Equation 3 on both IEMOCAP and CMU-MOSEI. As shown in Table 2, using our modified loss results in an improvement of 6-7$\%$ in both F1 score and MA.

Next, to motivate the necessity of checking the quality of signals from all the modalities and implementing corrective measures in the case of ineffectual features, we corrupt the datasets by adding white Gaussian noise with a signal-to-noise ratio of 0.01 to at least one modality in approximately 75% of the samples in the datasets. We then compare the performance of the various ablated versions of M3ER as summarized in Table 2 and detailed below.

M3ER – Modality Check Step – Proxy Feature Vector. This version simply applies the multiplicative fusion with the modified loss on the datasets. We show that this results in a drop of 4-12$\%$ in the overall F1 score and 9-12$\%$ in the overall MA from the non-ablated version of M3ER.

M3ER – Proxy Feature Vector. In this version, we perform the modality check step to filter out the ineffectual modality signals. This results in an improvement of 2-5$\%$ in the overall F1 score and 4-5$\%$ in the overall MA from the previous version. However, we do not replace the filtered out modalities with generated proxy features, thus having fewer modalities to work with. This results in a drop of 2-7$\%$ in the overall F1 score and 5-7$\%$ in the overall MA from the non-ablated version of M3ER.

Finally, with all the components of M3ER in place, we achieve an overall F1 score of 0.761 on IEMOCAP and 0.856 on CMU-MOSEI, and an overall MA of 78.2% on IEMOCAP and 85.0% on CMU-MOSEI. Additionally, we also show in Figure 7 that the mean L_2 norm distance between the proxy feature vectors regenerated by M3ER in and the ground truth data is around 0.01% of the L_2 norm of the respective data.

6 Conclusion, Limitations, and Future Work

We present M3ER, a multimodal emotion recognition model that uses a multiplicative fusion layer. M3ER is robust to sensor because of a modality check step that distinguishes between good and bad signals to regenerate a proxy feature vector for bad signals. We use multiplicative fusion to decide on a per-sample basis which modality should be relied on more for making a prediction. Currently, we have applied our results to databases with three input modalities, namely face, speech, and text. Our model has limitations and often confuses between certain class labels. Further, we currently perform binary classification per class; however, human perception is rather subjective in nature and would resemble a probability distribution over these discrete emotions. Thus, it would be useful to consider multi-class classification in the future. As part of future work, we would also explore more elaborate fusion techniques that can help improve the accuracies. We would like to extend M3ER for more than three modalities. As suggested in psychological studies, we would like to explore more naturalistic modalities like walking styles and even contextual information.

7 Acknowledgements

This research is supported in part by ARO grant W911NF-18-1-0313. We would also like to thank Abhishek Bassan and Ishita Verma for initial few discussions on this project.

References

Akputu, K. O.; Seng, K. P.; and Lee, Y. L. 2013. Facial emotion recognition for intelligent tutoring environment. In *IMLCS*, 9–13.

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *ICML*, 1247–1255.

Aviezer, H.; Trope, Y.; and Todorov, A. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338(6111):1225–1229.

Baltrusaitis, T.; Ahuja, C.; and Morency, L. 2017. Multimodal machine learning: A survey and taxonomy. *CoRR* abs/1705.09406.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335.

Castellano, G.; Kessous, L.; and Caridakis, G. 2008. Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in HCI*. Springer. 92–103.

Clavel, C.; Vasilescu, I.; Devillers, L.; Richard, G.; and Ehrette, T. 2008. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication* 50:487–503.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; and Taylor, J. 2001. Emotion recognition in human-computer interaction. *SP Magazine, IEEE* 18:32 – 80.

Ekman, P., and Friesen, W. V. 1967. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills* 24(3 PT 1):711–724.

Ekman, P. 1993. Facial expression and emotion. *American psychologist* 48(4):384.

Fernández-Dols, J.-M., and Ruiz-Belda, M.-A. 1995. Expression of emotion versus expressions of emotions. In *Everyday conceptions of emotion*. Springer. 505–522.

Gunes, H., and Piccardi, M. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications* 30(4):1334–1345.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika* 58(3):433–451.

Kim, Y.; Lee, H.; and Provost, E. M. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. *ICASSP* 3687–3691.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Knapp, R. B.; Kim, J.; and André, E. 2011. Physiological signals and their use in augmenting emotion recognition for human–machine interaction. In *Emotion-oriented systems*. Springer. 133–159.

Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; and Pantic, M. 2017. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65:23–36.

Lee, C. W.; Song, K. Y.; Jeong, J.; and Choi, W. Y. 2018. Convolutional attention networks for multimodal emotion recognition from speech and text data. *ACL 2018* 28.

Liu, K.; Li, Y.; Xu, N.; and Natarajan, P. 2018. Learn to combine modalities in multimodal deep learning. *arXiv:1805.11730*.

Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; and Poria, S. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems* 161:124–133.

Meeren, H. K.; van Heijnsbergen, C. C.; and de Gelder, B. 2005. Rapid perceptual integration of facial expression and emotional body language. *NAS* 102(45):16518–16523.

Navarretta, C. 2012. Individuality in communicative bodily behaviours. In *Cognitive Behavioural Systems*. Springer. 417–423.

Pantic, M.; Sebe, N.; Cohn, J. F.; and Huang, T. 2005. Affective multimodal human-computer interaction. In *International Conference on Multimedia*, 669–676. ACM.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Ringeval, F.; Sonderegger, A.; Sauer, J.; and Lalanne, D. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *FG*, 1–8. IEEE.

Russell, J. A.; Bachorowski, J.-A.; and Fernández-Dols, J.-M. 2003. Facial and vocal expressions of emotion. *Annual review of psychology* 54(1):329–349.

Sahay, S.; Kumar, S. H.; Xia, R.; Huang, J.; and Nachman, L. 2018. Multimodal relational tensor network for sentiment and emotion classification. *arXiv:1806.02923*.

Saragih, J. M.; Lucey, S.; and Cohn, J. F. 2009. Face alignment through subspace constrained mean-shifts. In *ICCV*, 1034–1041. IEEE.

Scherer, K. R.; Johnstone, T.; and Klasmeyer, G. 2003. Vocal expression of emotion. *Handbook of affective sciences* 433–456.

Shan, C.; Gong, S.; and McOwan, P. W. 2007. Beyond facial expressions: Learning human emotion from body gestures. In *BMVC*, 1–10.

Sikka, K.; Dykstra, K.; Sathyanarayana, S.; Littlewort, G.; and Bartlett, M. 2013. Multiple kernel learning for emotion recognition in the wild. In *ICMI*, 517–524. ACM.

Soleymani, M.; Pantic, M.; and Pun, T. 2011. Multimodal emotion recognition in response to videos. *TAC* 3(2):211–223.

Welling, M. 2005. Kernel canonical correlation analysis. *Department of Computer Science University of Toronto, Canada*.

Yoon, S.; Byun, S.; Dey, S.; and Jung, K. 2019. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP*, 2822–2826. IEEE.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory fusion network for multi-view sequential learning. *AAAI*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL (Volume 1: Long Papers)*, 2236–2246.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018c. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL (Volume 1: Long Papers)*, 2236–2246.