Anger Recognition in Spoken Dialog Using Linguistic and Para-Linguistic Information

Narichika Nomoto, Masafumi Tamoto, Hirokazu Masataki, Osamu Yoshioka, Satoshi Takahashi

NTT Cyber Space Laboratories, NTT Corporation, Japan

nomoto.narichika@lab.ntt.co.jp

Abstract

This paper proposes a method to recognize anger-dialog based on linguistic and para-linguistic information in speech. Anger is classified into two types; HotAnger (agitated) and ColdAnger (calm). Conventional prosody-features based on para-linguistic can reliably recognize the former but not the latter. To recognize anger more robustly, we apply other para-linguistic cues named dialog-features which are seen in conversational interactive situations between two speakers such as turn-taking and back-channel feedback. We also utilize linguistic-features which represent conversational emotional salience. They are acquired by Pearson's chi-square test by comparing the automatically-transcribed texts between angry and neutral dialogs. Experiments show that the proposed feature combination improves the F-measure of ColdAnger and HotAnger by 26.9 points and 16.1 points against a baseline that uses only prosody.

Index Terms: emotion recognition, anger, dialog speech, linguistic feature, call-center

1. Introduction

The goal of our study is recognizing "anger-dialog" in which one speaker is angry with another speaker in dialog speech; i.e. the customer (caller) is angry with the agent (callee) in a call-center. Emotion recognition in speech is a subject of many past studies. In particular, para-linguistic information has been a common topic and prosody-features have been used by many conventional studies. They report that these features are effective in analyzing emotion [1, 2, 3]. Meanwhile, anger is more correctly classified as either HotAnger or ColdAnger by how the anger is expressed [4]. HotAnger is evidenced by explosive voice patterns. ColdAnger is expressed in a severe tone and the voice is not raised. Prosody-features are effective in recognizing HotAnger, but not ColdAnger since its vocal characteristics are uncertain [5]. In addition, speech in a real environment is not as clean as the data tested on most conventional experiments; i.e. voice quality is actually distorted by the telephone band and Lombard effects are caused by the noisy environment. Thus, it has still remained difficult to recognize ColdAnger reliably from real speech data. However it is very important to reliably recognize ColdAnger, since ColdAnger is as common as HotAnger in the real world; self-controlled callers often express their anger calmly.

Our study is based on the perspective of *how emotion is expressed in spoken dialog*. We have already proposed "dialog-feature" as novel para-linguistic information to recognize both HotAnger and ColdAnger [6]. Dialog-features can capture interactive conversational situations in anger-dialogs that are independent of anger type; i.e. "angry speaker often speaks one-sidedly", "angry speaker does not respond to feedback-channel much", "the callee responds to feedback-channel often". They are affected by the temporal relation between utterances; utterance length, feedback-channel frequency, interval of turn-taking and ratio of utterance length. A statistical analysis and experiment showed that dialog-features are very effective in discriminating anger from neutral dialogs and are especially effective in recognizing ColdAnger.

This paper proposes a method using linguistic-features acquired automatically from the corpus. We focus on *conversational emotional salient words used in actual dialogs* and propose a method that uses them to recognize both anger types more robustly. Some studies have used linguistic-features from texts [7, 8, 9, 10]. Conventional linguistic-features are often based on emotional keywords; i.e. "disappointed = angry", "War = sad, angry". Emotional keywords are those that involve or arouse some emotion. These studies mostly targeted the written word. However, emotional keywords are not so used directly in actual spoken dialogs. Moreover, their usage is not uniform and the emotions that are roused depend on the context. In this study, without relying on conventional emotional keywords, conversational emotional salient word (CEMS), which are used as emotional expressions in actual spoken dialogs, are extracted by a statistical test from an extensive dialog corpus and used to recognize anger-dialog through a combination of prosody and dialog features.

The outline of the paper is as follows; Section 2 describes the proposed method, and explains the features used. The dialog corpus of angry and neutral dialogs is detailed in Section 3. An experiment and its results, and an analysis of CEMS acquired by the proposed method are shown in Section 4. Our conclusion is drawn in Section 5.

2. Proposed Method

2.1. Procedure to recognize anger-dialog

To recognize both types of anger reliably, we use linguistic, dialog and prosody features. Fig. 1 overviews the procedure of the proposed method to recognize anger-dialog in which the caller is angry with the callee. To detect even when the caller is angry in a part of the dialog, an analysis window of constant length is extracted from the dialog and shifted while judging each window.

We model dialog as a chain of utterances issued by two speakers in turn. The beginning time and the end time of each utterance are determined by Voice Activity Detection (VAD) etc. We define an utterance as the duration over which the turn is held. The continuous utterances by the same speaker are merged into one. An utterance that begins while the other

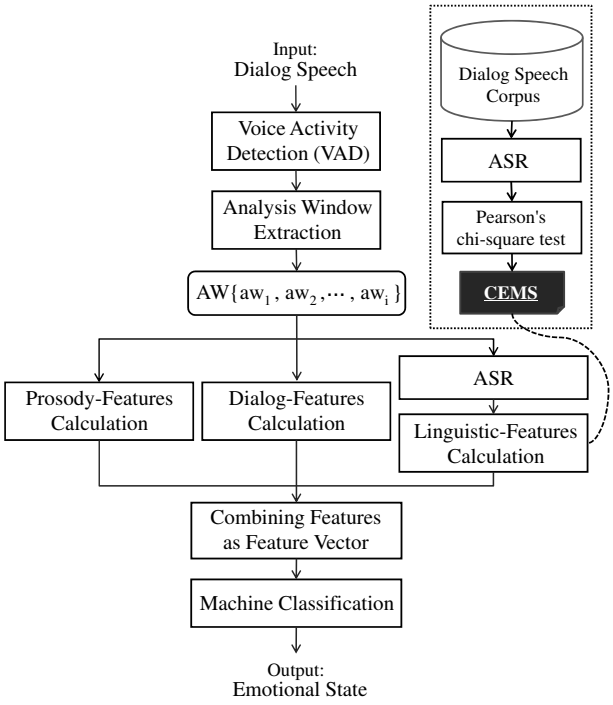

Figure 1: *Procedure of proposed anger-dialog recognition in a part of the dialog.*

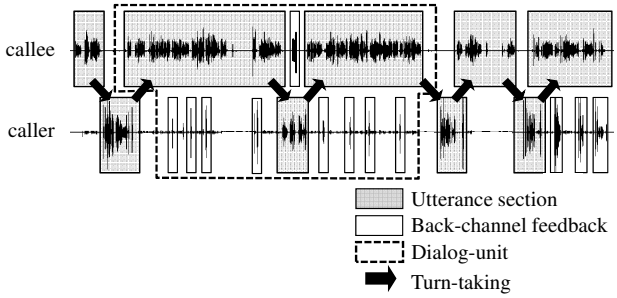

Figure 2: *Dialog model as chain of utterances issued by two speakers in turn and dialog-unit consisting of caller's and callee's utterance.*

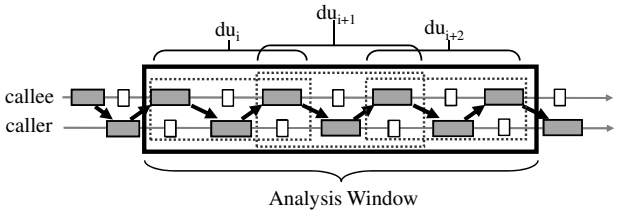

Figure 3: *Analysis windows consisting of three dialog-units.*

speaker is speaking and ends before the other ends is defined as back-channel feedback utterance, not turn-taking. From the dialog model defined above, we define a dialog-unit as a series of utterances. Each dialog-unit consists of the caller's utterance and the preceding and following utterance by callee. Dialog-unit is assumed to be the minimum unit of dialog. Fig. 2 shows a dialog model as a chain of utterances and a dialog-unit.

The proposed method uses long term tendencies to improve the recognition accuracy. We note that dialog-features can be reliably identified only if the analysis conducted over some long term. That is, each analysis window includes several dialog-units. Fig. 3 shows the example where the analysis window size is set to cover three dialog-units. The linguistic, dialog and prosody features in each analysis window are calculated.

From the data in each window, the mean and variance value of dialog and prosody features are calculated. Linguistic-features are the word frequency of CEMS and normalized by the number of words in each window. These features are merged to yield one feature-vector. Finally, the feature-vector is classified into anger or neutral by machine classification.

The next section details the three kinds of features used in this study.

2.2. Feature Extraction

2.2.1. Linguistic-features

We focus on "conversational emotional salient words (CEMS)" which appear remarkably often in angry and neutral dialogs. CEMS are automatically acquired from a dialog corpus by Pearson's chi-square test to determine word-frequency differences. Pearson's chi-square test is done to each dialog emotion class $E = \{e_k | HotAnger, ColdAnger, Neutral\}$ and speaker $S = \{s_l | caller, callee\}$; caller in HotAnger, ColdAnger vs. Neutral and callee in HotAnger, ColdAnger vs. Neutral. Word count $C_o(s_l, e_k, w_n)$ is calculated for the words in all utterances $U(s_l, e_k)$ in training corpus by $W = \{w_1, w_2, \ldots, w_n\}$. Then expectation of word count $C_e(s_l, e_k, w_n)$ is calculated as

$$C_w(s_l, e_k) = \sum_n C_o(s_l, e_k, w_n) \tag{1}$$

$$P(w_n|s_l) = \frac{\sum_k C_o(s_l, e_k, w_n)}{\sum_k \sum_n C_o(s_l, e_k, w_n)} \tag{2}$$

$$C_e(s_l, e_k, w_n) = C_w(s_l, e_k) P(w_n|s_l) \tag{3}$$

where $C_w(s_l, e_k)$ is sum of all words count by speaker s_l and emotion class e_k. $P(w_n|s_l)$ is the conditional probability that the word w_n is uttered by speaker s_l. The chi-square score of the word w_n is calculated by observed $C_o(s_l, e_k, w_n)$ and expectation $C_e(s_l, e_k, w_n)$ as

$$\chi^2(s_l, w_n) = \sum_k \frac{(C_o(s_l, e_k, w_n) - C_e(s_l, e_k, w_n))^2}{C_e(s_l, e_k, w_n)} \tag{4}$$

The words that appear with significant difference in term of frequency are extracted as CEMS. This approach yielded CEMS in the following six categories; caller in anger-dialog (Hot-/ColdAnger), caller in neutral-dialog, callee in anger-dialog (Hot-/ColdAnger), callee in neutral-dialog.

As the extracted word unit, we used unigram and 2-skip-bigram. 2-skip-bigram means word pair which allows for a one or two word gap; "A-B" bigram includes each appearance pattern of word that is "A-B", "A-*-B" and "A-*-*-B". The purpose using skip-bigram is to consider particular speech phenomena such as the insertion of filler and restating.

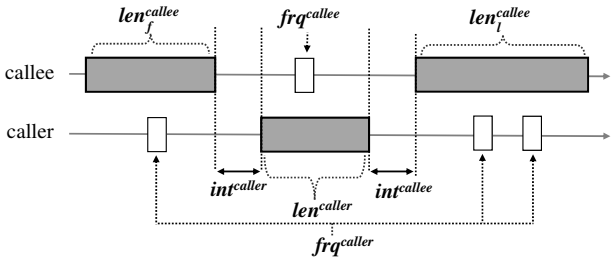

Figure 4: *Dialog-features in a dialog-unit.*

Finally caller's and callee's word frequency of CEMS and normalized by the number of words in each window are taken as the linguistic-features of the analysis window.

2.2.2. Dialog-features

We use four dialog characteristics each dialog-unit; utterance length, back-channel feedback frequency, interval of turn-taking, ratio of utterance length. Fig. 4 shows the relation between each dialog-unit and its characteristics. This paper defines seven features as dialog characteristics;

- len^{callee} : utterance length of callee [sec]
- len^{caller} : utterance length of caller [sec]
- frq^{callee} : back-channel feedback frequency of callee
- frq^{caller} : back-channel feedback frequency of caller
- int^{callee} : turn-taking interval of callee [sec]
- int^{caller} : turn-taking interval of caller [sec]
- rat : ratio of utterance length

The utterance length (len) is self-explanatory and indicates the strength of speaker's insistence. A dialog-unit has two utterances of the callee; the former utterance length (len_f^{callee}) and the latter (len_l^{callee}). Finally the utterance length of callee len^{callee} is computed as mean of them. The situation that the caller exhibits stronger insistence than the callee is expressed by this feature.

The back-channel feedback frequency (frq) is measure of the speaker's feeling of cooperation in the dialog. It expresses the situation that the caller exhibits little cooperation, unlike the callee.

The interval of turn-taking (int) is the time from the completion of one speaker's utterance to the commencement of speech by the other speaker; it shows the smoothness of the dialog. For instance, the value of int^{callee} 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 (rat) is computed by the following equation:

$$rat = \frac{len^{caller}}{len^{callee} + len^{caller}} \quad (5)$$

This indicates which speaker is currently driving the conversation. For instance, the larger the value of 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.

Table 1: *Summary of prosody features.*

f0	*max, min, mean, variance, dynamic-range*
energy	*max, mean, variance*
Δ*f0*	*max, min, mean*
Δ*energy*	*max, min, mean*

2.2.3. Prosody-features

Most studies use prosody to recognize emotion. It is well known that several emotional speech states can be well discriminated by the prosodic pattern [11]. It is reported that the levels of *f0* and *energy* rise with the speaker's anger. We use them as prosody-features. We also adopt their delta features (Δ*f0*, Δ*energy*) to capture the instantaneous variation, the voice becomes louder or higher. These features are extracted from the utterance section of the caller in each dialog-unit. Table 1 summarizes the prosody-features used in this study. The value of $dynamic\text{-}range$ is calculated by dividing max by min. The value of max, min and $variance$ are normalized by $mean$ in the utterance section. In all, fourteen prosody-features are used.

3. Corpus

Dialogs from a call-center are used as the test data. The dialogs ranged in duration from three to twenty minutes and consisted callee and caller speech. Every caller's 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.

4. Experiment

4.1. Experimental conditions

To validate the effectiveness of the proposed method we experimented that recognized analysis window into angry or neutral. Experimental conditions are described below.

Analysis windows in which all caller's utterances had the same label (HotAnger / ColdAnger / Neutral) were taken samples of HotAnger, ColdAnger, and Neutral, respectively. The size of analysis window was 10 dialog-units in accordance with the result of a prior examination. The result was that 800 anger windows (400 HotAnger, 400 ColdAnger) and 800 neutral windows were extracted. We performed a 10-fold cross-validation test for these data. The value of $f0$ is extracted by an estimation approach based on dominance spectrum [12]. The window length of the frame for $f0$ extraction was 42 milliseconds, and the shift length was 10 milliseconds. The value of $Energy$ is extracted by the RMS approach. 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. CEMS were preliminarily acquired from the corpus described in Section 3. In total, 1024 CEMS were extracted. Finally 4138 features were extracted from each analysis window and used to identify anger-dialog from neutral; dialog-features: 7 features $\times$ 2 (mean/variance), prosody-features: 14 features $\times$ 2, linguistic-features: 1024

features × 2 (TF/normalized-TF) × 2 (caller/callee). As the learning algorithm, support vector machine (SVM) was used. The SVM kernel function was taken to be the second of the polynomial kernels in accordance with the result of a prior examination. The evaluation metrics were calculated by Precision, Recall and F-measure.

4.2. Analysis of CEMS

Examples of extracted CEMS are described below. Though the personal pronoun (i.e. Japanese /wa-ta-shi/ (in English "*I*"), /a-na-ta/ ("*you*")) is often omitted in natural Japanese conversation, we found that the caller and callee tend to include the personal pronoun in anger-dialogs. This phenomenon of omitting the personal pronoun in spoken dialog can be described as a kind of linguistic strategy based on "negative politeness" by Brown and Levinson's Politeness Theory [13]. Negative politeness is a linguistic strategy to keep one's psychological distance from other people. It is one of the conversation strategies that seek to advance smoothly the conversation without violating the other party's area by specifying the speaker or listener. Intentionally specifying the speaker or listener is thought to occur when the caller tries to clarify the problem or the callee tries to politely explain more than the bare necessity. Fillers (i.e. /e-to/, /a-no/) are used a lot in speech and take the conversational role of softening the content of the utterance through "negative politeness". The statistical test showed that the angry caller does not tend to use fillers. Moreover, words that shows back-channel feedback (i.e. /ha-i/) were found to be used more often in the callee's utterances in anger-dialog and ending words of utterances that express the speaker's politeness (i.e. /de-su/) appear more in non-angry speaker' utterances than in angry ones. Thus we found that anger-dialog is expressed by not only emotional keyword but also non-emotional words in actual speech.

4.3. Result

Table 2 shows the results in each condition when using prosody- (pr: baseline), dialog- (dg), linguistic- (ln), prosody and dialog- (pr+dg), and all features combined (pr+dg+ln). If only one kind of features is used, linguistic-features yielded the highest F-measure for both HotAnger and ColdAnger (HotAnger: 71.2 points, ColdAnger: 70.6 points). This result presents that proposed linguistic-features based on CEMS without relying on emotional keywords are effective in anger estimation. Comparing the types of anger, the F-measure was better at identifying HotAnger than ColdAnger. It is considered that changes in callee's wording are more remarkable in HotAnger than ColdAnger. The F-measure of ColdAnger using only prosody-features was 52.6 points, which showed the difficulty of anger identification. By using dialog or linguistic features, the F-measure of ColdAnger improved. This shows that these features are robust in identifying ColdAnger. In addition, combination with all features, "pr+dg+ln", improves the F-measure of ColdAnger and HotAnger by 2.5 and 7.3 points against "pr+dg", respectively. Compared to baseline "pr", improves the F-measure of ColdAnger and HotAnger by 26.9 and 16.1 points, respectively.

Table 2: *Accuracy of recognizing HotAnger and ColdAnger in the analysis window.*

type	feature	Precision	Recall	F-measure
HotAnger	pr	.744	.580	.652
	dg	.867	.503	.637
	ln	**.821**	**.629**	**.712**
	pr+dg	.883	.637	.740
	pr+dg+ln	**.842**	**.786**	**.813**
ColdAnger	pr	.549	.505	.526
	dg	.959	.529	.681
	ln	**.756**	**.663**	**.706**
	pr+dg	.940	.652	.770
	pr+dg+ln	**.765**	**.827**	**.795**

5. Conclusions

This paper proposed a method to identify anger in spoken dialog; it uses linguistic and para-linguistic information. In addition to the dialog-features that have been proposed in para-linguistic research, we presented linguistic-features as conversational emotional salient words (CEMS) used in actual dialogs and proposed a method that automatically acquires them from dialog corpus using Pearson's chi-square test. The results of an experiment on call-center dialogs showed that the proposed linguistic-features are effective in identifying both types of Anger. The addition of linguistic and dialog features was shown to improve the F-measure by 26.9 points (ColdAnger) and 16.1 points (HotAnger) against the baseline (prosody-features). Thus the effectiveness of the proposed linguistic features was confirmed by the experiments.

6. References

[1] I. Fónagy, "A New Method of Investigating the Perception of Prosodic Features,", *Language and Speech*, 21, pp.34–49, 1978.

[2] I. Luengo, E. Navas and I. Hernáez, "Combining Spectral and Prosodic Information for Emotion Recognition in the Interspeech 2009 Emotion Challenge,", *in Proc. INTERSPEECH 2009*, pp.332–335, 2009.

[3] L. Devillers, C. Vaudable and C. Chastagnol, "Speech emotion recognition using hidden Markov models,", *Speech Communication*, Vol. 41, No. 4, pp.603–623, 2003.

[4] K. R. Scherer, "Vocal Affect Expression: A Review a Model for Future Research,", *Psychological Bulletin*, 99, pp.143–165, 1986.

[5] J. M. Montero, G. J. Arriola, J. Colas, E. Enriquez and J. M. Pardo, "Analysis and Modeling of Emotional Speech in Spanish,", *in Proc. ICPhS*, pp.957–960, 1999.

[6] N. Nomoto, H. Masataki, O. Yoshioka and S. Takahashi, "Detection of Anger Emotion in Dialog Speech Using Prosody Feature and Temporal Relation of Utterances,", *in Proc. INTERSPEECH 2010*, pp.494–497, 2010.

[7] C. M. Lee, S. S. Narayanan and R. Pieraccini,"Combining Acoustic and Language Information for Emotion Recognition,", *in Proc. ICSLP*, pp.873–876, 2002.

[8] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner and F. Metze, "Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features,", *in Proc. INTERSPEECH 2009*, pp.340–343, 2009.

[9] B. Schuller, G. Rigoll and M. Lang, "Speech Emotion Recognition. Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network. Architecture," *in Proc. ICASSP 2004*, pp.577–580, 2004.

[10] J. Tao, "Context Based Emotion Detection from Text Input,", *in Proc. INTERSPEECH 2004*, pp.1337–1340, 2004.

[11] K. R. Scherer, "How Emotion is Expressed in Speech and Singing,", *in Proc. ICPhS*, pp.90–96, 2005.

[12] T. Nakatani, T. Irino, and P. Zolfaghari, "Dominance Spectrum Based V/UV Classification and F0 Estimation,", *in Proc. EUROSPEECH*, pp.2313–2316, 2003.

[13] P. Brown. and S. Levinson, "*Politeness* : Some Universals in Language Usage", *Cambridge University Press*, 1987.