Learning Spontaneity to Improve Emotion Recognition in Speech

Karttikeya Mangalam, Tanaya Guha

Indian Institute of Technology, Kanpur

`mangalam@iitk.ac.in, tanaya@iitk.ac.in`

Abstract

We investigate the effect and usefulness of spontaneity (i.e. whether a given speech is spontaneous or not) in speech in the context of emotion recognition. We hypothesize that emotional content in speech is interrelated with its spontaneity, and use spontaneity classification as an auxiliary task to the problem of emotion recognition. We propose two supervised learning settings that utilize spontaneity to improve speech emotion recognition: a hierarchical model that performs spontaneity detection before performing emotion recognition, and a multitask learning model that jointly learns to recognize both spontaneity and emotion. Through various experiments on the well-known IEMOCAP database, we show that by using spontaneity detection as an additional task, significant improvement can be achieved over emotion recognition systems that are unaware of spontaneity. We achieve state-of-the-art emotion recognition accuracy (4-class, 69.1%) on the IEMOCAP database outperforming several relevant and competitive baselines.

Index Terms: Emotion Recognition, Spontaneous Speech, Multitask Learning.

1. Introduction

Recognizing human emotion is critical for any human-centric system involving human-human or human-machine interaction. Emotion is expressed and perceived through various verbal and non-verbal cues, such as speech and facial expressions. In the recent years, speech emotion recognition has been studied extensively, both as an independent modality [1, 2], and in combination with others [3].

The majority of work on speech emotion recognition follows a two step approach. First, a set of acoustic and prosodic features are extracted, and then a machine learning system is employed to recognize the emotion labels [4, 2, 5, 6, 7]. Although acoustic and prosodic features are more common, the use of lexical features, such as emotional vector, have been also shown to be useful [4]. For recognition, various methods have been proposed - starting from traditional hidden Markov models (HMM) [7] to ensemble classifiers [8], and more recently, deep neural networks [9, 10, 11]. Recently, Abdelwahab and Busso [2] proposed an ensemble feature selection method that addresses the problem of training and test data arising from different distributions. Zong et al. [5] introduced a domain-adaptive least squares regression technique for the same problem. Owing to the latest trends in machine learning, autoencoders [12] and recurrent neural networks (RNN) [9] have also been used for speech emotion recognition. Efforts to improve speech emotion recognition are primarily concentrated on building a better machine learning system. Although spontaneity, fluency and nativity of speech are well studied in the literature, their effect on emotion recognition tasks is not well studied. Work that addressed the problem of distinguishing between spontaneous and scripted speech include acoustic and prosodic feature-based classification [13, 14], and detecting target phonemes [15]. Dufour et al. [14] have also shown that spontaneity is useful for identifying speakers' role [14] by utilizing spontaneity information in their automatic speech recognition system. A recent work by Tian et al. [16] has established that emotional content is essentially different in spontaneous vs. acted speech (prepared, planned, scripted). As they compare emotion recognition in the two types of speech, they observe that different sets of features contribute to the success of emotion classification in spontaneous vs. acted speech. Another study on emotion recognition using convolutional neural network (CNN) has found that type of data (spontaneous or not) does affect the emotion recognition results [10]; however, this work does not use spontaneity information in emotion recognition task. A very recent work has used gender and spontaneity information explicitly in a long short term memory (LSTM) network for effective speech emotion recognition in an aggregated data corpus [11]. Our work differs from this work by providing a detailed analysis and insight towards the effect of spontaneity in emotion recognition in speech, and by proposing an SVM-based hierarchical and multitask learning framework.

In this work, we investigate the usefulness of spontaneity in speech in the context of emotion recognition. We hypothesize that emotional content is interrelated with the spontaneity of speech, and propose to use spontaneity classification as an auxiliary task to the problem of emotion recognition in speech. We investigate two supervised learning settings: (i) a *multilabel hierarchical model* that performs spontaneity detection followed by emotion classification, and (ii) a *multitask learning model* that jointly learns to recognize both spontaneity and emotion in speech and returns two labels. To construct the proposed models, we use a set of standard acoustic and prosodic features in conjunction with support vector machine (SVM) classifiers. We choose SVM because it has been shown to produce results comparable to long short term memory (LSTM) networks when the training dataset is not sufficiently large [16]. Through experiments on the IEMOCAP database [17], we observe that (i) recognizing emotion is easier in spontaneous speech than in scripted speech, (ii) longer context is useful in spontaneity classification, and (iii) significant improvement in emotion recognition can be achieved using spontaneity as an additional information, over spontaneity-unaware systems.

The rest of this paper is organized as follows: Section 2 describes the feature extraction process and the two supervised classification methods we have used for using spontaneity in emotion classification. Section 3 provides details on the experimental setup and results, followed by conclusion in Section 4.

2. Emotion Recognition using Spontaneity

In this section, we propose two models that utilize the spontaneity information in speech to improve emotion recognition: (i) a multilabel hierarchical model that performs spontaneity detection followed by emotion recognition, and (ii) a multitask learn-
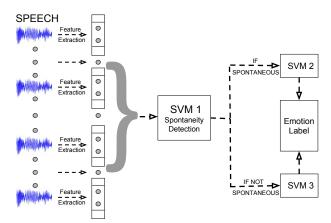
Figure 1: *Multilabel hierarchical emotion recognition using spontaneity*

Figure 2: *Joint emotion and spontaneity classification.*

ing model that jointly recognizes both spontaneity and emotion labels.

2.1. Feature extraction

We extract a set of speech features following the Interspeech2009 emotion challenge [18]. The feature set includes four low level descriptors (LLDs) - Mel-frequency cepstral coefficients (MFCC), zero-crossing rate (ZCR), voice probability (VP) computed using autocorrelation function, and fundamental frequency (F0). For each speech sample, we use a sliding window of length w with a stride length m to extract the LLDs. This generates a k dimensional local feature vector for each windowed segment. Each descriptor is then smoothed using a moving average filter, and the smoothed version is used to compute their respective first order delta coefficients. Appending the delta features, we obtain a local feature vector of dimension $2k$ for every windowed segment. To create a global feature for the entire speech sample, the local features are pooled temporally by computing 12 different statistics (e.g. mean, range, max, kurtosis) along each of the $2k$ dimensions, generating a global feature vector $\mathbf{f} \in \mathbb{R}^d$, $d = 24k$ for each data sample.

2.2. Multilabel hierarchical emotion recognition

Let us consider N training samples and their corresponding feature representations $\mathbf{F} = \{\mathbf{f}_j\}_{i=1}^N$, where $\mathbf{f}_j \in \mathbb{R}^d$. Each training sample with feature vector $\mathbf{f}_j$ is associated with two labels $y_j = \{y_j^s, y_j^e\}$, where $y_j^s \in Y^s$, $Y^s = 0, 1$ represents the binary spontaneity labels, and $y_j^e \in Y^e$, $Y^e = 0, 1, 2, 3$ denotes the emotion labels. Note that only four emotion labels are considered in this paper. We denote the entire label space as $Y = Y^s \times Y^i$.

In order to use the spontaneity information in speech, we propose a simple system which first recognizes if a speech sample is spontaneous or not. An emotion classifier is then chosen based on the decision made by the spontaneity classifier. We divide the entire training set Ω_{train} of N samples into two subsets: Ω_{train}^1 that contains all the spontaneous speech samples, and Ω_{train}^0 that contains all the scripted or the non-spontaneous samples. As shown in Figure 1, we train two separate support vector machine (SVM) classifiers for recognizing emotion using Ω_{train}^0 and Ω_{train}^1. Additionally, we train another SVM for spontaneity detection with sequence length ℓ (denotes the number of consecutive utterances in an input sample) using $\mathbf{F}$ on entire Ω_{train}. The sequence length ℓ is used to account for
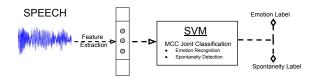
the context needed to recognize spontaneity, which is known to help in emotion recognition [19]. Later, in Section 3.2, we investigate the role of ℓ in spontaneity detection. Note that only the spontaneity classifier uses a sequence length of $\ell = 10$, but the emotion recognition is performed at utterance level.

2.3. Multitask learning for emotion and spontaneity

According to our hypothesis, spontaneity and emotional information in speech are interrelated. We perform the tasks of spontaneity detection and emotion recognition together in a multitask learning framework. Instead of focusing on a single learning task, a multitask learning paradigm shares representations among related tasks by learning simultaneously, and enables better generalization [20]. Following this idea, we jointly learn to classify both spontaneity and emotion. This is posed as a multilabel multioutput classification problem. The basic idea is presented in Fig. 2, where we train a single classifier that learns to optimize a joint loss function pertaining to the two tasks.

We define a weight matrix $\mathbf{W} \in \mathbb{R}^{|Y| \times d}$ containing a set of weight vectors $\mathbf{w}_{\{y^s, y^e\}}$ for classifying each of the $|Y|$ possible label tuples $\{y^s, y^e\}$, where $|\cdot|$ denotes the cardinality of the set. In order to jointly model spontaneity and emotion, we intend to minimize a loss function $\mathcal{L}(\mathbf{W}, Y, \mathbf{F})$ defined as follows.

$$\mathcal{L}(\mathbf{W}, Y, \mathbf{F}) = \frac{1}{2} \sum_{(y^s, y^e) \in Y} ||\mathbf{w}_{\{y^s, y^e\}}||^2 + \mathbf{C} \sum_{j=1}^N \zeta_j \quad (1)$$

The loss function $\mathcal{L}$ is sum of a a regularization loss term $||\mathbf{w}_{\{y^s, y^e\}}||$ and soft-margin loss term (optimization with slacks) i.e., ζ_j. The parameter $\mathbf{C}$ controls the relative balance of the two cost terms. The term ζ_j allows for misclassification of the near-margin training samples while penalizing $\mathcal{L}$ by imposing a loss term that varies on the degree of the misclassification. The optimal classifier weights $\mathbf{W}^*$ are then learned by minimizing the joint loss function $\mathcal{L}(\mathbf{W}, Y, \mathbf{F})$ as

$$\mathbf{W}^* \leftarrow \underset{\mathbf{W} \in \mathbb{R}^{|Y| \times d}}{\arg \min} \mathcal{L}(\mathbf{W}, Y, \mathbf{F}) \quad (2)$$

The classifier is trained i.e., $\mathbf{W}^*$ is learned using the entire Ω_{train} using the same set of features described earlier. Since emotion can vary between two consecutive recordings, the joint model uses a sequence length of $\ell = 1$.

3. Performance Evaluation

We perform detailed experiments on the IEMOCAP database [17] to demonstrate the importance of spontaneity in the context of emotion recognition, and to validate the proposed classification models.

3.1. Experimental setup

Database: We used the *USC-IEMOCAP database* [17] for performance evaluation. It comprises 12 hours of audiovisual data
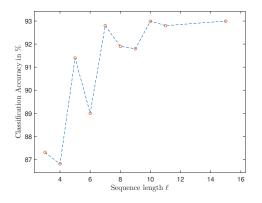
Figure 3: *Effect of varying context (ℓ) on spontaneity classification.*

Table 1: *Effect of features on spontaneity classification accuracy (in %)*

Feature(s) removed	$\ell = 5$	$\ell = 10$
None	91.4	93.0
ZCR	91.0	92.4
VP	90.6	92.6
F0	90.5	92.6
MFCC	83.4	85.5
VP, MFCC	80.7	83.8
F0, MFCC	83.2	84.9
ZCR, MFCC	78.8	82.3
VP, F0	90.6	91.5
VP, ZCR	90.2	92.1
F0, ZCR	90.6	92.2
VP, ZCR, F0	83.7	91.9
Any two, MFCC	< 76	< 80

Table 2: *Effect of keeping only the delta features on spontaneity classification accuracy (in %)*

Feature(s) removed	$\ell = 5$	$\ell = 10$
None	91.4	93.0
ZCR	91.0	92.4
VP	91.0	92.2
F0	90.9	92.7
MFCC	86.8	90.1

along with motion capture (mocap) recordings of face and text transcriptions. The data is collected in 5 different sessions, and each session contains several dyadic conversations. Altogether there are 151 conversations, which are labeled either improvised (spontaneous) or scripted. This serves as our spontaneity label y^s. There are almost equal number of scripted (52.2%) and spontaneous (47.8%) conversations in this database. Each conversation is further broken down to separate samples or utterances, which are organized speaker-wise in a turn-by-turn fashion. All samples are labeled by multiple annotators into one or more of the following six categories - neutral, joy, sadness, anger, frustration and excitement. A single sample may have multiple labels owing to different annotators. In such cases, the final label is chosen to be the label noted by the most annotators and randomly between all the leading labels in case of a tie. We used the four emotion categories: *anger*, *joy*, *neutral*, and *sadness*.

Parameter settings: The features described in Section 2.1 are computed using a sliding window of length $w = 25$ms with a stride of $m = 10$ms. This yields a local feature vector of dimension $k = 15$, and a global feature vector $\mathbf{f}$ of dimension $d = 360$ for each sample. The features are normalized to have values between -1 to 1. The SVMs use the radial basis function (RBF) kernel. All results reported are the average statistics computed over a 5-fold cross validation.

3.2. Understanding spontaneity

Sanity check: We started with the hypothesis that emotional content is different in the spontaneous vs. scripted speech. In order to check this experimentally, we trained an SVM (under the same experimental conditions described in the previous section) on Ω_{train} of the IEMOCAP database that can discriminate among the anger, joy, neutral and sadness. During the test phase, we computed the recognition accuracy for the spontaneous and scripted speech separately. We observe that while the overall accuracy (using the baseline SVM) of emotion recognition is 65.4%, recognition accuracy is higher for speech samples labeled spontaneous i.e. 73.0% and 56.8% for scripted speech (see Table 4). This basic result supports our assumption that emotional content is different in spontaneous vs. scripted speech. This observation is also consistent with the results reported in a CNN-based recent work [10].

Role of context in spontaneity: We performed spontaneity classification to understand the role of various parameters. We investigate the role of context and the contribution of various features in spontaneity detection. We train an SVM classifier with RBF kernel to distinguish between spontaneous and scripted speech using the features described in Section 2.1. At the utterance level, i.e. for $\ell = 1$, this system achieves an average accuracy of **80**%. In order to study the effect of context on spontaneity classification, we vary the sequence length ℓ. To account for longer context, we increase ℓ by concatenating consecutive utterances. Consequently, we concatenate their corresponding global features. This yields a feature vector $\mathbf{F}_\ell \in \mathbb{R}^d$. The variation of classification accuracy with different values of ℓ is shown in Fig. 3. The general trend is that the classification accuracy improves with the longer context (sequence length), and achieves **93**% accuracy for $\ell = 10$. This result can be intuitively explained by the fact that as longer parts of the conversation is used for classification, it becomes easier to detect spontaneity. The result also highlights that spontaneity can be detected with a fairly high accuracy and hence assures us that an additional spontaneity detection module would not harm the overall performance of a speech processing pipleine because of incorrect detection of spontaneity.

Role of features: We investigate the importance of each feature individually in spontaneity classification by performing an ablation study. We exclude one or more of the LLD features at a time, and record the corresponding spontaneity classification accuracy. From the results presented in Table 1, we observe that (i) MFCC features are the most important of all. (ii) Any single LLD feature can provide an accuracy of $\sim 75\%$ indicating that any LLD feature is well suited for the task of spontaneity classification. Moreover, comparing the accuracies achieved when removing both the delta and the actual features (as in Table 1) to removing the actual features but retaining the delta features

Table 3: *Emotion recognition results for individual classes in terms of weighted accuracy (in %).*

	Anger	Joy	Neutral	Sadness
SVM baseline	69.2	37.0	62.9	**76.9**
RF baseline	73.1	6.1	**78.8**	64.6
CNN-based [10]	58.2	**51.9**	52.8	66.5
Rep. learning [21]	53.5	36.9	52.6	64.3
Spontaneity-aware methods				
Hierarchical	**80.2**	37.5	65.9	73.3
Joint	71.2	13.1	75.9	76.3

Table 4: *Emotion recognition results for all classes together in terms of weighted accuracy (in %).*

	Scripted	Spontaneous	Overall
SVM baseline	56.8	73.0	65.4
RF baseline	62.1	66.0	64.1
CNN-based [10]	53.2	62.1	56.1
Rep. learning [21]	-	52.8	50.4
Spontaneity-aware methods			
LSTM [11]	-	-	56.7
Hierarchical	**64.2**	**74.0**	**69.1**
Joint	63.2	69.8	66.1

(see Table 2), we notice that the delta features play a more crucial role than the original features themselves for the task of spontaneity classification.

3.3. Emotion recognition results

To compare the gain from using the spontaneity information, we construct two baselines: an SVM-based emotion classifier, and a random forest (RF)-based emotion classifier. Both classifiers are trained to recognize emotion without using any information about the spontaneity labels. We also compare with two other recent work on emotion recognition: CNN-based emotion recognition [10] and representation learning-based emotion recognition [21]. Additionally, we compare our results with a recent LSTM-based framework [11] that uses gender and spontaneity information for emotion classification. The performances of the proposed spontaneity-aware emotion recognition methods (i.e., hierarchical and joint) along with those for the baselines and other existing methods are presented in Table 3 and Table 4. The proposed hierarchical SVM outperforms the baselines and all other competing methods by achieving an overall recognition accuracy of 69.1%. The joint SVM model achieves an accuracy of 66.1%. Comparing the performance of the baseline SVM with the proposed spontaneity aware SVM methods, we observe that even with the same features and classifier more than 3% improvement for the hierarchical method in overall emotion recognition accuracy is achieved just by adding spontaneity information (see Table 4). Looking at the improvements in individual classes, the class *anger* benefits the most

from spontaneity information. This is evident from a notable 11% improvement in recognition accuracy while using the hierarchical model over the SVM baseline (see Table 3. On the other hand, *neutral* shows 3% improvement, and *joy* is only slightly affected by spontaneity as we compare the hierarchical method to the SVM baseline. *Sadness* does not show any improvement when using spontaneity. The individual emotion recognition accuracies may possibly indicate that anger is a more spontaneous emotion (i.e., difficult to fake) than other emotions, such as sadness. Table 4 shows that recognition accuracy is always lower for scripted speech irrespective of the classification method used. This indicates that emotion is easier to detect in spontaneous speech, and this result is consistent with the observations made in an earlier work [10].

The proposed hierarchical classifier performs slightly better than our joint classifier possibly owing to the more accurate spontaneity classification. Recall that the spontaneity classifier for the hierarchical model used longer context ($\ell = 10$) while the joint model uses $\ell = 1$. Nevertheless, the joint classifier is still of practical use in the scenario when the temporal sequence of the recording is unknown, and hence the sequence length for spontaneity is necessarily constrained.

Clearly, spontaneity information helps emotion recognition. Our SVM-based methods could achieve better result than all competing methods by explicitly detecting and using spontaneity information in speech. The reason behind our SVM-based methods outperforming deep learning-based methods (e.g., CNN-based [10], Rep. learning [21]) is possibly the use of spontaneity and a longer context in the case of hierarchical method. The LSTM-based spontaneity aware method though uses the same four classes as ours, they use an aggregated corpus (using IEMOCAP and other databases) for training the LSTM network. Such training is different from our experimental setting.

4. Conclusion

In this paper, we studied how spontaneity information in speech can inform and improve an emotion recognition system. The primary goal of this work is to study the aspects of data that can inform an emotion recognition system, and also to gain insights to the relationship between spontaneous speech and the task of emotion recognition. To this end, we investigated two supervised schemes that utilize spontaneity to improve emotion classification: a multilabel hierarchical model that performs spontaneity classification before emotion recognition, and a multitask learning model that jointly learns to classify both spontaneity and emotion. Through various experiments, we showed that spontaneity is a useful information for speech emotion recognition, and can significantly improve the recognition rate. Our method achieves state-of-the-art recognition accuracy (4-class, 69.1%) on the IEMOCAP database. Future work could be directed towards understanding the effect of other meta information, such as age and gender.

5. Acknowledgement

6. References

[1] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Spoken Language, 1996. ICSLP 96. Proceedings.,*

Fourth International Conference on, vol. 3.   IEEE, 1996, pp. 1970–1973.

[2] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *ICASSP 2017*, 2017.

[3] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*.   ACM, 2004, pp. 205–211.

[4] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *ICASSP 2015*.   IEEE, 2015, pp. 4749–4753.

[5] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, 2016.

[6] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[7] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1.   IEEE, 2003, pp. I–401.

[8] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE trans. speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[9] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*.   IEEE, 2016, pp. 1–4.

[10] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Interspeech*, 2017.

[11] J. Kim, G. Englebienne, K. Truong, and V. Evers, *Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning*.   International Speech Communication Association (ISCA), 2017, pp. 1113–1117.

[12] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[13] R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linarès, "Spontaneous speech characterization and detection in large audio database," *SPECOM, St. Petersburg*, 2009.

[14] R. Dufour, Y. Estève, and P. Deléglise, "Characterizing and detecting spontaneous speech: Application to speaker role recognition," *Speech communication*, vol. 56, pp. 1–18, 2014.

[15] G. Mehta and A. Cutler, "Detection of target phonemes in spontaneous and read speech," *Language and Speech*, vol. 31, no. 2, pp. 135–156, 1988, pMID: 3256770.

[16] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Int. Conf. Affective Computing and Intelligent Interaction (ACII)*.   IEEE, 2015, pp. 698–704.

[17] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 12 2008.

[18] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[19] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*.   ACM, 2014, pp. 33–40.

[20] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.   ACM, 2004, pp. 109–117.

[21] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *INTERSPEECH*, 2016, pp. 3603–3607.