The Framework of the Speech Communication System with Emotion Processing

MASAKI KUREMATSU, JUN HAKURA, HAMIDO FUJITA
Faculty of Software and Information Science
Iwate Prefectual University
Takizawa aza sugo 152-52, Iwate
JAPAN
http://www.fujita.soft.iwate-pu.ac.jp/en/

Abstract: - In this paper, We proposed the speech communication system with emotion processing. Our system speaks a given document emotionally based on emotion extracted from speech and text. The difference point between our system and a speech dialogue system is to process emotion. When this system speaks given documents, it uses rules made from speech techniques and the analysis result of professional acting. In order to select these rules, the system extracts emotion from a document and speech. To extract emotion from a document, it uses a word emotion dictionary which shows similarity between a word and a basic emotion word in a concept hierarchy. To extract emotion from speech, it uses a speech emotion database based on the analysis result of speech. This system connects speech synthesize technique with emotion estimation technique. Now we are developing this system. After developing them, we should evaluate the effectiveness of these modules by experiments.

Key-Words: - Emotion, Speech Synthesize, Phrasing, Prominence, Extraction emotion from speech

1 Introduction

There are some speech dialogue systems like Galatea [1]. Most speech dialogue systems don't process user's emotion. Especially, speech recognition systems try to recognize only words in speech. But when we speak with other people, we try to guess their emotion by facial expression, gesture, speech and contents. We try to understand their thinking and react based on emotion we guessed.

So it is important for a computer system to process user's emotion, too. In order to process emotion, we consider facial expression, gesture and speech to express emotion. In this paper, we focus on speech, because we can express own emotion by speech consciously or unconsciously. For example, when we talk with our friend on the telephone, we have to express own emotion by speech and guess my friend's emotion in speech. Maybe, emotion expressed by facial expression is clearer than expressed by speech. So we think about facial expression, too. We describe about facial expression in other paper [2].

There are two big research themes about processing emotion in speech. One is how to speak more emotionally. Some researchers are studying about emotional speech. Most researches try to change speech features at speaking [3]. For example, Prevost et al [4] used contextual and syntactical information. The other theme is how to estimate emotion in speech. Most researchers in this field them paid attention to speech features, for example, fundamental frequency, speech rate, pitch and gain [3,5]. But we don't get the best speech future set to estimate emotion. Additionally, there are a few works to connect emotional speech with estimating emotion in speech. Speech and Estimation have strong relationship. So it is worth studying how to connect these techniques.

In this paper, we propose a framework of a speech communication system with emotion processing. Our system has two modules, "Speech Emotion Estimation Module" and "Emotional Speech Synthesize Module". First module tries to estimate emotion in speech based on speech features. In this paper, Speech Emotion means emotion in speech. Second module speaks texts given by users emotionally. In order to speak more emotionally, this module tries to change speed, tone, volume and pause based on emotion estimated form texts and speech and some speech techniques. This module selects speech techniques based on estimated emotion from texts and speech. Like this, our system connects the speech synthesize technique with speech emotion estimation technique. Figure 1 shows overview of our system.

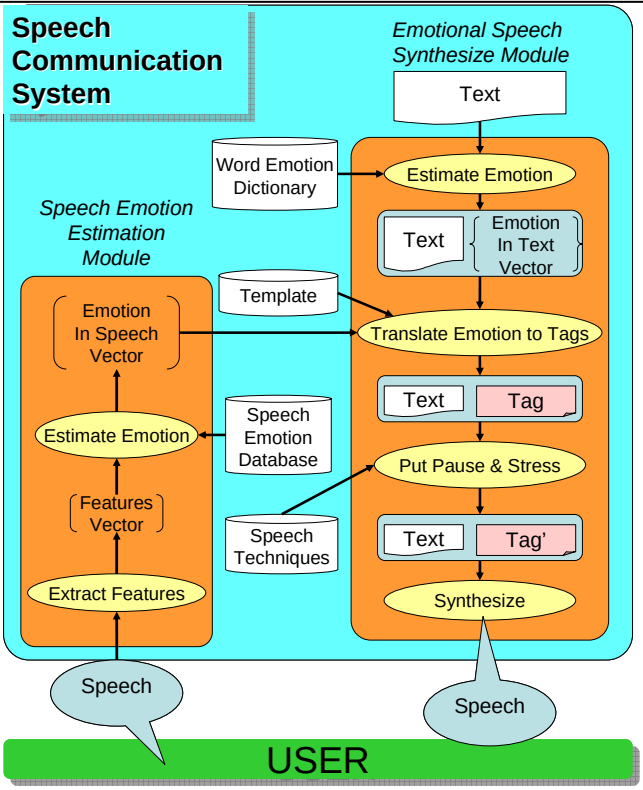

Figure1:Overview of the Speech Communication System

2 Speech Communication System

2.1 System Overview

Our system speaks texts given by users with emotion processing. In the system, emotion processing means that it estimate emotion in user's speech and texts and decides speech features, for example strength, pitch, using these emotions. So as to do, speech emotion estimation module tries to estimate emotion from speech and speech synthesize module estimates emotion in texts and decided how to speak texts.

In this paper, we define emotion based on Ekman's work[5]. He said that there are 6 basic emotions in common with people. They are Anger, Disgust, Fear, Joy, Sadness and Surprise. So we define emotion based on 6 basic emotions. Emotion in our system consists of 6 values which show the strength of each emotion.

2.2 Extract Speech Emotion Module

Some researchers have been trying to estimate emotion from speech. Most of them paid attention to speech features, for example fundamental frequency, speech rate, pitch and gain. These features depend on experiments, so we can not decide the best speech features set to extract emotion. These features are focused to recognize the content of speech. Maybe they are not fit to estimate emotion in speech. So we should add other viewpoints to estimate emotion. Then, we focus on pause in speech and change of sound wave pattern. Most researchers regard pause as a delimitation of speech and a chance for recognition process. However, we assume that people changes position and length of pause based on his emotion. So we focus on pause to estimate emotion in speech. Moreover, we guess that the change of speech pattern appears based on changing emotion. Based on this idea, we try to estimate by the appearance of change of speech by comparing wave patterns. They are different points between our research and other research. Maybe these patterns will be new viewpoints to estimate emotion from speech.

Based on above-mentioned idea, speech emotion estimation module tries to estimate emotion from speech as follow.

First, this module extracts speech features in every fixed time. Features the module extracts are Pitch, Power, Speed and Pause. The module extracts pitch, power and speed using FTT and gets maximum value, minimum value and average of these features. It extracts change of these features too. Now we show change of features as the number of up and down. Additionally, it extracts the number of pause and the rate of pause. The module makes a features vector using these values.

Next, the module matches the features vector to each record in a speech emotion database. A record in a speech emotion database consists of a features vector and an emotion vector. The structure of a features vector is same as features vector extracted from speech. An emotion vector consists of 6 values which show the strength of 6 basic emotions respectively. Each record shows that a features vector shows an emotion vector. We make this speech emotion database based on the analysis result of speech. We describe how to make it in section 2.4. The module calculates similarity between a features vector extracted from speech and in each recode of the speech emotion database using a cosine of an angle between them. It multiplies each value of an emotion vector in each record by the similarity.

Finally, the module returns the average of each emotion vectors and classifies into "very strong", "strong" and "weak" based on the value. We regard this emotion vector as emotion in speech in fixed time.

This module gives the emotion vector to speech synthesize module.

2.3 Emotional Speech Synthesize Module

Our system speaks some writer's works. So it tries to emulate a reciter. In order to speak more emotionally, we think viewpoints for speaking. There are following seven viewpoints, Accent, Articulation, Intonation, Phrasing, Prominence, Pause and Rhythm. Now, speech synthesis technique processes Accent and Articulation. So we add Phrasing and Prominence to our system. Phrasing means dividing a sentence to some phrases based on syntax and phonology. We put pauses in a document. These pauses are not written in the document. This system divides a sentence to some phrases and speaks each phrase with pauses. Prominence means putting stress on some phrases at speaking. Speaking with stress helps us to understand punch lines or keywords in his content. Pause improves the effect of stress. We put pauses and stress using some techniques for Phrasing and Prominence based on syntax and phonology. But there are some contradictions between them. In order to manage them, we focus on professional acting for speech. Professional acting means the action of reciter, actor and actress. We feel professional acting are more emotionally. The reason is that professional people change speech features based on documents. We think that professional people have management rules for changing features. We extract and use their rules to change speech features automatically.

Based on this idea, this module speaks through following 4 steps. After putting some tags on each sentence, it speaks a text. Tags show how to change volume, speed, tone and pitch at speaking or put pauses. The module uses emotion estimated from a text and speech and some speech techniques. Right side of figure 1 shows the overview of the speech synthesize.

Step1: Estimate emotion from a sentence

This module estimates emotion from a sentence. First, it translates a sentence to a morpheme line using a morphological analysis. Next, it gets the emotion vector for each morpheme using a word emotion dictionary. A word emotion dictionary shows the emotion vector for each word. Emotion vector for each word consist of 6 values which show the strength of 6 basic emotions. Each value means a similarity between the word and each basic emotion word. Basic emotion words are Anger, Disgust, Fear, Joy, Sadness and Surprise. They express basic emotion in Ekman's work. These values are gotten from a general concept hierarchy. Their values mean the distance between a word and each basic emotion word gotten by following equation.

$$dis(X,Y) = \frac{Dw}{Up + Dw}$$

Where *X* is a word and *Y* is a basic emotion word, *Up* is a number of concepts between *X* and *Z*, *Z* is a upper concept in common with *X* and *Y*, *Dw* is a number of concepts between *Y* and *Z*.

If there is not a morpheme in a word emotion dictionary, the value is 0.

Next, this module sums up the emotion vector for morphemes in a sentence. Additionally, the module gets other emotion vector whose value shows the quotients the sum of each emotion value divided by the number of words in both a word dictionary and a sentence. And the module classifies these values into "very strong", "strong" and "weak" based on these values. The module regards them as emotion in a sentence.

Step2: Translate emotion to tags using templates

The module puts some tags on a sentence based on the emotion vector calculated in Step.1 and given by the other module. To decide tags, it uses templates. Templates show that how to change values of speech features to express emotion. We make templates based on the experimental result about emotion in speech, for example, [7][8], and speech techniques. The module gets values of speech features from each templates using following equation. The equation means that the larger the difference between emotion in speech and text, the more it turns up the value.

$$V = Default + f(Te, Ue)$$

Where *V* is value of this speech features, *Default* is default value defined in templates, *Te* is emotion value estimated from Text, *Ue* is user's emotion value estimated from user speech. If it is very strong, then the value is 3. If it is strong, then the value is 2. If it is weak, then the value is 1. And *f(X,Y)* is a function that if X is greater than Y, it returns X minus Y and if X is less than or equal Y, it returns 0.

Finally, the module puts on tags which show speech features and the value of each feature.

Step3: Phrasing and Prominence
The module adds extra tags as follow.
First, the module puts pause and stress using speech techniques based on syntax and phonology. If the number of moras in a sentence is more than threshold, this module tries to divide it to some phrases. We defined threshold based on the number of moras in Tanka. Tanka is a short poem in Japanese and has good rhythm for hearing. It usually has 31 moras. We decided the threshold is 30.But this module does not divide conversation sentences. We think conversation sentences have been already divided into best phrases. This module divides a sentence to some phrases using rules based on syntax. Before dividing a sentence to some phrases, this system gets some blocks by parsing a sentence using dependency grammar. We made 5 simple rules for dividing a sentence based on Japanese grammar as following.

1. put pause after a block which last morpheme is the case particle
2. put pause after and before onomatopoeia
3. put pause after and before conversation sentences
4. put pause after a block which has more than 4 blocks from a block depend
5. put pause before stress

A rule consists of a condition part and a conclusion part. A condition part is defined by parsing results. A conclusion part of rules says where this module puts pause. The module tries to match the condition part of rules to the parsing result. If the condition part matches the parsing result, it puts pause on the point defined in a conclusion part. The module speaks documents in Japanese. So we make rules based on Japanese grammar. If the system speaks documents written by other language, we should make rules based on that language's grammar.

In order to put stress on a phrase, the module selects a phrase using rules based on syntax. It picks up onomatopoeia to put stress. Longman Dictionary says that Onomatopoeia is "the use of words that sound like the thing that they are describing". We think that onomatopoeia is an important point for hearing. So the module picks up onomatopoeia using a dictionary and puts stress on them. It picks up rheme phrase and theme phrase to put stress, too. Rheme and Theme are important phrases to understand what someone says. We put stress on these phrases at speaking. So we put stress on rheme and theme phrase. The module tries to find rheme and theme using rules based on syntax. The structure of these rules is same as the structure of rules for phrasing. But a conclusion part says that puts stress on a phrase defined in a condition part. We made 2 simple rules as following.

1. put stress on a phrase whose last morpheme is the case particle
2. put stress on verb, adverb, onomatopoeia and unknown words

Next, the module cuts extra pause and stress. Previous experiments [9] showed us that speaking with too much pause and stress are not good for audiences. So this module cuts extra pause and stress based on emotion estimated from a text and speech by using rules from professional acting. We think that professional people know a lot of speech techniques and select them to change speech features depend on a text and audience reaction. In order to cut extra stress and pause, we try to let the module to emulate professional acting. So we acquire some rules form a professional people by the method described in section 2.4. To use these rules, this module decides to keep or cut pause and stress used in speech like an expert.

Next, this module decides how to reproduce stress. There are some methods for expressing stress. When we want to put stress on speech, we turn up or down the volume of speaking, increase or decrease the speed of speaking and change the tone. We have to decide which speech feature and how much value we change. There is no explicit rule. People decide them by oneself depend on situation. So we decide them using above-mentioned rules acquired from professional people. These rules include what they change speech features for audience reaction. To use these rules, this module decides which speech feature and how much value it changes. For example, if a professional person turns up volume of his speech, when audience is board, this module turns up the volume, too. It is necessary for using these rules to estimate audience's emotion. So the speech emotion estimation module should try to estimate emotion from speech.

Step4: Speech Synthesize
This module translates a paragraph putted on pause and stress in Step.3 to Wave files by speech synthesize program and play it.

2.4 Making a Speech Emotion Database and Professional acting rules.

Our system needs a speech emotion database and professional acting rules. We make them from the

analysis result of data which recorded professional acting.

In order to make professional acting rules, we record a professional acting. This system speaks documents written by Kenji MIYAZAWA who is a famous Japanese writer. So we recorded that professional people read out a Kenji's work by a digital camera. In this paper, the professional people mean researchers about Kenji MIYAZAWA. After recording, a professional person puts emotion labels on own professional acting by himself. Emotion labels are same as basic emotion words used in our system. Next, we interviewed him why he changed speech features, which emotion he estimated from a text and what he do for audience reaction. To make rules, we extract points which make him to change his speech and feel emotion. These points are base of professional acting rules. To support rules, we record that people, who is not a professional, read out same work. We compare them with professional acting. The difference between them is a big point about professional acting. We give priority to the rule shows this point. We have already recorded the data of 12 people. There are 8 people in 20 ages and 4 people in 50 ages. We are analyzing the data and should record more.

In order to make a speech emotion database, we are recording another data. We want to get people's natural speech. So we are recording speech when people play a multi-players game whose name is CATAN [10]. CATAN is a game that players have to negotiate to other players to get some cards. So we thought this game fits to our request. Players mark their own speech by basic emotion labels after game. And we record that players speak some words or phrase with emotion consciously. These speech are not natural. But we assume that the pattern of these speech have similar features as natural voice. In order to pick up common features and patterns to speech, we compare two speech marked by same emotion label. We make speech pattern and emotion sets based on the result and keep them in a speech emotion database. We had recorded twice. 4 people played CATAN. The recorded total time is about 2 hours. We are doing experiments and analyzing player's speech.

3 Implementation and Evaluation

3.1 Implementation.

We are developing this system based on above-mentioned idea now. We use Juman[11] as a morphological analyzer, Cabocha [12] as a parser with a dependency grammar and SMARTTALK [13] as a speech synthesizer used in this prototype. Juman is a morphological analyzer for Japanese developed by Kyoto University. Cabocha is a Japanese dependency structure analyzer developed by Graduate School of Information Science, Nara Institute of Science and Technology. SMARTTALK is developed by OKI Electric Industry Corporation. It can specify the volume, the speed, the tone and the intonation of the sound by putting commands, like markup language, into a text to speak. The speech synthesize module translates stress to these commands based on rules and puts them on documents. When SMATTALK speaks a text with changing the volume, the tone, the speed and the intonation according to command, automatically.

3.2 Experiments about Extract Emotion from Text using a Word Emotion Dictionary.

This system is under construction. But we made some sub modules. We reported about it. The system estimates emotion in a text. To do it, we made a word emotion dictionary from EDR dictionary [14]. EDR is a one of well-known machine readable dictionary has a concept hierarchy. We compared the emotion estimated by the method described in section 2.3 and professional people's answer. Table 1 shows the result. In this table, Sum means the total of emotion values for words in a sentence and Quotients means sum divided by the number of words in both a sentence and a word emotion dictionary. Threshold means the value to classify into strong or weak. The matching rate means that the number of certain kind of sentences divided by the total number of sentences in a text. Certain kind of sentence means that it has one emotion estimated by an expert and the module. Emotion estimated by the module means that emotion value estimated by the module is not weak.

Table.1 The matching rate experts and our module

System	Threshold	Expert.A	Expert.B
Sum	3	0.30	0.15
	5	0.09	0.07
Quotients	0.1	0.49	0.22
	0.3	0.00	0.00

The result is Not good. The reason why is as follow. One is using a concept hierarchy is not good for our target. 6 basic emotions are sort of emotion. So the distance of these words is short. We should reconsider how to make a word emotion dictionary. Now we think using co-occurrence between words and 6

emotion words in some documents and/or the thesaurus.

3.3 Future Work

It is difficult to evaluate this system by theoretical. So we will evaluate it experimentally. The experiment has three steps. First, we compare the speech made by our module and exists speech synthesize module. These modules speak documents at random. If most of parts participants say more emotionally will be spoken by our module, we think our system can speak emotionally. Next, this system tries to estimate emotion in speech. After estimating, we compare output of this module and participant's true emotion. If the matching rate is good, we evaluate our system is good to estimate emotion from speech. Finally, our system tries to speak with changing speech features based on emotion extracted from speech. After speaking, we interview audiences about communication with this system.

4 Conclusion

We proposed the speech communication system with emotion processing. Our system speaks a given document emotionally based on emotion extracted from speech and text. The difference point between our system and a speech dialogue system is to process emotion. When this system speaks given documents, it uses rules made from speech techniques and the analysis result of professional acting. In order to select these rules, the system extracts emotion from a document and speech. To extract emotion from a document, it uses a word emotion dictionary which shows similarity between a word and a basic emotion word in a concept hierarchy. To extract emotion from speech, it uses a speech emotion database based on the analysis result of speech. This system connects speech synthesize technique with emotion estimation technique. Now we are developing this system. After developing them, we should evaluate the effectiveness of these modules by experiments.

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 Prof. Dr. Tamio SASAKI at Department of Social welfare, Iwate Prefectual University and Mrs. Natsumi SAWAI, Mr. Kouki NARITA and Mr. Seiji SASAKI who are senior students of Iwate Prefectual University and participants.

References:

[1] Helmut Prendinger, Mitsuru Ishizuka(Eds.) , *Life-Like Characters*, Springer

[2] Jun hakura, Mamoru Kashiwakura, Yuuichi Hiyama, Masaki Kurematsu and Hamido Fujita, *Facial Expression Recognition and Synthesis toward Construction of Quasi-Personality*, submitted this conference

[3] P.Y. Oudeyer, The production and recognition of emotions in speech: features and algorithms, *Human-Computer Studies 59*, pages 157-183, 2003

[4] Scott Prevost and Maek Steedman, Generating contextually appropriate Intonation, *In proceedings of the 6th Conference of the European Chapter of the Association Computational Linguistics*, pages 332-340, Uterchit, 1993

[5] R.W.Picard, *Affective Computing* , MIT Press.

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

[7]Emotional Speech Homepage, http:// wwwbox. uni-mb.si/ eSpeech/

[8]Expressive Speech, http://www.ai.mit.edu/projects /sociable/expressive-speech.html

[9] Hamido Fujita, Jun Hakura and Masaki Kurematsu, Virtual cognitive model for Miyazawa Kenji based on speech and facial images recognition, *WSEAS Transactions on Circuits and Systems*, *Issues 10, Vol.5*, 2006

[10] CATAN, http://www.kosmos.de/index.htm

[11]S.Kurohashi and M.Nagao, Japanese Morphological Analysis Systesm JUMAN version3.6, *Department of Information, Kyoto University*

[12] T.Kudo and Y.Matsumoto, Fast Methods for Kernel- Based Text Analysis, *ACL 2003* ,2003

[13] SMARTTALK, http://www.oki.com/jp/Cng/ Softnew/JIS/sm.html (in Japanese)

[14]EDR, http://www.iijnet.or.jp/edr/index.html