Acoustic Features of Anger Utterances during Natural Dialog

Yoshiko Arimoto[1], *Sumio Ohno*[2] *and Hitoshi Iida*[3]

[1]Graduate School of Bionics, Computer and Media Sciences,
Tokyo University of Technology, Tokyo, Japan
[2]School of Computer Science, Tokyo University of Technology, Tokyo, Japan
[3]School of Media Science, Tokyo University of Technology, Tokyo, Japan
ar@mf.teu.ac.jp, ohno@cc.teu.ac.jp, iida@media.teu.ac.jp

10.21437/Interspeech.2007-603

Abstract

This report focuses on an automatic estimation of speakers' anger emotion degree. Two kinds of pseudo-dialogs were held to collect spontaneous anger utterances during the natural Japanese dialog. In order to quantify the anger degree of utterances, a six-scale subjective evaluation was conducted to grade every utterance according to an anger emotion degree by twelve evaluators. With this data set, acoustic features of each utterance were examined to clarify what is the clue to estimate degree of anger utterances. To examine the possibility of automatic emotion estimation, we conducted experiment to estimate the degree of anger emotion automatically by multiple regression analysis using the acoustic parameters.

Index Terms: emotional speech, natural dialog, acoustic features

1. Introduction

Automatic speech recognition (ASR) systems are greatly demanded for customer-service systems. With advanced interactive voice response systems, human beings have more opportunities to have dialogs with computers. Current dialog systems process linguistic information, but do not process paralinguistic information. For that reason, computers can obtain less information from a speaker through a dialog than human listeners can. More appropriate reactions can be taken toward users if the computer will recognize the user's emotion conveyed by acoustic information. We aimed at automatic anger emotion estimation during natural Japanese dialog by its acoustic features manifesting anger utterance.

Several previous works have been done in the area of analyzing emotional speech such as [1], [2], [3], [4], and [5]. Our study differs from previous works in several ways. First, many previous studies have examined emotions of actors who had been instructed to read sentences which conveyed some particular emotions mainly for emotional speech synthesis. For emotion recognition, we specifically study spontaneous anger utterances that naturally occur during a dialog. Second, previous works have been aimed at a classification of utterances into several emotions categorized according to a psychological emotional model. Our study estimates the degree of one emotion according to a continuous emotional scale. Anger is our objective emotion for this study.

2. Speech data

This section introduces recorded speech data for an analysis and measurement method of speakers' degree of anger.

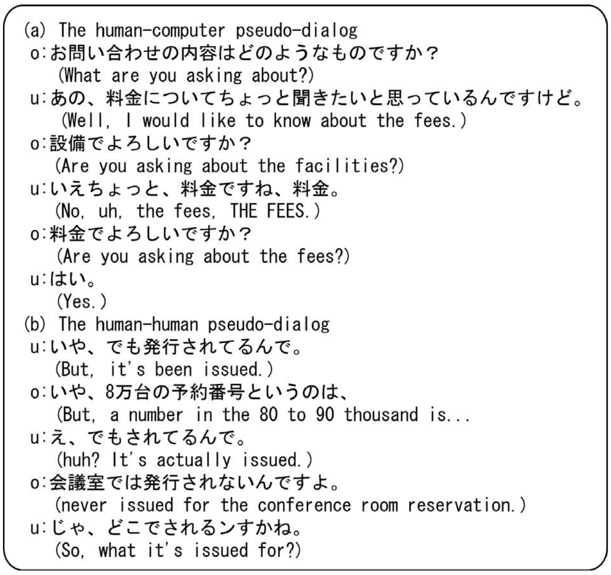
2.1. Recording

Human-computer and human-human pseudo-dialogs were recorded to collect anger utterances during a natural Japanese dialog. The two kinds of pseudo-dialogs simulated those of a telephonic reservation system and a customer-support contact center, respectively.

Speakers were 10 university students: 5 males and 5 females. Each speaker assumed the role of a user; one of the authors took the role of an operator. The speaker and the operator held several non-face-to-face dialogs. In those two kinds of dialogs, only minimal information on the pseudo-dialogs was given to the speakers to record spontaneous utterances following the operator's action.

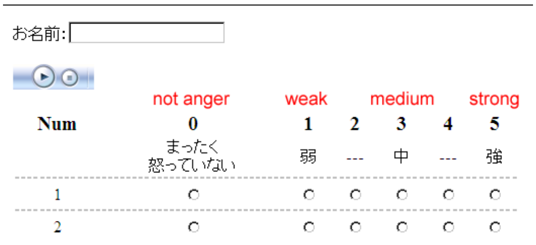
We adopted two means to induce speaker's anger emotion. To induce speaker's anger emotion, the operator feigned recognition failure or pretended to have some error and forced the speaker to make the same answer several times in the human-computer pseudo-dialog. In the human-human pseudo-dialog, the operator objected to the speaker's claim to induce the speaker's anger emotion, when the speaker made a complaint. Figure 1 shows two samples extracted from the recorded dialogs.

The number of recorded utterances was different for each speaker because of each speaker's characteristics. To reflect a proportion of the number of speaker's utterances, 1400 utterances were selected at random to use for statistical analyses. Those utterances were composed of 662 male utterances and 738 female utterances, from 6 (3 males and 3 females) of the 10 speakers.



(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？
(What are you asking about?)
u:あの、料金についてちょっと聞きたいと思っているんですけど。
(Well, I would like to know about the fees.)
o:設備でよろしいですか？
(Are you asking about the 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？
(Are you asking 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, a number in the 80 to 90 thousand is...
u:え、でもされてるんで。
(huh? It's 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、どこでされるンすかね。
(So, what it's issued for?)

Figure 1: *A part of recorded dialogs (o, the operator's utterance; u, the user's utterance).*



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 2: *Answer sheet for subjective evaluation.*

2.2. Measuring degree of expressed anger

To quantify the anger degree of each utterance, a six-scale subjective evaluation was conducted. Evaluators were 12 university students: 9 males and 3 females. The evaluators listened to each utterance through headphones to grade each of the 1400 utterances according to a scale from 0 (not anger) to 5 (strong). To clarify what acoustic features were helpful for the evaluators to grade isolated utterances according to the anger degree, each utterance was presented once in random order to the evaluators. Figure 2 shows the answer sheet for subjective evaluation.

As a result of the subjective evaluation, a mean of overall 12 evaluated values except outliers was calculated in every utterance as a score representing its anger degree. Table 1 shows utterances with the highest and lowest scores.

3. Acoustic parameters

We prepared 11 acoustic parameters characterizing acoustic features for each utterance with reference to previous works [1], [2], [3], [4], [5] and [6]. Table 2 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.

Table 1: *Utterances with the highest and lowest scores.*

score	utterance
highest	*daitai kikaidatte wakaranaijanaidesuka shocchu sohiuno okoruNde* (You can not say that the system is always secure because it often behaves like that.) *shisutemuni nanorette ittemo nanoranakattaNdesuyo* (The system did not respond when I asked it its name.)
lowest	*naNtoka* (Please!) *a, juhgojikara juhshichijide onegaishimasu* (from 15:00 to 17:00, please)

Table 2: *Acoustic parameters.*

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

3.1. Pitch parameters

As pitch parameters, F0mean, F0min, F0max, and F0stdv were prepared. For all pitch parameters, fundamental frequency (F_0) of each utterance was automatically extracted by STRAIGHT [7]. We prepared the gender-normalized parameters (F0mean, F0 min, F0max) to avoid the influence of a gender difference on an estimation. The gender-normalized parameters were calculated as an F_0 value for each utterance was simply subtracted from the average value of each gender datum.

3.2. Power parameters

As power parameters, Pstdv, Pmax and Pmag were prepared. Short-term power was calculated in a 20-ms window length at 5-ms intervals for those features. One parameter, Pmag, was a summation of RMS values of slopes of regression lines over 11 frames at every frame. It quantified the magnitude of changes of short-term power within a whole utterance.

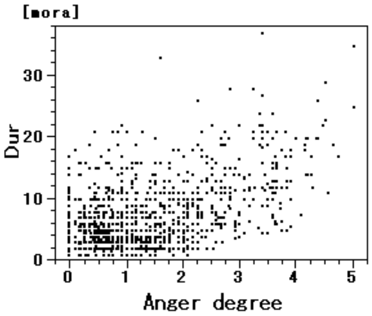


Figure 3: *Dur vs. anger degree*

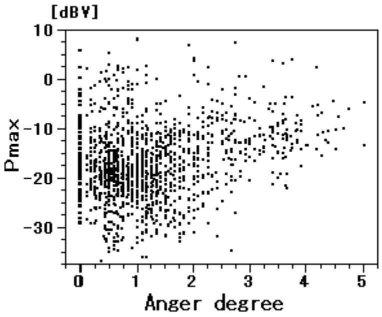


Figure 4: *Pmax vs. anger degree*

Table 3: *Selected Parameters in selected order.*

step	parameter	SPRC
1	Dur	0.4933
2	C1mean	0.1888
3	Rate	0.1795
4	Pmag	-0.1705
5	Pmax	0.1214
6	C1stdv	-0.0990
7	F0stdv	0.0878
8	F0min	0.0380

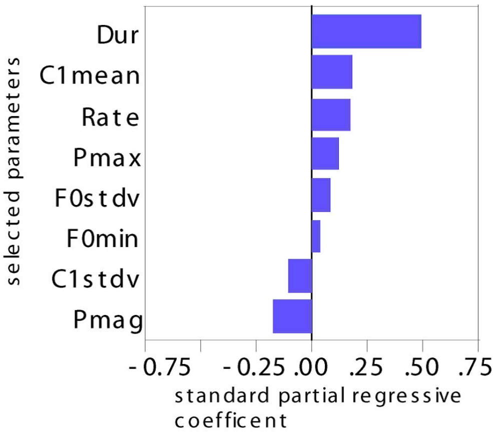


Figure 5: *Selected parameters in descending order.*

3.3. Duration and speaking rate parameters

As duration and speaking rate parameters, Dur and Rate were prepared. Dur was the average morae number within a breath group. Rate was the average speaking rate (mora/s) within a whole utterance.

3.4. Voice-quality parameters

As voice-quality parameters, C1mean and C1stdv were prepared. For those parameters, the first cepstral coefficient was adopted because it reflects an overall spectral tilt. The first cepstral coefficient was calculated only for voiced frames using FFT cepstrum method.

As a result of those parameter calculations, some utterances were removed from the dataset on account of missing values of some parameters. Consequently, a dataset with 1391 utterances was used for the following analysis and estimation experiment.

3.5. Distribution of parameters against scores of subjective evaluation

We examined individual tendencies of each parameter against the anger degree. Two tendencies are apparent in the distribution of acoustic parameters. Figures 3 and 4 show two examples of distributions between the parameters and the anger degree graded by subjective evaluation.

As for Dur in Fig. 3, when the utterance marked the higher score of subjective evaluation (the stronger anger degree), the parameter of its utterance shows the higher values. The tendency showed in Fig. 3 suggests that the parameter has individual estimation capability for the anger degree because it seems to correlate with the score of subjective evaluation.

On the other hand, as for Pmax in Fig. 4, a parameter like Pmax did not correlate with the score of subjective evaluation; it seems that it is difficult to estimate the anger degree individually. However, the parameter values fell within a limited range in the higher score of subjective evaluation, which suggests that all parameters, which show the same tendency as Pmax, have particular characteristics in the anger utterances.

4. Estimation experiment

The 1391 utterances were divided into a training set and a test set at random in the proportion of 2 to 1. Using this training set, an estimation experiment was 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 estimation experiment to clarify which parameters contribute to the estimation.

Standard errors of estimated values to scores of subjective evaluation for all utterances were calculated for an assessment of the result of the experiment.

5. Result and discussion

As a result of the experiment, the standard error was 0.78 for the training set, and 0.81 for the test set.

Table 3 and Fig. 5 show the selected parameters as a result of the experiment. In Table 3, the parameters are shown in selected order with its standard partial regression coefficient (SPRC). In Fig. 5, the parameters are shown in descending order of SPRC. The absolute values of SPRC show how much the selected parameters contribute to the anger degree estimation. An utterance with a high parameter value is estimated as a strong anger utterance when the value of SPRC of each parameter is positive. The utterance with a high parameter value is estimated as a weak anger utterance when it is negative.

As a result of the multiple regression analysis with forward selection, Dur, C1mean, Rate, Pmag, and Pmax were selected in the earlier steps. Those parameters get high absolute SPRC: greater than 0.1. Especially, Dur was more than 0.49. From these points, the features of anger utterance are as described below:

- longer duration and faster
- negative slope of overall spectral tilt
- louder and smaller magnitude of power changes

On the other hand, the pitch parameters were selected in the latter steps and those absolute SPRC were lower, which suggests that the pitch parameters contribute less for our estimation experiment. This result contradicts those of previous works, in which many researchers asserted a relationship between the pitch features and anger utterance.

The result of this experiment suggests that power features, the duration or speech rate features, and voice quality features might be more appropriate than pitch features for emotional degree estimation, particularly for assessing anger. Alternatively, the result might merely reflect the different qualities of emotional speech. Many previous works have specifically examined expressive emotions portrayed by professional actors. However, our speech data include emotions that occurred naturally during the dialog. To clarify whether this opposite result might have been caused by the different quality of emotional speech or might reflect the true nature of anger utterance during the natural dialog, it will be necessary to examine another pitch parameter that was not used in this report.

6. Conclusions

We examined an automatic estimation method to assess speakers' degree of anger during a natural Japanese dialog. Two kinds of pseudo-dialog were held to record spontaneous anger utterances. In order to quantify the anger degree of utterances, a six-scale subjective evaluation was conducted to grade every utterance according to an anger emotion degree by 12 evaluators. Using the dataset, acoustic features of each utterance were examined to clarify clues to detect anger utterances. Then, we conducted experiment to estimate the degree of anger automatically using multiple regression analysis with the acoustic parameters.

Results of the experiment demonstrate that the features of anger utterance are "longer duration and faster", "negative slope of overall spectral tilt" and "louder and smaller magnitude of power changes". However, the pitch parameters, which showed a relationship with anger utterances in the previous work, did not contribute to the estimation in our experiment.

For further research, we will examine parameters that are related to the time sequence of a dialog, such as a speaker's response time.

7. References

[1] Takeda, S. and Ohyama, G. and Tochitani, A. and Nishizawa, Y., "Analysis of prosodic features of "anger" expressions in Japanese speech", Journal of the Acoustical Society of Japan, Vol. 58, No. 9, pp. 561-568, 2002 (in Japanese).

[2] Ang, J. and Dhillon, R. and Krupski, A. and Shriberg, E. and Stolcke, A., "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog", Proc. Intl. Conf. on Spoken Language Processing, Vol. 3, pp. 2037-2040, 2002.

[3] Nagae, Y., "A Study on the Analysis and the Recognition of Speaker's Emotion Expressed in Utterances", Graduation Thesis of Faculty of Engineering, Utsunomiya University, 1997 (in Japanese).

[4] Cowie, R. and Cowie, E.D. and Tsapatsoulis, N. and Votsis, G. and Kollias, S. Fellenz, W. and Taylor J.G., "Emotion Recognition in Human-Computer Interaction", IEEE Signal Processing Magazine, Vol. 18, No. 1, pp. 32-80, 2001.

[5] Banse, R. and Scherer, K.R., "Acoustic Profiles in Vocal Emotion Expression", Journal of Personality and Social Psychology, Vol. 70, No. 3, pp. 614-636, 1996.

[6] Arimoto, Y. and Ohno, S. and Iida, H., "Emotion Labeling for Automatic Estimation of Speakers' Anger Emotion Degree", Oriental COCOSDA 2006, pp. 48-51, 2006.

[7] Kawahara, H. and Cheveigne, A. and Banno, H. and Takahashi, T. and Irino, T., "Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT", Proc. Interspeech2005, Lisboa, pp. 537-540, 2005.