A Extraction of Emotion in Human Speech Using Speech Synthesize and Each Classifier for Each Emotion

MASAKI KUREMATSU, JUN HAKURA, HAMIDO FUJITA
Faculty of Software and Information Science
Iwate Prefectual University
Takizawa aza sugo 152-52, Iwate
JAPAN

Abstract: - In order to estimate emotion in speech, most researchers make a classifier based on features in speech, for example power (sound pressure) and the fundamental frequency, gotten from human speech. But validity of estimation using the classifier is not good. In this paper, we discuss about how to enhance the estimation of emotion in speech. We think that making a classifier based on speech features is good for estimation. So we propose the 3 improvement points about the typical method. First point is that we use speech synthesize as training data for making a classifier. Second point is that we use not only mean and maximum but also Standard Deviation (SD), skewness and kurtosis of power and the fundamental frequency gotten from speech. Third point is that we make a classifier for each emotion and estimate emotion in speech by using them. In order to evaluate our approach, we did experiments. Experimental results show the possibility in which our approach is effective for improving typical method.

Key-Words: - Emotion, Speech Synthesize, Regression Tree, Linear Discriminant Analysis, Extraction emotion in speech

1 Introduction

Meaning of words in speech depends on speaker's emotion. People focus on not only words but also intonation, accent, speed and so on, when they talk with others. Because people change these features in speech based on own emotion. In order to enhance a speech dialog system, we should focus on and process user's emotion.

There are researches about estimation of emotions in human speech (see [1]). But research about estimation emotion in speech is still young as opposed to estimate of emotions with facial expression [2][3].

A typical method for estimation of emotion in speech has the following steps [1]. First, researchers collect a lot of human speech. When researchers collect them, they request speakers to express the specified emotion in speech. Speakers try to act like actors or actresses. Next, researchers get features from human speech. They get power and the fundamental frequency using frequency analysis as features in most researches. The length of each human speech is different. So they calculate mean and maximum value of these features. Finally, they make a classifier using a learning algorithm based on statistical value. Researchers use Regression trees, artificial neural network, support vector machine and so on. Emotion estimated by a classifier is one of some emotions specified by researches. Generally speaking, most researches try to estimate emotion in speech based on the method. But validity of estimation is not good. There are some problems in this method. For instance, there is no standard feature using estimation. It is a hard work for most people to express the specified emotion in speech. Most researchers and we should improve the method.

In this paper, we propose the following 3 improvement points about the typical method. First point is that we use speech synthesize as training data for making a classifier. Second point is that we use not only mean and maximum but also Standard Deviation (SD), skewness and kurtosis of power and the fundamental frequency gotten from speech. Third point is that we make a classifier for each emotion and estimate emotion in speech by using them. Next section presents our improvement points in detail. Section 3 describes experiments and results. Section 4 presents discussion about our approach according to the experimental results. And we show future works in Section 5.

2 How to Improve Estimation of Emotion in Speech

In order to improve a typical method for estimation of emotion in speech, we propose the following 3 methods: Using speech synthesize, Adding SD, skewness and kurtosis to features and Making a classifier for each emotion. Table 1 shows the difference between a typical method and our approach. We explain these methods more detail in this section.

Table.1: The Comparison between a Typical Method and Our Approach

Item	Typical method	Our Approach
Data	Human	Synthesized Voice
Features	Power Fundamental frequency	Power Fundamental frequency
Statistics value	Mean Maximum	Mean, Maximum, Standard Deviation, skewness, Kurtosis
Num. of Classifiers	1	Num. of Emotions

2.1 Using Speech Synthesize

We should collect human speech in a typical method. People should express the specified emotion in speech. It is not easy because it is difficult for most people to do like as actors or actress. In addition to, it is a hard work for researchers to collect a lot of human speech. In order to solve these problems, we use speech synthesize. There are some speech synthesize software, for example, Microsoft Speech SDK [4] and Festival [5]. So it is not difficult to synthesize various speeches. Most software, however, can not express emotion and we can not verify whether how to express emotion in other software is true. So we should add emotion labels on synthesized voice for making a classifier. We labeled emotion on synthesized voice based on evaluation by people. How to evaluate synthesized voice is the following. We synthesize some speeches. Each speech has same phrase and different parameters, for example, pitch speed and volume. People hear these speeches and answer which emotion expressed in speech. If more than half of them answer that a speech expresses emotions, we add the emotion labels on the speech. There are some strong points in this method. One is that data doesn't include noise. The other is that evaluating synthesized speech is easier than express the specified emotion in speech. So it is not difficult for researchers to collect a lot of data.

2.2 Adding SD, Skewness and Kurtosis

In typical method, we get power and the fundamental frequency from each speech as features using frequency analysis. And we calculate mean and maximum of these values. We use them to make a classifier. In our approach, we calculate SD, skewness and kurtosis, too. These values show the shape of frequency distribution of power and the frequency. Features showed by these values are different features showed by mean and maximum value. So we guess these values as useful for making a classifier. It is not a hard work to calculate these values. So the cost of process is not high.

2.3 Making a classifier for each emotion

We make only one classifier in typical method. This classifier tries to select one emotion from some emotions. We think that people express some emotions in speech. Features in speech are made of features based on each emotion. So we make a classifier for each emotion. Each classifier shows whether there is one emotion in human speech. We think the set of prediction values gotten by each classifier as emotion in speech. When we want to estimate only one emotion in speech, we estimate emotion whose prediction value is a maximum value. If there is not the prediction value shows an emotion, we don't estimate emotion in speech.

3 Experiments

3.1 Overview

In order to evaluate our approach, we did 2 experiments. In first experiment, we estimate emotion in acting human speech using a typical method and our approach. We evaluate our approach according to compare them. In this experiment, acting human speech is that people try to express the specified emotion in speech. In second experiment, we estimate emotion in natural human speech using a typical method and our approach. Natural human speech is that people speak freely. We similarly evaluate our approach with 1st experiment. We describe experiment in detail as following paragraphs.

Emotion (Emotion Labels) :We use "joyful", "anger", "hate", "fear", "sadness" and "surprise" as emotion in experiments. We selected them according to Ekman's woek [6].

Acting Human Speech : We recorded that a man in his twenties spoke 15 phrases with 6 above-mentioned emotions individually. The number of speech is 90.
Speech Synthesize: To synthesize speech, we use SMARTTALK [7] developed by OKI Software Corporation. We made the synthesized speech of 29 patterns. Each pattern used different parameter values. In order to label emotion on these speech, 4 women in her twenties and 5 men in his twenties evaluated them. More than half of evaluators answered that they felt some emotions in a speech, we picked up the speech as training data for each emotion.
Features : We get power and the fundamental frequency. We use Wavesurfer [8] to get these features in speech. This software is an open source developed by the School of Computer Science and Communication, the Royal Institute Technology in Sweden,
Learning Algorithm : We use making Regression Trees and Linear Discriminant Analysis (LDA) as learning algorithm to make a classifier. We use functions in R language [9] to execute these algorithms. We use Rpart (Recursive Partitioning and Regression Trees) function and Tree function to make a regression tree. Rpart function differs from the Tree function mainly in its handling of surrogate variables. To make a regression tree, we use Deviance in Tree function and Gini index in Rpart function individually.
Natural Human Speech : We recorded utterance of a man in his twenties, when he played a board game with a negotiation phase. In this case, we didn't request him to express emotion in speech intentionally. After recording, he labeled emotion his own utterance by himself. We regard utterances as natural human speech and use them in 2nd experiment.
We show the number of speech we use experiments in table.2.

Table.2 The number of Speech

	Acting Human Speech	Synthesized Speech	Natural Human Speech
Total	90	17	36
Joyful	15	2	14
Anger	15	5	4
Hate	15	2	7
Fear	15	1	3
Sadness	15	6	3
Surprise	15	1	5

Table.3: Patterns for making a classifier

ID	Training	Num. of	Statistical values of
A	Human	1	Mean, Maximum
B		1	Mean, Maximum SD,Skewness,Kurtosis
C		6	Mean, Maximum
D		6	Mean,Maximum, SD,Skewness,Kurtosis
E	Speech	1	Mean, Maximum
F		1	Mean,Maximum SD,Skewness,Kurtosis
G		6	Mean, Maximum
H		6	Mean,Maximum, SD,Skewness,Kurtosis

We estimate emotion in acting human speech using a typical method and our approach. There are 15 speech data for each emotion. We made 5 data sets from them. Each data set includes 12 speeches for each emotion. We use them to make a classifier. These speeches are training data. We use remaining speeches to evaluate a classifier. Remaining speeches are test data. That is, we make a classifier using 72 speeches and evaluate it using 18 speeches. On the other hand, we use synthesized speeches for making a classifier only. We use test data in data set to evaluate the classifier. We made classifiers on 8 patterns in Table.3. Pattern A is a typical method. So we compare the results using pattern A with others.

3.2 The Result of 1st Experiment

Table.4 shows the result of 1st experiment and figure.1 shows the comparison between these patterns. Pattern name is common in table.3, table.4 and figure.1. Rpart and Tree in table.4 and figure .1 mean regression tree algorithm. Validity in table.4 and figure.1 is the average of validity of estimation for each data set. Validity is the number of speech whose emotion matched emotion estimated by a classifier divided by the number of speech in a test data.
Validity of pattern B is higher than the validity of pattern A. Validity of pattern C and D are validity of pattern C and D are almost same pattern C. Validity of pattern E, F, G and H are lower then validity of pattern A.

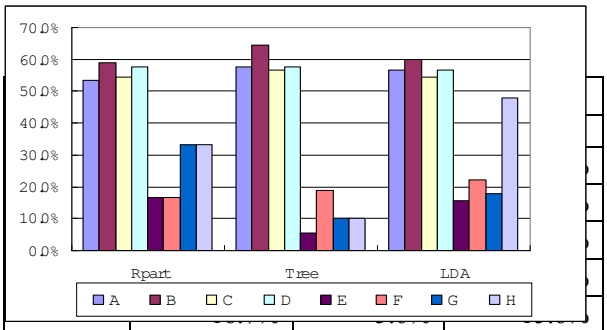

Figure.1: The Comparison of Validity in Acting Human Speech

F	16.7%	18.9%	22.2%
G	35.3%	10.0%	17.8%

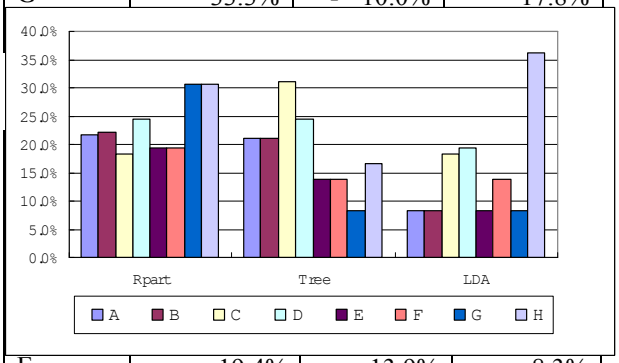

Figure.2 : The Comparison of Validity in Natural Human Speech

E	19.4%	13.9%	8.3%
F	19.4%	13.9%	13.9%
G	30.6%	8.3%	8.3%
H	30.6%	16.7%	36.1%

2.3 The Result of 2nd Experiment

We estimated emotion in natural human speech in 2nd experiment. The difference between 1st experiment and 2nd one is test data used to evaluate a classifier. Table.5 shows the result of 2nd experiment and figure.2 shows the comparison between these patterns. Pattern name is same in 1st experiment.

The result of 2nd experiment differs from the result of 1st experiment. The result depended on learning algorithm. Validity of pattern B is almost same as validity of pattern A. Validity of pattern C and D are almost higher than validity of pattern A. Validity of pattern E, F, G and H are higher than validity of pattern A expect using Tree function.

4 Discussion

4.1 Using speech synthesize

According to 1st experiment, validity of a classifier using synthesized speech (pattern E,F,G,H) are low. We cannot say that this improvement point is good for estimation emotion in speech. We guess that the number of data influences validity. The number of synthesized speech is smaller than the number of acting human speech. There is the polarization in the number of the synthesized speech for each emotion. In fact, the average of validity for "sadness" is about 0.8. On the other hand, the result of 2nd experiment shows that this improvement point is useful for estimation at using Rpart and LDA. We think the reason why this improvement point is better than typical method at using Rpart and LDA is similar to the way which people estimate emotion in speech. Though people are speakers in a typical method, people are listener in this improvement point. Our goal is that we make software to estimate emotion in natural human speech. So it is worth using synthesized speech to make classifiers. The other good point of this improvement is that using speech synthesize is easier than collecting human speech.

4.2 Adding SD, Skewness and kurtosis

Validity of using mean, maximum, SD, skewness and kurtosis (Pattern B,D,F,H) is higher than validity of using mean and maximum (Pattern A,C,E,G) in all case. So using SD, skewness and kurtosis is good for estimate emotion. We guess that these features show the shape of distribution and give good points to discriminate emotion. In order to improve this method, we should consider that we use other features, for example, range, difference and change to estimate emotion.

4.3 Making a classifier for each emotion

Validity of making a classifier for each emotion is higher a little than validity of making a classifier. Especially, the effect is big in the natural human speech. We think the reason as following. Human speech includes more than one emotion. Making a classifier for each emotion tries to estimate each emotion in speech. So this method matches the way which people estimate emotion in human speech. This experimental results show that our hypothesis is true. But validity is not high. We should improve this method to enhance validity.

4 Conclusion

To enhance a typical method for estimation of emotion in speech, we propose the following 3 method: using speech synthesize, adding SD, skewness and kurtosis to exist features, making classifiers for each emotion. The experimental results show the possibility in which

our approach is effective for improvement a typical method. Future works of our research are the following. We collect synthesized voice with emotion more and evaluate our approach. After analyzing the experimental results and classifiers, we reconsider which features we use to estimate emotion in speech. For instance, range, relational frequency distribution and so on. And we should consider some new knowledge, for example psychology-of-music, to enhance our approach.

Acknowledgements
This work is supported by a grant from Research and Regional Cooperation Division, Iwate Prefectural University, with which Hamido Fujita is the principal investigator. We would like to thank Ms. Natsumi SAWAI who is a master student of Iwate Prefectual University, Mr. Hiroshi NAKASATO who is a senior student of Iwate Prefectual University and people who attended our experiment.

[1] Pierre-Yves Oudeyer, The production and recognition of emotions in speech: features and algorithms, *International Journal of Human Computer Interaction*, Vol.59(1-2) ,pages.157-183, 2003
[2] A.Samal and P.Iyengar, Automatic recognition and analysis of human faces and facial expression: A survey, Pattern Recognition, Vol.25(1), Pages.65-77,1992
[3] L.T.Bosch, Emotions: What is Possible in the ASR method?, the Proceedings of the ISCA Workshop on Speech and Emotion, pages.189-194, 2000
[4] Microsoft Speech SDK, http://www.microsoft. com/speech/speech2007/default.mspx
[5] The Festival Speech Synthesis System, http://www.cstr.ed.ac.uk/projects/festival/
[6] P.Ekman and W. V. Friesen , Unmasking the Face, *Malor Books*
[7] SMARTTALK, http://www.oki.com/jp/Cng/ Softnew/JIS/sm.html
[8] Wavesurfer, http://www.speech.kth.se/wavesurfer/
[9] R-Project, http://www.r-project.org/