interspeech 2008 incorporating SST 08

Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems

Yoshiko Arimoto[1], *Hiromi Kawatsu*[2], *Sumio Ohno*[3], *and Hitoshi Iida*[4]

[1]Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology
[2]The National Institute for Japanese Language
[3]School of Computer Science, Tokyo University of Technology
[4]School of Media Science, Tokyo University of Technology
ar@mf.teu.ac.jp, kawa2@kokken.go.jp, ohno@cc.teu.ac.jp, iida@media.teu.ac.jp

10.21437/Interspeech.2008-96

Abstract

For the purpose of determining emotion recognition by acoustic information, we recorded natural dialogs made by two or three players of online games to construct an emotional speech database. Two evaluators categorized recorded utterances in a certain emotion, which were defined with reference to the eight primary emotions of Plutchik's three-dimensional circumplex model. Furthermore, 14 evaluators graded utterances using a 5-point scale of subjective evaluation to obtain reference degrees of emotion. Eleven acoustic features were extracted from utterances and analysis of variance (ANOVA) was conducted to assess significant differences between emotions. Based on the results of ANOVA, we conducted discriminant analysis to discriminate one emotion from the others. Moreover, the experiment estimating emotional degree was conducted with multiple linear regression analysis to estimate emotional degree for each utterance. As a result of discriminant analysis, high correctness values of 79.12% for Surprise and 70.11% for Sadness were obtained, and over 60% correctness were obtained for most of the other emotions. As for emotional degree estimation, values of the adjusted R square ($\hat{R}^2$) for each emotion ranged from 0.05 (Disgust) to 0.55 (Surprise) for closed sets, and values of root mean square (RMS) of residual for open sets ranged from 0.39 (Acceptance) to 0.59 (Anger).

Index Terms: natural dialog, spontaneous speech, emotional speech, prosody, discriminant analysis, multiple regression analysis

1. Introduction

As text chat is the main method of communication among many online game users, the users who are slow at typing may feel reduced enjoyment playing online games. Those users demand voice chat systems that enable hands-free communication. For example, Second Life[1], a prominent online game, has already introduced a voice chat system as one tool for user communication. However, some users are unwilling to use such system because their voice can convey personally identifying information to many unspecified users.

As one solution for this problem, anonymity-protected voice chat systems have been proposed that would enable anonymous communication with many unspecified users, by removing personally identifying information from the voice while retaining various vocal expressions. Voice conversion systems are one such solution, but personal information still remains in prosodic information of the voice. Our approach to achieve an anonymity-protected voice chat system is to make use of an automatic speech recognition system and an emotion recognition system to recognize the content and emotion of utterances, and an emotional speech synthesis to re-create the expressive speech with the recognized information.

The present report describes our methods for emotional discrimination and estimating emotional degree, and the capabilities of acoustic features. To realize the emotion recognition system for our project, we first created a large database of emotional speech extracted from natural dialog, not from reading materials or acted emotional speech. Dialogs among groups of two or three game players were held through voice chat over the Internet while all players were engaged in playing a massively multiplayer online role-playing game (MMORPG).

We considered that emotion involves continuous degrees of emotional elevation. We therefore tagged utterances not only with emotional categories, but also with emotional degree by subjective evaluation. Subjective evaluation was conducted to categorize approximately 10,000 recorded utterances into one emotion as a reference emotion and to grade utterances with a 5-point scale of emotional degree. We then conducted emotion discrimination experiments and emotional degree estimation experiments with acoustic features extracted from recorded player utterances to examine the potential of a model to recognize emotions and to estimate emotional degree using several acoustic features.

2. Recording voice chat during online game

2.1. Game players

Players were 13 university students (9 males, 4 females) with previous experience of the online game. The average amount of playing time per month for all players was 33 h (range, 0-100 h) and the average duration of experience was 38 months (range, 12-61 months). All players participated in a recording in a male group or a female group of two or three players. Group members played the online game together as one

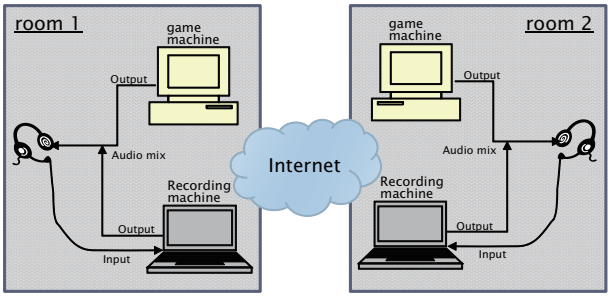

Figure 1: *Recording environment.*

Accepted after peer review of full paper

Table 1: *Definition of emotion.*

Emotion	Explanation
Joy	Feelings of gladness and thankfulness indicating intense satisfaction with something good
Acceptance	Feelings of positive welcoming and wanting to stay with something fascinating
Fear	Feelings of avoiding people or things that are harmful
Surprise	Feelings of loss of calm or judgment with strong upset about unexpected events
Sadness	Feelings of weeping or crying out as a result of irrevocable events such as meeting with misfortune
Disgust	Feelings of avoiding unacceptable states or acts
Anger	Feelings of irritation or annoyance with an unforgiven matter
Anticipation	Feelings of longing for achievement of a situation or presence of a favorable opportunity
Neutral	No apparent feelings at all
Others	Impossible to classify into the eight kinds of feelings, just too noisy, etc.

Table 2: *Number and rate of agreement between two annotators.*

Emotion	Number	Rate
Joy	287	38.2%
Acceptance	145	18.9%
Fear	82	23.5%
Surprise	345	43.6%
Sadness	112	27.8%
Disgust	172	24.9%
Anger	116	31.7%
Anticipation	236	30.0%
Neutral	334	27.5%
Others	113	24.7%
Total	1942	29.5%

Table 3: *Agreement of the result of grading emotional degree and number of utterances for estimation experiments.*

	Agreement			
Emotion	Kappa	Z-value	p-value	Number
Joy	0.33	42.19	0.00	207
Anticipation	0.31	36.44	0.00	215
Fear	0.28	33.15	0.00	181
Surprise	0.36	47.29	0.00	241
Disgust	0.26	26.73	0.00	226
Acceptance	0.29	26.99	0.00	233
Anger	0.33	38.62	0.00	188
Sadness	0.27	33.29	0.00	190
Total	0.30	107.22	0.00	1681

party and participated in game events while continuing discussion with each other. All players were required to speak with each other through the voice chat system and were prohibited from using the text chat function that came with the online game system except when talking to other players who were not participating in the recording.

2.2. Recording environment

Natural dialog during online gaming was recorded with the aim of constructing an emotional speech database. Two or three players were chatting with each other through the voice chat system, Skype[2]. Dialog between players was recorded using the voice-recording system for Skype, Tapur[3].

Each player was playing the online game in a remote location. Figure 1 shows our recording environment. Three online games (Ragnarok Online, Monster Hunter Frontier, and Red Stone) were utilized for recording scenarios.

Recorded dialogs were segmented into utterances for transcription. Any continuous speech segment between pauses exceeding 400 ms was regarded as a unit of utterance. For transcription, three kinds of tags for coughs, laughs, and noises and utterances that could not be transcribed were prepared. As a result, the total number of utterances in our database was 9114.

3. Annotation of emotional category and emotional degree

3.1. Selection and definition of emotional category

To tag one emotion to each of the 9114 utterances, evaluators shared the load of subjective evaluation for every utterance with each other. To remove gaps between the definition of emotions among evaluators, 8 emotions were selected with reference to the primary emotions in Plutchik's three-dimensional circumplex model[4]. In addition to Plutchik's 8 emotions, "Neutral" utterances and "Others" were added to our emotional categories for subjective evaluations. Moreover, all 10 emotional categories were defined with reference to a dictionary[5]. Table 1 shows our emotional categories and the definitions used for judgment.

3.2. Tagging emotional category

As the utterances of two players, 03_FMA and 02_MFM, displayed insufficient amplitude to be judged by ear and to be analyzed acoustically, 1009 utterances from these two players were excluded from the target utterances. An additional 1527 utterances with tags of cough, laugh or noises and utterances that could not be transcribed were excluded. As a result, target utterances for the annotation of emotional category excluded 2536 utterances.

Target utterances were divided into eight sets. Each evaluator tagged utterances with an emotion considered to represent the emotion present in the utterance. Evaluators were asked to judge utterances by those acoustic characteristics, not by the content of the utterance.

As result of emotional category tagging, agreement between two evaluation values was not particularly high (k-value, 0.20). Table 2 shows the number and rate of utterances for which the judgments of the two evaluators agreed. The 1829 utterances from 9 emotions in which judgments of both evaluators agreed and the 113 utterances of "Other" category were excluded were used in the subsequent discrimination analysis.

3.3. Grading emotional degree

Target utterances for emotional degree grading were extracted from the 1829 utterances annotated by emotional category in Section 3.2. If the number of utterances for an emotional category did not reach 250, utterances that one evaluator tagged

Table 4: *Acoustic features.*

Sign	Explanation
Fmini	Minimum value of gender-normalized F_0 value
Fmaxi	Maximum value of gender-normalized F_0 value
Fmean	Mean of gender-normalized F_0 value
Fstdv	Standard deviation of F_0 value
Dmora	Average number of morae in phrases separated by comma
Drate	Speaking rate (mora/s)
Pmaxi	Maximum value of short-term power
Pstdv	Standard deviation of short-term power
Pmagn	Magnitude of short-term power
Cmean	Mean of the first cepstral coefficient
Cstdv	Standard deviation of the first cepstral coefficient

with the emotion and the other tagged as "neutral" were added as emotional utterances of weaker degree. The total number of target utterances was 1854 and each emotional category comprised 250 utterances at most (Joy, Surprise, Disgust, Acceptance) and 197 utterances at least (Anger).

Evaluators graded each utterance on a Likert scale from 1 (weak emotion) to 5 (strong emotion). To obtain a continuous value for each utterance as an emotional degree, the mean of 13 evaluations was calculated for every utterance.

Table 3 shows the inter-evaluator agreements regarding grading of emotional degree for each emotion and the number of utterances for estimation experiments. The 1681 utterances which excluded the utterances with noise or the utterances apparently misjudged into the other emotion in the emotional category tagging were used in the subsequent estimation experiments.

4. Acoustic features

4.1. Extracting acoustic features

Eleven acoustic features were analyzed for all recorded utterances with reference to the previous work[6]. All 11 acoustic features could be classified into 4 groups: pitch features; power features; speaking rate / duration features; and voice-quality features. Values within 10% from the highest or lowest F_0 values were adopted as Fmaxi and Fmini, respectively. Some features could not be extracted from 8 utterances, because those utterances included only voiceless sound or were too short. These 8 utterances were thus excluded from the target utterances for analysis. Table 4 shows the extracted features and descriptions.

4.2. Correlations between features

The correlation analysis was conducted between every combination of the 11 features extracted from the 1821 utterances. As a result, strong correlations were seen in both "Fmini and Fmean", and "Fmaxi and Fmean". As a result, only Fmean was not used for the following experiments, because no relationship exists between "Fmini and Fmaxi", so that if Fmean were removed from the feature set, influences of strong correlations between features could be avoided while retaining the flexibility of discrimination or estimation capability of the feature set.

5. Discrimination of emotional category

5.1. Discrimination experiment method

ANOVA was conducted for every feature to determine which emotional categories could be discriminated from others. The 1821 utterances were divided into a closed set and an open set in the proportion of 3 to 1. Discriminant analysis was conducted with the closed set to create models to discriminate one emotion from the other emotions. For discriminant analysis, only features that were significant according to the results of ANOVA were adopted.

A few emotions were not discriminated from others. However, when those emotions were in a group, that group could be discriminated from others. ANOVA and discriminant analysis were conducted for those grouped emotions to identify features that could discriminate one emotion from others in the group.

5.2. Results and discussions

Table 5 shows the results of ANOVA and discriminant analysis. Cells for a significant feature that discriminates one emotion from the others were marked as "*". In the three columns from the right of Table 5, correctness values for the open set are shown.

Six emotions, Joy, Surprise, Sadness, Anger, Anticipation and Neutral, could be discriminated from other emotions using the features regarded as significant features according to the results of ANOVA. Some features could discriminate groups of "Sadness & Disgust", "Surprise & Anger", and "Acceptance & Fear & Disgust & Anticipation" from the others. Pmaxi was identified as the most effective for discriminating between emotions, as it shows the significant differences between many emotions according to the results of ANOVA.

As a result of discriminant analysis, high correctness values were obtained for Surprise and Sadness, at 79.1% and 70.1%. Correctness for almost all the other emotions was also high, at 60% or more, with the exception of Joy. As for discrimination experiments among the groups as previously stated, total correctness was 67.1% for Sadness and Disgust discrimination, whereas 76.2% correctness was obtained for the discrimination of Sadness against Disgust. Moreover, total correctness of Surprise and Anger discrimination was quite high at 79.0%, with 80.0% correctness for Surprise and 75.9% for Anger to discriminate each from the other. Discrimination accuracies for each emotion in the group of "Acceptance & Fear & Disgust & Anticipation" were 60.0% or more, with the exception of Fear. The correctness of Disgust was quite high, at 74.4% when utterances of Disgust were discriminated from others in the group.

6. Estimation of emotional degree

6.1. Estimation method

The 1681 utterances were divided into a closed set and an open set at random. Closed sets comprised 150 utterances from the total number of utterances for each emotion, while the open set comprised the remainder. The estimation experiment for each emotion was conducted using the closed set with multiple linear regression analysis based on least-squares method. Forward selection was applied for the estimation experiment to clarify which features contribute to the estimation of emotional degree.

6.2. Results and discussions

Values of the adjusted R square ($\hat{R}^2$) for each emotion ranged from 0.05 (Disgust) to 0.55 (Surprise) for the closed sets. The emotion with the second-highest $\hat{R}^2$ was Acceptance (0.49), followed by Joy (0.42), and then Anger (0.34). Values of the RMS of the residual for open sets ranged from 0.39 (Acceptance) to 0.59 (Anger). The second smallest RMS of the residual was 0.45 for Disgust, followed by 0.47 for Joy and Sadness. Degree of Acceptance was estimated well because a higher $\hat{R}^2$ was obtained and RMS of the residual was smaller. However the RMS of the residual of Disgust was smaller, degree of Dis-

Table 5: *Significant features by results of ANOVA and correctness by results of discriminant analysis.*

Emotion	Fmini	Fmaxi	Fstdv	Dmora	Drate	Pmaxi	Cstdv	Total	Target	Other
Joy						*		51.9%	59.2%	50.5%
Acceptance								-	-	-
Fear								-	-	-
Surprise	*	*		*				79.1%	71.8%	80.8%
Sadness						*		70.1%	64.3%	70.5%
Disgust								-	-	-
Anger		*					*	63.5%	65.5%	63.4%
Anticipation				*	*			67.0%	71.2%	66.4%
Neutral						*		62.0%	53.0%	64.0%
Sadness-Disgust	*	*						66.6%	70.4%	65.9%
Surprise-Anger			*			*		63.3%	66.7%	62.2%
Acceptance-Fear-Disgust-Anticipation						*		53.4%	47.8%	56.4%

Table 6: *Selected features according to multiple linear regression analysis.*

		Fmini	Fmaxi	Fstdv	Dmora	Drate	Pmaxi	Pmagn	Pstdv	Cmean	Cstdv
Joy	step	2			5	4	1	6		3	
	coeff	0.28			0.12	-0.22	0.59	0.08		-0.19	
Acceptance	step		4				1		5	2	3
	coeff		0.14				0.53		-0.1	-0.34	0.2
Fear	step	1			3		2				
	coeff	0.18			-0.12		0.21				
Surprise	step		3	5	4		1			2	
	coeff		0.34	-0.14	0.13		0.51			-0.26	
Sadness	step				4	2	5		3		1
	coeff				-0.13	-0.14	-0.11		-0.1		-0.24
Disgust	step				3		1	2			4
	coeff				-0.17		0.23	-0.14			0.12
Anger	step			2	3		1	4	5		
	coeff			0.23	0.12		0.5	-0.16	0.11		
Anticipation	step			3		4	1	5	6	2	
	coeff			-0.14		0.19	0.35	-0.27	0.24	-0.29	

gust could not be estimated well. Because its $\hat{R}^2$ was extremely low.

Table 6 shows the selected features in each emotion by the results of multiple regression analysis. "Step" denotes the selected order by forward selection and "coeff" denotes the standard partial regression coefficient for each selected feature. As for the acoustic parameters, Pmaxi was selected at the first or second step of forward selection for models of the 7 emotions. In addition, absolute values of standard partial regression coefficients were high, from 0.11 (Sadness) to 0.59 (Joy). Cmean was selected at the second step for models of Surprise, Anticipation and Acceptance. Dmora was also selected at the third, fourth or fifth step for models of Surprise, Joy, Anger, Fear, Disgust and Sadness. We found that standard partial regression coefficients of some features showed different signs in each emotion. For example, coefficients of Pmaxi showed a positive sign for Surprise or Joy, but a negative sign for Sadness. Coefficients of Dmora also showed positive signs for Surprise, Joy and Anger, whereas coefficients for Fear, Disgust and Sadness were negative. This means that different acoustic characteristics were present for expressed emotional elevation in each emotion.

7. Summary

For the purposes of emotion recognition according to acoustic information, emotion discrimination experiments and estimation of emotional degree experiments were conducted using ANOVA, discriminant analysis and multiple linear regression analysis. The discrimination experiment yielded high correctness values, with 79.12% for Surprise and 70.11% for Sadness, and values over 60% correctness were obtained for every emotion. The estimation experiment showed values of the adjusted R square ($\hat{R}^2$) for each emotion from 0.05 (Disgust) to 0.55 (Surprise) for closed sets, and values of root mean square of the residual for open sets ranged from 0.39 (Acceptance) to 0.59 (Anger). Moreover, we found that different acoustic characteristics were present for expressed emotional elevation in each emotion.

As further research, we will analyze the linguistic information of emotional utterances appearing in natural Japanese dialog to determine whether any emotion-specific words or expressions are present.

8. References

[1] Second Life, http://secondlife.com/

[2] Skype, http://www.skype.com/intl/ja/.

[3] Tapur, http://www.tapur.com/jp/.

[4] Robert Plutchik, "Emotion - A Psycho-evolutionary Synthesis", Harper, Row, 1980.

[5] Yamada, T., Shibata, T., Kuramochi, Y., Yamada, A., "Shin-Meikai Japanese Dictionary 6th Edition", Sanseido, 2005 (in Japanese).

[6] Arimoto, Y., Ohno, S. and Iida H., "Acoustic Features of Anger Utterances During Natural Dialog", Proc. of INTERSPEECH 2007, pp. 2217-2220, 2007.