ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Emotional Intensity Level Analysis of Speech Emotional Intensity Estimation

Megumi Kawase and Minoru Nakayama

Tokyo Institute of Technology - School of Engineering
Meguro, Tokyo - Japan

Abstract. An estimation procedure using three models to determine the appropriate emotional intensity from the 10 listed emotional intensity classes for utterances has been developed in order to support better communication between humans and machines. In order to improve estimation performance, utterances were divided into segments and an estimated emotional intensity and its probability were produced as outputs. Two feature vectors were produced from the outputs and these features were used for the utterance-level classification using Support Vector Machine and Random Forest techniques. In the results, the accuracy of emotional intensity estimation in two out of three models was improved using the procedure proposed. In addition, features which contributed to the estimations were analyzed.

1 Introduction

Intensity estimation has a lot of potential applications for human-robot interaction, patient monitoring, security surveillance and entertainment. If it is not possible to read the intensity of emotions during speech input, the possibility that responses given when humans and machines communicate are greatly misunderstood cannot be eliminated. In order to avoid such situations, research on how to better estimate emotional intensity has been conducted.

In Han et al.'s research [1], speech data was divided into segments, and the emotional probability of each segment was estimated using the acoustic features extracted from the segment. In our previous study [3], we estimated the emotional intensity of a Japanese speech corpus using deep learning, and found that based on the segment-level emotional intensity it was possible to estimate emotional intensity with an accuracy of 52.4% using serial estimation. However, the number of segments varies greatly by utterance and thus the method of estimating utterance-level emotional intensity which uses the most frequent value of segment-level emotional intensity may be inappropriate. It is necessary to propose a new method of utterance-level classification based on the segment-level emotional intensity. In this study, estimation probability is considered to be useful information, an estimation method which takes these into account is anticipated to improve performance. The following topics will be addressed in this paper.

Usefulness of both the class estimation probability vector obtained for each segment and the class frequency vector obtained for each utterance is determined.

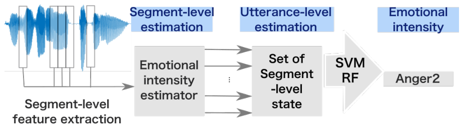

Segment-level estimation
Utterance-level estimation
Emotional intensity
Emotional intensity estimator
Set of Segment -level state
SVM RF
Anger2
Segment-level feature extraction

Fig. 1: Individual estimation

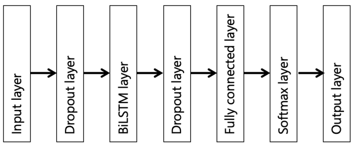


Fig. 2: The constructed network

2 Method

2.1 Speech Corpora

Speech data from an online gaming voice chat corpus with emotional labels (OGVC) [2] is used as training and test data for a deep learning program. The duration of the speech data is 218ms to 7393ms. Since the sampling frequency for the emotion-labeled speech corpus EmoDB[4] was 16 kHz, the sampling frequency was reduced from 44.1 kHz to 16 kHz using downsampling.
“Joy”,“Sadness” and “Anger” considered to be independent, and are relatively evenly spaced around the origin in Russell’s circumplex model [5]. Therefore, the speech data of 4 voice actors (2 males and 2 females) was performed using emotional intensities of “ 0 (Neutral), ” “ 1 ” , “ 2 ” and “ 3 ” respectively, and in total 992 of 248 different utterances were used. All speech data with “ 0 ” intensity was regarded as “ Neutral ” . As a result, all recorded speech was classified into 10 categories: 3 emotions × 3 intensities + Neutral. Data augmentation is a method of increasing the amount of speech data by adding Additive White Gaussian Noise to the training data. The amount of speech data after augmentation is 5 times greater that the amount of the original data.

2.2 Models

The following three models were created to classify utterances.
Individual estimation (shown in Fig.1): This model considers each emotion and its intensity independently, and classifies them into 1 of 10 categories.
Parallel estimation: Emotion and intensity estimators are created separately, and the emotional intensity is determined by integrating the result obtained from each of the two estimators into one.
Serial estimation: Emotions are first estimated and classified into one of four classes of emotions. After that, the emotional intensity of each class of emotion except “ Neutral ” were classified.

2.3 Algorithm for estimating emotional intensity

Individual estimation was used to explain the details of the algorithm. Figure 1 shows the overview of the approach.
Segment level feature extraction : The input signal was converted into

frames using a 30 ms hamming window shifting 15 ms per frame. A 41-dimensional feature vector of each frame was extracted. The features are as follows: GTCC(13), GTCCDelta(13), MFCCDelta(13), spectralCrest(1), and pitch(1) The subscript indicates the number of dimensions. The features were chosen experimentally. A segment-level feature vector was formed by stacking the features in the neighboring frames. According to [6], a speech segment of about 250 ms contains sufficient information to estimate the emotion. In this experiment, the segment-level feature is set to 15 frames. Therefore, the total length of each segment is 15 ms × 15 + (30 - 15) ms = 240 ms.
Segment level estimation: A network which consisting of two dropout layers, a bidirectional LSTM layer and a softmax layer, was constructed. The network was trained to predict the probability of each emotional intensity (the class estimation probability vector) using segment-level features. The number of input units was set according to the vector size of the segment-level features. The output size was set according to the level of emotional intensity of each of the 10 classes. The dropout rate number and the hidden units were chosen using Bayesian optimization.
Utterance level estimation: Class mean vectors and class frequency vectors were used for the utterance-level emotional intensity classification. The details are described in Section 2.4.

2.4 Class mean probability vector and class frequency vector

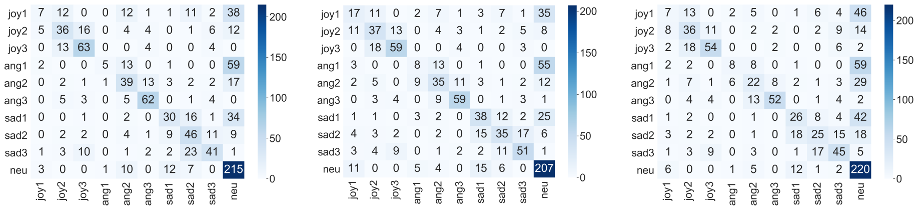
A vector consisting of the probability of the segment belonging to each class (class estimation probability vector) and the emotional intensity with the highest probability (segment label) are obtained as outputs. Using these vectors, the following utterance-level features were created.

$$F_1^k = \frac{1}{N}\sum_{i\in U} p_i(e_k), F_2^k = \frac{n(e_k)}{N}$$

where N is the number of segments in the utterance, $p_i(e_k)$ is the probability that the ith segment is of the kth class, $n(e_k)$ is the number of segments in the utterance that are of the kth class, F_1^k represents the utterance average of the class estimation probability vector for each segment (class mean probability vector), and F_2^k represents the percentage of segment labels of each class in the utterance (class frequency vector). It is thought that the creation of such utterance-level features is not affected by the fact that the number of segments differs between utterances. In this paper, we compare the performance of the three methods of utterance-level classification using the following utterance-level features. 1.F_1^k, 2.F_2^k, 3.F_1^k and F_2^k.We used these features as the features for Support Vector Machine (SVM) and Random Forest (RF) learning methods and compared the results of classification using the models which were produced.

Table 1: Comparison of the accuracy (%) of each class of learners

		Individual	Parallel	Serial
Previous method (mode)		49.9	**52.1**	52.4
F_1^k	SVM	**56.5**	46.9	56.7
	RF	55.1	46.8	**57.1**
F_2^k	SVM	53.8	44.2	55.4
	RF	54.0	45.3	54.9
F_1^k & F_2^k	SVM	55.0	46.6	56.6
	RF	54.8	47.7	56.0



	joy1	joy2	joy3	ang1	ang2	ang3	sad1	sad2	sad3	neu
joy1	7	12	0	0	12	1	1	11	2	38
joy2	5	36	16	0	4	4	0	1	6	12
joy3	0	13	63	0	0	4	0	0	4	0
ang1	2	0	0	5	13	0	1	0	0	59
ang2	0	2	1	1	39	13	3	2	2	17
ang3	0	5	3	0	5	62	0	1	4	0
sad1	0	0	1	0	2	0	30	16	1	34
sad2	0	2	2	0	4	1	9	46	11	9
sad3	1	3	10	0	1	2	2	23	41	1
neu	3	0	0	1	10	0	12	7	0	215

(a) RF

	joy1	joy2	joy3	ang1	ang2	ang3	sad1	sad2	sad3	neu
joy1	17	11	0	2	7	1	3	7	1	35
joy2	11	37	13	0	4	3	1	2	5	8
joy3	0	18	59	0	0	4	0	0	3	0
ang1	3	0	0	8	13	0	1	0	0	55
ang2	2	5	0	9	35	11	3	1	2	12
ang3	0	3	4	0	9	59	0	1	3	1
sad1	1	1	0	3	2	0	38	12	2	25
sad2	4	3	2	0	2	0	15	35	17	6
sad3	3	4	9	0	1	2	2	11	51	1
neu	11	0	0	5	4	0	15	6	0	207

(b) SVM

	joy1	joy2	joy3	ang1	ang2	ang3	sad1	sad2	sad3	neu
joy1	7	13	0	2	5	0	1	6	4	46
joy2	8	36	11	0	2	2	0	2	9	14
joy3	2	18	54	0	0	2	0	0	6	2
ang1	2	2	0	8	8	0	1	0	0	59
ang2	1	7	1	6	22	8	2	1	3	29
ang3	0	4	4	0	13	52	0	1	4	2
sad1	1	2	0	0	1	0	26	8	4	42
sad2	3	2	2	0	1	0	18	25	15	18
sad3	1	3	9	0	3	0	1	17	45	5
neu	6	0	0	1	5	0	12	1	2	220

(c) Previous method

Fig. 3: The confusion matrix using an individual model

3 Results

3.1 Performance comparison by model

All results are reported using unweighted accuracy, so that they are equal to the average of the per class recall. They were measured and compared using *5*-fold cross-validation. Table 1 shows the accuracy obtained by classification of features using SVM and RF. The table also shows the accuracy of the utterance-level classification using the most frequent value segment-level emotional intensity. For individual and serial estimation, the highest accuracy of each model was achieved when only F_1^k was used. Especially for serial estimation using RF, an accuracy of emotional intensity estimation of 57.1% resulted. Therefore, the probability information which was eliminated when outputting segment-level emotional intensity from the class estimation probability vectors is considered necessary for utterance-level classification. On the other hand, in parallel estimation, there was no improvement in the accuracy using either SVM or RF.

3.2 Comparison of SVM and RF for individual estimation

Figure 3 shows the confusion matrices when only F_1^k is used for individual estimation. The accuracy of " Neutral " decreased when each method was used. This may be due to the fact that the number of neutral misclassified utterances was corrected. SVM can discriminate between " Neutral " and " Intensity 1 " or " Intensity 2 " better than the previous methods were able to. However, RF can successfully discriminate between “ Intensity 2 ” and “ Intensity 3 ” , but not between “ Intensity 2 ” and “ Neutral ” or “ Intensity 1 ” .

4 DISCUSSION

4.1 Utterance-level classification

We proposed a method of utterance-level classification using SVM and RF with class mean probability vectors and class frequency vectors of each utterance, and compared the results. In these results, we found that both the individual and serial estimation models improved the accuracy when only F_1^k was used by 6.6% and 4.7% respectively, while parallel estimation did not improve the accuracy at all. This may be due to the fact that, unlike the other two models, the parallel estimation method does not take into account the difference in emotions. This suggests that the effect of emotional intensity on feature values is not common nor linear across emotions, and that it is necessary to distinguish between them.

4.2 Contribution of evaluation vectors

Since the class estimation probability vector was considered to be effective for utterance-level classification described in Section 3.1 above, the evaluation vectors were calculated. These vectors were determined according to [7] in order to investigate which kind of information is effective.

$$\begin{aligned} f_1^k &= \frac{1}{N}\sum_{i\in U} p_i(e_k), f_{2-5}^k = \frac{n\{p_i(e_k) > \hat{\theta}\}}{N}, \hat{\theta} \in \{0.1, 0.2, , ..., 0.4\} \\ f_6^k &= \sum_{i=1}^{N} \frac{1}{2}\{p_i(e_k) + p_{i+1}(e_k)\} \end{aligned}$$

where $n\{p_i(e_k) > \hat{\theta}\}$ is the number of segments where $p_i(e_k)$ exceeds the threshold $\hat{\theta}$. The f_1^k is the same as F_1^k as defined in Section 2.4, and f_{2-5}^k represents the percentage of segments whose estimated probability exceeds the threshold $\hat{\theta}$. Also, f_6^k represents the areas formed between the estimated probability lines and the horizontal axis in Fig. 4. An evaluation vector for f_{1-6}^k is obtained from each utterance. The class with the highest value for each evaluation vector was used for utterance-level classification.

Table 5 shows the results of a comparison of the accuracy of the utterance-level classifications using the highest evaluation vector values for evaluation vectors f_1^k to f_6^k. Compared with the accuracy when using the previous method, all models of accuracy improved when evaluation vectors were used. In particular, when average probability was used, a stable improvement in the accuracy was observed, so it is considered to be an effective evaluation vector for all models. In addition, the evaluation vectors which yielded high accuracy differed between models.

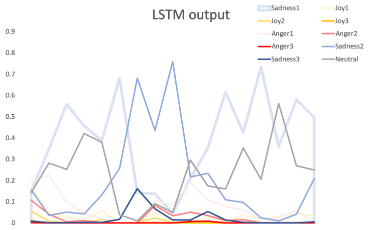


Fig. 4: Estimated probability of each emotional intensity

	Individual	Parallel	Serial
Previous (mode)	49.8	52.0	51.2
Average	**52.6**	**52.6**	52.3
Threshold (0.1)	50.7	48.2	**53.0**
Threshold (0.2)	50.8	51.2	51.8
Threshold (0.3)	51.9	51.5	51.7
Threshold (0.4)	51.6	52.4	51.4
Area	52.0	52.2	52.1

Fig. 5: Comparison of the accuracy of each feature

5 Conclusions

In this study, we used deep learning to estimate the emotional intensity of a Japanese speech corpus. We created features from the segment label sets and the class estimation probability vectors for each segment output, and proposed a method for determining utterance labels based on the features. As a result, we obtained the following conclusions for the purpose mentioned in the foreword.
1. Class mean probability vectors are more useful than label frequency vectors in determining utterance labels, and in particular, when SVM is applied to individual estimation, up to 6.6%.
2. The evaluation vectors created from the class estimation probability vectors are effective in determining the utterance label, and their contribution to the utterance label determination is significant, especially since the class mean probability vector shows a stable improvement in the accuracy. By combining the evaluation vectors, it is expected that an even higher level of accuracy may be obtained. Discussion about how to achieve this is future work.

References

[1] Han Kun, Yu Dong, Tashev Ivan: “Speech emotion recognition using deep neural network and extreme learning machine”, In INTERSPEECH-2014, 223-227.

[2] Speech Resources Consortium, “ Online gaming voice chat corpus with emotional label (OGVC) ”, http://research.nii.ac.jp/src/OGVC.html

[3] Megumi Kawase, Minoru Nakayama,”Acoustic features of a Japanese speech corpus for emotional intensity estimation”, IEICE Tech. Report, vol.120, no.306, pp.55-60, Dec.2020.

[4] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walther F. Sendlmeier and Benjamin Weiss, “ A Database of German Emotional Speech”, Proc. Interspeech 2005

[5] Russell James A, “ A circumplex model of affect”, Journal of Personality and Social Psychology, 1980, 39, 1161-1178

[6] M.Shuiyang, P.C Ching, L. Tan. ”Deep Learning of Segment-Level Feature Representation with Multiple Instance Learning for Utterance-Level Speech Emotion Recognition”, INTERSPEECH2019,September 15-19, 2019, Graz, Austria.

[7] Mao, S., Ching, P., Lee, T. (2019) ”Deep Learning of Segment-Level Feature Representation with Multiple Instance Learning for Utterance-Level Speech Emotion Recognition.” Proc. Interspeech 2019, 1686-1690, DOI: 10.21437/Interspeech.2019-1968.