Automatic Recognition of Self-Reported and Perceived Emotion: Does Joint Modeling Help?

Biqiao Zhang
University of Michigan
Ann Arbor, MI, USA
didizbq@umich.edu

Georg Essl
University of Wisconsin-Milwaukee
Milwaukee, WI, USA
essl@uwm.edu

Emily Mower Provost
University of Michigan
Ann Arbor, MI, USA
emilykmp@umich.edu

ABSTRACT

Emotion labeling is a central component of automatic emotion recognition. Evaluators are asked to estimate the emotion label given a set of cues, produced either by themselves (*self-report label*) or others (*perceived label*). This process is complicated by the mismatch between the intentions of the producer and the interpretation of the perceiver. Traditionally, emotion recognition systems use only one of these types of labels when estimating the emotion content of data. In this paper, we explore the impact of jointly modeling both an individual's self-report and the perceived label of others. We use deep belief networks (DBN) to learn a representative feature space, and model the potentially complementary relationship between intention and perception using multi-task learning. We hypothesize that the use of DBN feature-learning and multi-task learning of self-report and perceived emotion labels will improve the performance of emotion recognition systems. We test this hypothesis on the IEMOCAP dataset, an audio-visual and motion-capture emotion corpus. We show that both DBN feature learning and multi-task learning offer complementary gains. The results demonstrate that the perceived emotion tasks see greatest performance gain for emotionally subtle utterances, while the self-report emotion tasks see greatest performance gain for emotionally clear utterances. Our results suggest that the combination of knowledge from the self-report and perceived emotion labels lead to more effective emotion recognition systems.

CCS Concepts

•Computing methodologies → Artificial intelligence; Multi-task learning;

Keywords

Audio-visual emotion recognition, self-reported emotion, perceived emotion, multi-task learning

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2993173

1. INTRODUCTION

Expressions of emotion convey information about the underlying state of an individual. This information can be partially masked either intentionally or unintentionally, leading to variability in the labels associated with emotional displays. This variability in the label space is one of the main differences between emotion recognition and other machine learning tasks. As a result, emotion labeling experiments must clearly identify the purpose of a given set of labels: will the labels capture the *felt sense* of the individual who produces the emotion, will they instead capture how that person believes others will perceive his/her emotional display (*self-report label*), or will they instead capture how others actually do perceive the display (*perceived label*)?

All three types of labels are important in real use-cases. For example, emotion recognition systems that are designed for applications are often focused on correctly identifying felt-sense and/or self-report labels (e.g., intelligent personal assistants and computer games [20]). The recognition of perceived emotion labels has important application in the monitoring and treatment of disease associated with emotion perception deficits [9,14]. However, emotion recognition systems have focused only on a single type of label traditionally, rather than leveraging the potentially complementary information conveyed by the separate strategies. This paper explores the impact of jointly modeling both an individual's self-report and the perceived labels of others.

Individuals differ in their ability to convey emotion. Therefore, the patterns of emotion expression can vary across individuals, resulting in difficulty in transferring models learned from a set of speakers to a new speaker when using self-reported emotion, as observed in [34]. The emotion labels provided by others can act as stabilizers to reduce fluctuations caused by individual differences. On the other hand, the varying patterns (e.g., intensity of cues) of emotion expression can result in different levels of difficulty for observers, as found in [17]. The emotion labels provided by the speakers themselves can work as a stabilizer to explain how a single individual expresses a range of emotions. Therefore, our motivating hypothesis is that if we can control for the manner in which others perceive emotion and how one perceives one's own emotion, we will see improvement in both tasks. In addition, we hypothesize that we can get complementary improvement by better capturing the complexity inherent in the interactions between multimodal cues. We ask the following research questions: (1) can joint modeling lead to better performance across both types of labels; (2) can the same performance gain be achieved through com-

plex feature learning; and (3) is the performance gain from joint modeling and complex feature learning additive?

We conduct an experiment on the IEMOCAP dataset [6] using a subset that contains both perceived and self-report labels. We construct the emotion recogntion problem as binary one-against-rest classifications to account for the fact that an utterance can be labeled with multiple emotions. We use linear support vector machines (SVMs) as the baseline method. We propose a multi-task learning method that jointly models self-report and perceived emotion (each label type is a task). We also explore the influence of non-linear feature learning using deep belief networks (DBNs). Finally, we analyze the combined impact of both components.

Our experimental results suggest that joint modeling is able to utilize the complementary knowledge presented in both self-report and perceived labels and that the combination of non-linear feature learning and joint modeling results in more effective emotion recognition systems. The novelty of this work includes: (1) the first attempt to jointly learn self-reported and perceived emotion; (2) an exploration of the influence of feature learning, using DBN feature pretraining, on multi-task learning.

2. RELATED WORKS

2.1 Emotion Expression and Perception

Brunswik's functional lens model is a theory of human perception [5]. Adaptations of this model have been used to study emotional communication [3, 18, 28, 29]. In this model there are two entities: an encoder (e.g., performer or speaker) and one or more decoders (e.g., listeners or evaluators). The encoder produces a message that conveys his/her communicative goals, accompanied by various paralinguistic properties (e.g., emotion, age, and gender). This message is encoded in a set of cues that are expressed over multiple channels (e.g., the face and the voice), called distal indicators. The cues are transmitted to the decoders, who perceive and interpret the information, referred to as proximal percepts. The proximal percepts contain redundancy and through the combination of multimodal percepts, the decoders are able to arrive at a higher-level judgment of both the communicative intent and the paralinguistic information. Laukka et al. studied the link between distal indicators and proximal percepts [22]. They found that there are a wide variety of cues, such as sound level, rhythm, tempo and timbre, are associated with both the intended and perceived emotion. It shows that intended emotion and perceived emotion are linked through cues, which provides support for jointly modeling two types of labels.

Researchers in psychology and cognitive science have found support for the idea that the expression and perception of emotion can be influenced by cultural, gender and individual differences [24, 37]. Matsumoto et al. investigated the display rules of emotion on participants from five different countries and found that there were culturally-specific display rules. The works of Elfenbein et al. [10, 11] found that people from the same national, ethnic, or regional group can recognize emotion more accurately than people from different backgrounds. Further, people may depend on different cues when perceiving emotions. Yuki et al. [37] found that individuals from cultures that control for emotional display depend more heavily on the eyes, compared to the mouth, while people from cultures that encourage emotional display focus more on the mouth than the eyes. Emotion expression and perception are also heavily influenced by individual factors. Hall et al. found that women are better at conveying emotion through facial expression, compared to men [15]. Rotter et al. found that females can recognize emotion expression of both males and females more accurately in general [27]. Moreover, Martin et al. found that emotional sensitivity, represented by the minimum time required to recognize pleasant vs. unpleasant emotion given images of emotional faces, is different across individuals [23]. Their findings suggest that there are differences in the interpretation of proximal percepts and the production of distal indicators, resulting in variability in the label space of both self-report and perceived emotion.

2.2 Automatic Recognition of Self-Reported and Perceived Emotion

Research in affective computing has compared the automatic recognition of self-reported and perceived emotion. Truong et al. investigated the agreement rate between self-reported and perceived emotion labels [35]. They found that the agreement increased when data from multiple modalities were provided (a finding also supported by Busso et al. [7]), and that the agreement between self-rated and perceived labels were lower than the agreement among the perceived label evaluators (inter-evaluator agreement). Busso et al. [8] compared the self-report and perceived labels across categorical and dimensional descriptors, and found that there are discrepancies between self-reported and perceived emotion labels. They further found that the use of self-reported labels significantly lower the inter-rater agreement and that self-reported labels tend to have more extreme values for dimensional descriptors. In Truong et al. [34], the authors trained regressors for recognizing valence (pleasant vs. not pleasant) and activation (energetic vs. calm). Separate models were built for self-reported and perceived emotions. They found that perceived emotion was harder to predict.

The work of Busso et al. supports the notion that there are differences between self-reported and perceived emotion labels [8]. The work of Truong et al. has demonstrated the challenges in predicting self-report labels [34]. The works in Section 2.1 supported the notion that there is variation in the production of distal indicators and in the interpretation of proximal percepts. The combination of this body of research leads us to hypothesize that we can improve the accuracy of emotion recognition systems if we can: (1) capture the complexity of distal indicators (feature learning) and (2) better understand the relationship between how an encoder interprets his/her own distal indicators and how decoders interpret the associated proximal percepts (multi-task learning). We posit that the combination of these two approaches will lead to a more stable and robust system.

3. DATA

We experiment on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [6]. IEMOCAP contains five sessions of dyadic interactions between pairs of male-female actors. The emotional behaviors are elicited using scripted and improvised scenarios. The dataset includes 12 hours of data across three modalities: audio, video, and motion-capture (referred to as "mocap"). The mocap recording was made over a single actor at a time. Consequently, only half of the data have matched audio-visual and mocap

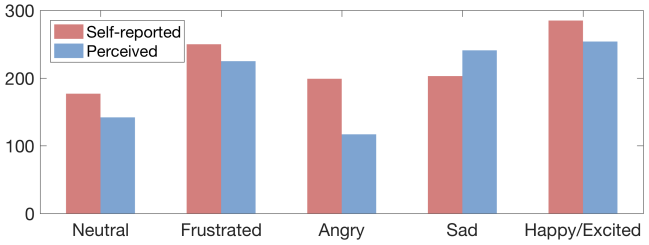

Figure 1: The number of utterances in each emotion.

recordings (see [6] for details about the recording setup).

The data were segmented into speaker turns (*utterances*) and were evaluated at the utterance-level. The evaluations include both categorical and dimensional labels; in this paper, we focus only on the categorical labels. The categorical labels were evaluated by at least three evaluators. The labels were chosen from the set of {angry, happy, neutral, sad, frustrated, excited, disgusted, fearful, surprised, other}. We merge the classes of happiness and excitement as in [25]. There was no limitation on the number of labels an evaluator could select for a given utterance. This subset, which we refer to as the *original data*, contains 5,042 utterances.

Six out of ten actors were asked to self-report the emotional content of their own recordings of the improvised scenarios. They used the same evaluation paradigm as the evaluators, described above. This subset contains 1,184 utterances from the *original data* that have self-reported emotion labels and matched audio and mocap data. The data in this subset have two labels: (1) perceived emotion and (2) self-reported emotion. The perceived emotion labels are a vector that describe the emotions perceived by the evaluators. We define the perceived emotion ground truth as any emotion label noted by at least two (out of three) evaluators. For example, the perceptual evaluations of three evaluators for $utterance_i$ may be distributed as [2 0 0 2 1 0 ... 0], where two evaluators noted anger, two noted sadness, and one noted frustration (evaluators were not restricted to the number of emotions selected). Therefore, $utterance_i$ would be associated with the perceived emotions of anger and sadness. The perceived ground truth label for each emotion is a binary vector that describes the presence of each label (e.g., for the example above the final label would be [1 0 0 1 0 ... 0]). The self-report label is also a binary vector that marks the presence or absence of a given label. We downsample the *self-evaluation* subset to include only utterances with at least one perceived emotion and one self-reported emotion from the set of {angry, happy/excited, neutral, sad, and frustrated}. This results in 967 utterances. We refer to this data as the *self-evaluation subset*.

On average, each utterance in the *self-evaluation subset* has 1.15 ± 0.39 self-reported emotion labels and 1.01 ± 0.11 perceived emotion labels. We compare the distribution of self-reported emotion and perceived emotion in Figure 1. There are differences between the two distributions, notably for the class of anger. We compute the Hamming similarity between the two types of emotion, defined as the proportion of instances that have the same label in self-report and perceived emotions, given an emotion class. The similarity for neutral, frustrated, angry, sad, and happy/excited are 0.87, 0.83, 0.88, 0.91 and 0.90, respectively, and the average over all classes is 0.88.

We use both the *original* and *self-evaluation* sets of data.

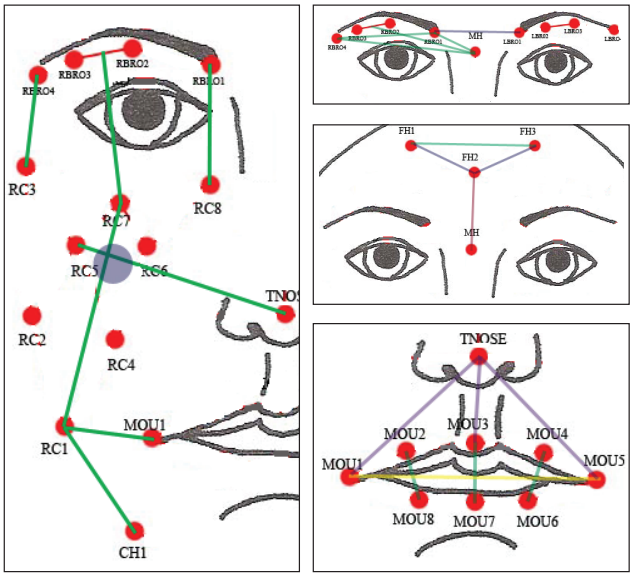

Figure 2: The positions of the markers and the distance features (only shown on right side of face). Image courtesy E. Mower, M. J. Mataric, and S. Narayanan [25]. Each marker position is represented in three-dimensional vector coordinates (x, y, z).

The *original* data are used for unsupervised feature learning (see details in Section 4.2). The *self-evaluation* data are used for supervised classification (see details in Section 4.3-4.4).

4. METHODOLOGY

4.1 Feature Extraction

4.1.1 Acoustic Features

We extract the Interspeech 2009 Emotion Challenge features [30] using openSMILE [12]. We use a relatively small feature set due to the limited size of the data. The feature set contains 16 frame-level Low-Level Descriptors (LLDs), including zero-crossing-rate, root mean square energy, pitch frequency, harmonics-to-noise ratio, and Mel-Frequency Cepstral Coefficients (MFCC) 1-12. Twelve statistics are applied to the frame-level LLDs and the first-order delta coefficient of the LLDs to generate the 384 utterance-level features. The statistics are: max, min, range, the position of the maximum and minimum value, arithmetic mean, standard deviation, the slope and onset of the linear approximation of the contours, quadratic error (between actual contour and the linear approximation), skewness and kurtosis.

4.1.2 Visual Features

We extract visual features using the 3D motion-capture markers. The mocap features are the Euclidean distances between the (x, y, z) coordinates of the markers. The positions of the markers and the distances calculated are shown in Figure 2. These features were introduced [25] and used in [19, 25]. They capture movements associated with emotional facial expressions. For example, the distance between TNOSE and MOU1/MOU5 changes as a function of smiles and frowns. We apply five statistics to the frame-level distance features, including mean, variance, quantile maximum, quantile minimum and quantile range. This results

Table 1: Number of units in each layer of the DBN models for audio, mocap and combined features.

Modality	Input	Layer 1	Layer 2	Layer 3
Audio	384	600	600	{100,200}
Mocap	540	800	800	{200,300}
Combined	924	1400	1400	{300,400,500}

in 540 utterance-level features. We exclude missing data in our utterance-level calculations.

We perform speaker-dependent z-normalization on each feature. The normalization is applied separately for the *original* and *self-evaluated* data. In this way, the initial input for classification is identical for models that do and do not use feature learning, allowing for a more direct comparison.

4.2 Feature Learning

We use the pretraining of deep belief networks (DBNs) [16] for feature learning. DBNs are formed by stacking Restricted Boltzmann Machines (RBMs) [31], which are undirected neural networks that only have inter-layer connections. RBMs learn the posterior probability of the output (often binary, referred to as "hidden units") given the inputs (binary or Gaussian, referred to as "visible units"). We select DBN feature learning because: (1) it can capture complex non-linear interactions between features; (2) its unsupervised nature makes the learned features task-independent; (3) it has been shown to be effective for reducing dimension and can outperform traditional feature selection methods, such as Information Gain and Principal Feature Analysis [19].

We train three DBN models on the *original* data, one each for audio, mocap, and both modalities, using the implementation in [33]. We set the number of hidden layers to 3, as in [19]. We choose Gaussian-Bernoulli RBM (GBRBM) as the first layer, since it takes Gaussian visible units, and is suitable for our real-valued features. The second and third layers of the models are Bernoulli-Bernoulli RBMs (BBRBMs), where both the visible units and hidden units are binary. It is suggested in [4] that it is often more beneficial to have an over-complete first layer (i.e. number of hidden units > number of visible units), compared to an under-complete first layer (i.e. number of hidden units < number of visible units). In addition, previous work [21] found that networks that have the same number of hidden units for each layer generally outperform networks that have increasing or decreasing numbers of hidden units at each layer. We use these insights and set the number of hidden units in the first and second layer to be approximately 1.5 times over-complete of the original input features. We decreased the size of the final layer to be in line with prior work on this dataset [19]. The number of units for each layer are shown in Table 1. The number of units in the final layer is selected in cross-validation (Section 4.4). We fix the size of the mini-batches to 32, according to [4]. and set the learning rate to 0.004 for the GBRBM layer and 0.02 for other BBRBM layers based on empirical re-construction error.

4.3 Classification Models

We form the recognition of neutral, frustrated, angry, sad and happy/excited as five one-against-rest binary classification problems and train five separate models. This is because each utterance can be labeled with multiple emotions.

The main question we ask in this work is: can jointly modeling self-reported and perceived emotion lead to better performance for both types of emotions? Therefore, we propose two approaches: independent modeling (denoted as IM) and joint modeling (denoted as JM). The models are compared on the *self-evaluation* data. In IM, we train separate classifiers, one each for self-report and perceived emotion. We use linear Support Vector Machine (SVM) for the IM baseline. When training on the original features, we adopt L_1-regularization to serve as a "built-in" feature selection in addition to the commonly used L_2-regularization, since it can enforce sparsity of the features. We weight the cost of error in the positive class and negative class differently during training to deal with unbalanced data, as suggested in [36]. The per-class weight is calculated by the reciprocal of the proportion of that class in the training data.

In JM, we model self-report and perceived emotion in a single classifier using multi-task learning, with each emotion type as a task. We use the multi-task feature learning (MTFL) algorithm of [1, 2]. This method is based on the hypothesis that task 1 through T share a common feature representation. Therefore, the weight vectors w_1 through w_T for the tasks can be jointly learned. The weight matrix W, defined as $[w_1, w_2, ..., w_T]$, can be rewritten as UA, where U is an orthogonal matrix for feature transformation, and $A = [a_1, a_2, ..., a_T]$ is the weight matrix on the new space.

$$\min_{U,A} \sum_{t=1}^{T} \sum_{i=1}^{m} c_{t(y_{ti})} \max(0, 1 - y_{ti} \langle a_t, U^T x_i \rangle) + \gamma \|A\|_{2,1}^2 \quad (1)$$

Equation (1) shows the objective function of MTFL used in this work. Here, m is the number of training instances, $y_{ti} \in \{-1, 1\}$ is the label of the i-th training instance in task t, x_i is the i-th training instance, $<>$ stands for inner product, $c_{t(y_{ti})}$ is the cost for error in task t for the class y_{ti} belongs to, and γ is the regularization parameter. We use hinge loss ($\max(0, 1 - y_{ti} \langle a_t, U^T x_i \rangle)$) to match the linear SVM. MTFL encourages sparsity of the transformed features and couples the tasks by regularizing on A using the $L_{2,1}$-regularizer. In the general case, U and A are both learned from the data. However, if we force $U = I$, the regularization would be directly imposed on W, in which case the "feature learning" problem reduces to a "feature selection" problem [1, 2]. The convex variants of Equation (1) can be solved by iteratively performing a supervised task-specific step and an unsupervised task-independent step. The former step becomes solving linear SVM with a variable transformation process when hinge loss is used [38]. More details about the algorithm can be found in [1, 2]. In this work, we use both the general setting and the special case where $U = I$ as the multi-task equivalent of L_2 and L_1-regularization when training on the original features. Liblinear is used to solve the linear SVMs [13].

We ask an additional question in this work: can IM and JM be improved by operating on a feature space learned through DBN? We investigate whether the advantages of JM are diminished given a non-linear feature preprocessing step. We also train a single task linear SVM model and multi-task MTFL model on the DBN pretrained features. When feature pretraining is applied, we use only the L_2 regularization for linear SVM and the general case of U for MTFL. The reason we are not selecting the input features by enforcing sparsity is that the DBN feature learning has already played the part of dimensionality reduction.

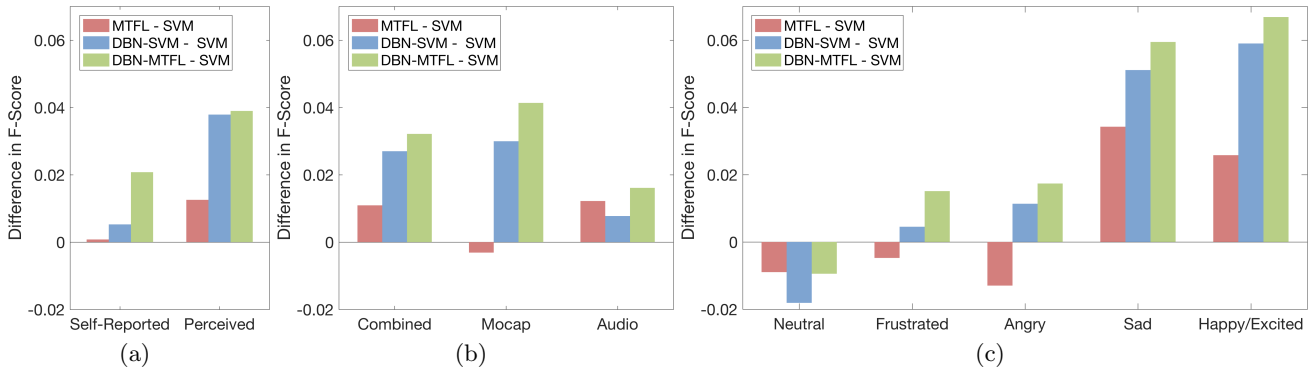

Figure 3: Average differences in F-score between the baseline SVM and MTFL/DBN-SVM/DBN-MTFL. The average is taken for (a) self-reported and perceive emotion, across modality and emotion classes; (b) combined modality, mocap and audio, across types of emotion and emotion classes; (c) each emotion class, across types of emotion and modalities.

4.4 Cross-Validation and Model Selection

It is important to make the reader aware that we use F-score as a performance measure, rather than the common metric of unweighted recall, to account for the multi-label binary classification problem. The F-score is defined as the harmonic mean of the precision and recall of the positive class (i.e. the presence of the emotion), as in [32]. We report the leave-one-speaker-out cross-validation F-score for each model. At each round, data from one speaker is left out as the test set, while data from other speakers are used for training. In the DBN pretraining, the data of the test speaker are also excluded.

We compare four different settings: modeling self-reported emotion and perceived emotion individually or jointly, on the original features or on the DBN pretrained features. This leads to four models: original-SVM (SVM), original-MTFL (MTFL), DBN-SVM and DBN-MTFL, with SVM being the baseline. There are at least two versions for each model to be selected from: L_1 vs. L_2-regularization for SVM, learned U vs. $U = I$ for MTFL and different number of final hidden units (input to classifiers) for DBN-SVM and DBN-MTFL. We select the version and the hyper-parameters by optimizing the cross-validation F-score on the training set only, where cross-validation is also performed in a leave-one-training-speaker-out way. The range of the regularization parameter γ (in all models) is $\{10^{-4}, 10^{-3}, ..., 10^{5}\}$ and the range of the permutation parameter ϵ (in MTFL and DBN-MTFL) is $\{10^{-8}, 10^{-7}, ..., 10^{-1}\}$. Note that γ is equivalent to the cost of error C for linear SVM, and $C = 1/(2 \times \gamma)$.

5. RESULTS AND DISCUSSION

5.1 Performance of Classification Models

We compare the performance of the SVM, MTFL, DBN-SVM, and DBN-MTFL on the task of recognizing perceived and self-reported emotion labels. On average, all models outperformed the baseline SVM model, in the order SVM < MTFL < DBN-SVM < DBN-MTFL (Table 2). MTFL performs better than SVM in the majority of the cases, excepting the prediction of the self-reported emotion label given the unimodal mocap data. The improvement in performance from SVM to MTFL suggests that jointly predicting self-reported and perceived emotion is beneficial with respect to performance on both tasks. We find that the DBN feature learning increases the performance for both SVM and the MTFL (DBN-SVM and DBN-MTFL, Table 2). The DBN-MTFL model produces the highest accuracy overall (exception: perceived emotion from unimodal acoustic features). This suggests that the individual benefits of the non-linear feature learning and the joint modeling are additive.

Table 2: Average F-score of SVM, MTFL, DBN-SVM and DBN-MTFL. The best results in each combination of modality and emotion type is underlined. S: self-report, P: perceived emotion, Combined: both modalities.

Model	Combined		Mocap		Audio		Avg.
	S	P	S	P	S	P	
SVM	0.574	0.534	0.532	0.475	0.477	0.492	0.514
MTFL	0.579	0.551	0.513	0.487	0.493	<u>0.500</u>	0.521
DBN-SVM	0.578	0.584	0.533	0.533	0.487	0.497	0.535
DBN-MTFL	<u>0.588</u>	<u>0.585</u>	<u>0.555</u>	<u>0.534</u>	<u>0.502</u>	0.499	<u>0.544</u>

The statistical significance are assessed using repeated measure ANOVA on the F-score of each emotion. Model and modality are treated as the two within-subject factors. We find that the influence of model and the interaction between modality and model are significant for perceived emotion ($p = 0.011$ and 0.005, respectively), but not for self-reported emotion. We compare the difference in F-score between each pair of models over the 5 emotions $\times$ 3 modalities for perceived emotion using paired t-test. We find that DBN-SVM is significantly better than SVM and MTFL ($p = 0.006$ and 0.020, $t = 3.26$ and 2.64, respectively), and DBN-MTFL is significantly better than SVM and MTFL ($p = 0.004$ and 0.010, $t = 3.44$ and 2.99, respectively).

We compare the performance of the models in Figure 3, assessing the influence of self-reported vs. perceived label (Figure 3a), modality (Figure 3b) and emotion class (Figure 3c). In each figure, we treat SVM as the baseline model and assess the change in F-score as a function of the model types (MTFL, DBN-SVM, DBN-MTFL). We find that the overall performance gain is higher for the perceived labels, and that the advantage of non-linear feature learning is more obvious for perceived labels, compared to the self-reported labels.

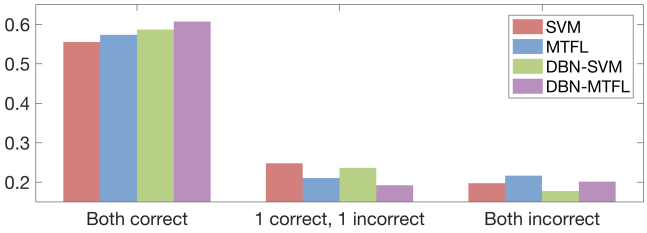

Figure 4: Percentage of utterances that have self-report and perceived labels correct/incorrect.

In the self-reported emotion problem, joint modeling and DBN feature learning, by themselves, show little improvement. However, we see that the combined influence of the two approaches is greater than simple addition.

We find that, of the two unimodal systems (mocap and audio), the mocap modality benefits more from the DBN feature learning. In fact, only using MTFL harms the performance for mocap, yet the combination of feature learning and MTFL leads to a large gain in average performance. On the contrary, the performance gain for joint modeling is higher for audio than mocap, and the addition of DBN feature learning introduces a relatively smaller gain.

We find that the increase in performance is the highest for the emotions of sadness and happiness/excitement. The system performs worse, across all model types, for neutrality compared to the baseline SVM system. In addition, MTFL has lower performance, compared to baseline SVM, for the emotions of frustration and anger. The emotion-specific results mirror the trend in similarity from Section 3. The self-reported and perceived emotion labels for the classes of sadness and happiness/excitement are the most similar, compared to those of neutrality, frustration, and anger. This may suggest joint modeling on the original feature space is most effective when the discrepancies between self-evaluation and perception are smaller. However, the performance gain of DBN-MTFL over DBN-SVM is quite consistent in all the five emotions, indicating feature learning increases the robustness of joint modeling.

For each utterance, there are three different situations for prediction: (1) both self-report and perceived label are correct, (2) one label is correct, the other is incorrect, and (3) both labels are incorrect. We present the three situations (averaged over emotion class and modality) for SVM, MTFL, DBN-SVM and DBN-MTFL in Figure 4. We find that the proportion of the co-occurrence of success mirrors the overall performance of the models, namely SVM $<$ MTFL $<$ DBN-SVM $<$ DBN-MTFL. Interestingly, we find that DBN feature learning decreases both (2) and (3) (SVM vs. DBN-SVM, MTFL vs. DBN-MTFL), yet joint modeling only contributes to reducing (2), not (3) (SVM vs. MTFL, DBN-SVM vs. DBN-MTFL). The fact that joint modeling increases the co-occurrence of both success and error gives support to our hypothesis that joint modeling leverages the knowledge carried in both labeling methods.

5.2 Prototypical vs. Non-Prototypical Emotion

The performance of emotion recognition systems is often assessed as a function of subtlety, described in terms of prototypicality. Prototypicality is defined as complete agreement between evaluators, while non-prototypicality is defined as only majority vote agreement. Previous works have found that it is harder to classify utterances with non-prototypical emotions, compared to utterances with prototypical emotions [19, 26]. In this study, we compare the performance of the proposed techniques as a function of the prototypicality over the perceived emotion label only.

Table 3: Average F-score of self-reported and perceived emotion on prototypical and non-prototypical utterances. The best results in each column are underlined. The performance gain of feature pretraining + joint modeling is shown in the bottom line. S: self-report, P: perceived emotion.

Model	Prototypical		Non-Prototypical	
	S	P	S	P
SVM	0.541	0.511	0.490	0.452
MTFL	0.538	0.517	0.494	0.470
DBN-SVM	0.542	0.535	0.498	0.504
DBN-MTFL	<u>0.559</u>	<u>0.536</u>	<u>0.512</u>	<u>0.505</u>
Performance Gain	0.017	0.025	0.022	0.053

In the *self-evaluation* subset, 52% of the data are prototypical. The Hamming similarity (averaged over five emotions) between two types of labels are 0.92 and 0.83 for prototypical and non-prototypical data, respectively. This suggests a larger discrepancy between self-report and perception for non-prototypical data.

We present the average performance of each model on prototypical and non-prototypical data for self-reported emotion and perceived emotion in Table 3. Similar to previous works, we also find that results on prototypical data consistently outperform results on non-prototypical data for both self-reported and perceived emotion. DBN-MTFL achieved the highest average performance, on both prototypical data and non-prototypical data. We compare DBN-MTFL with the baseline SVM on the bottom line. We find that the performance gain for non-prototypical data is higher than for prototypical data, especially for perceived emotion.

We assess the performance change of the models over the prototypical data and non-prototypical data, again using repeated measure ANOVA. We find that the influence of model and the interaction between modality and model are significant on the non-prototypical data for perceived emotion ($p = 0.023$ and 0.004, respectively), but not on prototypical data or for self-reported emotion. Comparing the models in a pairwise manner for the perceived emotion of non-prototypical data using paired t-test, we find that MTFL is significantly better than SVM ($p = 0.016$, $t = 2.75$), DBN-SVM is significantly better than both SVM and MTFL ($p = 0.013$ and 0.043, $t = 2.85$ and 2.23, respectively), and DBN-MTFL is significantly better than both SVM and MTFL ($p = 0.008$ and 0.029, $t = 3.11$ and 2.43, respectively).

5.3 Mixed vs. Clear Emotion

The definition of prototypicality used in Section 5.2 does not extend well to self-reported labels because they are derived from a single evaluator (the actor). Instead, we describe subtlety in self-reported labels in terms of the number of labels provided by the actor. We use the term "mixed" when the actor describes his/her data with multiple labels and "clear" when only one label is provided.

In the *self-evaluation* subset, 14% of the utterances are mixed. The average Hamming similarity between self-report and perceived labels for mixed and clear emotions is 0.75

Table 4: Average F-score of self-reported and perceived emotion on utterances with mixed self-reported emotion (>1 labels) and clear self-reported emotion (=1 labels). The best results in each column are underlined. The performance gain of feature pretraining + joint modeling is shown in the bottom line. S: self-report, P: perceived emotion.

Model	Mixed		Clear	
	S	P	S	P
SVM	0.618	0.404	0.492	0.505
MTFL	0.603	0.438	0.497	0.514
DBN-SVM	0.564	0.447	0.510	0.540
DBN-MTFL	0.582	0.428	0.525	0.543
Performance Gain	-0.036	0.024	0.033	0.038

and 0.90, respectively. This suggests that when an individual notes variability in his/her performance, it is more likely that the self-report and perceived emotion will disagree. However, this does not automatically lead to a designation of non-prototypicality; only 54% of mixed emotions are non-prototypical. This highlights a difference between the perception of variability for self and for other evaluators.

We present the average performance of each model on mixed and clear data for self-reported and perceived emotion in Table 4, again listing the difference in performance between DBN-MTFL and the baseline SVM on the bottom line. We find the largest performance gain for perceived emotion labels from the clear subset. Joint modeling and DBN pretraining actually have negative influence on the self-reported emotion of mixed data. Interestingly, we notice that the performance of perceived emotion on the mixed data is lower than on the clear data. This indicates that when emotion expression is considered subtle by the producer, it is indeed harder for both the classifier (Table 4) and the human evaluator (Hamming similarity result) to accurately recognize it, although this subtlety itself may not be fully captured by variation in evaluation.

The repeated measure ANOVA shows that for data with mixed emotion, the influence of model on self-reported emotion is significant ($p = 0.009$). Pairwise comparison using paired t-test indicates that SVM is significantly better than DBN-SVM and DBN-MTFL ($p = 0.002$ and 0.015, $t = 3.72$ and 2.77, respectively), and so is MTFL ($p = 0.004$ and 0.024, $t = 3.44$ and 2.54, respectively). This suggests that non-linear feature learning has a negative effect in this case. For data with clear emotion, the influence of model is significant for both self-reported and perceived emotion ($p = 0.038$ and 0.019, respectively), and the combined influence of model and modality is significant for perceived emotion ($p = 0.020$). Pairwise model comparison shows that DBN-MTFL is significantly better than SVM ($p = 0.002$, $t = 3.76$), MTFL ($p = 0.002$, $t = 3.74$) and DBN-SVM ($p = 0.021$, $t = 2.59$) for self-reported emotion. For perceived emotion, DBN-SVM is significantly better than both SVM and MTFL ($p = 0.011$ and 0.022, $t = 2.94$ and 2.57, respectively), and DBN-MTFL is significantly better than both SVM and MTFL ($p = 0.005$ and 0.007, $t = 3.29$ and 3.16, respectively).

6. CONCLUSION AND FUTURE WORKS

In this paper, we explore the impact of jointly predicting self-reported emotion and perceived emotion in addition to non-linear feature learning. We hypothesize that joint modeling using multi-task learning leads to performance increases for both kind of labels, and the performance gain of joint modeling and DBN feature learning is complementary. We experiment on IEMOCAP using a multi-label classification paradigm to test this hypothesis.

Our results show that overall, DBN feature learning and joint modeling together produce the highest performance, suggesting the individual benefits of the two approaches are additive. The performance gain is higher for the perceived labels, compared to the self-reported labels, yet we notice that the combined influence of the two approaches is greater than simple addition for self-reported labels. We find that while DBN feature learning does not show preference over different kinds of error, joint modeling increases the co-occurrence of both success and error, and decreases the mismatch of correct and incorrect predictions for self-report and perceived labels. Our findings suggest that joint modeling is able to leverage the potentially complementary information conveyed by both individual labeling strategies, and combining non-linear feature learning with joint modeling leads to more effective emotion recognition systems.

The Brunswik Lens model discusses how communicative cues produced by an encoder (distal indicators), are altered by transmission (proximal percepts), and interpreted by a decoder. Our results suggest that when an individual produces an emotional message that he/she believes to be clear, there is benefit to capturing variability in the distal indicators (feature learning) and variability due to transmission (multi-task learning). Interestingly, when an individual does not believe his/her emotion to be clear, this approach is ineffective. This suggests that additional research is needed to understand how to automatically interpret ambiguous emotional expressions.

A limitation of this work is the size of the data for the supervised emotion classification task. We plan to conduct experiments on additional datasets to further test our proposed methods. In addition, we will explore the impact of these methods on the prediction of dimensional labels.

7. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, volume 19, 2007.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] T. Bänziger, S. Patel, and K. R. Scherer. The role of perceived voice and speech characteristics in vocal emotion communication. *Journal of Nonverbal Behavior*, 38(1):31–52, 2014.

[4] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. 2012.

[5] E. Brunswik. Representative design and probabilistic theory in a functional psychology. *Psychological review*, 62(3):193, 1955.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *ICMI*, pages 205–211, 2004.

[8] C. Busso and S. S. Narayanan. The expression and perception of emotions: comparing assessments of self versus others. In *INTERSPEECH*, pages 257–260, 2008.

[9] G. Csukly, P. Czobor, E. Szily, B. Takács, and L. Simon. Facial expression recognition in depressed subjects: the impact of intensity level and arousal dimension. *The Journal of nervous and mental disease*, 197(2):98–103, 2009.

[10] H. A. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[11] H. A. Elfenbein and N. Ambady. When familiarity breeds accuracy: cultural exposure and facial emotion recognition. *Journal of personality and social psychology*, 85(2):276, 2003.

[12] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM MM*, pages 1459–1462, 2010.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[14] R. C. Gur, R. J. Erwin, R. E. Gur, A. S. Zwil, C. Heimberg, and H. C. Kraemer. Facial emotion discrimination: II. behavioral findings in depression. *Psychiatry research*, 42(3):241–251, 1992.

[15] J. A. Hall. *Nonverbal sex differences: Accuracy of communication and expressive style*. 1990.

[16] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[17] P. N. Juslin and P. Laukka. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, 1(4), 2001.

[18] A. Kappas, U. Hess, and K. R. Scherer. 6. voice and emotion. *Fundamentals of nonverbal behavior*, 1991.

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

[20] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[21] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, 10:1–40, 2009.

[22] P. Laukka, T. Eerola, N. S. Thingujam, T. Yamasaki, and G. Beller. Universal and culture-specific factors in the recognition and performance of musical affect expressions. *Emotion*, 13(3):434, 2013.

[23] R. A. Martin, G. E. Berry, T. Dobranski, M. Horne, and P. G. Dodgson. Emotion perception threshold: Individual differences in emotional sensitivity. *Journal of Research in Personality*, 30(2):290–305, 1996.

[24] D. Matsumoto, S. Takeuchi, S. Andayani, N. Kouznetsova, and D. Krupp. The contribution of individualism vs. collectivism to cross-national differences in display rules. *Asian Journal of Social Psychology*, 1(2):147–165, 1998.

[25] E. Mower, M. J. Mataric, and S. Narayanan. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070, 2011.

[26] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan. Interpreting ambiguous emotional expressions. In *ACII*, pages 1–8, 2009.

[27] N. G. Rotter and G. S. Rotter. Sex differences in the encoding and decoding of negative facial emotions. *Journal of Nonverbal Behavior*, 12(2):139–148, 1988.

[28] K. Scherer. Emotion in action, interaction, music, and speech. *Language, music, and the brain: A mysterious relationship*, pages 107–139, 2013.

[29] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256, 2003.

[30] B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, pages 312–315, 2009.

[31] P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In *Parallel distributed processing: explorations in the microstructure of cognition*, volume 1, pages 194–281, 1986.

[32] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[33] M. Tanaka and M. Okutomi. A novel inference of a restricted boltzmann machine. In *ICPR*, pages 1526–1531, 2014.

[34] K. Truong, D. van Leeuwen, M. Neerincx, and F. de Jong. Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion. In *INTERSPEECH*, 2009.

[35] K. P. Truong, M. A. Neerincx, and D. A. Van Leeuwen. Assessing agreement of observer-and self-annotations in spontaneous multimodal emotion data. In *INTERSPEECH*, pages 318–321, 2008.

[36] K. Veropoulos, C. Campbell, N. Cristianini, et al. Controlling the sensitivity of support vector machines. In *IJCAI*, pages 55–60, 1999.

[37] M. Yuki, W. W. Maddux, and T. Masuda. Are the windows to the soul the same in the east and west? cultural differences in using the eyes and mouth as cues to recognize emotions in japan and the united states. *Journal of Experimental Social Psychology*, 43(2):303–311, 2007.

[38] B. Zhang, E. Mower Provost, and G. Essl. Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In *ICASSP*, 2016.