An Investigation of Emotion Change Detection from Speech

Zhaocheng Huang\textsuperscript{1,2}, Julien Epps\textsuperscript{1,2}, Eliathamby Ambikairajah\textsuperscript{1,2}

\textsuperscript{1} School of Electrical Engineering and Telecommunications, UNSW Australia
\textsuperscript{2} ATP Research Laboratory, National ICT Australia (NICTA), Australia
zhaocheng.huang@student.unsw.edu.au, j.epps@unsw.edu.au, ambi@ee.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 this paper, we report on an initial investigation into detecting the instant of emotion change using Gaussian Mixture Models (GMM) based methods, either without or with prior knowledge of emotions: the Generalized Likelihood Ratio and Emotion Pair Likelihood Ratios, together with a novel normalization scheme to improve emotion change detection accuracy. Experimental results based on the IEMOCAP corpus are presented that demonstrate a promising baseline. Despite the challenging nature of the problem, this work provides a path towards systems that detect and understand emotion changes, and also presents very interesting questions for further investigation.

Index Terms: Emotion change detection, Gaussian mixture model, generalized likelihood ratio

1. Introduction

Emotion recognition based on speech has been extensively studied since becoming regarded as an important part of Human Computer Interaction (HCI) [1]. To achieve this, many frameworks were proposed, for example [2-5]. Most of them aim to recognize emotions individually in a categorical manner (e.g. neutral, anger, etc) or a dimensional manner (e.g. arousal, valence and dominance). However, one interesting question that has received less attention is the automatic detection of the instant of emotion change, which might be as important as emotion recognition for some applications such as emotional intelligence, detecting task change in real time HCI, surveillance [6] 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, which is a major research area of contemporary affective science [7, 8]. It also facilitates understanding of task transition, where cognitive load may change quickly between tasks of different types [9].

By correctly detecting emotion change points, not only can we realize the aforementioned applications, but also provide a path for further study on the nature of emotion changes. Nevertheless, to the best of our knowledge, few researchers have focused specifically on localising the point in time at which emotions transition from category to the next. Among the most relevant previous work about emotion changes is [10, 11], in which Niedenthal et al studied the emotion transitions between happiness, sadness and neutral in order to understand why smiles drop, based on facial expression. In facial-related emotion research, researchers believe that facial changes can mirror changes in emotions, referring to onset and offset of facial expressions [12]. These representations can be seen in some studies where either facial or speech information is used [11, 13]. However, how to correctly detect locations where emotion changes occur remains unanswered.

In this paper, we investigate the problem of detecting transition points from one emotion to another, not only drawing from methods proposed for speaker change detection, but also proposing a novel likelihood based method. This is a challenging task because emotional effects seem to be weaker than speaker effects [14, 15], however some prior knowledge of the emotions of interest might be brought to bear on the problem. As part of this work, we investigate system settings and configurations for effective emotion change detection and the performance for different types of methodologies.

2. Related Work

Some studies have hinted at the possibilities of detecting emotion change. For example, Wei et al [13] proposed a very interesting framework for emotion recognition based on speech. The evolution of temporal expression, namely the onset, offset and apex of emotional intensity in speech were used in a Hidden Markov Model (HMM) context, where emotional sub-states (e.g. neutral, happiness, etc) were modelled in each isolated utterance, which is similar to language modelling in speech recognition. 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 is the use of HMM, which can capture the temporal information of changes in emotions, although in this paper we focus on detecting changes rather than modelling durations. Apart from this, some have investigated classification of speech styles [16, 17] using variance vectors of GMM supervectors, a somewhat similar problem. However, like most emotion recognition systems to date, these focused on classifying utterances rather than detecting the change points.

Turning our attention to the speaker change detection literature, Chen et al proposed a classic method based on Bayesian Information Criterion (BIC), a metric-based method for model selection [18] popular in speaker diarisation [18]. BIC was used within a 1 second sliding window based on an assumption that the window can be well modelled as two Gaussian distributions rather than one if there is a change point inside, and be better modelled as one Gaussian otherwise.

Another method proposed for speaker change detection is to use two consecutive fixed-length windows, model each by GMMs adapted from a universal background model, and find differences between them. Many distance-based methods, such as the Generalized Likelihood Ratio (GLR) [19], Kullback-
Knowledge of Emotions
knowledge about emotions, simply differentiating changes and
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
[14].

3. Emotion Change Detection

3.1. Proposed System

Motivated by the ideas from Section 2, the general system
proposed for emotion change detection, shown in Fig. 1,
with the tolerance region.

![Figure 1: General system diagram for emotion change detection, optionally using emotion-specific models.](image)

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. 2). 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.

![Figure 2: Emotion change detection dual sliding window. The window center represents a candidate change point. During each window, features from multiple frames will be extracted.](image)

3.2. Generalized Likelihood Ratio – Without Prior Knowledge of Emotions

In [19], a generalized likelihood ratio (GLR) based distance
was applied to segregate speech from speakers given a speech
sequence. In the proposed context, GLR requires no prior
knowledge about emotions, simply differentiating changes and
non-changes via likelihood ratios, shown as in (1). Normally,
one multivariate Gaussian model is used to separately fit a
previous window, a current window and the entire dual
window (previous and current) respectively, and their log
likelihoods are calculated and combined as a score:

\[
D_{GLR} = LL_{prev} + LL_{curr} - LL_{prev+curr}
\]  

(1)

where \( LL_{prev} \), \( LL_{curr} \) and \( LL_{prev+curr} \) are the likelihoods
of the three independent models. According to equation (1), if
there is a change point, the dual window tends to be better
modelled as two separate Gaussians for the \( prev \) and \( curr \)
windows, which leads to a higher \( LL_{prev} \) and \( LL_{curr} \). If there
is no change, then the Gaussian modelling the entire dual
window tends to be a better model, leading to a higher \( LL_{prev+curr} \). GLR can be used to detect emotion change of any kind. However it is computationally expensive, especially when the window shift is small, because three Gaussian models need to be estimated for every dual window. This motivates the additional investigation of faster methods.

3.3. Emotion Pair Likelihood Ratio-based Scoring – With Prior Knowledge of Emotions

The second method proposed herein is to apply likelihood
ratios between different pairs of GMM-based emotion models
for emotion change detection. The likelihood ratios are
referred to Emotion Pair Likelihood Ratios (EPLRs). One
assumption behind this is that if prior information of emotions
is available, the performance might be improved over methods
with no prior information. Another assumption is that
likelihood is a key measure in emotion recognition. If there is
a change in emotion, this might be reflected through variations
in likelihoods.

Assuming that emotion-specific models \( \lambda_{e_i} \) are available
for a set \( E \) of emotions of interest, let \( LL_{i}^{prev} \) denote the log
likelihood of emotion \( i \) in the previous window. Then the log
likelihood ratio for the transition from emotion \( e_i \) to \( e_j \) during
the dual window can be expressed as

\[
LLR_{e_i \rightarrow e_j} = LL_i^{curr} - LL_i^{prev} \\
= \sum_{y_j \in \mathbb{X}_{curr}} \log p(y_j | e_i) - \sum_{y_i \in \mathbb{X}_{prev}} \log p(x_i | e_j)
\]  

(2)

Where \( x_i \) and \( y_j \) are frame level features in the previous and
current window respectively. Since for any pair of emotions \( e_i \)
and \( e_j \), the transition can occur in either direction, a simple
way to combine the log likelihood ratios into a single score is

\[
LLR_{e_j \rightarrow e_i} = |LLR_{e_i \rightarrow e_j} + LLR_{e_j \rightarrow e_i}|
\]  

(3)

3.4. Proposed Inter-Emotion Likelihood Normalization

According to equation (3), EPLRs are essentially a measure of
how much the \((LL_i^{curr} + LL_i^{prev})\) changes from a previous to a
current window. Nevertheless, likelihoods might vary
dramatically between windows, and are highly dependent
on the phonetic and speaker content as well as emotion. Indeed,
upon closer examination of the curve of EPLRs vs ground
truths, we found that both emotion-specific likelihoods and
EPLRs change fairly consistently across all previous/current
window boundaries, which is presumably due to phonetic
variations between particular windows and the global
similarity between emotion models, i.e. the movement of
likelihoods tend to be synchronized over time across all
emotions. As a result, the between-emotion differences of likelihood levels are more salient than changes in $(LL_i + LL_j)$ of emotion pairs between two windows. To minimize this likelihood synchronization, we performed window-based normalization to remove the global similarity between likelihoods from different emotions.

Given a dual window that has $N$ frames in total, $R_{ij}$ is used to represent the likelihood of emotion $i$ in the $n$-th frame. The window-based normalization can be formulated as:

$$LL_i^{\text{prev}} = LL_i^{\text{prev}} - \mu_i^{\text{prev}}$$
$$LL_i^{\text{curr}} = LL_i^{\text{curr}} - \mu_i^{\text{curr}}$$

where

$$\mu_i = \frac{\sum_{n=1}^{N} \sum_{e \in E} LL_{n,i}}{|E|}$$
$$\mu_{curr} = \frac{\sum_{n=N+1}^{N} \sum_{e \in E} LL_{n,i}}{|E|}$$

where $|E|$ denotes the number of elements in $E$. Then the normalized EPLRs are written as:

$$EPLR_{i \rightarrow ej} = \frac{LL_{i \rightarrow ej} + LL_{ej \rightarrow e_i}}{LL_i^{\text{prev}} + LL_i^{\text{curr}} - LL_i^{\text{prev}}}$$

4. Evaluation

4.1. Database

All work presented in this paper is based on the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database [25], which comprises 12 hours of emotional speech from ten speakers in scripted or spontaneous spoken conversational scenarios. This database is reasonably large compared to existing English databases publicly available in the emotion recognition community. Also, short utterances with 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. We address emotion changes among four emotions, namely neutral, anger, sadness and happiness, where happiness includes happiness and excitement utterances. Only utterances with majority consensus were used [26]. By concatenating all scripted and spontaneous utterances for every speaker, an initial database with 986 change points in total was built. However, the result was not very natural, since emotion changes may occur within less than 1 second. A further modification was to remove utterances less than 8 seconds in length, producing a final database with 199 change points for our experiments. Based on the same scheme, a data subset of anger and neutral only utterances was also constructed, with 55 change points.

4.2. Experimental Settings

In this work, we began by replicating the GMM baseline emotion recognition system from [26] to validate a conventional GMM system on the full IEMOCAP database. Using a voicing probability threshold of 0.55, 13 MFCCs and their first derivatives were extracted from 25ms frames (10ms overlapping) with the OpenSMILE [27] toolkit, and similar emotion classification accuracies to [23] (53%) were observed. Thereafter, the emotion change database with 199 change points in total was employed for all emotion change detection experiments. 10-fold leave-one-speaker-out speaker independent GMMs with 64 mixtures were constructed. For the GLR method, one Gaussian with diagonal covariance was used. Detection Error Trade-off (DET) and Equal Error Rate (EER) were used as the evaluation criteria, to assess how well the proposed methods operate across the range of threshold values.

4.3. Emotion Change Detection Parameter Settings

Firstly, we considered two parameters to be tuned for the GLR and the EPLR methods: window size and window shift. Since EPLR can only employ a single pair of emotion models, after preliminary experiments on different pairs, only $LLR_{\text{ang} \rightarrow \text{ang}}$ was adopted. In general similar performances were observed for other emotion pairs, although they were not as effective as the anger-neutral pair. Note that except in Section 4.4, we detected all kinds of transitions between four emotions using EPLR based on anger and neutral (other emotion pairs gave similar insights). During these experiments, the tolerance region width was fixed to 1 second and 0.2 second for the 0.4 second and 0.1 second window shift conditions respectively, allowing a 3-score tolerance region in each case for fair comparison. EER results for Generalized Likelihood Ratio (GLR) and $LLR_{\text{ang} \rightarrow \text{ang}}$ can be seen in Figure 3.

Figure 3: Equal error rates for GLR-based emotion change detection systems, comparing frame length and frame shift parameter choices

It was found (Fig. 3) that the best EER for the GLR method was obtained when a 3.5 second window size was adopted. On the contrary, a 0.7 second small window size gave best EER for $LLR_{\text{ang} \rightarrow \text{ang}}$. This may be reasonable, since the GLR method needs a larger amount of data for reliable Gaussian training, whereas the EPLRs are more informative around change points and will be affected by more variability in likelihoods if large window sizes are considered. With respect to the window shift, 0.4 second window shift and 1 second tolerance width performed generally better for the two methods and these settings are employed for all subsequent experiments.

Somewhat surprisingly, although EPLR contains prior knowledge of emotions, it is outperformed by the GLR method when detecting general emotion transitions. This might be due to phonetic and speaker variability. Attempting to detect all kinds of emotion changes using only likelihood ratios between two emotions seems not to provide better performance over a general model selection-type approach (GLR). One might expect that EPLR would detect changes only between anger and neutral more effectively, which motivated the next experiment.
4.4. Emotion Pair Change Detection

Referring to Figure 3, we selected 0.7s and 3.5s window sizes and examined the EERs with a tolerance region of 1 second duration for the data subset containing changes between only neutral and anger, as seen in Table 1.

Table 1: EERs for detection of changes between only neutral and anger, for different window size settings

<table>
<thead>
<tr>
<th></th>
<th>GLR</th>
<th>EPLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7s</td>
<td>25.6%</td>
<td>21.8%</td>
</tr>
<tr>
<td>3.5s</td>
<td>23.9%</td>
<td>20.2%</td>
</tr>
</tbody>
</table>

When detecting only emotion transitions between neutral and anger, LLR_{norm C+ang} is somewhat more accurate than GLR, for both small and large window sizes – the prior emotion information is helpful, although perhaps less than might be expected. Being more easily distinguished than other emotion pairs, anger and neutral benefited from larger windows during change detection using EPLR.

4.5. Comparison of Proposed Methods

4.5.1. Detection Error Trade-off

In this experiment, the proposed methods were compared for emotion transition detection among four emotions using the best settings from Section 4.2 and a tolerance region of 1 second duration, as shown in Fig.4. The result suggests that GLR, which requires no prior knowledge of emotions and speaker identity, outperforms EPLR (anger-neutral) and normalized EPLR (anger-neutral), both of which take into account prior knowledge of emotions. However, the normalization method provides a modest improvement for EPLR and has a roughly comparable performance to the GLR method. EPLR is much more computationally efficient and operates on only a short window length (0.7s) during online detection, compared with GLR (3.5s). The fusion of GLR and normalized EPLR, using linear combination, also provided a further improvement in EER, of around 2%.

5. Conclusion

This paper has presented an initial investigation into emotion changes, emphasizing on localizing emotion change points. To address this problem, within a dual sliding window approach, a GMM-based EPLR method using prior knowledge of emotions and a normalization method are proposed, and a GLR method, originally proposed for speaker change detection, has been applied. Experimental results suggest that the normalized EPLR can provide roughly comparable performance while being more computationally efficient and lower delay in real-time online applications. Even though detecting emotion changes remains a challenging problem, these first steps help pave the way towards understanding emotion changes and detecting the instant of change accurately.

Our work does present some possibilities regarding this interesting and novel research area. Firstly, a database containing a larger number of changes and more realistic change points is required for further investigation. The former might be approached by shuffling existing utterances within the IEMOCAP corpus. In addition to emotion categories, numerical ratings are potentially helpful for understanding emotion changes, and may be relevant to investigating the type as well as the instant of emotion change. Also, for the dual windowing pattern, there are a number of methods proposed for speaker change detection that might be investigated, such as KL divergence, CLLR, one class SVM, etc. Apart from the dual windowing framework, it may be possible to conceive other frameworks capable of detecting emotion changes. Moreover, 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. It is also crucial to develop normalization methods to better remove variability in terms of speaker identity and phonetic content.

6. Acknowledgements

This work was partly funded by the US Army International Technology Center (Pacific). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.
7. References


