Speaker Identification using Normalized Features in Emotional Environment

Taro Togawa^{a,*}, Sayuri Kohmura^a, Jun Takahashi^a

^aFUJITSU LABORATORIES LTD., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, 211-8588, Japan

*Corresponding Author: togawa.taro@fujitsu.com

Abstract

This paper proposes a speaker identification method using normalized features that is robust to the speaker's emotional changes. In general, the voice speaker identification method is based on the assumption that the voice does not contain emotions such as happiness, anger, sadness, and fear. However, acoustic features related to personality such as accents and intonations change with emotional changes, making it difficult for individual differences in speech to appear. Therefore, this paper proposes a speaker identification method that is robust to various emotional changes. In the proposed method, the differences between the acoustic features of emotional speech and neutral speech are learned in advance, and the normalized features compensating for the emotional changes in the input speech are calculated. We evaluated the performance of speaker identification accuracy for five types of emotional speeches. As a result, speaker identification accuracy was improved by approximately 10% compared with the previous method under the assumption that the emotion to the input speech was known, and the effectiveness of the proposed method was confirmed.

Keywords: speaker identification, speaker diarization, emotional environment.

1. Introduction

In recent years, with the improvement in the accuracy of speech recognition technology, there has been an increase in efforts to improve business efficiency in companies by recording the voices of people in meetings with a microphone, converting the recorded data into text data and utilizing the data to support minutes preparation.

In general, meetings involve utterances by multiple participants, such as through information sharing, questions, and discussions. In a face-to-face meeting, the speech of a plurality of speakers is often mixed up in recorded data. Therefore, there is a need for a speaker identification technology that identifies whose speech is contained in the recorded data. Such speaker identification technology is expected to be used in a variety of fields such as in the security field, where personal authentication by voice is performed, and the customer relationship management (CRM) field where optimal services are provided for each customer.

2. Related Work

2.1 Common Speaker Identification Technologies

This section describes previous studies on speaker identification technologies. First, the positioning of the speaker identification technology in the speaker recognition field is confirmed. There are two types of technologies utilized to recognize a speaker based on their voice. One technology is speaker verification, which determines whether or not the input speech is of the registrant, and the other is speaker identification, which identifies which person is speaking among a plurality of persons whose input speech is registered. Furthermore, operational procedure divides speaker identification into “text dependent,” where the same text is uttered by the speaker during the training and testing phases, and “text independent,” where different texts are uttered by the speaker during the training and testing phases. Speaker identification is often used in the aforementioned meeting minutes application. The speaker recognition evaluation task performed by the National Institute of Standards and Technology (NIST SRE) is mainly a text-independent type

of speaker recognition task. In addition, since the technology application range is wide, text-independent speaker recognition has become the mainstream type used in research. This study also utilizes text-independent speaker identification.

In text-independent speaker identification, a machine learning technique such as the support-vector machine (SVM) has been used often since the 1990s because (1) an audio signal can be treated as a set of feature vectors, and (2) it is necessary to improve the accuracy for an unlearned speaker. The method constructing a high-dimensional vector that connects the average vectors of speaker models expressed by a Gaussian mixture model (GMM), which is a generation model, and identifies the speaker using a SVM⁽¹⁾, has been a cornerstone of speaker identification research since the 2000s. Furthermore, there have been many studies using deep learning using a deep neural network (DNN) as an identification model.

An acoustic feature quantity; a feature quantity related to vocal cord characteristics such as a logarithmic power spectrum, mel-frequency cepstrum coefficients (MFCC), and line spectral frequencies (LSF); and a combination of these feature quantities are often used in speaker identification.

2.2 Speaker Identification in Emotional Environment

Speaker identification in an emotional environment such as that of happiness, anger, sadness, and fear is one of the more challenging research areas⁽²⁾. This is because the acoustic features related to personality, such as accents and intonation changes due to emotional changes, make it difficult for individual differences in speech to appear.

There have also been many studies on emotion recognition based on speech. For example, a method of estimating emotions, such as anger, happiness, and sadness, using a discriminator based on long term modulation spectrum (LTMS) or bidirectional-LTMS network configuration using acoustic features such as a logarithmic power spectrum has been proposed^(3, 4).

Regarding speaker identification in an emotional environment, Jawaker⁽⁵⁾ demonstrated the effect of improving speaker identification accuracy in an emotional environment using the judgment results of an identifier based on a plurality of acoustic feature quantities such as MFCC and LSF. Shahin⁽⁶⁾ proposed an emotion-based

speaker identification method that identified input speech emotions using a GMM identifier and learned a discrimination model with a DNN for each identified emotion.

2.3 Problem of the Previous Method

These previous methods simultaneously dealt with two kinds of variable factors: emotional change and individuality, and the model for identification had a high degree of freedom. For these previous methods, it was assumed that all speakers to be identified could be learned using emotional speech databases. Therefore, when there was a speaker whose speech contained only a small amount of a certain emotion, the identifier could not be sufficiently learned, and the accuracy of the speaker was often greatly reduced. Moreover, since most open speech corpus and speech obtained in real life are neutral, it is costly to collect various emotional speech data for every speaker, which was a practical problem in previous research.

Therefore, developing a speaker identification method corresponding to various emotional changes, even when there is only a small amount of emotional speech data, has been desired.

3. Proposed Approach

3.1 Point of Approach

In the context of the emotion research, the Darwinian school of psychology represented by Ekman continues to assert the existence of universal, culturally independent basic emotions. Based on this idea, it is thought that the basic emotions contain some characteristics common to various speakers and characteristics particular to an emotion appear in a greater quantity when the voice changes from neutral to the other basic emotion.

If it is possible to model the common emotional changes from the neutral emotion in the feature space of the speech, there would be less emotional voice variation and the model for identification would have lower degree of freedom. As a result, it is expected that the accuracy of the speaker is improved even when learning using only a small amount of emotional speech data.

3.2 Modeling of Emotional Change

We investigated the differences between the features of neutral speech and those of emotional speech using a portion of emotional speech corpus. It was found that feature change vectors showed a similar tendency regardless of speaker differences. Figure 1 visualizes the distribution of neutral and emotional speech (sadness and anger) for three speakers. In Figure 1, the features, consisting of 128 dimensions of the logarithmic power spectrum of these speeches, are dimensionally compressed to two feature dimensions by the principal component analysis. Figure 1 shows that, although there are variations in the emotions of sadness and anger among the individual utterances, there is a similar in the change from neutral to one of the emotions regardless of the speaker.

3.3 Outline of the Proposed Method

Therefore, in this study, we propose a method to improve speaker identification accuracy for various emotions, even with only a small amount of learning data, by using the speech property that features tend to change similarly due to emotion regardless of the speaker. The proposed method learns a change vector of a features due to emotional change in advance, normalizing the feature changes due to the emotion by applying a learned change vector according to the emotion of a given input speech and identifying a speaker based on the normalized features. The framework of the proposed method is shown in Figure 2.

3.4 Learning Method of Normalized Vectors

A learning method for feature normalized vectors in the learning phase is described below. For each emotion, a normalized vector Y_{em} is calculated by taking the difference between the average value of the frame-based features $X_{em}(i)$ of the emotional speech learning data set and the average value of the frame-based features $X_{neutral}(i)$ of the neutral speech learning data set.

$$Y_{em} = \frac{1}{n} \sum_{i=1}^n X_{em}(i) - \frac{1}{m} \sum_{i=1}^m X_{neutral}(i) \quad (1)$$

where n is the total number of frames in the emotional speech learning data set, and m is the total number of frames in the neutral voice learning data set.

3.5 Emotion Estimation

In the evaluation phase, the emotion estimation can be realized through the same structure utilized in previous research ^(3, 4). In this study, to determine whether it is

Fig. 1. Distribution of the acoustic feature vectors.

Fig. 2. The framework of the proposed method.

appropriate to normalize the features for each emotion, it is assumed that the emotion estimation is ideal, and the emotion to the input speech is treated as if it is known.

3.6 Features Normalization for Input Speech

In the evaluation phase, the method for normalizing the feature quantity is described below. The normalized features $X'_{em}(i)$ are calculated by subtracting the normalized vector Y_{em} for each emotion obtained by Equation (1) from the frame-based features $X_{em}(i)$ for the input speech.

$$X'_{em}(i) = X_{em}(i) - Y_{em} \quad (2)$$

If the input speech is a neutral emotion, normalization is not performed.

3.7 Speaker Classifier

An SVM is used as the speaker classifier, which is one of the identifiers frequently used in previous studies.

In the learning phase, to reduce the influence of determination bias based on a speaker having a larger amount of learning data, the weight corresponding to the

reciprocal of the ratio of the amount of data for each speaker is first applied, and then the classifier is learned.

In the evaluation phase, the speaker identification results estimated for each frame are totaled for the entire utterance, and the speaker identified most frequently by the majority decision is estimated as the speaker for the entire utterance.

4. Experiment and Discussion

4.1 Speech Database

In this research, we used a Berlin database⁽⁷⁾ (Emo-DB) composed of seven emotions and ten speakers (five men and five women) as speech data. These data are 1 s to 2 s sentence utterances sampled at 16 kHz and uttered in specific emotions through acting. The speech data was divided into two sets: one for model training and the other for evaluation. Of the seven emotions, two emotions were excluded because the number of sentences per speaker was extremely small, and the remaining five emotions (420 utterances in total) were used in the evaluation experiment. Table 1 lists the number of sentences in the evaluation data.

Table 1. Number of sentences

Emotion	Training data	Test data
Neutral	40	39
Happiness	36	35
Anger	64	63
Sadness	31	31
Boredom	41	40
Total	212	208

4.2 Conditions

Because the speech data includes silences of several hundred milliseconds each, silence intervals were excluded in preprocessing. Further, a frame was formed in units of 256 samples (25 ms), and the feature vectors were calculated. In previous research, the feature vectors were often used by combining multiple features of vocal cords, but in this research, the logarithmic power spectrum, which is a basic feature, is used. In the training phase, the normalized vectors of the features and the speaker classifier were learned using the training data set of Table 1. In the evaluation phase, the normalized feature was obtained by applying the normalized vector to the evaluation data set, the non-training data, and the speakers were classified by the classifier.

(a) Without normalization

(b) With normalization

Fig. 3. Distribution of acoustic feature vectors.

4.3 Experiment 1

First, when normalization using the proposed method is applied, it was investigated whether the feature vectors for emotional speech are collectively distributed for each speaker. In this experiment, the feature vectors are 128-dimensional logarithmic power spectra. For visualization, Figures 3 (a) and (b) show a two-dimensional compression with t-SNE⁽⁸⁾. In Figure 3 (a), without normalization, it can be confirmed that even if speech from the same speaker is spoken, there is a large difference in the distribution region between the neutral, sadness, and anger emotions. In contrast, in Figure 3. (b), where the normalization by the proposed method is applied, the feature distributions for the speech of the same speaker tend to be collective for each speaker regardless of the emotion. Based on this result, it was confirmed that it is appropriate to compensate emotional change with the same normalized vector regardless of the speaker.

4.4 Experiment 2

Next, the proposed and conventional speaker identification accuracy methods were evaluated. For comparison, two previous methods are used. Previous Method 1 involves dividing training and test data for multiple emotions, and multiple classifiers are trained and applied for each emotion. For Previous Method 2, a single classifier is trained and applied from training and test data that are mixed with a plurality of emotions. Table 2 summarizes the results of speaker identification evaluation for each method.

When the classifier is trained using all the training data sets, the proposed method and Previous Method 1 have the same identification accuracy. In contrast, when trained with half of the training data, the proposed method had the highest identification accuracy, approximately 10% higher than Previous Method 1. Therefore, the proposed method for normalizing the features of each emotion was confirmed to be effective under the condition that the amount of training data is small.

Table 2. Speaker identification accuracy (10 speakers)

Method	Trained by all data		Trained by half data	
	Per frame	Per utterance	Per frame	Per utterance
Proposed	42.2%	85.6%	39.2%	75.0%
Previous 1	41.7%	85.6%	33.4%	65.4%
Previous 2	39.0%	69.2%	35.9%	60.1%

4.5 Discussion

The relationship between the amount of emotional speech training data and an appropriate speaker identification method is discussed. In Table 2, the identification accuracy of the proposed method is approximately 15% higher than that of Previous Method 2. The proposed method reduces the influence of the emotional change by using the normalized vector.

In contrast, for Previous Method 1, while the influence of the change in the feature amount due to the emotion can be eliminated by dividing the classifier for each emotion, the training data amount per classifier is reduced. Therefore, it is presumed that the mechanism for switching the speaker identification method according to the training data amount is effective, such as using the proposed method under the condition of a small training data amount and using Previous Method 1 as the training data amount increases.

5. Conclusions

This paper proposed a speaker identification method that is robust to the speaker emotional changes and uses normalized features. We evaluated the performance of speaker identification accuracy for five types of emotional speech. The speaker identification accuracy was improved by approximately 10%, compared with Previous Method 1, when using the proposed method under the assumption that the emotion to the input speech was known, confirming the effectiveness of the proposed method.

In the future, a combination of features other than those of the logarithmic power spectrum will be examined, and an emotion estimation model trained using a machine learning method will be introduced to verify the performance of the hybrid operation of emotion estimation and speaker identification.

References

- (1) D.A. Reynolds and R.C. Rose: "Robust text independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process., Vol. 3, pp. 72-83, 1995
- (2) S. Furui: "Speaker-dependent-feature extraction, recognition and processing techniques," Speech Communication, Vol. 10, Issues 5-6, pp. 505-520, 1991
- (3) S. Ghosh, E. Laksana, L.P. Morency and S. Scherer: "Representation Learning for Speech Emotion Recognition," Interspeech, ISCA, San Francisco, USA, pp. 3603-3607, 2016
- (4) A. Satt, S. Rozenverg, and R. Hoory: "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," Interspeech, ISCA, Stockholm, Sweden, pp. 1089-1093, 2017
- (5) N.P. Jawarkar, R.S. Holambe, and T.K. Basu: "Text-independent Speaker Identification in Emotional Environments: A Classifier Fusion Approach," Frontiers in computer education, Vol. 133 of the series advances in intelligent and soft computing, pp. 569-576, 2012
- (6) I. Shahin, A. Bou Nassif, and S. Hamsa: "Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments," Neural Computing and Applications, pp. 1-13, 2018
- (7) F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier,

and B. Weiss: "A database of German emotional speech," Ninth European Conference on Speech Communication and Technology, 2005

- (8) L. Maaten, and G. Hinton: "Visualizing data using t-SNE," Journal of machine learning research 9, pp. 2579-2605, 2008.