Emotion recognition method based on normalization of prosodic features

Motoyuki Suzuki* Shohei Nakagawa† Kenji Kita†
* Faculty of Information Science and Technology, Osaka Institute of Technology, Japan
E-mail: moto@m.ieice.org
† Institute of Technology and Science, The University of Tokushima, Japan
E-mail: nakagawa-shohei@iss.tokushima-u.ac.jp, kita@is.tokushima-u.ac.jp

***Abstract*—Emotion recognition from speech signals is one of the most important technologies for natural conversation between humans and robots. Most emotion recognizers extract prosodic features from an input speech in order to use emotion recognition. However, prosodic features changes drastically depending on the uttered text.**

In order to solve this problem, we have proposed the normalization method of prosodic features by using the synthesized speech, which has the same word sequence but uttered with a "neutral" emotion. In this method, all prosodic features (pitch, power, etc.) are normalized. However, nobody knows which kind of prosodic features should be normalized.

In this paper, all combinations of with/without normalization were examined, and the most appropriate normalization method was found. When both "RMS_Energy" (root mean square frame energy) and "VoiceProb" (power of harmonics divided by the total power) were normalized, emotion recognition accuracy became 5.98% higher than the recognition accuracy without normalization.

I. Introduction

Emotion recognition from speech signals is one of the most important technologies for natural conversation between humans and robots. If a robot can recognize an emotion of a user, the robot can make his own emotion, and change his behavior depending on the emotion. It is very natural for humans. Therefore, emotion recognition is a key technology for the next generation of human-robot interaction.

Many emotion recognition systems have been developed (e.g. [1], [2]). In most of these systems, prosodic features of an input speech are frequently used for recognition. It is well known that prosodic features are strongly related to the speaker's emotion. However, as a matter of course, prosodic features also strongly depend on the uttered text. As a result, prosodic features of speech change drastically with the content of what is being said, even if spoken with the same emotion.

In order to improve emotion recognition performance, prosodic features should be robust against differences of uttered text. We have developed[3] the normalization method of prosodic features by using synthesized speech. In this method, prosodic features extracted from an input speech are normalized by using the prosodic features extracted from the synthesized speech, which consists of the same word sequence, but uttered with "neutral" emotion. In other words, the method uses a difference of prosodic features between an emotional speech and a neutral speech.

Experimental results[3] showed that the normalized method gave higher recognition performance than the emotion recognition without normalization. However, the effects were not so high. One of the problem of the method is that the normalization is carried out for all prosodic features except MFCC. MFCC is affected by many factors, not only phonetic information but also speaker and emotion. In the normalization method, there is no difference about phonetic information, but speaker and emotion are different. In general, difference between speakers is bigger than difference between emotions. Therefore, normalization of MFCC causes deterioration of emotion recognition.

The other prosodic features are normalized, but it may be better that some features are not normalized. In this paper, appropriate normalization method are searched for each prosodic feature, and then emotion recognition performance is improved.

II. Overview of the normalization method by using synthesized speech

A. Basic structure

Figure 1 shows an overview of the emotion recognizer[3]. It has the same basic structure as conventional emotion recognizers. First, prosodic features are extracted from an input speech signal, and then feature vectors are input to a statistical recognizer. The statistical recognizer is trained by using training samples in advance, and outputs estimated emotions.

One of the most important differences is that the emotion recognizer requires a reference speech. This speech consists of the same text as the input speech, but uttered with neutral emotion. The prosodic features of an input speech are normalized by using the prosodic features extracted from the reference speech.

The reference speech is made by a speech synthesizer because it is impossible to record the reference speech at the same time. First, an input (emotional) speech is input to a speech recognizer in order to recognize an uttered text. Then, the reference speech is created by a speech synthesizer. In general, most speech synthesizers cannot produce output with emotion. Therefore, the output speech given by a speech synthesizer can be regarded as a reference speech uttered with neutral emotion.

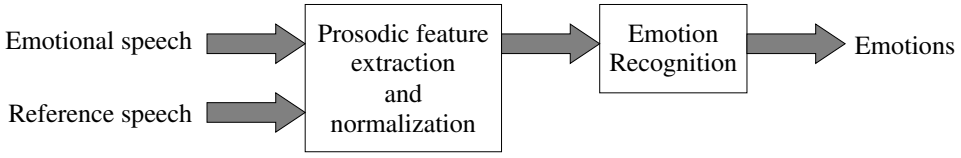

Fig. 1. Block diagram of the proposed method

Note that in the experiments described in this paper, transcripts of input speech data were given to the speech synthesizer instead of recognized text in order to avoid recognition errors.

B. Normalization of prosodic features

In the paper[3], two normalization methods — frame-level normalization and vector-level normalization — were proposed. The vector-level normalization is simple and easy to calculate, but it showed little improvement. Therefore, the frame-level normalization method is employed in this paper.

In the first step, low-level features (pitch, power, MFCC, and so on) are extracted for both an input speech signal and a reference speech. Then, the correspondence between frames is calculated by using the Dynamic Time Warping (DTW) method. MFCC and ΔMFCC parameters are calculated for both speech signals, and the correspondence between frames is calculated based on these parameters. Finally, all frames of low-level features except MFCC are normalized by subtracting the feature values of the corresponding frames.

Let $L^{(e)}(x)$ be a feature value of the x-th frame of a low-level feature calculated from the input speech signal, and $L^{(n)}(x)$ be a feature value of the x-th frame calculated from the reference speech. $c(x)$ denotes a set of frame numbers of the reference speech corresponding to the x-th frame of the input speech (Fig. 2). This correspondence $c(x)$ is automatically calculated by the DTW method. The normalized feature values $\hat{L}(x)$ are calculated by

$$\hat{L}(x) = L^{(e)}(x) - \frac{1}{|c(x)|} \sum_{i \in c(x)} L^{(n)}(i) \tag{1}$$

After normalization, statistical parameters are calculated from $\hat{L}(x)$ in the second step.

III. Appropriate normalization method for each prosodic feature

In the previous research[3], all prosodic features were normalized. We used several kinds of features, such as pitch, power, MFCC, zero-crossing rate, harmonics-to-noise ratio, and delta coefficients of these features. It cannot be said that all prosodic features should be normalized.

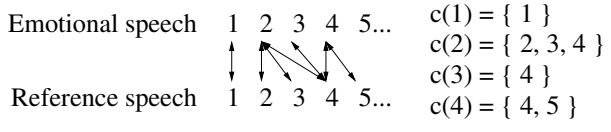

Fig. 2. Example of a correspondence parameter $c(x)$

The other problem is that which normalization operator should be used, subtraction or division. It is reasonable that power and pitch should be normalized by division because these features are perceived in log-scale by human. However, how about zero-crossing rate? harmonics-to-noise ratio? and delta coefficients of features? We cannot say which normalization operator should be used theoretically.

In order to solve these problems, we search an appropriate normalization method for each prosodic feature. All combinations of three normalization methods (subtraction, division, and none) are tested, and the most appropriate combination are found by checking the performance of emotion recognition.

A. Detailed of normalization method

Three normalization methods (division, subtraction, and none) are defined as:

subtraction

$$\hat{L}(x) = L^{(e)}(x) - \bar{L}^{(n)}(x) \tag{2}$$

division

$$\hat{L}(x) = \frac{L^{(e)}(x)}{\bar{L}^{(n)}(x)} \tag{3}$$

none

$$\hat{L}(x) = L^{(e)}(x) \tag{4}$$

where, $\bar{L}^{(n)}(x)$ denotes the average of $L^{(n)}(i)$ corresponding to the x-th frame of the input speech. In the division method, if $\bar{L}^{(n)}(x) = 0$ then $L^{(e)}(x)$ is used as $\hat{L}(x)$ instead of ∞.

IV. Experiments

A. Definition of emotions

There are two types of definition of emotions. One is emotion categories such as "joy," "anger," and "sadness." In this case, emotion categories are defined by hand, and a statistical classifier is used as a statistical recognizer. The other type is the position in an emotional space. The emotional space is defined in advance, and a statistical recognizer outputs a location in the space. Two or three dimensions are frequently used for representing an emotion space (e.g. [1], [4], [5]). In this case, a statistical regression model is used as a statistical recognizer. In this paper, we employed the latter type of emotion definition.

TABLE I
PROSODIC FEATURES

features	# dimensions
(Δ)RMS_Energy	2
(Δ)ZCR	2
(Δ)VoiceProb	2
(Δ)Pitch	2
(Δ)MFCC	24
Total	32

TABLE II
EMOTION RECOGNITION PERFORMANCE

Normalization	Average distance	
Baseline	0.9338	—
Division	0.9313	(+0.27%)
Subtraction	0.9331	(+0.08%)
Best	0.8780	(+5.98%)

B. Database

The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB)[6] was used for all experiments. It consists of natural dialogue, and 3,706 speech data uttered by 12 females. All speech data were manually labeled by six evaluators. In this database, the six-dimensional emotional space was defined; the dimensions are: pleasantness (pleasant—unpleasant), arousal (aroused—sleepy), dominance (dominant—submissive), credibility (credible—doubtful), interest (interested—indifferent), and positivity (positive—negative). The first two dimensions are related to personal emotional state, the third and fourth dimensions are related to interpersonal relationships, and the last two dimensions are related to attitude. Therefore, the first two dimensions are the most important to recognize a user's emotion.

C. Prosodic features

We used the standard prosodic features set, defined in the INTERSPEECH 2009 Emotion Challenge[7], and the openSMILE toolkit[8], [9] was used for extraction. Table I shows prosodic features. In detail, root mean square (RMS) frame energy, zero-crossing rate, voice probability (power of harmonics divided by the total power), pitch frequency, and 12-dimensional MFCC.

Delta coefficients were also calculated for each parameter, and then 12 statistical parameters (average, standard deviation, maximum, minimum, range, etc.) were calculated for all features. As a result, a 384-dimensional vector was calculated for one utterance.

D. Evaluation criterion

All speech data have manually-labeled six-dimensional emotion vectors. In this experiment, we focused on the first and second dimensions (pleasantness, arousal). Therefore, the system output two-dimensional emotion vectors, and the evaluation was carried out by calculating distance between two vectors calculated by Eq. (5):

$$d(\vec{e}, \vec{c}) = |\vec{e} - \vec{c}| \tag{5}$$

where, $\vec{c}$ denotes an emotional vector written in the UUDB, and $\vec{e}$ denotes an estimated one given by the system.

E. Other experimental setups

OpenJTalk[10], which is one of the HMM-based speech synthesizer, was used, and female speech signals were created. This synthesizer sometimes gave unnatural prosody, but we did not correct these manually.

We employed Support Vector Regression (SVR) as a statistical emotion recognizer. SVR is the same framework as Support Vector Machine, but it can output continuous values. The libSVM[11] was used in the experiments.

Training and testing were carried out by using the speaker-independent cross-validation framework. One speaker was selected for testing, and the other 11 speakers were used for training of SVR. Twelve experiments were carried out for each speaker, and the average of distance was used for evaluation. Because prosodic features related to emotions may be different among speakers, it is better that only the testing speaker's utterances are used for training. However, it is difficult for practical use to collect many emotional speech uttered by the testing speaker. Therefore, in this experiment, we used the speaker-independent (many speakers were used for training, but the testing speaker was not included in the training speakers) setup.

F. Results and discussion

Table II shows experimental results. In this table, "Baseline" denotes the results without normalization. "Division" and "Subtraction" denote the results with the normalization of all prosodic features except MFCC, and "Best" denotes the best results of all combinations of normalization. In this experiments, five kinds of features were used. Therefore, the number of combinations of normalization was $3^5 = 243$. "Best" shows the best performance of 243 experiments. Percentages in parentheses are improvement rates compared with the "Baseline."

The results showed that both "Division" and "Subtraction" increased the performance slightly, but "Best" combination improved the performance about 6.0%. It means that the appropriate normalization method should be selected for each feature. The best combination was effective for both pleasantness and arousal dimension. Especially, arousal dimension was improved about 9.4% (from 0.6505 to 0.5896) compared with the "Baseline."

Table III shows the best combination of normalization method for each feature. The best combination was that "RMS_Energy" was normalized by subtraction, "VoiceProb" was normalized by division, and other features are not normalized. The "RMS_Energy" is not represented by log domain.

TABLE III
BEST COMBINATION OF NORMALIZATION METHOD

Features	Normalization
(Δ)RMS_Energy	Subtraction
(Δ)ZCR	None
(Δ)VoiceProb	Division
(Δ)Pitch	None
(Δ)MFCC	None

Therefore, it should be normalized by division, but the best normalization was "Subtraction."

Table IV shows improvement rates compared with the "Baseline" for several combinations of normalization methods. In this table, a row indicates a prosodic feature, and a column indicates a normalization method. The percentage located in the "A" row and the "B" column is the improvement rate given by the setting that the prosodic feature "A" is normalized by the method "B" and the other features are normalized by the same method as the best combination. Bold-faced percentages indicates the result of the best combination.

For the "RMS_Energy," the performance was drastically decreased without normalization, but there was little difference between "Division" and "Subtraction." In other words, a kind of normalization method (division or subtraction) is not important for "RMS_Energy." On the other hand, both division and subtraction methods decreased the performance for "Pitch" and "MFCC."

Figure 3 shows improvement rates given by fixing the normalization method for "Pitch." For example, the normalization method for "Pitch" was fixed to subtraction. Then all combination patterns became $3^4 = 81$. 81 improvement rates were sorted, and written as the "subtraction" line in Fig. 3. In this figure, the "None" line is located higher position than the other two lines. It means that the normalization method "None" is effective for emotion recognition for the "Pitch."

Both "RMS_Energy" and "MFCC" gave similar results to Fig. 3. As a result, it is important for these three features whether normalized or not. "RMS_Energy" should be normalized, but "Pitch" and "MFCC" should not be normalized.

On the other hand, there was little difference among three normalization methods for "ZCR" and "VoiceProb." You can see the same conclusion from Figure 4. These two features may not contribute to the emotion recognition performance.

TABLE IV
IMPROVEMENT RATES FOR SEVERAL COMBINATIONS OF NORMALIZATION

Features	None	Division	Subtraction
RMS_Energy	+2.63%	+4.78%	**+5.98%**
ZCR	**+5.98%**	+4.97%	+5.96%
VoiceProb	+3.44%	**+5.98%**	+3.54%
Pitch	**+5.98%**	+1.40%	+1.83%
MFCC	**+5.98%**	−0.88%	−0.54%

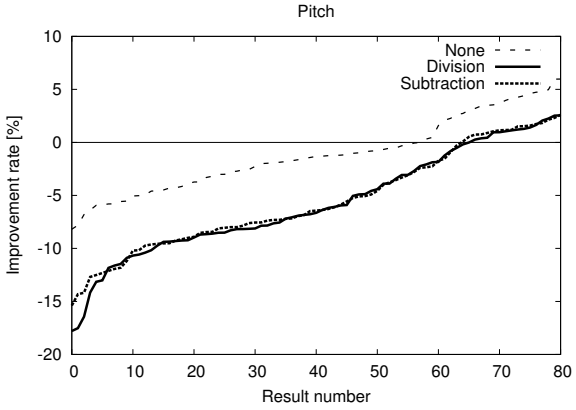

Fig. 3. Improvement rates given by fixing the normalization method for pitch

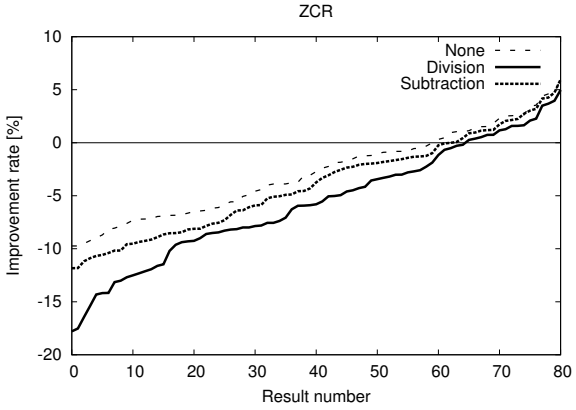

Fig. 4. Improvement rates given by fixing the normalization method for ZCR

G. Analysis for each utterance

Some speech data were improved, but others were deteriorated. In order to find what kind of utterances were improved, we have checked relationship between improved rate of each utterance and following three parameters; speaking speed, number of words in an utterance, and length of emotional vector. However, any relationships could not be found. Correlation coefficients were 0.09, 0.10, and 0.02 respectively.

In the experiments explained above, leave-one-speaker-out cross validation was carried out. It means that one speaker was selected as testing speaker, and other 11 speakers were used for training. Finally average performance of 12 experiments were calculated. In this section, the recognition performance for each speaker was also checked.

Table V shows the recognition performance for each speaker (before calculating the average). It can be said that the proposed method was effective for several speakers such as the speaker 5, 6, and 8, but it caused deterioration of recognition performance for other several speakers such as the speaker 1, 3, 4, and 12.

The speaker 6 and 8 spoke clearly, but the speaker 5 did not so clear. The speaker 3 and 4 spoke unclearly, but the speaker

TABLE V
RECOGNITION PERFORMANCE FOR EACH SPEAKER

Speaker	Baseline	Best normalization	
Speaker 1	0.8928	0.9194	(-2.98%)
Speaker 2	0.9178	0.9093	$(+0.93\%)$
Speaker 3	0.8705	0.8803	(-1.13%)
Speaker 4	0.8549	0.8664	(-1.34%)
Speaker 5	1.3373	1.1912	$(+10.93\%)$
Speaker 6	0.8886	0.7590	$(+14.58\%)$
Speaker 7	0.8897	0.8807	$(+1.01\%)$
Speaker 8	1.1726	0.8827	$(+24.72\%)$
Speaker 9	0.9353	0.9005	$(+3.72\%)$
Speaker 10	1.2769	1.2281	$(+3.82\%)$
Speaker 11	0.8417	0.7805	$(+7.27\%)$
Speaker 12	0.7435	0.7824	(-5.23%)

1 and 12 did clearly. Average pitch frequency was not related to the recognition performance, either. We could not find the reason why the proposed method is effective for only several speakers. We leave exploring this problem to future work.

V. CONCLUSIONS

Most of emotion recognizers extract prosodic features from an input speech in order to use emotion recognition. However, prosodic features changes drastically depending on the uttered text. Therefore, it is difficult to recognize an emotion by using prosodic features directly. In order to solve this problem, we have proposed the normalization method[3] of prosodic features by using the synthesized speech, which has the same word sequence but uttered with a "neutral" emotion. In this method, all prosodic features except MFCC are normalized. However, the best normalization method for prosodic features is not known.

In this paper, all combinations of with/without normalization were examined, and the most appropriate normalization method was found. Five kinds of prosodic features were used, and three normalization methods (subtraction, division, and none) were applied to each feature independently. Totally $3^5 = 243$ combinations were examined, and the best combination were selected. When both "RMS_Energy" and "VoiceProb" were normalized, emotion recognition accuracy improved 5.98% compared with the recognition accuracy without normalization.

Analysis of experimental results said that "RMS_Energy" should be normalized, but "Pitch" and "MFCC" should not be normalized. Normalization method (division or subtraction) is not important for normalization of "RMS_Energy." On the other hand, there was no difference with/without normalization for "ZCR" and "VoiceProb." It may mean that these two features slightly contribute to the emotion recognition performance.

We could find the best combination of normalization, but we do not know why the combination is the best, and the proposed method is effective for which kind of utterances. We have to investigate the experimental results more deeply. It is one of the future work.

ACKNOWLEDGMENT

A part of this work was supported by JSPS KAKENHI Grant Number 21300036.

REFERENCES

[1] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10–11, pp. 787–800, 2007.

[2] I. Luengo, E. Navas, and I. Hernaez, "Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge," in *Proc. INTERSPEECH*, 2009, pp. 332–335.

[3] M. Suzuki, S. Nakagawa, and K. Kita, "Prosodic feature normalization for emotion recognition by using synthesized speech," in *Proc. 16th Annual Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, 2012, pp. 306–313.

[4] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford University Press, 1989.

[5] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[6] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.

[7] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. INTERSPEECH 2009*, 2009, pp. 312–315.

[8] F. Eyben, M. Wollmer, and B. Schuller, "openEAR — Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009)*, vol. I, 2009, pp. 576–581.

[9] ——, "openSMILE — Speech and music interpretation by large-space extraction." [Online]. Available: http://opensmile.sourceforget.net/

[10] Nagoya Institute of Technology, "The Japanese TTS system OpenJTalk." [Online]. Available: http://open-jtalk.sourceforge.net/

[11] C.-C. Chang and C.-J. Lin, "libSVM—a library for Support Vector Machines." [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm/