An Investigation of Emotion Changes from Speech

Zhaocheng Huang

1 School of Electrical Engineering and Telecommunications, UNSW Australia
2 ATP Research Laboratory, National ICT Australia (NICTA), Australia

zhaocheng.huang@student.unsw.edu.au

Abstract

Emotion recognition based on speech plays an important role in Human Computer Interaction (HCI), which has motivated extensive recent investigation into this area. However, current research on emotion recognition is focused on recognizing emotion on a per-file basis and mostly does not provide insight into emotion changes. In my research, emotion transition problem will be investigated, including localizing emotion change points, recognizing emotion transition patterns and predicting/recognizing emotion changes. As well as being potentially important in applications, the research delving into emotion changes paves the way towards a better understanding of emotions from engineering and potentially psychological perspectives.

Index Terms: Emotion changes, emotion change detection, emotion change modeling, speech based affective computing.

1. Introduction

Speech based emotion recognition, as an important part of Human Computer Interaction (HCI) has been extensively studied [1] and many frameworks were proposed [2, 3]. Most of them either emphasize on correctly differentiating emotion categories (e.g. neutral, anger, etc) or to predict emotion primitives (e.g. arousal, valence and dominance). However, these do not provide insights into emotion changes, which might be as important as emotion recognition for some applications such as emotional intelligence [4], detecting task change in real time HCI, surveillance [5] and detecting the onset of emotional outbursts in large speech databases. Changes among emotions are informative for understanding emotions and consequently self-regulating emotions and behavior [4]. They also potentially facilitate understanding of task transition, where cognitive load may change quickly between tasks of different types.

Among the most relevant previous work about emotion changes, Niedenthal et al. [6] studied the emotion transitions between happiness, sadness and neutral at the desire to understand why smiles drop and emotion congruence, based on facial expression in 2001. Subsequently, An interesting HMM-like mental state transition framework describing transition probabilities between and within emotions was proposed by Jiang et al. [7]. Recently, although for emotion recognition, emotional sub-state language model for capturing temporal evolution of emotional expressions was shown to be potentially interesting and useful [8]. The temporal evolution in the paper refers to onset, offset and apex of emotional intensity in speech, which is not uncommon in facial expression research where it is believed that facial changes can mirror changes in emotions (i.e. onset, offset and apex) [9].

Employing a similar temporal evolution pattern, Van Der Wal et al [10] first classified increase, decrease and no change with respect to emotional intensity for 16 emotions, albeit on very small database. In parallel with the increasing popularity of emotion dimensions within emotion recognition community, there is a small but increasing amount of work done tracking or detecting emotion changes on arousal and valence [11-13]. This is, in part, due to the fact that dimensional representation of emotions offers a measure of emotional intensity and a good way to avoid confusions in emotion categories.

It can be discerned that there has been increasing interest in emotion change research. Nevertheless, there has not been a systematic investigation of emotion transition in depth. In my research, there are two main research questions with respect to emotion changes, which will be discussed in greater details in Section 3:

- Can we correctly localize emotion change points in time?
- Can we model emotion changes to achieve reliable emotion change-related classification and regression? This potentially includes recognizing emotion change patterns, identifying “delta emotion” and predicting emotion changes.

2. Related Work

The first phase in my research is emotion change detection. To the best of our knowledge, few researchers have focused specifically on localizing the point in time at which emotions transition from category to the next. A similar problem setting can be found when turning our attention to the speaker change detection literature, where people aim to segregate homogenous segments for each speaker from a meeting or conversation by finding speaker change points. In 1998, Chen et al proposed a classic method based on Bayesian Information Criterion (BIC), a metric-based method for model selection [14] popular in speaker diarization. BIC was used within a 1 second sliding window based on an assumption that the window can be well modeled as two Gaussian distributions rather than one if there is a change point inside, and be better modeled as one Gaussian otherwise. Another method proposed for speaker change detection is to use two consecutive fixed-length windows, modelling each by GMMs adapted from a universal background model, and finding differences between them. Many distance-based methods, such as the Generalized Likelihood Ratio (GLR), Kullback-Leibler (KL) divergence, and Cross Log Likelihood Ratio (CLLR) have been investigated [15]. A number of interesting methods such as one class SVM [16], Variational Bayes free energy [17] and factor analysis [18] have also been used. However, these require large databases to be effective. If there is a change
point, the value of distance-based measures will be higher than in locations where there are no change points. This is a reasonable assumption, as changes in speaker identity lead to changes in models. However, this may be more challenging for detecting emotion change, as confounding factors such as speaker and phonetic variability will degrade performance. Moreover, in a more general context, some nonparametric methods such as statistical test (e.g. Martingale test [19]) are also available for the change-point detection task.

Outside of the initial investigation done by Niedenthal [6], there has been very little work done in investigation of emotion transition until last few years. An search of literature about emotion changes shows that there has been an increasing interests regarding emotion changes either in emotion categories [5-8] or in dimensional primitives [10-13]. However, they focused on emotion changes from different perspectives. For example, In [7], the Probabilistic HMM-like structure for emotion transitions is somehow inspiring for emotion change modeling in the sense that within the framework, dependencies between various types of emotion changes can be modeled. However, their psychological experiment for collecting the data of emotion transition was not very reliable, as people who were under certain emotional state were asked to fill questionnaires by just using their imagination. Wei et al. [8], employed a very interesting emotional sub-state language model for emotion recognition based on speech, which apart from modeling emotional states, temporal evolution of emotions expressions including onset, offset and apex of emotional intensity were taken into consideration and modeled by a bigram language model. This framework is attractive for two reasons: One is the idea of onset, offset and duration with respect to emotional intensity in speech, which enables us to investigate emotion from a change perspective. The second attraction is the use of HMM, which can capture the temporal information of changes in emotions and could be potentially combined with the mental state transition network in [7]. The aforementioned research does hint at some possibilities of emotion change modeling via dynamic probabilistic modeling methods such as HMM for emotion change classification and delta emotion recognition.

Unlike emotion changes in categorical emotions, some studies investigated emotion changes in dimensional attributes such as arousal, valence and dominance. Examples of such investigation include [11, 12]. Metallinou and others [11] investigated emotion change tracking in arousal, valence and dominance using body language and prosodic cues. They show that tracking emotion changes is a challenging problem. Lade et al were among the first group of people to apply topic model into emotion recognition and emotion change detection [12]. Latent Dirichlet Allocation (LDA) based topic models have been extensively studied in text mining, where a document will be represented as a set of probabilities of being associated with different topics. A topic is a set of words that co-occur in documents. These topics change when the emotion changes and significant changes in attributes can be detected. Despite of lack of ground truth of changes and clarity of the time step they used, this method is useful for capturing the temporal information when emotion changes and also has the potential to be used in emotional categories. In [13], the authors studied continuous emotion regression from a newer perspective, where they introduced trend loss using order loss into their cost function, which is optimized by Gradient Descent algorithm with Neural Network as the model. As the cost function is a trade-off between prediction error and trend loss, the reasonable improvement in preserving trends of dimensions is at the expense of higher prediction error. Despite this, the study and those above are potentially interesting for predicting emotion changes/trends.

3. Proposed Research

3.1. Overview

![Image of Emotion Change System](image-url)

Figure 1: Overview diagram of Emotion Change System. Emotion change detection and Emotion change Modeling constitutes the main part of my research.

An overview of the proposed emotion change system can be seen in Figure 1. My research will be mainly focusing on emotion change detection and emotion change modeling, as shown in the red block. Detecting salient emotion transition points could potentially be investigated using emotion categories (neutral, anger, happiness, etc) or dimensions (high vs low in arousal or valence), i.e. two key representation methods for emotion research. This second part is in general called emotion change modeling, as we want to have a close look at the internal structure between changes among various emotions, constituting of three possibilities, namely:

- **Emotion Change Recognition:** Recognizing emotion change patterns such as (neutral to anger or high to low in arousal), rather than signal emotion recognition/prediction.
- **Delta emotion recognition:** Delta emotion can be a description of changing emotions such as if a person is becoming more happy or more angry, or his/her arousal is increasing.
- **Emotion change/trend prediction:** Predicting emotion change values in arousal and valence. It would be interesting to compare emotion change prediction and absolute emotion prediction.

3.2. Methodology

3.2.1. Databases

In my research, the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database [20] and the SEMAINE database [21] were selected, because these two databases are reasonably large compared to existing English databases publicly available in the emotion recognition community. For IEMOCAP, it comprises 12 hours of emotional speech from ten speakers in scripted or spontaneous spoken conversational scenarios. Also, short utterances with utterance level emotional labels and numerical ratings are available, which can be used to build an emotion change database for initial investigation by concatenating same-speaker emotional utterances, as shown in Fig. 2. In the SEMAINE database, continuous numerical ratings of emotion dimensions, annotated by 6-8 raters are available in solid SAL recordings that are transcribed. This could be used for emotion change...
3.2.2. Features

Frame-level features such as prosodic, spectral and voice quality features, and utterance-level functionals (e.g. mean, percentiles, etc) have proven to be effective for emotion recognition [2]. In our initial investigation of emotion change detection [22], 13 dimensional MFCCs and their derivatives were used to achieve promising results due to their popularity, effectiveness and stable acoustic property when it is combined with Gaussian Mixture Models (GMMs). It is a fair assumption that by employing a wider variety of features, including prosodic and voice quality features alongside MFCCs, the performance might be further improved [2]. Moreover, functionals are potentially useful in the dual windowing framework.

3.2.3. Emotion Change Detection

For emotion change detection task, we used a sliding dual windowing framework, as shown in Fig. 3, comprising previous and current fixed-length windows.

![Figure 3: Emotion change detection dual sliding window.](image)

The window center represents a candidate change point. During each window, features from multiple frames will be extracted.

Within these two windows, spanning multiple frames, features are extracted on a per-frame basis. During the detection stage, a score representing the difference between the two windows will be compared with a threshold to produce a decision. If a score above the threshold is located in the tolerance region around the true change point, a change is correctly detected (as seen in Fig. 3). Other scores above the threshold outside tolerance regions are considered false alarms. To employ this paradigm for detecting changes in emotions, three parameters are required: (a) window size, (b) window shift and (c) the length of the tolerance region.

3.2.4. Emotion Change Modeling

A hypothesis about emotion changes is that the internal structure of emotion change patterns could be modeled and used for classification. The candidate modeling methods are Hidden Markov Models (HMM) [8] and Dynamic Topic Models [23]. This pattern could either be change patterns such as changes from neutral to anger or delta emotions such as more angry or an increase in arousal/valence. In my preliminary investigation, it was found that features around emotion change points are informative for change detection. These features can be investigated around these points to understand whether they underlie some emotion change patterns, towards recognizing them. This could be achieved by either using conventional probabilistic models and temporal features or employing dynamic probabilistic methods alongside with emotion-related features. However, an issue facing us is the limitation of databases with enough emotion changes. Likewise, delta emotion recognition in emotional intensity also requires large databases. A reasonable compromise, at the beginning, is to investigate delta emotion recognition in emotion dimensions, where the task can be regarded as a three-class classification, i.e. increase, decrease and unchanged in arousal/valence, mainly because the IEMOCAP and SEMAINE databases may not be large enough for the emotion change classification and delta emotion recognition in emotion categories.

The Third challenge is to predict emotion changes/trends, meaning predicting delta arousal/valence either in frame-level (changes) or in utterance level (trends), which can be interesting to compare with continuous absolute emotion prediction. This problem setting, at the same time, will make the most of the two databases, as IEMOCAP can provide changes in utterance level, while SEMAINE has frame-level ratings. The initial investigation into this problem will start with conventional regression methods such as Support Vector Regression (SVR) and Long Short Term Memory (LSTM). Moreover, as changes are highly related to the previous and future information, some techniques that are able to learn this information are especially attractive for this problem. Examples of these methods include Output Associative Relevance Vector Machine (OA RVM) and Bidirectional Long Short Term Memory (BLSTM).

4. Progress

![Figure 4: EER vs Tolerance lengths for the general emotion change detection task (Four emotions)](image)
As well as empirically determining the three emotion change detection parameters, the initial investigation employed the proposed dual-windowing scheme to detect emotion change points among four emotions using the well-known Generalized Likelihood Ratio (GLR) and proposed a new Emotion Pair Likelihood Ratio (EPLR) measure [22]. Our results show that emotion change points can be effectively detected when the proposed methods are combined, gave an EER as low as 20.8%. Currently, I am moving towards next phase of my research, which is emotion change modeling, trialing on the IEMOCAP database and the SEMAINE database.

5. Expected Contributions

To address currently unanswered problems, my research is expected to be the first systematic investigation of emotion changes, compared with existing emotion recognition from speech research. This will contribute to the affective computing research community in terms of new insights towards emotion change problems. Also it can benefit a range of research areas such as task transition where cognitive load is interfered by emotion changes, and change quickly [24], emotion regulation where timing of recognizing emotion change is essential [4] and robust emotion recognition where temporal information facilitates recognizing emotions [8]. Practically, emotion change research can be applied in a range of applications. An example of these is that if emotion change points can be correctly detected, it might be much more computationally inexpensive than conventional emotion recognition, especially in spontaneous data where neutral emotion constitutes nearly 90% of the emotions, as unwanted continuous recognition of the same emotion will be replaced by emotion change detection. Another example is that if emotion change modeling could be achieved, Human Computer Interaction (HCI) will be more effective by informing people the changes so that they can react correspondingly, which is more intelligent and user-friendly.

6. Reference