Automatic Emotional Degree Labeling for Speakers' Anger Utterance during Natural Japanese Dialog

Yoshiko Arimoto[†], Sumio Ohno[††], Hitoshi Iida[†††]

[†]Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology,
[††]School of Computer Science, Tokyo University of Technology,
[†††]School of Media Science, Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo, Japan, 192-0982
ar@mf.teu.ac.jp, ohno@cc.teu.ac.jp, iida@media.teu.ac.jp

Abstract

This paper describes a method of automatic emotional degree labeling for speaker's anger utterances during natural Japanese dialog. First, we explain how to record anger utterance appeared in natural Japanese dialog. Manual emotional degree labeling was conducted in advance to grade the utterances by a 6 Likert scale to obtain a referencial anger degree. Then experiments of automatic anger degree estimation were conducted to label an anger degree with each utterance by its acoustic features. Also estimation experiments were conducted with speaker-dependent datasets to find out any influence of individual emotional expression on automatic emotional degree labeling. As a result, almost all the speaker's models show higher adjusted R^2 so that those models are superior to the speaker-independent model in those estimation capabilities. However, a residual between automatic emotional degree and manual emotional degree (0.73) is equivalent to those of speaker's models. There still has a potential to label utterances with the speaker-independent model.

1. Introduction

With great advance of automatic speech recognition (ASR) system, a voice command system or an interactive voice response system such as an automotive navigation system and a customer service system are demanded to be more sensitive and communicative to users. These systems currently process linguistic information, but not process nonlinguistic information or paralinguistic information which users present during a dialog with a computer. For that reason, computers can obtain less information on a speaker through a dialog than human listeners can. If computers would recognize user's emotions conveyed by acoustic information, more appropriate reactions could be taken toward users. A large emotional speech corpus could be required for machine learning of speaker's emotion to realize an emotion recognition system. However manual emotional labeling for a large corpus is troublesome and time-consuming task. Our approach is to automatically label whole utterances in emotional speech corpus as a certain emotion by acoustic features in order to design large emotional corpus used for machine learning of speaker's anger emotion.

Several related works (Ang et al., 2002; Cowie et al., 2001; Banse and Scherer, 1996) have been done in the area of analyzing emotional speech. Our study differs from the former studies in several ways. First, many former studies have recorded emotional speech of actors who had been instructed to read sentences which conveyed some particular emotions. We specifically recorded natural dialogs contained spontaneous anger utterances that naturally occur during a dialog for emotion recognition. Second, the former studies have been classified recorded emotional speech into several certain emotions categorized according to a psychological emotional model. We labeled an anger degree with each utterance according to a continuous emotional scale.

2. Recording

Human-computer and human-human pseudo-dialogs were recorded to collect anger utterances during a natural Japanese dialog. The human-computer pseudo-dialog simulated a dialog with a telephonic reservation system and the human-human pseudo-dialogs simulated a dialog taken place when user phoned to a customer-support contact center.

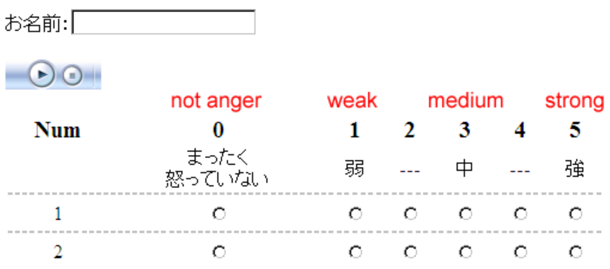
Speakers were 10 university students, 5 males and 5 females. Each speaker assumed the role of a user, while one of the authors took the role of an operator. The speaker spoke in a soundproof box to the operator outside through a headset microphone to make a non-face-to-face conversation. In the two kinds of pseudo-dialogs, only minimal information to proceed the dialogs was given to the speakers to record spontaneous emotional utterances following the operator's action.

To induce speaker's anger emotion, the operator forced the speaker to make the same answer several times in the human-computer pseudo-dialog, by feigning recognition failure or pretending to have some system errors. The operator objected to the speaker's claim when the speaker made a complaint in the human-human pseudo-dialog, for recording the speaker's anger emotion. Table 1 shows two samples extracted from the recorded dialogs. A symbol "O : " in the table shows the operator's utterance and a symbol "U : " in the table shows user's utterance. In (a) the human-computer pseudo-dialog, a user irritatedly repeated the same word in the user's second utterance against the operator's wrong recognition. In (b) the human-human pseudo-dialog, a user broke into the operator's utterance in anger with his second utterance before the operator finished her first utterance.

The recorded users' speech were cut into the utterances, regarding continuous speech segment between pauses more

(a) The human-computer pseudo-dialog	
O :	お問い合わせの内容はどのようなものですか？ (How could I help you?)
U :	あの料金についてちょっと聞きたいと 思ってるんですけど。 (Well, I would like to know about the fees.)
O :	設備でよろしいですか？ (About our facilities?)
U :	**いえちょっと、料金ですね、料金。** **(No, uh, the fees , THE FEES.)**
O :	料金でよろしいですか？ (About the fees?)
U :	はい。 (Yes.)
(b) The human-human pseudo-dialog	
U :	いや、でも発行されてるんで。 (But, it's been issued.)
O :	いや、８万台の予約番号というのは (But the reservation number of 80 thousand is ...)
U :	**え、でもされてるんで。** **(Huh? It's been actually issued!)**
O :	会議室では発行されないんですよ。 (never issued for the conference room reservation.)
U :	じゃ、どこでされるんスかね。 (So, where would this number be issued on earth?)

Table 1: Sample recorded dialogs which the operator induced user's anger.



お名前:

	not anger	weak		medium		strong
Num	0	1	2	3	4	5
	まったく 怒っていない	弱	---	中	---	強
1	○	○	○	○	○	○
2	○	○	○	○	○	○

Figure 1: The answer sheet for subjective evaluation.

labeler	*kappa*	*z-score*	*p-value*
EF01	0.27	16.35	0.00
EF02	0.07	4.73	0.00
EF03	0.35	20.74	0.00
EM01	0.23	10.14	0.00
EM02	0.27	15.06	0.00
EM03	0.19	7.61	0.00
EM04	0.45	25.10	0.00
EM05	0.37	20.11	0.00
EM06	0.34	18.42	0.00
EM07	0.39	20.48	0.00
EM08	0.37	20.99	0.00
EM09	0.45	24.10	0.00
ALL	0.12	38.80	0.00

Table 2: The kappa between labelers and mode.

than 200 ms as a unit of utterance. 1160 utterances were selected randomly for statistical analysis from the utterances of 5 speakers (3 males and 2 females). Any adjustment of the numbers of each speaker's utterances were never made owing to make a dataset which reflect a proportion of actual speaker's amount of utterances. The dataset were composed of 661 male utterances and 499 female utterances.

3. Manual anger degree labeling as a referencial emotional degree

As a referencial anger degree, manual labeling for all 1160 utterances was conducted. A mean of labelers' graded values for each utterance were adopted as a referencial emotional degree for following automatic labeling experiment.

3.1. Manual labeling methodology

A six-scale subjective evaluation was conducted to grade each utterance on how angry it was heard. Labelers were 12 university students, 9 males and 3 females. 12 labelers listened to all 1160 utterances which were presented once at random labeler by labeler to avoid the influence of presentation order on labeling. The labelers graded each utterance on a scale of 0 (not anger) and from 1 (weak anger) to 5 (strong anger). They were asked to grade each utterance by acoustic characteristics of the utterance, not by a meaning of the utterance to clarify what acoustic features were contribute for the labelers to distinguished the different grade of anger utterances. Figure 1 is an example of the answer sheet for subjective evaluation.

3.2. Result and evaluation of manual labeling

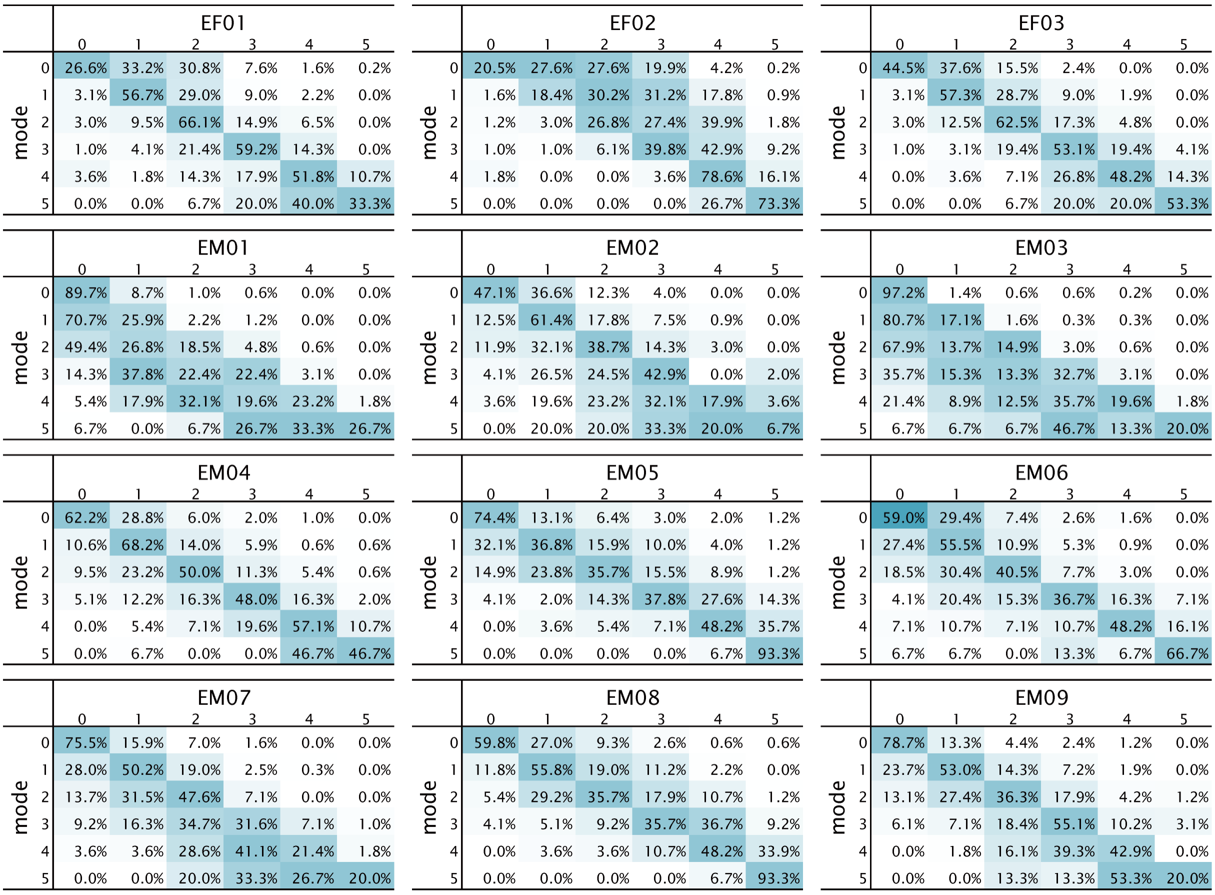
As a result of the subjective evaluation, the agreement and correlation between inter-labelers were analyzed. Table 2 shows kappa coefficient beween each labeler's value and mode value by 12 labelers. Figure 2 are confusion matrices of each labeler's value and mode value of 12 labelers. More dark color was painted on a cell of a higher percentage of agreement for each confusion matrix.

As for Table 2, although a kappa coefficient of all labeler's value (0.12) is extremely low, the highest kappa coefficients between each labeler's value and mode value of 12 labelers is 0.45 of both EM04 and EM09, and the second highest is 0.39 of EM07.

There are the strongest correlation, 0.68, between EM05 and EM08, and the weakest correlation, 0.33 between EF01 and EM02. The average inter-labeler correlation is 0.52. On the other hand, the confusion matrices of Fig. 2 show that many labelers agreed with the mode value. But the figure also shows that many labelers judged the utterances as the adjacent scale to the mode value of 12 labelers.

In our previous work (Arimoto et al., 2005) on a classification of anger utterance into a group of anger degree in a discrete scale, the result also showed that a higher classification accuracy was obtained when making allowance for classification of each datum into the adjacent clusters. That was caused by the utterances located close to the boundary between adjacent clusters, because those utterances misclassified into the adjacent clusters.

These results of inter-labeler agreement and correlation and



EF01

mode	0	1	2	3	4	5
0	26.6%	33.2%	30.8%	7.6%	1.6%	0.2%
1	3.1%	56.7%	29.0%	9.0%	2.2%	0.0%
2	3.0%	9.5%	66.1%	14.9%	6.5%	0.0%
3	1.0%	4.1%	21.4%	59.2%	14.3%	0.0%
4	3.6%	1.8%	14.3%	17.9%	51.8%	10.7%
5	0.0%	0.0%	6.7%	20.0%	40.0%	33.3%

EF02

mode	0	1	2	3	4	5
0	20.5%	27.6%	27.6%	19.9%	4.2%	0.2%
1	1.6%	18.4%	30.2%	31.2%	17.8%	0.9%
2	1.2%	3.0%	26.8%	27.4%	39.9%	1.8%
3	1.0%	1.0%	6.1%	39.8%	42.9%	9.2%
4	1.8%	0.0%	0.0%	3.6%	78.6%	16.1%
5	0.0%	0.0%	0.0%	0.0%	26.7%	73.3%

EF03

mode	0	1	2	3	4	5
0	44.5%	37.6%	15.5%	2.4%	0.0%	0.0%
1	3.1%	57.3%	28.7%	9.0%	1.9%	0.0%
2	3.0%	12.5%	62.5%	17.3%	4.8%	0.0%
3	1.0%	3.1%	19.4%	53.1%	19.4%	4.1%
4	0.0%	3.6%	7.1%	26.8%	48.2%	14.3%
5	0.0%	0.0%	6.7%	20.0%	20.0%	53.3%

EM01

mode	0	1	2	3	4	5
0	89.7%	8.7%	1.0%	0.6%	0.0%	0.0%
1	70.7%	25.9%	2.2%	1.2%	0.0%	0.0%
2	49.4%	26.8%	18.5%	4.8%	0.6%	0.0%
3	14.3%	37.8%	22.4%	22.4%	3.1%	0.0%
4	5.4%	17.9%	32.1%	19.6%	23.2%	1.8%
5	6.7%	0.0%	6.7%	26.7%	33.3%	26.7%

EM02

mode	0	1	2	3	4	5
0	47.1%	36.6%	12.3%	4.0%	0.0%	0.0%
1	12.5%	61.4%	17.8%	7.5%	0.9%	0.0%
2	11.9%	32.1%	38.7%	14.3%	3.0%	0.0%
3	4.1%	26.5%	24.5%	42.9%	0.0%	2.0%
4	3.6%	19.6%	23.2%	32.1%	17.9%	3.6%
5	0.0%	20.0%	20.0%	33.3%	20.0%	6.7%

EM03

mode	0	1	2	3	4	5
0	97.2%	1.4%	0.6%	0.6%	0.2%	0.0%
1	80.7%	17.1%	1.6%	0.3%	0.3%	0.0%
2	67.9%	13.7%	14.9%	3.0%	0.6%	0.0%
3	35.7%	15.3%	13.3%	32.7%	3.1%	0.0%
4	21.4%	8.9%	12.5%	35.7%	19.6%	1.8%
5	6.7%	6.7%	6.7%	46.7%	13.3%	20.0%

EM04

mode	0	1	2	3	4	5
0	62.2%	28.8%	6.0%	2.0%	1.0%	0.0%
1	10.6%	68.2%	14.0%	5.9%	0.6%	0.6%
2	9.5%	23.2%	50.0%	11.3%	5.4%	0.6%
3	5.1%	12.2%	16.3%	48.0%	16.3%	2.0%
4	0.0%	5.4%	7.1%	19.6%	57.1%	10.7%
5	0.0%	6.7%	0.0%	0.0%	46.7%	46.7%

EM05

mode	0	1	2	3	4	5
0	74.4%	13.1%	6.4%	3.0%	2.0%	1.2%
1	32.1%	36.8%	15.9%	10.0%	4.0%	1.2%
2	14.9%	23.8%	35.7%	15.5%	8.9%	1.2%
3	4.1%	2.0%	14.3%	37.8%	27.6%	14.3%
4	0.0%	3.6%	5.4%	7.1%	48.2%	35.7%
5	0.0%	0.0%	0.0%	0.0%	6.7%	93.3%

EM06

mode	0	1	2	3	4	5
0	59.0%	29.4%	7.4%	2.6%	1.6%	0.0%
1	27.4%	55.5%	10.9%	5.3%	0.9%	0.0%
2	18.5%	30.4%	40.5%	7.7%	3.0%	0.0%
3	4.1%	20.4%	15.3%	36.7%	16.3%	7.1%
4	7.1%	10.7%	7.1%	10.7%	48.2%	16.1%
5	6.7%	6.7%	0.0%	13.3%	6.7%	66.7%

EM07

mode	0	1	2	3	4	5
0	75.5%	15.9%	7.0%	1.6%	0.0%	0.0%
1	28.0%	50.2%	19.0%	2.5%	0.3%	0.0%
2	13.7%	31.5%	47.6%	7.1%	0.0%	0.0%
3	9.2%	16.3%	34.7%	31.6%	7.1%	1.0%
4	3.6%	3.6%	28.6%	41.1%	21.4%	1.8%
5	0.0%	0.0%	20.0%	33.3%	26.7%	20.0%

EM08

mode	0	1	2	3	4	5
0	59.8%	27.0%	9.3%	2.6%	0.6%	0.6%
1	11.8%	55.8%	19.0%	11.2%	2.2%	0.0%
2	5.4%	29.2%	35.7%	17.9%	10.7%	1.2%
3	4.1%	5.1%	9.2%	35.7%	36.7%	9.2%
4	0.0%	3.6%	3.6%	10.7%	48.2%	33.9%
5	0.0%	0.0%	0.0%	0.0%	6.7%	93.3%

EM09

mode	0	1	2	3	4	5
0	78.7%	13.3%	4.4%	2.4%	1.2%	0.0%
1	23.7%	53.0%	14.3%	7.2%	1.9%	0.0%
2	13.1%	27.4%	36.3%	17.9%	4.2%	1.2%
3	6.1%	7.1%	18.4%	55.1%	10.2%	3.1%
4	0.0%	1.8%	16.1%	39.3%	42.9%	0.0%
5	0.0%	0.0%	13.3%	13.3%	53.3%	20.0%

Figure 2: The confusion matrices of mode vs. each labeler.

speakerID	*med.* = 0	*med.* > 0	*Total*
FSM	37	99	136
FTM	117	246	363
MIA	50	221	271
MMR	9	124	133
MTK	103	154	257
TOTAL	316	844	1160

Table 3: The number of utterances

our previous finding might reflect that each emotional degree is not in a discrete-scale, but in a continuous-scale.

To obtain a continuous value for each utterance as score of its anger degree, a mean of overall 12 evaluated values except outliers was calculated in every utterance. Also, utterances with extremely low evaluation value (median=0) were removed from a dataset to avoid influence of expression of other emotion. Table 3 shows the number of utterances of our dataset.

Figure 3 shows histograms on anger degree in each speakers. The histograms of 4 out of 5 speakers are like lognormal distribution, but the histogram of FSM shows different tendency from the others. It shows the second peak of the density of stronger anger degree utterances. To avoid an influence of the tendency of FSM anger utterances, two kinds of datasets were prepared, one is a speaker-independent dataset and the other is a speaker-dependent dataset for a experiment of automatic anger degree labeling. The speaker-independent dataset composed of whole 844 utterances and the speaker-dependent datasets composed of different number of utterances according to the speakers.

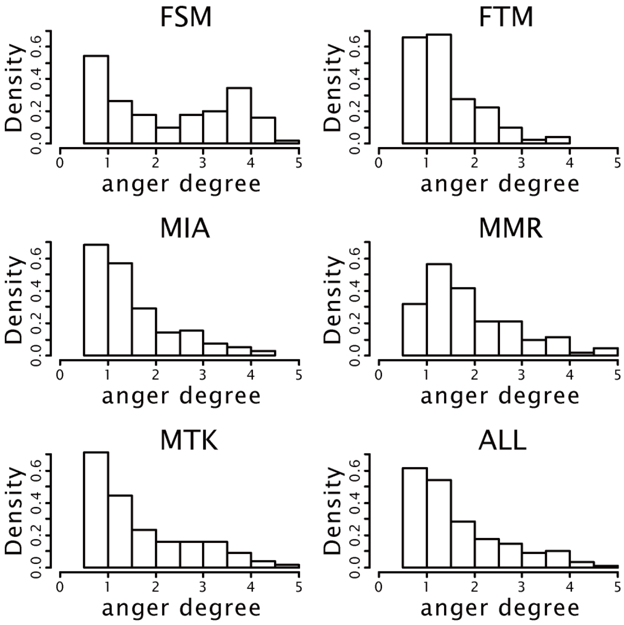


Figure 3: The histograms of anger degree in each speaker.

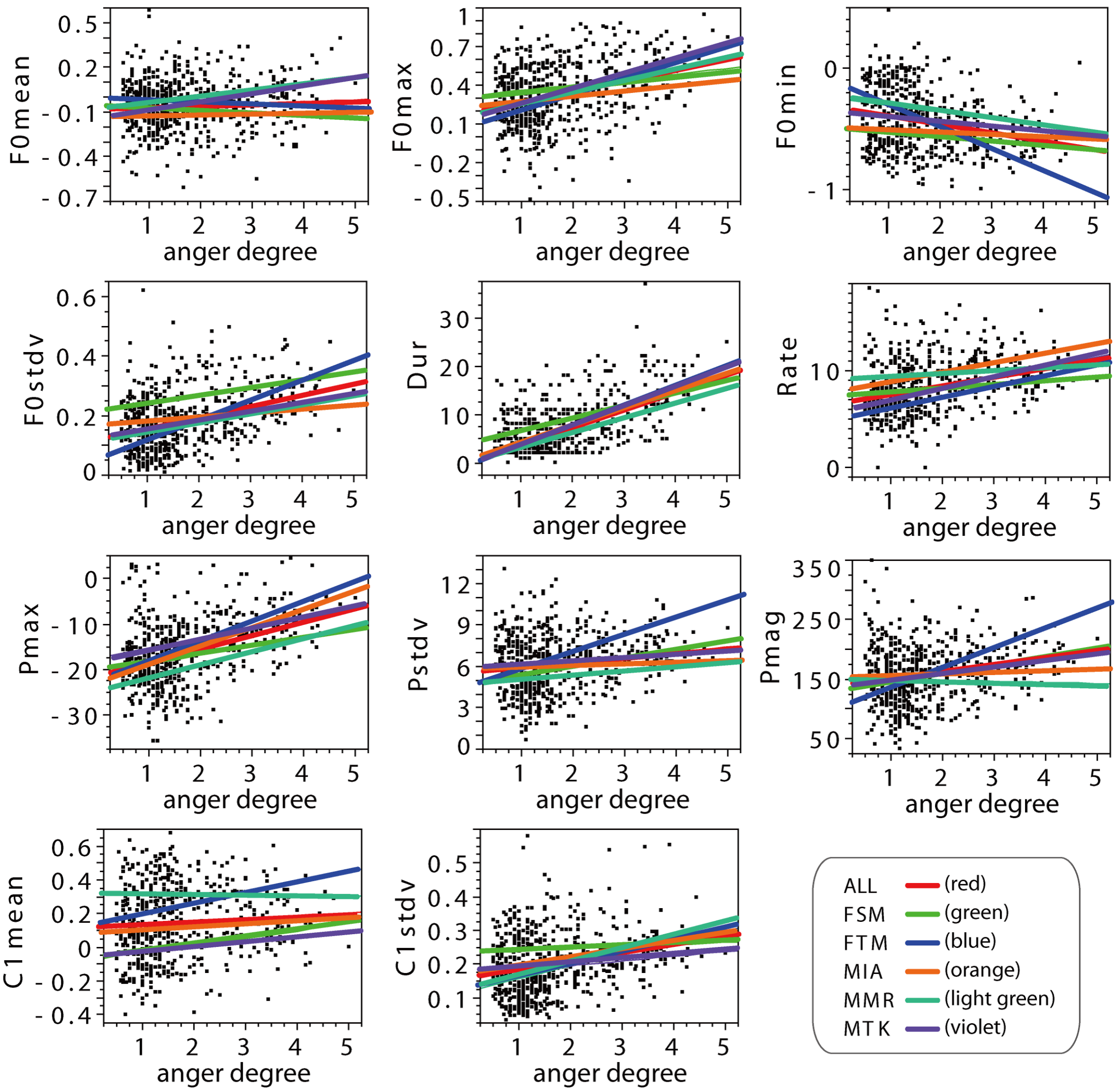


Figure 4: The distributions of each parameter vs anger degree.

parameter	*description*
F0mean	speaker-normalized F_0 mean
F0min	speaker-normalized F_0 min
F0max	speaker-normalized F_0 max
F0stdv	standard deviation of F_0
Dur	average mora number within a breath group
Rate	speaking rate (mora/s)
Pstdv	standard deviation of short-term power
Pmax	short-term power max
Pmag	magnitude of short-term power changes
C1mean	average of the first cepstral coefficient
C1stdv	standard deviation of the first cepstral coefficient

Table 4: The acoustic parameters

4. Acoustic parameters

We prepared 11 acoustic parameters with reference to former studies (Ang et al., 2002; Cowie et al., 2001; Banse and Scherer, 1996; Arimoto et al., 2007). Table 4 shows all adopted parameters and their descriptions. Every parameter has the representative value of a whole utterance such as a mean or a standard deviation.

For the parameters calculated with F_0 value, speaker-normalized values were prepared by simply subtracting the average of all data of each speakers, to remove the effect of the differences between each speakers.

There are not strong correlations among every combinations of each parameters from the training set. The strongest is 0.69 between Pstdv and Pmag, the weakest is 0.00 between Rate and C1mean.

Figure 4 shows the distribution between each parameters and anger degree. 6 colored lines on each panel are the regression lines speaker by speaker. Almost all the speakers show the same tendencies against the anger degree, but some speakers show more strong inclining or declining tendencies with the paramters such as F0min, Pstdv and Pmag of FTM.

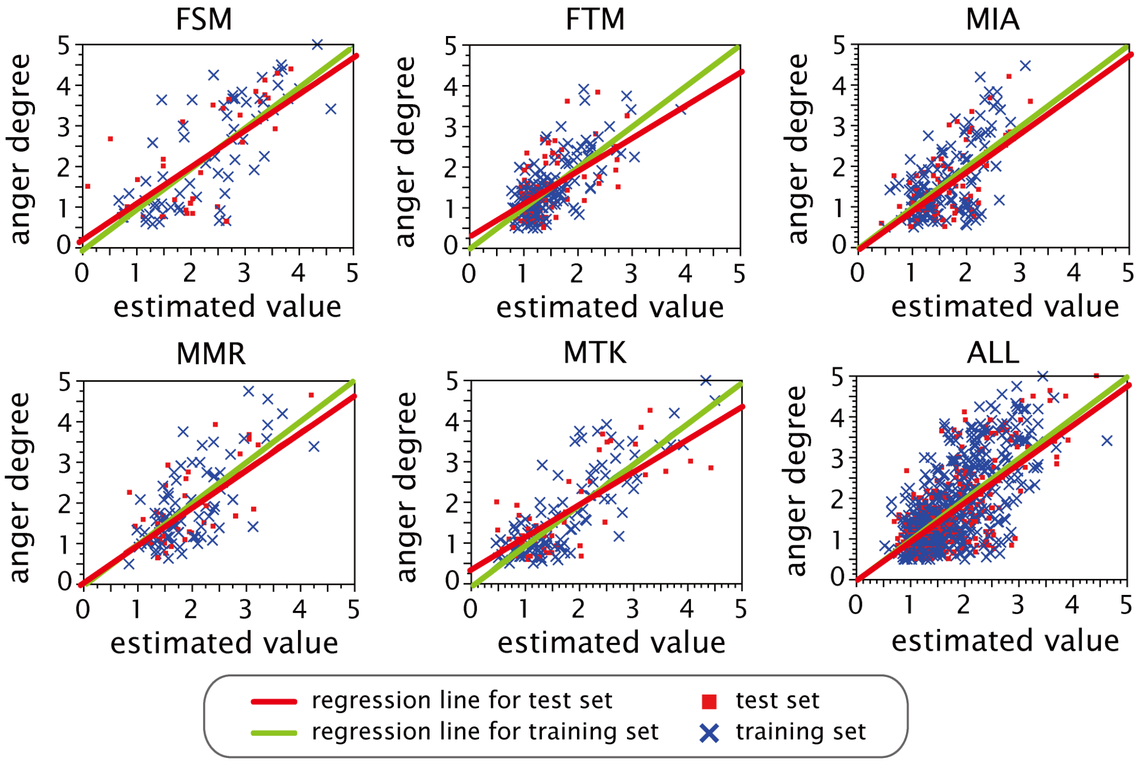


Figure 5: The distributions of anger degree vs. estimated value.

model	*R*	$\hat{R}^2$	*F-value*	*p-value*	*AIC*
FSM	0.76	0.54	13.82	9.76E-10	-9.46
FTM	0.71	0.49	40.83	1.66E-23	-222.68
MIA	0.61	0.35	13.91	2.27E-12	-98.30
MMR	0.69	0.45	18.04	1.59E-10	-52.27
MTK	0.81	0.65	64.84	2.28E-23	-91.87
ALL	0.67	0.44	64.78	3.26E-68	-366.11

Table 5: Multiple correlation coefficient (R), adjusted R square ($\hat{R}^2$) , F-value, p-value and AIC of each model

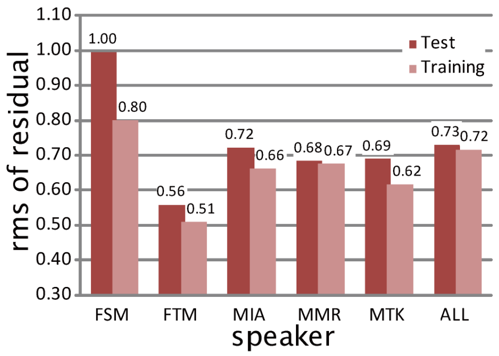


Figure 6: Root mean square of residual between automatic labeling value and manual labeling value.

5. Automatic anger degree labeling

5.1. Automatic labeling method

Each dataset was divided into a training set and a test set at random in the proportion of 2 to 1 for an experiment of automatic anger degree labeling. Using this training set, labeling experiments were conducted to estimate the anger degree of each utterance using multiple linear regression analysis based on least-square method. Forward selection was applied for the labeling experiment to clarify which parameters contribute to the anger degree labeling. Also n-fold cross validation (n=3) was conducted to make each speaker-dependent model because its dataset size was too small to make a model with 2/3 of speaker-dependent dataset.

5.2. Result

Figure 5 shows that the distributions between anger degree and estimation value calculated by the linear regression model of each dataset. Table 5 shows multiple correlation coefficient (R), adjusted R square ($\hat{R}^2$) , F-value, p-value and AIC of each model of the training set. Because n-fold cross validation made three models for one speaker-dependent model, only the highest $\hat{R}^2$ model was showed in Table 5.

Root mean square of residual between automatic anger degree labeling value and manual anger degree labeling value for each dataset were calculated for an evaluation of automatic labeling method.

The $\hat{R}^2$s of all speaker's models are between 0.35 (MIA) and 0.65 (MTK), and those root mean square of residual are between 0.51 (the training set of FTM) to 1.00 (the test set of FSM) for both the test and training sets. The speaker-independent model does not show a high $\hat{R}^2$ (0.44) between automatic labeling values and manual labeling values, and its root mean square of residual is under 0.73 for both the test and training sets, comparing with each

paramter	*FSM*	*FTM*	*MIA*	*MMR*	*MTK*	*ALL*
F0mean		- -	++			
F0min	-		- -	-	-	
F0max					+++	
F0stdv	++	++		+++	-	+
Dur	++	+++	+++	+++	+++	+
Rate	++	+	+++	+++		+
Pmax	+++	+++	+++		+	+
Pmag	++		- - -		-	-
Pstdv		-	- -		-	-
C1mean	+++	++	-	+++		+
C1stdv	-	- -	-		- -	

Table 6: The selected parameters of all the models

speaker-dependent model.

6. Discussions

Comparing every speaker-dependent model to the speaker-independent model in the Table 5, four of five speaker's model (FSM, FTM, MMR, and MTK) shows stronger correlations and higher $\hat{R}^2$ values than the speaker-independent model. Also three of five speaker's model (MIA, MMR, and MTK), show approximately equivalent root mean square of residual to the speaker-independent model and FTM's model shows remarkably lower root mean square of residuals than the speaker-independent model. On the other hand, FSM's model shows extremely larger residual than the speaker-independent model.

From this result, almost all the speaker's model more accurately estimates individual anger degree than the speaker-independent model. But MIA's model, which shows weak correlation and lower $\hat{R}^2$ value than the speaker-independent model, and FSM's model, which shows rather large residual than the speaker-independent model, could not be more suitable for anger degree estimation than the speaker-independent model.

According to the Fig. 3, the FSM's histogram of manual anger degree shows the two peaks of utterance number, one is around anger degree 1 and the other is anger degree 4, while the others' show one peak and gradually reduce the number of utterance toward the strong anger. This indicates that FSM's anger could be more distinctive than the others'. Taking account for FSM's anger degree tendency, FSM's residual is rather large with linear regression analysis.

From the distribution between anger degree and estimated values calculated by MIA's model in Fig. 5, we could find that the manual labeling (anger degree) has the utterance near anger degree 5, but the estimation of all utterances is under 3.5. This could be caused by the less utterance number of stronger anger. Many lower anger utterance in dataset influence on the estimation of the model and caused its weaker correlation and lower $\hat{R}^2$ value.

Table 6 shows selected parameters for the speaker-independent model and the speaker-dependent models. In three times of multiple linear regression analysis based on n-fold cross validation (n=3), different sets of parameters were adopted for each models. When selected parameters showed negative standard partial regressive coefficients, the symbol“ - ” was put on a cell of its parameter. When selected parameters were showed positive standard partial regressive coefficient, a symbol“ + ” was put on a cell of its parameter.

Table 6 indicates that some speaker's models adopted parameters which were not adopted by speaker-independent model. Also some parameters of each speaker's model shows opposite sign to those of the speaker-independent model. For example, the Pmag of FSM shows positive sign ("++") while that of the speaker-independent model shows negative one. It was considered that the speaker-dependent model has is own acoustic features of anger speech, and the speaker-independent model could not cover all the speaker's acoustic features of anger speech.

However, Figure 6 indicates that almost all speaker's models show approximate equivalent residual to speaker-independent model. So there is some possibility to estimate the utterance of open data on speakers by the speaker-independent anger degree model.

7. Conclusion

We have studied a method of automatic emotional degree labeling for speaker's anger utterances during natural Japanese dialog. The experiments of automatic anger degree estimation were conducted to label an anger degree with each utterance by its acoustic features. Also the estimation experiments were conducted with speaker-dependent datasets to find out any influence of individual emotional expression on automatic emotional degree labeling. As a result, we found that almost all the speaker-dependent estimation model was superior to the speaker-independent model in its estimation capabilities. But the differences of those estimation capabilities are rather small, there still has a potential for the speaker-independent anger degree estimation.

8. References

J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 3, pages 2037–2040.

Y. Arimoto, S. Ohno, and H. Iida. 2005. A method for discriminating anger utterances from other utterances using suitable acoustic features. In *Proc. of SPECOM 2005*, pages 613–616.

Y. Arimoto, S. Ohno, and H. Iida. 2007. Acoustic features of anger utterances during natural dialog. In *Proc. of INTERSPEECH 2007*, pages 2217–2220.

R. Banse and K.R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

R. Cowie, E.D. Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, and W. Fellenz J.G. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18:32–80.