An Extraction of Emotion in Human Speech Using Speech Synthesize and Classifiers for Each Emotion

MASAKI KUREMATSU, JUN HAKURA, HAMIDO FUJITA
Faculty of Software and Information
Iwate Prefectual University
Takizawaazasugo Takizawa, 152-52, Iwate
JAPAN
{kure,hakura,issam}@soft.iwate-pu.ac.jp http://www.fujita.soft.iwate-pu.ac.jp/en/

Abstract: - The typical method of estimation of emotion in speech has the 3 steps. First, researchers collect a lot of human speech. Next, researchers get speech features from human speech using frequency analysis and calculate the statistical value of them. Finally they make a classifier from the statistical value using a learning algorithm. Most researchers consider the collection of human speech, feature selection and learning algorithm to increase the validity of estimation. But the validity of estimation is not high. In this paper, we propose the 3 new methods to enhance the typical method of estimation of emotion in speech. First method is that we use synthetic speech to make a classifier. Second method is that we use not only mean and maximum but also Standard Deviation (SD), skewness and kurtosis to make a classifier. Third method is that we use the classifier for each emotion. In order to evaluate our approach, we did experiments. Experimental results show the possibility in which our approach is effective for improving the former method.

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

1 Introduction

When people talk with others, people focus on not only words but also intonation, accent, speed and so on. These features show speaker's emotion which affects the meaning of speech. Psychologists show the relationship between emotion and speech features [1].

In order to enhance man machine interface using speech, we should focus on user's emotion. Researchers try to estimate emotions in human speech For example, Murray et al. searched for general qualitative acoustic correlates of emotion in speech[2]. Picard shows the correlation of human speech effect and emotion factor[3]. In the Sociable Machines Project of MIT[4], they show the effect of emotions on human voice. This project tries to develop an expressive anthropomorphic robot called Kismet that engages people in natural and expressive face-to-face interaction. Oudeyer chose to focus on speaker-dependant emotion recognition and his team tries to estimate emotion in speech using the statistical value of features and some data mining techniques.

A typical method of estimation of emotion in speech has the 3 steps [5]. First, researchers collect a lot of human speech. When researchers collect them, they request speakers to express a certain emotion in speech. Speakers try to act like actors or actresses. Next, researchers get features from human speech using frequency analysis. They usually get power and the fundamental frequency as features and calculate the statistical value of them, for example, mean and maximum value. The reason why they calculate the statistical value is that the length of each human speech is different. Finally, they make a classifier from the statistical value using a learning algorithm. Regression trees, artificial neural network, support vector machine, principal component analysis are well-known learning algorithms. Emotion estimated by a classifier is one of emotions defined by researches. In order to the validity of emotion, researches change the collection of human speech, feature selection, learning algorithms and so on. But the validity of estimation is not high. There are some problems in this method. For instance, we hadn't known which features are useful for estimation of emotion in speech, yet. It is a hard work for most examinee to express the certain emotion in speech. In order to improve the validity of estimation and the ability of human computer interaction, we should enhance the method.

As mentioned above, there are some researches for estimation of emotion in speech. But research for estimation of emotion in speech is still young as opposed to estimate of emotions with facial expression [6][7].

In this paper, we propose the following 3 new methods to the typical method. First method is that we use synthetic speech to make a classifier. Second

method is that we use not only mean and maximum but also Standard Deviation (SD), skewness and kurtosis as the statistical value of speech features. In this method, we focus on power and the fundamental frequency in speech as speech features for estimation like the typical method. Third method is that we use the classifier for each emotion.

Next section describes our approach in detail. Section 3 describes experiments and results. Section 4 presents discussion about our approach according to experimental results. And we show future works in Section 5.

2 The Approach to Enhance the Typical Method of Estimation of Emotion in Speech

In order to enhance the typical method of estimation of emotion in speech, we propose the following 3 methods: Using speech synthesize, Adding SD, skewness and kurtosis to the statistical value of features and Using the 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
Training Data	Human Speech	Synthetic Speech
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 need human speech data in a typical method. It is a hard work for researchers. For example, human speech includes noise. Researchers need a lot of time to collect speech. And, it is difficult for most examinee to express certain emotion like as actors or actress. In order to reduce the load of this work, we use speech synthesize. We can use some speech synthesize software like Microsoft Speech SDK [8] and Festival [9]. So it is not difficult to synthesize various speeches. Most software, however, don't have an emotion expression facility. But, in order to make a classifier, we need to know emotion expressed in each speech. So, we put emotion labels on synthetic speeches based on evaluation by people using the following method. We synthesize some speeches from same phrase. But, parameters like pitch, speed and volume are different each other. People estimate emotion in speech and answer it. If an emotion is estimated by more than half of judges, we put the emotion label on the speech. There are some strong points in this method. One is that data doesn't include noise. The other is that evaluating synthetic speech is easier than expressing the certain 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 features. We use them to make a classifier. In our approach, we calculate SD, skewness and kurtosis in addition to mean and maximum value. These values show the shape of frequency distribution of features. The viewpoint that SD, skewness and kurtosis show differs from the viewpoint that mean and maximum value show. So we guess these statistical values as useful for making a classifier. It is not a hard work to calculate these values.

2.3 Using the classifier for each emotion

We use only one classifier for every emotion in typical method. This classifier tries to estimate one of emotions. We think that people express some emotions in speech at once. Speech features mix the feature based on each emotion. The relation between features and emotion is very complex. It is difficult to estimate emotion using one classifier. So we make the classifier for each emotion. Each classifier is specialized one emotion and shows whether there is the emotion in human speech. We regard the set of prediction values gotten by classifiers as emotion in speech. When we want to estimate only one emotion in speech, we estimate emotion which has the maximum value of the prediction value. If we can't select one emotion based on the prediction value, 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 the typical method and our approach. We evaluate our approach based on the comparison of them. In this paper, acting human speech means that people try to express the certain emotion in speech. In second experiment, we estimate emotion in natural human speech using the typical method and our approach. In this paper, natural human speech means that people speak without

requests. We evaluate our approach by same method in 1st experiment. We describe experiment in detail as following paragraphs.

We talk about parameters of these experiments.
Emotion (Emotion Labels) :We use "joyful", "anger", "hate", "fear", "sadness" and "surprise" as emotion in experiments. We selected them according to Ekman's work [10]
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 [11] developed by Oki Electric Industry Co., Ltd. We made the synthetic speech of 29 patterns. Each pattern used different parameter values. In order to put emotion labels on these speeches, 4 women and 5 men in twenties evaluated them. We picked up speech that more than half of them answered emotion as training data for making a classifier.
Features : We get power and the fundamental frequency using Wavesurfer [12]. 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 a learning algorithm to make a classifier. We use functions in R language [13] 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 putted emotion labels on his own utterance by himself. We regard utterances as natural human speech and use them in 2nd experiment.
We show the number of speeches we use experiments in table.2.

We estimate emotion in acting human speech using the 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 synthetic speeches for making a classifier only. We use above-mentioned test data to evaluate the classifier. We have 8 Patterns to make a classifier. Table.3 shows them. Pattern A is the typical method. So we compare the results using pattern A with others. The shading in the table.3 indicates using our approach.

Table.2 The Number of Speech in Our Experiments

	Acting Speech	Synthetic Speech	Natural 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 Data	Num. of Classifier	The Statistical value of Features
A	Human	1	Mean, Maximum
B		1	Mean, Maximum SD,Skewness,Kurtosis
C		6	Mean, Maximum
D		6	Mean,Maximum, SD,Skewness,Kurtosis
E	Synthetic Speech	1	Mean, Maximum
F		1	Mean,Maximum SD,Skewness,Kurtosis
G		6	Mean, Maximum
H		6	Mean,Maximum, SD,Skewness,Kurtosis

3.2 The Result of 1st Experiment

Table.4 shows the result of 1st experiment. Pattern name is common in table.3 and table.4. Rpart and Tree in table.4 are function names in R language for making a regression tree. The validity in table.4 is the average of the validity of estimation for each data set. The validity is the number of speech whose emotion matched emotion estimated by a classifier divided by the number of speech in test data.
The validity of pattern B is higher than the validity of pattern A. The validity of pattern C and D are almost same pattern A. The validity of pattern E, F, G and H are lower than the validity of pattern A.

Table.4: The Validity of Estimation in 1st Experiment

Pattern (ID)	Learning Algorithm		
	Rpart	Tree	LDA
A	53.3%	57.8%	56.7%
B	58.9%	64.4%	60.0%
C	54.4%	56.7%	54.4%
D	57.8%	57.8%	56.7%
E	16.7%	5.6%	15.6%
F	16.7%	18.9%	22.2%
G	33.3%	10.0%	17.8%
H	33.3%	10.0%	47.8%

Table.5: The Validity of Estimation in 2nd Experiment

Pattern (ID)	Learning Algorithm		
	Rpart	Tree	LDA
A	30.59%	28.88%	10.54%
B	32.49%	28.72%	7.19%
C	27.22%	33.36%	17.06%
D	24.21%	34.94%	15.33%
E	16.67%	23.61%	10.32%
F	16.67%	11.90%	19.05%
G	33.33%	22.82%	10.32%
H	33.33%	10.71%	41.27%

3.3 The Result of 2nd Experiment

We estimated emotion in natural human speech in 2nd experiment. The difference between 1st and 2nd experiments is test data for evaluation a classifier. Table.5 shows the result of 2nd experiment. Pattern name is same in 1st experiment.

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

4 Discussion

4.1 Using speech synthesize

According to 1st experiment, the validity of a classifier using synthetic speech (pattern E,F,G,H) are low. We cannot say that this method is useful for estimation of emotion in speech. We guess that the number of data influences the validity. The number of synthetic speech is smaller than the number of acting human speech in 1st experiment.

On the other hand, the result of 2nd experiment shows that this method is useful for estimation of emotion in speech at using Rpart and LDA. We think that the reason why this method is better than typical method at using Rpart and LDA is the following. This method is similar to the way which people estimate emotion in speech. Although a classifier tries to estimate emotion in speech form the viewpoint of the speaker in the typical method, a classifier tries to do form the viewpoint of the listener in this method. Our goal is that we develop a system to estimate emotion in natural human speech. So it is worth using synthesized speech to make a classifier. The other good point of this improvement is that using speech synthesize is easier than collecting human speech. To verify the usefulness of synthetic speech in estimation of emotion in speech, we need a lot of synthetic speech.

4.2 Adding SD, Skewness and kurtosis

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

4.3 Using the classifier for each emotion

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

4.4 Analysis Experimental Results

Table.6 shows the validity of estimation for each emotion in 1st and 2nd experiment. There is a difference between the validity of estimation of every emotion. We guess that the validity depends on the number of training data for a classifier. So we calculated a correlation coefficient between the validity and the number of data. But this value say the correlation between them is weak expects using Speech Synthesize and Tree function.

The shading in the table.6 indicates Maximum value of the validity of estimation of emotion. These values are more than 60% expects Joyful and Surprise in 2nd

experiment. The result shows that we can improve the validity of estimation by using combination speech features and learning algorithms for each emotion. We think that experimental results show our approach is useful for estimation of emotion in human speech. But the number of data in experiments is small. We should collect data more and consider combination features in speech and learning algorithms for each emotion.

4.5 Next Step for Estimate Emotion

These experimental results show that synthetic speech evaluated by people is useful for making a classifier. The reason is that synthetic speech is constructed from the viewpoint of the listener. In exists method, training data is constructed from the viewpoint of the speaker and a classifier is evaluate from the viewpoint of the listener. There is a difference between the listener and the speaker. So the validity of estimation is not high. On the other hand, in case we use synthetic speech, training data is constructed from the viewpoint of the listener and a classifier is evaluated from the viewpoint of the listener. So, synthetic speech is useful for estimation of emotion in speech.

Our goal is estimation of emotion in speech in real time. But, our proposed method is insufficient to this goal. People try to estimation emotion by paying attention speech features in short and long time. In other hand, people focus on features in syllable and utterance. Our proposed method focuses on only features in utterance. We should develop the new method focuses on features in syllable and unite it with the proposal method. In order to do, we should apply the method for estimation of emotions with facial expression to estimate emotion in speech.

5 Conclusion

To enhance the typical method of estimation of emotion in speech, we propose the following 3 methods: Using speech synthesize, Adding SD, skewness and kurtosis as statistical value of features to make a classifier, Using the classifier for each emotion. Experimental results show the possibility in which our approach is effective for enhancement the typical method. Future works of our research are the following. We collect synthetic speech and put emotion labels on them. We reconsider how to estimate emotion in speech based on the results of experiments. For example, which features do we focus on? Which combination features and learning algorithms do we use? We have to reconsider evaluation of our approach and do it, too. And we have to consider emotion that we try to process, again. People express and estimate more than one emotion in human speech. So we should think processing multi emotions in speech to develop better human computer interaction.

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 experiments.

Reference

[1] Thomas Goldbeck, Frank Talkmitt and Klaus R. Scherer, EXPERIMENTAL STUDIES ON VOCAL COMMUNICATION, *FACCTES OF EMOTION RECENT RESEARCH*,pages.119-138, 1988

[2] I.R.Murray and J.L Arnott, Toward a simulation of emotion in synthetic speech : a review of the literature on human vocal emotion, *Journal of Acoustical Society of America*, Vol.93(2),pages 1097-1108

[3] R.W.Picard, *Affective Computing*, The MIT Press,1997

[4] C.Breazeal, B.Scassellati, Infant-like Social Interactions Between a Robot and a Human Caretaker, *Adaptive Behavior*, vol.8, Pages.49-74,2000, http://www.ai.mit.edu/projects/sociable/expressive-speech.html

[5] 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

[6] 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

[7] L.T.Bosch, Emotions: What is Possible in the ASR Framework, *the Proceedings of the ISCA Workshop on Speech and Emotion*, pages.189-194, 2000

[8] Microsoft Speech SDK, http://www.microsoft.com/speech/speech2007/default.mspx

[9] The Festival Speech Synthesis System, http://www.cstr.ed.ac.uk/projects/festival/

[10] P.Ekman and W. V. Friesen , *Unmasking the Face*, Malor Books, 2003

[11] SMARTTALK, http://www.oki.com/jp/Cng/Softnew/JIS/sm.html

[12] Wavesurfer, http://www.speech.kth.se/wavesurfer/

[13] R-Project, http://www.r-project.org/

Table.6: The Validity of Estimation for Each Emotion

		1st Experiment								2nd Experiment							
*	**	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
Training Data		Human				Synthetic Speech				Human				Synthetic Speech			
# of Classifiers		1	1	6	6	1	1	6	6	1	1	6	6	1	1	6	6
Mean		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Maximum		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
SD		N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Skewness		N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Kurtosis		N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
jo	RT	20%	27%	20%	**60%**	0%	0%	0%	0%	11%	7%	6%	**34%**	0%	0%	0%	0%
	TR	47%	**60%**	40%	47%	0%	0%	0%	0%	13%	16%	10%	29%	0%	0%	0%	0%
	LDA	33%	33%	47%	53%	0%	53%	0%	40%	3%	4%	19%	24%	0%	14%	0%	14%
an	RT	40%	40%	40%	53%	0%	0%	**100%**	**100%**	10%	20%	10%	10%	0%	0%	**100%**	**100%**
	TR	47%	40%	53%	40%	20%	0%	47%	27%	5%	5%	5%	10%	75%	0%	75%	50%
	LDA	47%	73%	47%	47%	0%	0%	0%	7%	0%	0%	0%	5%	0%	0%	0%	0%
ha	RT	47%	60%	73%	47%	**100%**	**100%**	**100%**	**100%**	11%	17%	14%	14%	**100%**	**100%**	**100%**	**100%**
	TR	47%	73%	60%	53%	0%	73%	0%	0%	11%	14%	17%	26%	0%	71%	29%	14%
	LDA	47%	67%	47%	60%	13%	7%	27%	67%	6%	23%	17%	20%	29%	0%	29%	**100%**
fe	RT	**100%**	**100%**	93%	67%	0%	0%	0%	0%	**100%**	**100%**	87%	73%	0%	0%	0%	0%
	TR	**100%**	**100%**	73%	87%	0%	0%	0%	0%	**100%**	**100%**	87%	80%	0%	0%	0%	0%
	LDA	87%	80%	80%	80%	0%	47%	0%	7%	0%	0%	7%	7%	0%	**100%**	0%	**100%**
sa	RT	33%	53%	13%	40%	0%	0%	0%	0%	27%	27%	27%	13%	0%	0%	0%	0%
	TR	20%	40%	40%	33%	13%	33%	13%	33%	20%	33%	33%	33%	**67%**	0%	33%	0%
	LDA	53%	40%	33%	33%	80%	27%	80%	**100%**	27%	0%	20%	0%	33%	0%	33%	33%
su	RT	80%	73%	**87%**	80%	0%	0%	0%	0%	24%	24%	20%	0%	0%	0%	0%	0%
	TR	**87%**	73%	73%	87%	0%	7%	0%	0%	24%	4%	**48%**	32%	0%	0%	0%	0%
	LDA	73%	67%	73%	67%	0%	0%	0%	67%	28%	16%	40%	36%	0%	0%	0%	0%

* jo=Joyful, an=anger, ha=hate, fe=fear, sa=sadness, su=surprise ** RT=Rpart, TR=Tree LDA = Linear Discriminant Analysis