INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression

Ting Dang[1,2], *Vidhyasaharan Sethu*[1], *Julien Epps*[1,2], *Eliathamby Ambikairajah*[1,2]

[1] School of Electrical Engineering and Telecommunications, UNSW, Sydney, Australia
[2] DATA61, CSIRO, Sydney, Australia

ting.dang@student.unsw.edu.au, v.sethu@unsw.edu.au, j.epps@unsw.edu.au, e.ambikairajah@unsw.edu.au

Abstract

Existing continuous emotion prediction systems implicitly assume that prediction certainty does not vary with time. However, perception differences among raters and other possible sources of variability suggest that prediction certainty varies with time, which warrants deeper consideration. In this paper, the correlation between the inter-rater variability and the uncertainty of predicted emotion is firstly studied. A new paradigm that estimates the uncertainty in prediction is proposed based on the strong correlation uncovered in the RECOLA database. This is implemented by including the inter-rater variability as a representation of the uncertainty information in a probabilistic Gaussian Mixture Regression (GMR) model. In addition, we investigate the correlation between the uncertainty and the performance of a typical emotion prediction system utilizing average rating as the ground truth, by comparing the prediction performance in the lower and higher uncertainty regions. As expected, it is observed that the performance in lower uncertainty regions is better than that in higher uncertainty regions, providing a path for improving emotion prediction systems.

Index Terms: continuous emotion prediction, inter-rater variability, uncertainty, Gaussian Mixture Regression, probabilistic model, pattern recognition

1. Introduction

One approach by which a person's state of emotion can be described is in terms of a few affective attributes, which use numerical values to indicate the type and degree of the emotions. The most widely adopted set of these affective attributes are arousal and valence, since the space constructed by these two dimensions can cover almost all subtle and complex emotion states [1]. Recently, there has been growing interest in continuous emotion prediction that can predict how these affective attribute values change with time over the duration of the utterance, since affect can play a key role in human-computer interaction scenarios and in platforms for mental illness detection such as depression [2, 3].

The emotion prediction problem is generally viewed as a regression problem, where speech waveforms are labeled with a specific numerical value for each affective attribute indicating the short-term emotion intensity. The numerical labels of the speech frames are generally achieved by averaging multi-rater evaluations as perceived by several raters listening to the speech (and watching associated videos if available). However, differences in perception among raters may result in time-varying inter-rater variability. Taking the average of these individual ratings to produce a 'gold standard' forces discrepancies between raters to be ignored. Several studies [4-7] have showed the importance of taking information from multiple raters into account for both categorical emotion recognition and continuous emotion prediction systems, or have argued that emotion attributes should be ranked instead of trying to predict absolute ordinal values. The latter follows since it has been reported that humans are better at rating emotions in relative rather than absolute terms [8-10]. Some of these studies [4, 5] claim that hard labels may not be able to model natural emotion variability. Others [6, 7] consider a multi-task learning framework, which learns each rater's individual track, or adds additional training targets comprising of inter-rater standard deviation. On the other hand, higher performance was achieved when dealing with the relative labels using preference learning compared to conventional classification systems [10]. One shortcoming of these methods is treating the multi-rater information as point estimation instead of a distribution, which comprehensively represents the inter-rater variability.

The overall distribution of the inter-rater variability is able to reflect the uncertainty of speech frames to some extent, meaning a high inter-rater variability indicates a high uncertainty of the speech frame. Incorporating this kind of information in the model for uncertainty prediction can give us insights into the natural variability in human emotion expressions.

The key challenges in estimating the prediction uncertainty directly from annotated speech are finding a probabilistic model for the posterior distribution, which provides the means to estimate the uncertainty information, and to incorporate the inter-rater variability in the model. In terms of the first question of a probabilistic model, commonly used Support Vector Machines [11-13] and Long Short-Term Memory-Neural Networks [14, 15] are not able to handle this problem, since they can only find a point estimate based on one specific structural risk minimization. However, probabilistic models such as Relevance Vector Machines (RVMs) [16-18] and GMR [19] are capable of incorporating a probabilistic description of the target labels and both of them have been shown to be effective in predicting emotions [16, 19]. Considering the incorporation of the inter-rater variability in the model, the multivariate RVM [20] does not perform well when multi-rater evaluations of arousal and valence are modelled as multi-task learning. Moreover, RVMs are constrained to modelling the distribution of the target only as a Gaussian distribution. In this paper, we adopt GMR as the regression model, which can flexibly incorporate the inter-rater variability in the feature concatenation phase by concatenating multi-ratings with the features on a frame basis.

http://dx.doi.org/10.21437/Interspeech.2017-512

2. Methodology

2.1. Conventional GMR

Let $\boldsymbol{X}_n = [\boldsymbol{x}_n^{\mathrm{T}}, \Delta\boldsymbol{x}_n^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{Y}_n = [\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$ represent the features and labels consisting of the static and dynamic information at frame n, where delta is calculated as [19]:

$$\Delta\boldsymbol{x}_n = \frac{\boldsymbol{x}_{n+1} - \boldsymbol{x}_{n-1}}{2};\ \Delta\boldsymbol{y}_n = \frac{\boldsymbol{y}_{n+1} - \boldsymbol{y}_{n-1}}{2} \tag{1}$$

The training features and labels are represented as $\boldsymbol{X} = \left[\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}}, \cdots \boldsymbol{X}_N^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^{\mathrm{T}}, \boldsymbol{Y}_2^{\mathrm{T}}, \cdots \boldsymbol{Y}_N^{\mathrm{T}}\right]^{\mathrm{T}}$, where N represents the total number of frames. The Gaussian Mixture Model (GMM) $\lambda^{[\boldsymbol{Z}]}$ of the joint probability distribution of features and labels is trained using all the joint features $\boldsymbol{Z}_n = [\boldsymbol{X}_n^{\mathrm{T}}, \boldsymbol{Y}_n^{\mathrm{T}}]^{\mathrm{T}}$ by the EM algorithm as in [21].

$$\lambda^{(\boldsymbol{Z})} = \sum_{m=1}^{M} w_m\, N([\boldsymbol{X}, \boldsymbol{Y}]; \begin{bmatrix}\boldsymbol{u}_m^{(\boldsymbol{X})} \\ \boldsymbol{u}_m^{(\boldsymbol{Y})}\end{bmatrix}, [\begin{matrix}\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})}\end{matrix}]) \tag{2}$$

where m is the mixture number and w_m is the weight for each mixture. $\boldsymbol{u}_m^{(\boldsymbol{X})}$ and $\boldsymbol{u}_m^{(\boldsymbol{Y})}$ represent the mean vectors of the m_{th} mixture component for features and labels respectively. The matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{YY})}$ represent the covariance of the m_{th} mixture for features and labels. $\boldsymbol{\Sigma}_m^{(\boldsymbol{XY})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}$ are the cross-covariance matrices of the m_{th} mixture for features and labels. Full covariance matrices are employed to better capture statistical properties of the features and labels.

In order to find label $\boldsymbol{Y}$, the conditional probability is estimated as [21]:

$$P(\boldsymbol{Y}|\boldsymbol{X}, \lambda^{(\boldsymbol{Z})}) = \prod_{n=1}^{N}\sum_{m=1}^{M} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) P(\boldsymbol{Y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}) \tag{3}$$

$$P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) = \frac{w_m N(\boldsymbol{X}_n; \boldsymbol{u}_m^{(\boldsymbol{X})}, \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})})}{\sum_{k=1}^{M} w_k N(\boldsymbol{X}_n; \boldsymbol{u}_k^{(\boldsymbol{X})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{XX})})} \tag{4}$$

$$P(\boldsymbol{Y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}) = N(\boldsymbol{Y}_n; \boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}, \boldsymbol{D}_m^{(\boldsymbol{Y})}) \tag{5}$$

$$\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})} = \boldsymbol{u}_m^{(\boldsymbol{Y})} + \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})-1}(\boldsymbol{X}_n - \boldsymbol{u}_m^{(\boldsymbol{X})}) \tag{6}$$

$$\boldsymbol{D}_m^{(\boldsymbol{Y})} = \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} - \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})-1}\boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \tag{7}$$

It can be seen that $P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})$ is also a GMM for each frame n [21]. The time sequence $\hat{\boldsymbol{y}} = [\boldsymbol{y}_1^{\mathrm{T}}, \boldsymbol{y}_2^{\mathrm{T}}, \cdots \boldsymbol{y}_N^{\mathrm{T}}]^{\mathrm{T}}$ is estimated based on maximizing the function in (3) over consecutive frames and the EM algorithm is generally applied [21].

A good approximation to the EM algorithm [21, 22] with the dominant mixture sequence $\hat{\boldsymbol{m}}$ has been shown to be effective in voice conversion systems [21]. The likelihood in (3) can be approximated with a mixture component sequence for each frame as:

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{argmax} P(\boldsymbol{m}|\boldsymbol{X}, \lambda^{(\boldsymbol{Z})}) \tag{8}$$

Then the label $\hat{\boldsymbol{y}}$ can be estimated based on the mixture component sequence shown as:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{argmax} P(\hat{\boldsymbol{m}}|\boldsymbol{X}, \lambda^{(\boldsymbol{Z})})\, P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda^{(\boldsymbol{Z})}) \tag{9}$$

Instead of considering $P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})$ as a GMM, the approximate algorithm takes the approach of adopting the dominant Gaussian mixture component as the posterior probability. Our preliminary experimental results indicate that the approximate algorithm gives comparable results to the EM algorithm, and consequently the approximate algorithm is utilized in this paper. The replacement of a GMM with a single Gaussian distribution provides a convenient means to estimate the uncertainty as the standard deviation of the dominant mixture component for each frame n.

2.2. Proposed multi-rater GMR

2.2.1. Incorporating multi-rater variability

To incorporate multi-rater variability into the model, each individual rating $\boldsymbol{Y}_{nk}$ is concatenated with the frame-based features $\boldsymbol{X}_n$, where k is the rater number. For each frame n, all the K ratings per frame are utilized to construct the K sets of joint features as:

$$\boldsymbol{Z}_{nk} = [\boldsymbol{X}_n^{\mathrm{T}}, \boldsymbol{Y}_{nk}^{\mathrm{T}}]^{\mathrm{T}} \tag{10}$$

All the training joint vectors are given as:

$$\boldsymbol{Z} = [\boldsymbol{Z}_{11}^{\mathrm{T}}, \boldsymbol{Z}_{12}^{\mathrm{T}}, \cdots \boldsymbol{Z}_{1K}^{\mathrm{T}}, \boldsymbol{Z}_{21}^{\mathrm{T}}, \boldsymbol{Z}_{22}^{\mathrm{T}}, \cdots, \boldsymbol{Z}_{NK}^{\mathrm{T}}]^{\mathrm{T}} \tag{11}$$

The joint GMM is trained similarly to (2) using the new joint features $\boldsymbol{Z}$. By doing this, the joint model $\lambda^{(\boldsymbol{Z})}$ is able to capture the variability introduced by all the possible ratings of each frame in the label dimension. The conditional posterior $P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})$ can similarly be estimated as in (3)-(7) where $\boldsymbol{Y}_n = [\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$. $P(\boldsymbol{Y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})})$ in (5) can be also rewritten as:

$$\begin{aligned} P(\boldsymbol{Y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}) &= N(\boldsymbol{Y}_n; \boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}, \boldsymbol{D}_m^{(\boldsymbol{Y})}) \\ = N\left([\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}; \begin{bmatrix}\boldsymbol{u}_m^{(\boldsymbol{y}_n)} \\ \boldsymbol{u}_m^{(\Delta\boldsymbol{y}_n)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n\boldsymbol{y}_n)} & \boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n\Delta\boldsymbol{y}_n)} \\ \boldsymbol{\Sigma}_m^{(\Delta\boldsymbol{y}_n\boldsymbol{y}_n)} & \boldsymbol{\Sigma}_m^{(\Delta\boldsymbol{y}_n\Delta\boldsymbol{y}_n)}\end{bmatrix}\right) \end{aligned} \tag{12}$$

Since the aim of the multi-rater system is to find the uncertainty information related to the label $\boldsymbol{y}_n$ instead of $\boldsymbol{Y}_n = [\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$, we can obtain $P(\boldsymbol{y}_n|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})$ by marginalizing $P([\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})$ over $\Delta\boldsymbol{y}_n^{\mathrm{T}}$ as:

$$\begin{aligned} P(\boldsymbol{y}_n|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) &= \int_{\Delta\boldsymbol{y}_n} P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})\, d(\Delta\boldsymbol{y}_n) \\ = \sum_{m=1}^{M} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) \int_{\Delta\boldsymbol{y}_n} P([\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})})\, d(\Delta\boldsymbol{y}_n) \\ = \sum_{m=1}^{M} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) P(\boldsymbol{y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}) \end{aligned} \tag{13}$$

where

$$P(\boldsymbol{y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}) = N(\boldsymbol{y}_n; \boldsymbol{u}_m^{(\boldsymbol{y}_n)}, \boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n\boldsymbol{y}_n)}) \tag{14}$$

The parameters of $\boldsymbol{u}_m^{(\boldsymbol{y}_n)}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n\boldsymbol{y}_n)}$ can be calculated using the $\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}$ and $\boldsymbol{D}_m^{(\boldsymbol{Y})}$ terms in (6) and (7).

2.2.2. Uncertainty estimation and evaluation

During the test phase, $P(\boldsymbol{y}_n|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})$ is calculated for each frame based on the approximated algorithm described earlier. For each frame n, the suboptimal mixture component $\hat{m}_n$ is firstly estimated as:

$$\hat{m}_n = arg\max_{m_n} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) \tag{15}$$

This then allows for the estimation of $\hat{\boldsymbol{y}}_n$, the expected value of $\boldsymbol{y}_n$, and its standard deviation $\hat{\sigma}_n$ as a time-varying indicator of the prediction uncertainty in $\hat{\boldsymbol{y}}_n$ for each frame as:

$$\hat{\boldsymbol{y}}_n = \boldsymbol{E}[\boldsymbol{y}_n|\boldsymbol{X}_n, \hat{m}_n, \lambda^{(\boldsymbol{Z})}] = \boldsymbol{u}_{\hat{m}_n}^{(\boldsymbol{y}_n)} \tag{16}$$

$$\hat{\sigma}_n = \boldsymbol{D}[\boldsymbol{y}_n|\boldsymbol{X}_n, \hat{m}_n, \lambda^{(\boldsymbol{Z})}] = \boldsymbol{\Sigma}_{\hat{m}_n}^{(\boldsymbol{y}_n\boldsymbol{y}_n)} \tag{17}$$

It should be noted that the covariance mixture $\boldsymbol{D}_m^{(\boldsymbol{Y})}$ in (7) is fixed for each mixture m across all frames and does not vary with the frame-based test features $\boldsymbol{X}_n$. Consequently, the standard deviation $\hat{\boldsymbol{\sigma}}_n$ will only take one of M distinct values as we adopt the dominant mixture in each frame n. This will result in a quantized standard deviation $\hat{\boldsymbol{\sigma}}_n$ for the uncertainty prediction. The quantization can be improved by increasing GMM mixture number, but this is constrained by limited training data. The RECOLA database used in this paper contains 90 minutes of speech and can only be used to reliably train GMMs of 16 mixtures or less.

This new paradigm aims to predict the uncertainty of emotional speech, as opposed to conventional point estimation that indicates the exact emotion intensity. Incorporation of the overall distribution of inter-rater variability into the model and the exploration of GMR for uncertainty prediction has not yet been investigated in continuous emotion systems. The proposed framework reveals the time-varying nature of human emotional expressions by explicitly modelling the level of prediction certainty.

3. Database

The RECOLA database [23], used in the experiments reported in this paper, is a multimodal database in French. It contains audio, video and physiological signals, where acted and spontaneous interactions are collected in a remote collaborative framework. Speech data from 18 speakers was equally divided into training and development partitions with 9 speakers in each partition, which is identical to the partitions used in the Audio-Visual Emotion Recognition Challenge (AV+EC 2016) [24]. Test partitions were not utilized since the affective attributes were not released. All the experiments were conducted using the training dataset to train the model and evaluated on the development set.

The annotation was performed every 40ms by six gender-balanced raters for arousal and valence with values between -1 and 1. The ground truth used to compare with the multi-rater system was the average value computed over six raters.

4. Experimental Results

4.1. Experimental settings

65 low-level descriptors (LLDs) and their first-order derivatives were extracted using OpenSMILE [25], using the same LLDs and delta features as [26]. Two second windows with 40 ms shift were used to compute the statistical features by applying five functionals: maximum, minimum, mean, standard deviation, and range. Dynamic features and labels were calculated as in [19]. PCA was used to conduct dimensionality reduction in the feature space from 650 dimensions to 40 dimensions, preserving approximately 85% of the data variance in the training dataset [19, 27]. There were 82 final feature dimensions. The first 80 features consisted of 40 dimensional static features $\boldsymbol{x}_n$ after PCA and 40 dimensional dynamic features $\Delta\boldsymbol{x}_n$ calculated on $\boldsymbol{x}_n$, and the final 2 were label features $[\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$. The reason for using 80 feature dimensions was to preserve enough feature variability in the training dataset and to provide a sufficiently high dimensionality to train the system with a large number of parameters for GMM. Delays of 4 s for arousal and 2 s for valence were applied during the training phase, based on a previous study [17]. The delay thus introduced in the predicted uncertainty was compensated by removing the corresponding frames. GMMs with different numbers of mixture components, ranging from 4 to 32, were tested using HTK.

4.2. Performance of uncertainty estimation

Under the assumption that a high inter-rater variability $\tilde{\sigma}_n$ will result in a high predicted uncertainty $\hat{\sigma}_n$, we aim to investigate the positive correlation between the predicted standard deviation $\boldsymbol{\sigma} = [\hat{\sigma}_1, \hat{\sigma}_2, \cdots \hat{\sigma}_N]$ and the multi-rater standard deviation $\tilde{\boldsymbol{\sigma}}$ calculated from six ratings. Pearson's correlation coefficient (CC) was computed on 9 evaluation utterances individually. The final performance measure was mean correlation coefficient averaged over 9 evaluation utterances. The results are shown in Table 1. A moving averaging filter was utilized to smooth the predicted standard deviation for each utterance with the optimal window size determined from [100, 800] experimentally. It should be mentioned that we do not have a specific baseline for a direct comparison owing to the new paradigm. Therefore, we mainly aim to reveal the essence of the uncertainty.

Table 1: *Mean CC computed between* $\boldsymbol{\sigma}$ *and* $\tilde{\boldsymbol{\sigma}}$

		No Smoothing		Smoothing	
		Arousal	**Valence**	**Arousal**	**Valence**
Mixture	4	0.3530	0.0461	0.5173	0.0890
	8	***0.4097***	0.0937	***0.5684***	0.1322
	16	0.3998	0.0457	0.5350	0.0745
	32	0.3872	0.0476	0.4410	0.0881

The best performance of 0.4097 and 0.5684 were achieved with 8 mixture components before and after smoothing, and there was not much variation between different mixture numbers for arousal. This indicates the positive correlation between the predicted $\boldsymbol{\sigma}$ and the inter-rater $\tilde{\boldsymbol{\sigma}}$ to some extent.-. However, the performance of valence is worse for all mixture numbers, agreeing with previous studies [17] that showed valence is well predicted by video signal. Therefore, we mainly focus on analyzing the uncertainty prediction for arousal.

In order to have a comprehensive understanding of the predicted uncertainty, we investigated the scatter plot of the predicted $\boldsymbol{\sigma}$ and the inter-rater $\tilde{\boldsymbol{\sigma}}$ over the entire test dataset as shown in Figure 1. We can observe a positive correlation between the predicted $\boldsymbol{\sigma}$ in x axis and the inter-rater $\tilde{\boldsymbol{\sigma}}$ in y axis. It indicates a relatively strong correlation between the predicted $\boldsymbol{\sigma}$ and the inter-rater $\tilde{\boldsymbol{\sigma}}$, which means a high inter-rater variability will result in a high uncertainty in prediction.

In addition, one segment of the ratings from 6 raters of speaker 2 is shown in Figure 2, which displays the uncertainty of the emotion prediction changing over time. Only one speech segment from speaker 2 is shown in order to reduce clutter but the predicted uncertainty was observed to be generally consistent with the inter-rater variability across all speakers and speech segments. The grey error bar shows $\pm\hat{\sigma}_n$, the predicted standard deviation from the expectation value $\hat{\boldsymbol{y}}_n$ for each frame n as shown in (16)-(17). Six colored lines indicate the individual raters respectively. It can be seen that the speech segments with higher inter-rater variability are associated with higher variability in predicted estimates and vice versa.

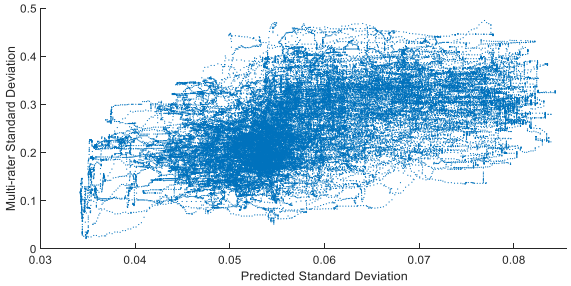

Figure 1*: Scatter plot of predicted* $\boldsymbol{\sigma}$ *and inter-rater* $\tilde{\boldsymbol{\sigma}}$.

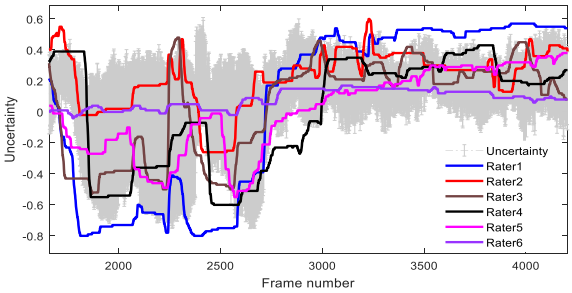

Figure 2: *Uncertainty plot for speaker 2 using 8 mixtures.*

4.3. Correlation between uncertainty and conventional emotion prediction for arousal

In order to gain an in-depth understanding of this paradigm for support of the speech based prediction for arousal, we also analyzed the performance of conventional emotion prediction systems that use the average rating as the ground truth, by taking the uncertainty information into account. Generally, low inter-rater variability, as represented by low predicted uncertainty, indicates that raters were more in agreement about the emotion expressed in those speech segments. Thus the prediction should be easier to make accurately. Given the inter-rater variability or the predicted uncertainty information, we have analysed the conventional emotion prediction performance for arousal by segmenting the speech frames into two regions: low variability regions where speech segments are easier to predict, and high variability regions, where speech segments are harder to predict.

In terms of defining low and high variability regions, the predicted uncertainty should be considered instead of the inter-rater variability since the latter is generally not accessible during the test phase in real scenarios. We propose using the percentiles based on the histogram of predicted $\hat{\sigma}_n$ to determine the low and high variability regions based on the performance given in Table 1. Speech segments with $\hat{\sigma}_n$ smaller than the value of ρ percentiles and higher than the value of $(100-\rho)$ percentiles were clustered as low and high variability regions respectively. Five thresholds of $\rho \in [10, 50]$ with step increase of 10 were investigated. However, when compared to the ground truth of the truly low and high variability regions similarly defined by the histogram of inter-rater variability, we found that the percentage of correctly predicted uncertainty frames in low variability regions is relatively low when compared to the truly low variability regions similarly defined by ground truth as shown in Figure 3. Here, the black line indicating the percentage of correctly predicted uncertainty frames in low variability regions, is relatively low when using $\rho = 10$ and $\rho = 20$. Increasing the threshold results in more correctly predicted uncertainty frames, showing that the predicted uncertainty is more condensed in the middle region. Thus, segmenting low and high variability regions only based on the predicted $\hat{\sigma}_n$ may result in unreliable comparison of the conventional emotion systems. Therefore, the truly low and high regions defined by inter-rater variability were used as a reference for the predicted $\hat{\sigma}_n$ to define low and high variability regions, serving as an initial analysis of the correlation between the uncertainty and conventional point estimation.

To avoid misleading speech segments clustered to low and high variability regions, only accurately predicted uncertainty segments are selected, by finding the intersection segments that appear in both the truly low/high and the predicted low/high variability region. Finally, the conventional point estimation and average ground truth of the selected segments are concatenated in low and high variability regions respectively, which are used to compute CC in low and high variability regions shown in Table 2.

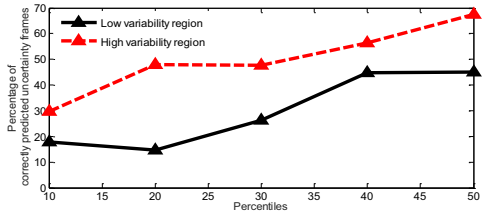

Figure 3*: Percentage of correctly predicted uncertainty frames in low and high variability regions with 8 mixtures.*

Table 2: CC *on low and high variability regions based on the intersection of ground truth* $\tilde{\boldsymbol{\sigma}}$ *and predicted* $\boldsymbol{\sigma}$

Region	Percentiles				
	10th	20th	30th	40th	50th
Low	***0.8013***	***0.7055***	***0.6891***	***0.6905***	0.6885
High	0.3787	0.3837	0.4829	0.6125	0.6905

As expected, it can be seen that the performance in the lower variability regions is in general much better than in the higher variability regions when they are defined using 10th-40th percentiles. In addition, the conventional arousal prediction system had a CC of 0.7990 computed over the entire test dataset when using the mean of the labels as the ground truth using 8 mixtures. It should be noted that the comparison between the conventional performance using entire test dataset and Table 2 is not a direct comparison, since a different number of speech frames is used to compute the CC for each.

5. Conclusion

This paper proposes a novel paradigm that is able to incorporate the uncertainty information of speech frames by explicitly accounting for multi-rater variability in the system. The results of this investigation show the effectiveness of the proposed method for uncertainty prediction. The second interesting finding is the high correlation between the uncertainty and the emotion prediction achieved by using the average value over multiple raters. The predictions are more reliable in the lower inter-rater variability regions than that in the higher inter-rater variability regions. As the first study to analyze the uncertainty related to emotion prediction, these findings provide more insights into the time-varying variability introduced in emotion prediction by multiple raters, and opens a new avenue for improving emotion prediction. However, further studies need to be carried out to address the quantisation of uncertainty predictions.

6. References

[1] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 827-834: IEEE.

[2] S. A. Langenecker, L. A. Bieliauskas, L. J. Rapport, J.-K. Zubieta, E. A. Wilde, and S. Berent, "Face emotion perception and executive functioning deficits in depression," *Journal of Clinical and Experimental Neuropsychology,* vol. 27, no. 3, pp. 320-333, 2005.

[3] C. A. Mazefsky and D. P. Oswald, "Emotion perception in Asperger's syndrome and high-functioning autism: The importance of diagnostic criteria and cue intensity," *Journal of autism and developmental disorders,* vol. 37, no. 6, pp. 1086-1095, 2007.

[4] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Proc. Interspeech*, 2016.

[5] E. Mower *et al.*, "Interpreting ambiguous emotional expressions," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1-8: IEEE.

[6] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems (TiiS),* vol. 2, no. 1, p. 6, 2012.

[7] F. Ringeval *et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters,* vol. 66, pp. 22-30, 2015.

[8] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, 2015, pp. 574-580: IEEE.

[9] G. N. Yannakakis and H. P. Martínez, "Ratings are Overrated!," (in English), *Frontiers in ICT,* Mini Review vol. 2, no. 13, 2015-July-30 2015.

[10] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!," *IEEE transactions on affective computing,* vol. 5, no. 3, pp. 314-326, 2014.

[11] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV-1085-IV-1088: IEEE.

[12] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home,* vol. 6, no. 2, pp. 101-108, 2012.

[13] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *INTERSPEECH*, 2003: Citeseer.

[14] M. Wöllmer *et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech*, 2008, vol. 2008, pp. 597-600.

[15] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing,* vol. 31, no. 2, pp. 153-163, 2013.

[16] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing,* vol. 30, no. 3, pp. 186-196, 2012.

[17] Z. Huang *et al.*, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 41-48: ACM.

[18] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of machine learning research,* vol. 1, no. Jun, pp. 211-244, 2001.

[19] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 2288-2291: IEEE.

[20] A. Manandhar, K. D. Morton, P. A. Torrione, and L. M. Collins, "Multivariate Output-Associative RVM for Multi-Dimensional Affect Predictions," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 10, no. 3, pp. 439-446, 2016.

[21] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.

[22] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, vol. 1, pp. I/9-I12 Vol. 1: IEEE.

[23] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 2013, pp. 1-8: IEEE.

[24] M. Valstar *et al.*, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3-10: ACM.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462: ACM.

[26] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[27] H. Khaki and E. Erzin, "Continuous emotion tracking using total variability space," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.