Classification of Emotional Speech in Anime Films by Using Automatic Temporal Segmentation

Yutaro Hara
Graduate School of Computer and Information Sciences
Hosei University
3-7-2 Kajino-cho, 184-8584 Koganei, Japan
Email: yutaro.hara.h2@gs-cis.hosei.ac.jp

Katunobu Itou
Faculty of Computer and Information Sciences
Hosei University
3-7-2 Kajino-cho, 184-8584 Koganei, Japan
Email: itou@hosei.ac.jp

***Abstract*—This paper describes emotional speech classification in anime films. An emotional speech corpus was constructed by using data collected over 8 h. The corpus consists of emotional speech material of a total of 984 utterances. Five emotions, namely, joy, surprise, anger, sadness, and the neutral case, were labeled and divided into training and test data. In a previous study, Attack and Keep and Decay were adopted as parameters to describe temporal characteristic of the power transition. This paper proposed an improved method of A-K-D unit, and evaluated it. As a result, acoustic features of the proposed method were more effective than the conventional method when we used for GMM.**

***Keywords*—Emotion, animated film, Speech analysis, Pattern classification.**

I. Introduction

A number of “anime” films have been produced in Japan. More than 140 anime films were produced in 2008; these include children’s film as well as other genres such as sci-fi, horror, and sports. They are often produced as a drama series. Anime films involve typical directorial techniques and a voice acting technique called “anime tone” in Japan. In this acting technique, the speaking style and emotional expression involve a characteristic sound and prosody, which have not been thoroughly studied.

In recent years, many studies have been conducted on emotional classification. However, emotional classification is complicated because it depends on clear acoustic features that may not be determined, and hence, depends on the word length and speaker. This study deals with the emotional speech of voice actors/actresses in a cartoon film; as yet, few studies have been conducted on cartoon films. A previous study [1] focused on emotional speech classification of voice actors/actresses in an animated movie that is “The Incredibles” [2]. Automatic classification enables the maintenance of an emotional speech corpus by automatically applying to it the label of emotions of the corpus of some other cartoon film. In addition, a support of the performance exercise of the emotional expression is enabled if it apply to an automatic classification system of the emotional expression such as the cartoon film and will open possibility of some production of the cartoon film work.

Section II introduces emotions and some samples used in this paper. Section III presents an explanation of temporal structure of emotional speech; it also introduces acoustic features and the proposed method of A-K-D unit estimation. Section IV describes 4 classifiers used in this paper. Section V describes Feature Subset Selection (FSS) and Sequential Forward Floating Selection (SFFS). In Section VI, we present a result of the recognition rate in each emotion. In Section VII, we discuss the result and effectiveness of the proposed method. Section VIII describes the general summary and some future issues.

II. Speech Database of Emotions

Emotional speech has been dealt with a spontaneous speech and a conscious speech. The spontaneous speech is difficult to specify the emotion of the moment of a speech, and it is easy to be dependent on a mental condition and environment. Otherwise, the conscious speech such as acting emotional speech can detect the emotion by the power and tone of voice, and it is easy to judge a emotion for some observers subjectively. Furthermore, if the voice acting is mastered like some voice actors/actresses, the emotional expression is easy to accord with a subjective evaluation with only a pitch. We use acting emotional speech of voice actors/actresses because of the superior ability that it is difficult to be dependent on a mental condition and the environment and the ease of a subjective emotional classification by some observers.

In this study, emotional speech is extracted from DVD-Video of “Honey and Clover” a Japanese anime film. This film is based on everyday school life and consists of 24 episodes. This film contains a lot of emotional speech samples because 4 male and 3 female important characters appeared in the film.

Emotional speech samples of 4 male voice actors and 2 female voice actresses were extracted manually with a sampling frequency of 48 kHz and a bit rate of 16 bits from 24 episodes (total film length was approximately 8 h) of this anime drama. The samples were classified into seven categories such as “Joy”, “Surprise”, “Anger”, “Sadness”, “Neutral”, “Others” and “No emotion”. We selected the previous 5 emotions because they’re used most [1][3][4][5]. We discarded speech samples that have any overlap of plural speech and were annotated with different emotional labels by two annotators. A details of the emotional categories of the sample are listed in Table 1.

“Other” in Table 1 shows that the samples which did not include 5 emotions meet some condition, and which did not

 ISBN: 978-1-61208-110-6

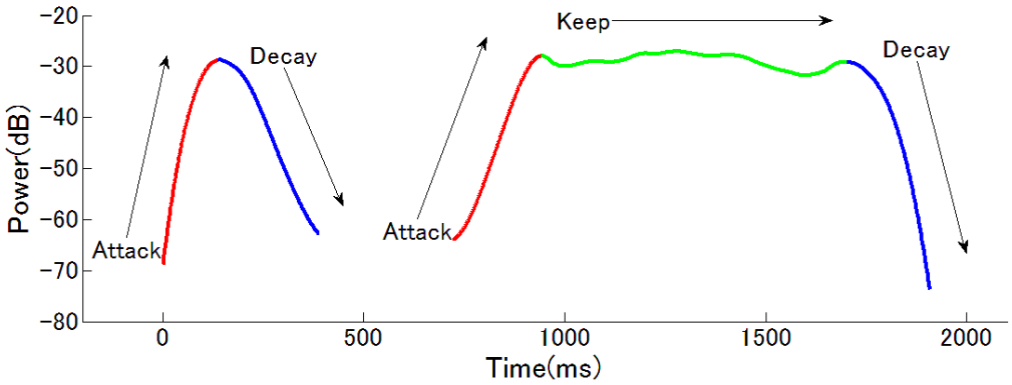

Figure 1. A-K-D unit of emotional speech

TABLE 1. Number of samples of emotional speech

	Joy	Surprise	Anger	Sadness	Neutral	Other
Training	167	75	128	196	195	/
Test	35	42	27	57	62	/
Total	202	117	155	253	257	1671

accord in the subjectivity emotional classification. In addition, it is assumed that 5 emotions do not include plural emotions.

III. Acoustic Feature Analysis of Emotional Speech

A. Temporal structure of emotional speech

Emotional speech has characteristic temporal structures in power transition and pitch transition such as three-layerd models that were modeled by F0 contour,power envelope, spectrum [6], and Support Vector Machine (SVM) and Gaussian Mixture Model (GMM) and K-nearest neighbor method (k-NN) that were modeled by energy mean of fall-time and Energy mean of rise-time and so on [4]. Some temporal characteristics of a power transition and a F0 transition were considered as derivatives (delta features) of whole segments [4][6]. Mitsuyoshi [5] proposed a temporal structure model of utterance based on power transition. In this model, an utterance is divided into three parts; “Attack”, which lasts from the beginnings of the utterance to the peak in power domain, “Keep”, which lasts during keeping the power level, and “Decay”, which begins decreasing the power. Emotional speech, especially, has characteristics in Attack and Decay. In Japanese anime, the beginnings of “Joy” utterance and “Anger” utterance have high pitch, so that the mean or the maximum of F0 of Attack unit is higher than other emotional types. For “Sadness” utterances, duration of Keep tend to be shorter, and their power and pitch do not vary so much. In Decay unit of sadness utterances, therefore, magnitude of derivation of F0 and/or power tend to be small. Mitsuyoshi classified a speech into A-K-D for a unit and it was called “ A-K-D unit ”. This study performed A-K-D unit detection, and modeled acoustic features in each unit (Attack-Keep-Decay). Figure 1 shows examples of A-K-D unit of emotional speech.

In the study [5], in Attack unit and Decay unit, inclination and maximum value and continues length in power were used. In Keep unit, continues length and power average and Δpower average and, power variance were used. With these acoustic features, emotional classification was performed for the spontaneous speeches and the conscious speeches. The structural modeling based on the A-K-D unit is promising, but, there are some problems to be solved. In this paper, we improved the detection algorithm of A-K-D unit. We also examined more acoustic features, such as ΔF0 and minimum value of power and F0 based on the A-K-D unit. Further, when the A-K-D unit is estimated, we do not consider F0 only the power transition.

B. Acoustic features

In emotional speech analysis, F0 and power have been widely used as acoustic features [1][4][5][6][9]. In this study, both F0 and power are used as acoustic features. In addition, 77 dimension acoustic features shown in Table 2 were used in total. F0 represents a pitch of a speech, and it was extracted using STRAIGHT [7] in this study. Power is calculated as ratio of a total of power spectra in 70 msec segment to a silent section (the power level of background noise). MFCC(Mel-Frequency Cepstrum Coefficient) represents a frequency response of a human vocal tract, and have been used in emotional speech recognition and speech recognition [8]. In addition, we normalized all acoustic features by using the average of acoustic features of Neutral data as a standard of each speaker.

C. A-K-D unit estimation

1) Conventional method of A-K-D unit estimation: When the Δp in equation (1) crossed the threshold, Attack begins.

$$\Delta p = p_n - p_{n-1} (n = 1, 2, 3, 4, ...) \quad (1)$$

Attack is defined to last during the segment where $\Delta p > 0$. Keep is defined to last while an absolute value of Δp keeps under the threshold. When it crosses the threshold, Keep finishes and Decay begins. The Decay slope was defined by $\Delta p \leq 0$. Decay finished when $\Delta p \geq 0$ was detected, which means one A-K-D unit ended, and next A-K-D unit estimation begins. The conventional method of estimating A-K-D unit consists of the above-mentioned items.

 ISBN: 978-1-61208-110-6

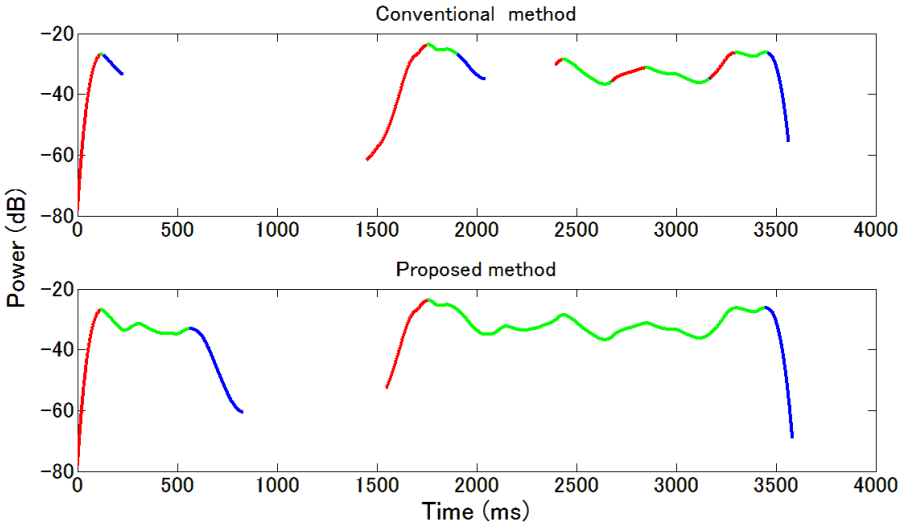

Figure 2. A-K-D unit of conventional method and proposed method (red line is Attack, green line is Keep, blue line is Decay)

TABLE 2. Acoustic features

Extraction segment	Acoustic feature
A-K-D Unit	F0 maximum
	F0 minimum
	F0 range
	F0 mean
	F0 median
	F0 Standard deviation
	F0 inclination
	Power maximum
	Power minimum
	Power range
	Power mean
	Power median
	Power Standard deviation
	Power inclination
	Continuous length
Whole Speech	F0 maximum
	F0 minimum
	F0 range
	F0 mean
	F0 median
	F0 Standard deviation
	ΔF0 mean
	ΔF0 maximum
	ΔF0 minimum
	ΔF0 mean of positive incline
	ΔF0 mean of negative incline
	Power maximum
	Power minimum
	Power range
	Power mean
	Power median
	Power Standard deviation
	ΔPower mean of positive incline
	ΔPower mean of negative incline
	ΔPower median of positive incline
	ΔPower median of negative incline
	12 dimensions MFCC

The conventional method, however, the power transition have a single peak as left plot of Figure1, a unit that should be recognized as Decay may be misrecognized as Keep.

The other way around, the case of the power transition does not have a single peak, When $|\Delta p|$ crossed the threshold, the unit that should be recognized as Keep may be misrecognized as Attack or Decay. In this study, we proposed a method to improve the above-mentioned point.

2) Proposed Method of A-K-D unit estimation: As preprocessing of A-K-D unit estimation, as a first, we performed the Voice Activity Detection (VAD). The V/UV decision of STRAIGHT was used for VAD. If there is an interval between a voiced segment and the next voiced segment within 100msec, it was treated as one voiced segment. After the VAD, we performed A-K-D unit estimation based on the power transition in the voiced segment. In addition, the power was smoothed by using Savitzky-Golay Filter with 3 degree and a window width of 201 msec.

When voiced segment begins, Attack begins. When the equation (2), Attack finishes and Keep begins.

$$|R^n_{Attack}| > 0, R^{n+1}_{Attack} < 0 \tag{2}$$

Here, R is an inclination of the power that is calculated every 10 msec.

Next, we decide a rough shape of the power transition. Here, R_{Attack} is the inclination of the power from Attack began to Attack finished, and R_{KD} is the inclination of the power from Keep began to voiced segment finished.

If $|R_{KD}| < R_{Attack}$

- There is not a single peak like left plot of Figure 1, that is to say, there is not Keep unit.
- Attack began equals Decay begins.
- When the voiced segment finishes, Decay finishes.

Else

- There is a single peak like right plot of Figure 1, that is to say, there is Keep unit.

$$R^n_{Keep} > 0, R^{n+1}_{Keep} < 0 \tag{3}$$

$$\left|R^{n+1}_{Keep}\right| > max(|R_{Keep}|) \tag{4}$$

- When the equation (3) and (4), Keep finishes and Decay begins.

- When the voiced segment finishes, Decay finishes.

When one A-K-D unit ended, next A-K-D unit estimation begins This study decided one A-K-D unit by the above-mentioned algorithm. Figure 2 shows a result of A-K-D unit estimation by conventional method and proposal method. As the threshold in the conventional method, we used the average of Δp of the whole speech.

In Figure 2, the proposal method can estimated A-K-D unit more precisely than the conventional method. The region that should be recognized as Keep was misrecognized to be Attack and Decay, and there was the region where Decay unit is not found in after Keep unit in the conventional method of Figure 2. But, these points were improved by the proposal method.

IV. EMOTIONAL SPEECH MODEL

Emotional speech classification has been performed with various classifiers in some previous studies [3][4][9][10]. In this study, acoustic features of emotional speech of some voice actors/actresses were modeled by K-nearest neighbor method (K-NN), Probabilistic Neural Network (PNN), Support Vector Machine (SVM),Gaussian Mixture Model (GMM). We performed Feature Subset Selection (FSS) in each model, and performed Gaussian Mixture Size Selection (GMSS) in GMM.

A. K-nearest neighbor method (K-NN)

One of the method that is a standard in pattern recognition is neighbor method. When new data are given, nearest neighbor method calculate distance with the other data and classify it in a category same as data in the neighborhood most. On the other hand, K-NN refers to not only the nearest data but also the K unit data of the neighborhood, and it classifies the class where most learning patterns belong to.

B. Probabilistic Neural Network (PNN)

K-NN refers to the K unit data of the neighborhood, but PNN refers to data in the distance to decide a category. PNN has a high recognition precision. Because PNN approximated precisely to the relation of true probability density distribution between each category, by putting a kernel function formed from a sample pattern on top of one another. Therefore, if the number of sample patterns increases, the recognition rate nears an ideal value according to Bayesian statistics, and the classifier with high recognition rate can be realized.

C. Support Vector Machine (SVM)

SVM is one of the classifier with supervised learning. SVM is a method to constitute a classifier of two classes with a linear threshold element, and there are three main characteristics.

1) It can be expect a high generalization ability, because it decide an identification plane by the margin maximization.
2) The learning is resulted in quadratic programming problem by Lagrange multiplier method, and a local optimal solution become a global optimal solution by all means.
3) It performs linear identification on the feature space by defining the feature space that reflected prior knowledge for the space of the identification object. And, It is not necessary to show a conversion to the feature space explicitly by defining the kernel function that expressed dot product on the feature space.

Because of these characteristics, SVM shows high recognition rate for the unlearning data. In addition, SVM shows such characteristics because a optimization is performed for both the recognition error and the generalization in learning. In this study, we implemented SVM by using LIBSVM[11].

D. Gaussian Mixture Model (GMM)

A mixture Gaussian distribution is expressed in probabilistic model called the mixture distribution by piling of some single Gaussian distribution. GMM expresses arbitrary consecutive density functions by coordinating a weight coefficient, and a mean of each distribution, and covariance. There is the case that it is not caught a distribution even if maximum likelihood estimation is used in single Gaussian distribution, but when maximum likelihood estimation is used by linear combination of some Gaussian distribution, GMM can catch the distribution. GMM can express in the next equation:

$$p(x) = \sum_{k=1}^{k} \pi_k N(x \mid \mu_k, \Sigma_k) \tag{5}$$

Here, π_k is the mixture coefficient, and $N(x \mid \mu_k, \Sigma_k)$ is the mixture factor. Each Gaussian distribution have a peculiar mean μ_k and a covariance Σ_k. In addition, the mixture Gaussian distribution is determined by parameters such as the weight coefficient, a mean, and a covariance. These parameters are determined by using maximum likelihood estimation. A function that a likelihood function becomes maximum can demand by using the maximum likelihood estimation.

1) Gaussian Mixture Size Selection (GMSS): The number of Gaussian distributions is an important element in GMM. In a previous study [12], a BIC-based method called Gaussian Mixture Size Selection (GMSS) has been proposed; this method can be used to control the complexity of the speaker model and to determine the number of Gaussian distributions in GMM. In a previous study [9], GMSS was used with the Akaike Information Criterion (AIC) [13]. BIC and AIC are the most commonly used an evaluation standard in an information standard. In this study, we are used for GMSS with AIC.

Let $X = \left\{x_j \in R^d : j = 1, \ldots, N\right\}$ be the training data set, $\lambda = \{\lambda_i : i = 1, \ldots, K\}$ be the candidates for the parameter of models. AIC of GMM is given by the following equation:

$$AIC_i = logP(X \mid \lambda_i) - (2d + 1) \tag{6}$$

Here, $logP(X \mid \lambda_i)$ is the logarithm of the likelihood of training data X by GMM, d is the number of acoustic features.

The mixture size of GMM is determined by evaluating the following:

$$\Delta AIC = AIC_M - AIC_{2M} \tag{7}$$

The mixture size is doubled if ΔAIC is negative. Otherwise, it is represented as M. Thus, the number of Gaussian

 ISBN: 978-1-61208-110-6

distributions can be set according to the training data; when the training data is sparse, the mixture size is expected to be small. In this study, we find the number of the mixture distributions that is most suitable for every class, and make a high performance GMM.

V. Feature Subset Selection (FSS)

The database used for pattern recognition is represented by some examples of acoustic features and generally consists of a high-dimensional vector. Therefore, the calculation cost may increase due to the many features and classes. Furthermore, the best classification cannot be achieved due to noise. However, the recognition rate can be improved by using FSS [14].

By FSS, a candidate with the most effective acoustic feature set for classification is selected from a given the acoustic feature set; then, a subset s consisting of d features containing the information required for classification is identified from a feature set S with D features. Note that since the output value of the evaluation function is high, it is a good feature set.

A. Sequential Forward Floating Selection (SFFS)

There is various ways in Feature Subset Selection. A method that till a certain standard is satisfied by increasing features or reducing features is generally used. A representative method involves the use of a search algorithm called forward type that increases the number of features from 0 to higher values. Sequential forward selection (SFS) proposed by Whitney is a representative of the forward type [15]. Another method involves the use of a search algorithm called backward type that determines the best feature set by reducing the number of features from the total number of features. Sequential backward selection (SBS) proposed by Marill and Green is a representative of the backward type [16]. These algorithms can be easily used in various applications. However, it may not be demanded the best feature combination because they are one-direction search algorithm and cannot carry it out about the all possible feature combination. Therefore, there is SFFS (Sequential Floating Forward Search) suggested as the algorithm that improved SFS and SBS by Pubdil [17]. SFFS is the Floating type algorithm which put Forward type algorithm and Backward type algorithm together. And, SFFS is the higher performance than SFS and SBS. SFFS algorithm is shown in Figure 3.

In many emotional speech classification, SFFS is used as feature subset selection [9][18]. In this study, we use SFFS to demand the high performance feature combination. In addition, SFFS is performed by PNN which is used a error rate as evaluation function.

VI. Result

We examined with open test. The 3 feature combinations were compared by four classifiers which are K-NN and PNN and SVM and GMM. A result is shown in Table 3. In Table 3, a set A is Whole Speech features, a set B is Whole Speech and A-K-D unit (conventional method) features, a set C is Whole Speech and A-K-D unit (proposal method) features.

Initialize:
$Y = \{ y_j \mid j = 1,\cdots,D \}$ *//Input and available measurements//*
$X = \{ x_j \mid j = 1,\cdots, k, x_j \in Y \}$, $k = 0,1,\cdots,D$ *// ouput//*
$X_0 = 0$, $k = 0$
Execute:
Repeat
Step1(Inclusion)
$x^+ = \text{argmax } J(X_k + x)$
$X_{k+1} = X_k + x^+$
$k = k+1$
Step2(Conditional Exclusion)
$x^- = \text{argmax } J(X_{k-1} - x)$
if $J(X_{k-1} - x^-) > J(X_{k-1})$
$X_{k-1} = X_k - x^-$
$k = k - 1$
goto Step2
else
goto Step1
until k is the number of features required

Figure 3. SFFS algorithm

Each set were performed Feature Subset Selection. The feature combinations of each set is shown in Table 5.

TABLE 3. Recognition rate in each classifier

	kNN	PNN	GMM	SVM
set A	63%	64%	64%	60%
set B	64%	64%	65%	62%
set C	64%	64%	67%	62%

TABLE 4. the highest recognition rate in GMM with set C

	Joy	Surprise	Anger	Sadness	Neutral	Total
GMM	60%	45%	44%	86%	77%	67%

In K-NN and SVM and GMM, the set B and the set C which are used A-K-D features are a little higher recognition rate than set A which is not used A-K-D features. In addition, a best model is GMM with set C. The Table 4 shows the recognition rate for each emotion in GMM with set C.

VII. Discussion

The set B and the set C which are used A-K-D features showed the higher recognition rate than set A which is not used A-K-D features. Furthermore, it may be said that the acoustic features in A-K-D unit are effective for emotional speech classification, because the acoustic features of each region of Attack and Keep and Decay were selected by Feature Subset Selection in Table 3. In addition, the highest recognition model involved A-K-D features which are estimated by the proposed method. But, because there is not a difference in recognition rate with the classifiers except GMM, it cannot be said that the proposed method is better unconditionally. However, it may be said that the acoustic features of the proposed method is more effective than the conventional method when we use a classifier which can express a feature distribution in detail like GMM.

In this study, an error rate was used as an evaluation function when Feature Subset Selection was performed. But, using the error rate as the evaluation function may be dependent on the classifier and the number of samples of test data or

 ISBN: 978-1-61208-110-6

TABLE 5. Feature combinations for optimal solutions

set A (Whole Speech)
F0 (max,min,range,mean,median), ΔF0 (mean,maxi,mean of pos.incline), Power (min,mean,median), MFCC (7th)
set B (A-K-D unit (Conventional method) + Whole Speech)
Attack F0 (max,mean,median), Keep Power (median), Keep F0 (mean), Decay Power (min), Decay F0 (median), MFCC (1st,2nd), F0 (max,min,mean,median), ΔF0 (mean), Power (range)
set C (A-K-D unit (Proposal method) + Whole Speech)
Attack Power (max,median), Keep F0 (median), Decay Power (max,min,mean,median), Decay length, F0 (max,mean,median), ΔF0 (mean of neg.incline), Power (max,mean), MFCC (1st)

training data. Because the number of samples was uneven with every emotion, there is a possibility that the best Feature Subset Selection was not possible. Therefore, in future, it is necessary to perform the best Feature Subset Selection by using the evaluation function which is difficult to depend on the classifier and the number of samples or equalizing the number of samples.

In addition, Anger is misrecognized a lot by surprise, and surprise is misrecognized a lot by anger. Sadness is got high recognition rate, but it was easy to be misrecognized neutral, and neutral is easy to be misrecognized by sadness. The pair of those emotions resembled in non-language information, and the subjective emotional classification depended on language information are considered as one of the reasons. In future, it is necessary to solve this problem by performing the subjective emotional classification does not depended on language information or speaker oneself.

VIII. Conclusion

This paper focused on A-K-D unit, and we proposed the improved method of A-K-D unit estimation. As a result, acoustic features of the proposed method were more effective than the conventional method when we used for GMM. However, the acoustic features in this study cannot be used to classify all emotions precisely. Therefore, the acoustic features effective for emotional speech classification have to be examined. In addition, because we did not inspect a precision of automatic detection / estimate of the A-K-D unit, it will be necessary to inspect the precision in future. In a related study [8], the Teager energy operator (TEO) is used as an acoustic feature that does not depend on the word length and speaker. By using this operator, the recognition rate may be improved.

And, if this study is used as an application system, it is necessary for a accorded rate in the subjective emotional classification with a speaker or a third person to be inspected. In addition, this paper is not considered a influence of language information in the subjective emotional classification. Therefore, we think that it is necessary to consider how much influence language information has on human subjective emotional classification, by performing the subjective emotional classification for non-language information.

References

[1] N.Amir and R.Cohen, "Characterizing emotion in the soundtrack of an animated film: Credible or incredible?", Affective Computing and Intelligent Interaction, vol.4738/2007, pp.148-158, 2007.

[2] B.Bird (Director), "The Incredibles [motion picture]", United States: Walt Disney Pictures, 2004.

[3] W.J.Yoon and K.S.Park, "A Study of Emotion Recognition and Its Applications", Modeling Decisions for Artificial Intelligence, vol.4617/2007, pp.455-462, 2007.

[4] 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", ICASSP '04, vol.1, pp.I-577-80, 2004.

[5] S.Mitsuyoshi, F.Ren, Y.Takana, and S.Kuroiwa, "NON-VERBAL VOIVE EMOTION ANALYSIS SYSTEM", IJICIC, vol2, pp.819-830, 2006.

[6] C.F.Huang and M.Akagi, "A three-layered model for expressive speech perception", Speech Communication, vol50(10), pp.810-828, 2008.

[7] K.Hideki, Z.Parham, A.D.Cheveign, and R.D.Patterson, "FIXED POINT ANALYSIS OF FREQUENCY TO INSTANTANEOUS FREQUENCY MAPPING FOR ACCURATE ESTMATION OF F0 AND PERIODICITY ", Proc.EUROSPEECH'99, vol.6, pp.2781-2784, 1999.

[8] N.Katsuya et al."Speech Emotion Recognition with the Teager Energy Operator [in Japanese]", IEICE technical report. Speech 105(572), pp.1-6, 2006.

[9] D.Ververidis and C.Kotropoulos, "Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm", ICME, pp.1500-1503, 2005.

[10] W.Ser, L.Cen, and Z.L.Yu, "A Hybrid PNN-GMM Classification Scheme for Speech Emotion Recognition ", ICPR, pp.1-4, 2008.

[11] C.W. Hsu, C.C.Chang, and C.J.Lin, "A Practical Guide to Support Vector Classification ", 2010.

[12] M.Nishida and T.Kawahara, "Speaker Model Selection Selection Based on the Bayesian Information Criterion Applied to Unsupervised Spervised Speaker Indexing", IEEE TRANSAP, vol.13, no.4, pp.583-592, 2005.

[13] H.Akaike."A new look at the statistical model identification", IEEE Trans. Automatic Control, vol.19, no.6, pp.716-723, 1974.

[14] I.Guyon."Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, vol.46, pp.389-422, 2002.

[15] A.W.Whitney."A Direct Method of Nonparametric Measurement Selection", IEEE Transactions on Computers, vol.20, pp.1100-1103, 1971.

[16] T.MARILL and D.M.GREEN, "On the effectiveness of receptors in recognition systems", Trans. on IT, vol.9, pp.1-17, 1963.

[17] P.Pudil, J.Novovicova, J. Kittler, "Floating search methods in feature selection", Pattern Recognition Letters, vol.15, pp.1119-1125, 1994.

[18] D.Ververids and C.Kotropoulos, "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition", Signal processing, vol188(12), pp.2956-2970, 2008.

 ISBN: 978-1-61208-110-6