Detection of anger emotion in dialog speech using prosody feature and temporal relation of utterances

NOMOTO Narichika, MASATAKI Hirokazu, YOSHIOKA Osamu and TAKAHASHI Satoshi

NTT Cyber Space Laboratories, NTT Corporation, Japan

Abstract

This paper proposes a novel feature for detecting anger in dialog speech. Anger is classified into two types; loud HotAnger and calm ColdAnger. Prosody can reliably detect the former but not the latter. We analyze both types of anger dialog in the two-party setting, and discover that they exhibit some differences in the temporal relation of utterances from neutral dialog. We create a dialog feature that reflects these differences, and investigate its effectiveness in detecting both types of anger. Tests show the proposed feature combination improves the F-measure of Cold and HotAnger by 24.4 points and 8.8 points against baseline technique that uses only prosody.

Index Terms: emotion detection, HotAnger, Cold Anger, dialog feature, dialog speech

1. Introduction

A lot of research has tried to analyze and extract various bits of information from human voice. Emotion detection from speech is a subject of many past studies. The Interspeech2009 Emotion Challenge [1] was held in order to “help bridge the gap between the excellent research on human emotion recognition from speech and the low compatibility of results”. Recent studies have targeted the identification of customer’ emotion from contact center dialog speech data [2][3]. One study splits anger into HotAnger and ColdAnger [4]. The former is evidenced by explosive speech patterns in contrast to the latter. Conventional studies mostly use prosody-features to detect emotion [5]-[7]. They reported that features such as pitch and energy are effective in detecting anger. It was also mentioned that prosody-feature is unreliable in detecting ColdAnger [8]. However, it is important to be able to detect ColdAnger. In the real world, ColdAnger is as common as HotAnger. Moreover, it is thought that the customer who expresses an opinion with ColdAnger is as important as one that offers HotAnger.

We introduce here a method that can detect both HotAnger and ColdAnger in dialog speech between two parties as is common in contact center calls. We focus on some common conversation situations in which one speaker, the caller, is angry when talking to the second speaker, the callee. Our research treats dialog speech between humans and introduces features that allow the two types of anger to be discriminated. After listening to a lot of angry dialog, we identified the following characteristics.

- The caller speaking one-sidedly.
- The callee doesn’t respond to back-channel feedback much.
- The callee frequently offers back-channel feedback, and takes the part of the listener.
- The callee responds to back-channel feedback often.
- The dialog flow is not smooth.
- The silence is caused frequently.

These characteristics are associated with the temporal relationship between utterances. In this paper, we extract the temporal relationships of utterances as “dialog-features”, and use them to create a comprehensive anger detection method. Since the prosody-features of ColdAnger are insufficient to differentiate it from other speech, but this becomes possible by using dialog-features, the accuracy with which these two forms can be identified is improved by using dialog-features in combination with prosody-features.

The outline of the paper is as follows; Section 2 describes the proposed method, and explains the features used. Dialog data of angry and neutral dialogs is detailed and analyzed in Section 3. An experiment and its results are shown in Section 4. Our conclusion is drawn in Section 5.

2. Proposed Method

To detect both types of anger reliably we use dialog-features and prosody-features. We start by describing the features and then the proposed method.

2.1 Dialog as Chain of Utterances

To analyze the relation between two speakers (in this paper caller and callee) we created a dialog model.

![Fig.1: Dialog as a chain of utterances](image)

We modeled dialog as a chain of utterances issued by two speakers in turn. Fig.1 shows dialog as a chain of utterances. The beginning time and the end time of each utterance were calculated by Voice Activity Detection (VAD) etc. We define an utterance as the duration over which the utterance right is held. The continuous utterances by the same speaker are merged into one utterance. An utterance that begins while the other speaker is speaking and ends before the other ends is not defined as a transfer of utterance right. Such an utterance is defined as a back-channel feedback utterance.

2.2 Dialog-Features

We define a dialog-unit as a series of utterances and use four dialog characteristics each dialog unit. Each dialog unit consists of the presumed target speaker” (caller) utterance and the preceding and a follow up utterance by the other speaker (callee). This paper defines seven features as dialog characteristics; utterance length, back-channel feedback frequency, utterance
right transfer interval, ratio of utterance length. Fig.2 shows the relationship between $i_{th}$ ($i=1,2,\ldots$) dialog-unit (du) and its features.

- $\text{len}^{\text{calle}}$: utterance length of callee [sec]
- $\text{len}^{\text{caller}}$: utterance length of caller [sec]
- $\text{frq}^{\text{calle}}$: back-channel feedback frequency of callee
- $\text{frq}^{\text{caller}}$: back-channel feedback frequency of caller
- $\text{int}^{\text{calle}}$: utterance right transfer interval of callee [sec]
- $\text{int}^{\text{caller}}$: utterance right transfer interval of caller [sec]
- $\text{rat}^{\text{calle}}$: ratio of utterance length

\[ \text{len}^{\text{caller}} = \frac{\text{len}^{\text{calle}} + \text{len}^{\text{caller}}}{2} \]  
(1)

The situation that the caller exhibits stronger insistence than the callee is expressed by this feature.

The back-channel feedback frequency ($\text{frq}$) is measure of the speaker’s feeling of cooperation in the dialog. Accordingly, this expresses the situation that the caller exhibits little cooperation, unlike the callee.

The utterance right transfer interval ($\text{int}$) is time from the completion of speaking by the one speaker to commencement of speech by the other speaker; it shows smoothness of the dialog. For instance, the value of $\text{int}^{\text{caller}}$ is the period from when the caller finishes speaking to when the callee begins to speak. A value greater than zero indicates pause; a negative value indicates that the listener starts to speak before the current speaker halts. When this feature is extremely large or the value of variance is large, the conversation doesn’t go well, and an awkward situation is indicated.

The ratio of utterance length ($\text{rat}$) is computed by the following equation:

\[ \text{rat}^{\text{calle}} = \frac{\text{len}^{\text{calle}}}{\text{len}^{\text{caller}} + \text{len}^{\text{calle}}} \]  
(2)

This indicates which speaker is currently driving the conversation. For instance, the larger the value of $\text{rat}$ is, the longer the caller speaks and thus drives the conversation. It allows us to identify one-sided dialogs.

In this study, the conversation situation is captured by using these four characteristics, seven features.

### 2.3. Prosody-Features

Most studies on identifying emotion use prosody. It is well known that several emotional speech states can be well discriminated by the prosodic pattern [9]. It is reported that the level of pitch and energy rises with the speaker’s anger. We use their pitch and energy terms as prosody-features. We also adopt their delta features to capture the alternation, the voice becomes louder or higher in pitch, instantaneously. These features are extracted from the utterance section of the presumed speaker (caller) in the dialog-unit; the terms include maximum ($\text{max}$), minimum ($\text{min}$), average ($\text{ave}$), variance ($\text{var}$) and dynamic-range ($\text{ran}$) value. The dynamic range is calculated by dividing $\text{max}$ by $\text{min}$. Pitch is extracted by an estimation approach based on dominance spectrum [10]. Energy is extracted by the RMS approach. Maximum, minimum and variance value are normalized by the mean value in the utterance section. In all, fourteen prosody-features are used. Table.1 summarizes the prosody-features used in this study.

<table>
<thead>
<tr>
<th>Prosody-feature</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{pitch}$</td>
<td>maximum, minimum, average</td>
</tr>
<tr>
<td>$\Delta \text{pitch}$</td>
<td>variance, dynamic range</td>
</tr>
<tr>
<td>$\text{energy}$</td>
<td>maximum, average, variance</td>
</tr>
<tr>
<td>$\Delta \text{energy}$</td>
<td>maximum, minimum, average</td>
</tr>
</tbody>
</table>

### 2.4 Combing Dialog- and Prosody-Features

Fig.3 shows the processing flow of the proposed method which uses both dialog- and prosody-features. We propose a method that can detect the interval of dialog in which the caller is angry.

From the data in each window, the mean and variance value of each feature are calculated. In all, 42 features (dialog-features: 7 times 2, prosody-features: 14 times 2) are extracted from each analysis window and used to identify angry dialog from neutral. We use support vector machine (SVM) as the learning algorithm. SVMs have been well studied for binary classification tasks. They are based on strong theoretical underpinnings and can be used in a variety of domains.
3. Database and Analysis

3.1 Dialog Data: ground truth
Dialogs in contact center were used as the test data. The dialogs ranged in duration from three to twenty minutes and consisted of agent (callee) and customer (caller) speech.

Every utterance was tagged with an emotion state; HotAnger, ColdAnger, Neutral, by two subjects. If the label was anger, the subjects were required to assign one of three levels (low-medium-strong) to reflect the perceived strength of the caller’s anger. We assumed that a “low” level of anger reflected merely a complaint so only medium or strong levels of anger were taken to be instances of either HotAnger or ColdAnger. If the subjects assigned different Anger labels, the utterance was taken to be HotAnger. This process yielded, from 108 dialogs, 1159 HotAnger utterances, 1394 ColdAnger utterances, and 3462 Neutral utterances.

3.2 Result of Analysis
We investigated the relationships between dialog type and speech features. Table.2 shows the mean value of each dialog-feature. The utterance length of the callee is short in HotAnger and long in ColdAnger while that of the caller is long in HotAnger and ColdAnger. This result shows the callee tends to takes the part of listener. The back-channel feedback frequency of the callee increased in HotAnger and ColdAnger, especially in ColdAnger. This shows the tendency that the callee strongly responds by generating back-channel feedback to pass his attitude to the caller. The back-channel feedback frequency of the caller decreased in HotAnger and ColdAnger, especially in the latter. This shows that the caller takes the conversation initiative and doesn’t cooperate in communicating with the callee. We noticed that the utterance right transfer time (caller to callee) was short and indeed the callee tended to speak over the caller’s utterances in HotAnger. This indicates the dialog situation in which the callee tends to respond early with feeling of being rushed. In ColdAnger, the utterance right transfer time (caller to callee) was long. This indicates the dialog situation in which the callee tends to be slow in responding since he is at a loss for a ready answer. The utterance right transfer time (callee to caller) was long in HotAnger and ColdAnger, especially in the latter. This shows that the dialog flows are not smooth and that the caller often expresses anger by silence. The callee’s utterance time ratio increased in HotAnger and ColdAnger. This shows that the caller takes the conversation initiative over long periods. Thus, it was confirmed that anger yields unique dialog-features as we originally hypothesized.

Table.2: Mean value of dialog-features

Table.3 and Table.4 shows the mean value of each prosody-feature. Each feature is normalized by the value of Neutral. Past studies on prosody features have shown that pitch is effective in detecting anger. The maximum value of an utterance tended to be higher in HotAnger and ColdAnger, especially in the former. The minimum value of an utterance tended to be lower in HotAnger and ColdAnger, especially in the latter. The average of an utterance tended to be highest in HotAnger, whilst in ColdAnger it was lower than in neutral. Regarding delta pitch, the maximum value of an utterance tended to be higher in HotAnger and ColdAnger, especially in the latter. The minimum value of an utterance tended to be lower in HotAnger and ColdAnger, especially in the latter. As for the average, ColdAnger yielded the lowest value, while HotAnger yielded the highest value. The results of the other features broadly mirrored the tendencies noted in past studies.

Table.3: Mean value of pitch: prosody-feature

Table.4: Mean value of energy: prosody-feature

4. Experiment
4.1 Experimental Conditions
Analysis windows in which all utterances had the same label (HotAnger / ColdAnger / Neutral) were taken samples of HotAnger, ColdAnger, and neutral, respectively.

After a consideration of the volume of data available, we randomly selected 400 of the 416 HotAnger windows, 400 of the 443 ColdAnger windows, and 800 of the 3367 neutral windows, to balance the volume of each data type. The HotAnger and ColdAnger samples, total of 800 samples, were studied without distinction between the two types. Learning and evaluation were used 10-fold cross-validation.

The window length of the frame for pitch extraction was 42 milliseconds, and the shift length was 10 milliseconds. The window length of the frame for energy extraction was 16 milliseconds, and the shift length was 10 milliseconds. To calculate delta, we used three preceding and three following frames (seven frames in total so 70ms) from the frame in question. We used the SVMlight software package [11][12]. The SVM kernel function was taken to be the second of the polynomial kernels in accordance with the result of a prior examination. The precision and recall were calculated by the following mathematical expressions.

Precision = \frac{\text{number of windows tagged correctly as anger}}{\text{number of window presumed as anger}} \times 100

Recall = \frac{\text{number of windows tagged correctly as anger}}{\text{number of correct window labeled anger}} \times 100

F-measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
4.2 Result and Discussion

Table 4 shows the results gained in each condition when using prosody-feature (baseline), dialog-feature (propose1), and using prosody and dialog-feature (propose2). “chance” is the probability of the evaluation data containing an anger sample. The precision and recall of ColdAnger by using prosody-features are both about 50%, which shows the difficulty of detection. By using dialog-features, the precision of ColdAnger improves to 95.9%. In addition, in combination with prosody-features, recall also improves, to 63.7%. The result for HotAnger shows that using F-measure as the only dialog-feature yields nearly the same performance as using prosody-feature; F-measure is about 65 points, and with both features this improves to 74.0 points.

Table 4: Presumed results of analysis window

<table>
<thead>
<tr>
<th>feature</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HotAnger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chance</td>
<td>33.3</td>
<td>(100.0)</td>
<td>(50.0)</td>
</tr>
<tr>
<td>baseline</td>
<td>74.4</td>
<td>58.0</td>
<td>65.2</td>
</tr>
<tr>
<td>propose1</td>
<td>86.7</td>
<td>50.3</td>
<td>63.7</td>
</tr>
<tr>
<td>propose2</td>
<td>88.3</td>
<td>63.7</td>
<td>74.0</td>
</tr>
<tr>
<td>ColdAnger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chance</td>
<td>33.3</td>
<td>(100.0)</td>
<td>(50.0)</td>
</tr>
<tr>
<td>baseline</td>
<td>54.9</td>
<td>50.5</td>
<td>52.6</td>
</tr>
<tr>
<td>propose1</td>
<td>95.9</td>
<td>52.9</td>
<td>68.2</td>
</tr>
<tr>
<td>propose2</td>
<td>94.0</td>
<td>65.2</td>
<td>77.0</td>
</tr>
</tbody>
</table>

Precision versus recall plots are shown in Fig. 5. These plots clearly show that using both prosody-feature and dialog-feature improves the ability of detecting Hot/ColdAnger in dialog speech. The improvement is most significant for ColdAnger.

5. Conclusion

This paper proposed a novel approach to the detection of anger in dialog speech. The proposed method uses dialog-features in addition to prosody-features. At first, dialog was modeled as a chain of utterances, and four dialog-features were defined from the time relationships of utterances. We then introduced a method for integrating those features in each analysis window. Our analysis confirmed that anger yields unique dialog-features. The results of an experiment on contact center dialog showed that the proposed method was especially effective in detecting ColdAnger (precision: 94.0% recall: 65.2%). The addition of prosody-features was shown to improve the F-measure by 24.4 points (ColdAnger) and 8.8 points (HotAnger) from performance of the baseline technique (prosody-features only).

6. Acknowledgement

We are grateful to our project manager, Mr. Haruhiko KOJIMA of NTT Cyber Space Laboratories, for giving us the opportunity to pursue this work. We also wish to thank the members of Spoken Dialog System Group.

7. References