Cutomer Satisfaction Classification Based on Face Emotion Recognition and Speech Emotion Recognition

Duman Care Khrisne[1], I Made Arsa Suyadnya[1],I Putu Elba Duta Nugraha[1], and Theresia Hendrawati[2]

[1]Department of Electrical Engineering
Faculty of Engineering, Udayana University
Denpasar, Indonesia

[2]Informatics Study Program
Faculty of Technology and Informatics, Institut Bisnis dan Teknologi Indonesia
Denpasar, Indonesia

duman@unud.ac.id

Abstract Nowadays, customer satisfaction is very important for businesses and organizations. Manual methods such as surveys and distributing questionnaires to customers are considered less fast to get feedback from customers. Today businesses and organizations are looking for a quick way to get effective and efficient feedback on customer satisfaction, to get potential customers faster. In this study, we propose a new method for detecting customer emotion, Face Emotion Recognition (FER) and Speech snippets Emotion Recognition (SER) to identify customer satisfaction using deep learning techniques. The application is built using the Deep Learning method with CNN architecture for FER and with the ResNet architecture model for SER, the application has good performance in recognizing emotions from facial expressions and emotions that arise from speech. This can be seen from the results of the test data which produces quite good accuracy values, 73% for FER (face) and 77% for SER (speech snippets) respectively. In theory, this value indicates good classification performance.

***Index Terms*— Customer Satisfaction, Face Emotion Detection, Speech Emotion Detection.**

I. INTRODUCTION

In monitoring consumer satisfaction with the business of a service, business owners usually provide a suggestion box or distribute questionnaires as a data collection medium. In the digital era, customer satisfaction has an even more important role than before. Ratings can tell us whether a product or service has the level of security or reliability required by customers. Giving a rating (such as giving a star) does make it easier for users to rate a product or service, but a rating cannot capture the feelings or satisfaction of the customer.

This is because customers have difficulty expressing emotions through written comments or just by rating them. One of the techniques used to better demonstrate customer satisfaction with goods is video testimonials [1]. However, video testimonials have a problem in the analysis process, a video provides more insight but requires more energy and resources for the analysis process.

Automatic emotion recognition based on facial expressions is an interesting area of research, which has been presented and applied in several fields such as safety, health and human machine interfaces [2]. Researchers in this field are interested in developing techniques to interpret, code facial expressions, and extract these features to get better predictions by computers. Until now, the use of facial recognition is a solution to assess customer satisfaction that is effective and efficient with respect to time [3]. Facial emotion recognition (FER) tries to correlate the movement of facial features with consumer satisfaction, and finally detects patterns in facial expressions that can predict the level of consumer satisfaction. Most customer satisfaction is emotional. This happens at the level of consciousness and the human subconscious. Through the experiences they go through while using the product, consumers experience positive and negative emotional reactions

Another approach that can be used as a reference in emotion recognition is the acoustic feature of sound. Several studies have shown that several statistical

parameters have a high correlation between speech and the speaker's emotional state [4]. Those parameters are pitch, energy, articulation and spectral shape. For example, the emotion of sadness has a low standard deviation of pitch and a slow rate of speech, while the emotion of anger usually has a higher standard deviation of pitch and speaks quickly [5].

There are many studies on human emotion recognition. However, most of them are concentrated on one object of study either through facial expressions or emotion recognition through pronunciation (voice). Several studies have caught our attention as a recent review of this research, including the following. Research using intelligent systems and conventional methods [6]–[9], using available speech datasets as a standard for how well a system performs emotion recognition. Many emotion databases are provided by several institutions abroad such as SUSAS (Speech under Simulated and Actual Stress), Berlin database of emotional speech (Emo-DB), Surrey Audio-Visual Expressed Emotion (SAVEE) Database, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and many more. Previous research also provides information indicating that there are several parameters that show a strong relationship between speech and the emotions being felt. These parameters are pitch, wave, articulation and spectral shape. Speech Emotion Recognition (SER) in general, it starts by extracting features from a set of speech signals and labeling them based on the existing emotion classes. The combination of features and labels is then used to train a classification model which is generally an artificial neural network (ANN) architecture [10]. Some frequently used feature extraction methods include MFCC, chromogram, spectral contrast features, and Tonnetz [11]–[13]. Neural-based SER models typically utilize N-dimensional convolutional neural networks (CNNs) [14], recurrent neural networks (RNNs) [15], Long short-term memory net works (LSTMs) [16], or as a combination and variation of those techniques [17]. From the state-of-the-art review that has been presented in this study, we decided to take some of the advantages of previous studies, we decided to build an application that can recognize activities in videos to detect emotions from facial expressions and voices captured from video testimonials, using a computer vision and deep learning approach.

This study tries to correlate the movement of facial features and the spectrogram produced by the voice when giving testimonials with consumer satisfaction, and finally detect patterns in facial expressions and speech sound spectrograms that can predict the level of consumer satisfaction. Most customer satisfaction is emotional. This happens at the level of consciousness and the human subconscious. Through the experiences they go through while using the product, consumers experience positive and negative emotional reactions. Emotions in this study are described based on research conducted by Paul Ekman. Emotions are divided into 7 class, namely happy, surprised, neutral, fear, sad, angry, and disgusted. The classification results can be used to predict consumer satisfaction. Prediction results will be compared with ground truth from testimonials video.

II. Research Method

A. Face Emotion Recognition

Before performing the Face Emotion Recognition process, it is first necessary to obtain a region that shows the position of the face in the video clip (image). To obtain the position of the face in this study, it was done by utilizing the Haar Cascade for the face (face haar cascade) which has been trained on the OpenCV library [18].

The artificial neural network model for Face Emotion Recognition (FER) is obtained by training the model (Convolutional Neural Network - CNN). To do this, training data is needed in the form of facial images for the target class, which in this study are divided into 7 classes of facial emotions (Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprised). The dataset used is 43,596 images. This number is divided again into images to train the model and test the model, 28,702 images are used to train the model (training data) and 7,171 image data are used to carry out the testing process. Data obtained through the dataset provider site kagle.com. The dataset used is CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), added with images obtained from the google image search engine with image types .jpg, jpeg and .png for images with a maximum size of 1000 × 1000 pixels.The architecture for face classification in FER can be seen in Figure 1.

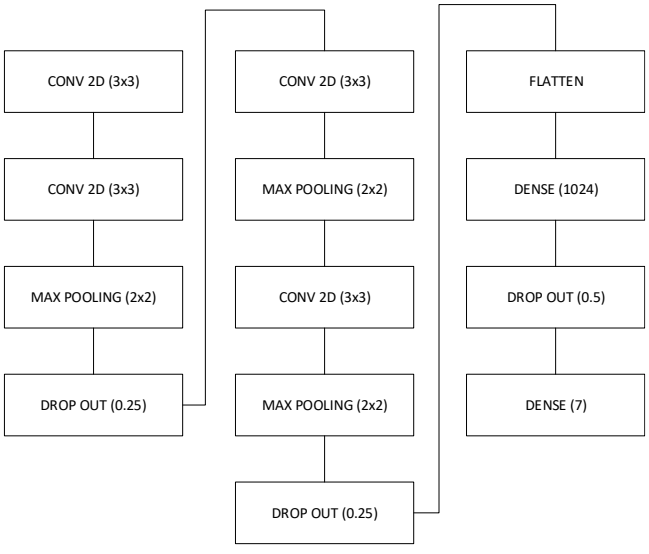

Fig. 1. FER CNN Architecture for Image Classification.

B. Speech Emotion Recognition

After completing FER, the next step in this research is to create Speech Emotion Recognition (SER). Voice data is also divided into 7 classes similar to FER, data obtained

through the dataset provider site kagle.com. The dataset used is CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), The collected voice data amounted to 7,436 voice clip data. Since the dataset collected for SER is in the form of sound clips and the approach that will be taken in this study is to use CNN (ResNet 50) to recognize sound, we must change the sound signal so that it can be input for the convolution process. One approach taken in this study is to change the sound signal into an image form (tensor with length, width and color depth). One way to change sound into a tensor is to change it into a spectrogram. The spectrogram used in this study is a spectrogram in the decibel domain (not frequency) Figure 2 shows some example of spectogram. In this study, fine tuning of Resnet-50 was carried out which had been trained using imagenet taken from the tensorflow keras library. The Resnet-50 finetune architecture can be seen in Figure 3.

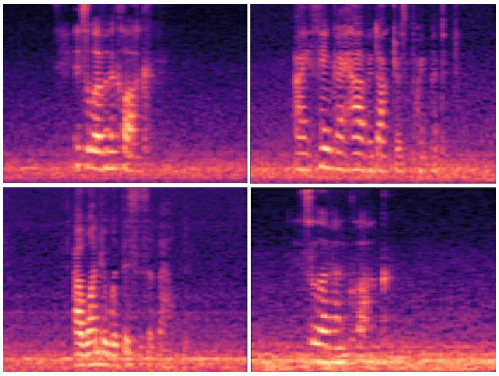

Fig. 2. Spectogram Example.

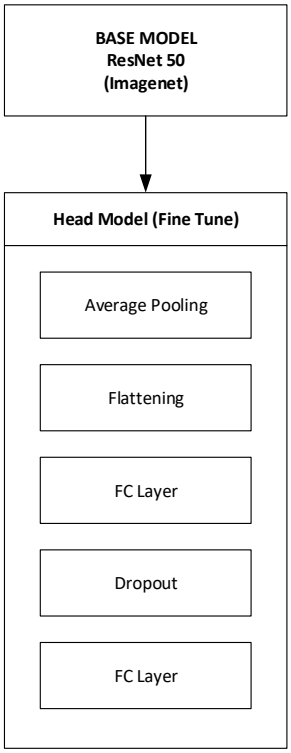

Fig. 3. SER Resnet Architecture for Image (Spectogram) Classification.

III. RESULT

FER is built by training an artificial neural network (CNN) using data on the collected dataset. The training process is carried out with the number of epochs of 100, a learning rate of 10e-5 using the Adam optimizer. The training accuracy and training loss plots in the training process can be seen in Figure 4. Figure 4 shows decreasing loss and increasing accuracy, as the training progresses, indicating that CNN can classify facial emotions well. The validation results also show that there is no overfitting from the training conducted. After getting the neural network model from the training process, the neural network model can be used to recognize faces found in the video (using Haar Cascade), after getting the face position/image (Haar Cascade Result), the face image is classified using CNN. Some of the recognition results can be seen in Figure 5.

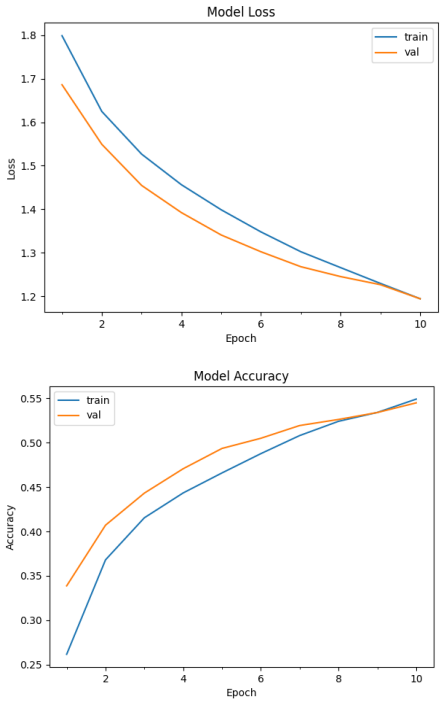

Fig. 4. FER Training history shows Model Loss and Model Accuracy during Training Epoch

Fig. 5. FER result example

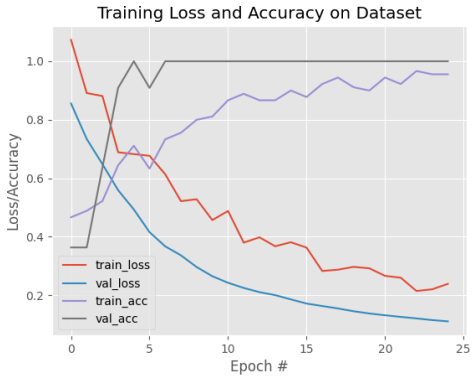

Fig. 6. SER Training history shows Model Loss and Model Accuracy during Training Epoch

In addition to training the model for FER, we also train the model for SER, because the model for SER is trained using spectrograms that are very similar between one class and another. We cannot use the same model as FER (CNN), for that in this study we use the ResNet-50 fine tune technique which has been trained using Imagenet. Figure 6 shows the results of the Resnet-50 model fine tune training. From the graph it can be seen that after training the model has achieved low loss and high accuracy. The validation data also shows that the model does not experience overfitting. After obtaining the artificial neural network model from the training process carried out, the artificial neural network model can be used to recognize speech found in videos. Figure 7 shows several FER recognition results that have been combined with SER results.

Fig. 7. FER (blue) merged with SER (green) Example Result

IV. Evaluation

After getting the results of FER and SER recognition in this study, it is necessary to analyze the results obtained by FER and SER. For FER, the accuracy search process is carried out by labeling the previously prepared datatest. The number of data sets for the test is 7,178 images divided into 7 emotion classes. After the trial process, the results are shown in Table 1. From Table 1 we see that the average overall FER accuracy is 73%, with the largest accuracy contribution in the sad class at 90% and the lowest accuracy in the fearful class with an accuracy value of 55%. For SER the same dataset was used and produced results as shown in Table 2. From Table 2, we see that the average overall SER accuracy is 77%, with the largest accuracy contribution in the fear class at 92% and the lowest accuracy in the disgusted class with an accuracy value of 58%.

TABLE I
FER Test Result

No	Class	No of Image	Correct	Wrong	Accuracy[a]
1.	Angry	958	630	328	66%
2.	Disgusted	111	82	29	74%
3.	Fearful	1024	555	469	55%
4.	Happy	1774	1280	494	72%
5.	Neutral	1233	914	319	75%
6.	Sad	1247	1114	133	90%
7.	Surprised	831	687	144	82%
	Total	7178	5262	1916	
	Average Accuracy				73%

TABLE II
SER Test Result

No	Class	No of Image	Correct	Wrong	Accuracy[a]
1.	Angry	958	690	268	72%
2.	Disgusted	111	64	47	58%
3.	Fearful	1024	943	81	92%
4.	Happy	1774	1419	355	80%
5.	Neutral	1233	1048	185	85%
6.	Sad	1247	907	340	72%
7.	Surprised	831	640	191	77%
	Total	7178	5262	1916	
	Average Accuracy				77%

V. Conclusion

This study attempts to build FER and SER to recognize emotions from images and audio clips in videos, FER is built using Haar Cascade to determine the position of the face and CNN for emotion classification on the face in SER, the process of converting sound in the video into a spectrogram and carrying out the classification process using the ResNet architecture model. The results of the study showed good performance in recognizing emotions from facial expressions and emotions that arise from voice/speech. This can be seen from the results of the data test which produced accuracy values that were close to quite good, 73% for FER (face) and 77% for SER (speech) respectively. In theory, this value indicates good classification performance.

Acknowledgment

The Faculty of Engineering at Udayana University and the Institute for Research and Community Service Udayana University funded this work with grant number B/78.888/UN14.4.A/PT.01.03/2022.

REFERENCES

[1] O. Appiah, "Rich media, poor media: The impact of audio/video vs. Text/picture testimonial ads on browsers'evaluations of commercial web sites and online products," *J. Curr. Issues Res. Advert.*, vol. 28, no. 1, pp. 73–86, 2006, doi: 10.1080/10641734.2006.10505192.

[2] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," *Procedia Comput. Sci.*, vol. 175, pp. 689–694, 2020, doi: 10.1016/j.procs.2020.07.101.

[3] P. A. Riyantoko, Sugiarto, and K. M. Hindrayani, "Facial Emotion Detection Using Haar-Cascade Classifier and Convolutional Neural Networks," *J. Phys. Conf. Ser.*, vol. 1844, no. 1, 2021, doi: 10.1088/1742-6596/1844/1/012004.

[4] B. Heuft, T. Portele, and M. Rauth, "EMOTIONS IN TIME DOMAIN SYNTHESIS," in *International Conference on Spoken Language*, 1996, pp. 1974–1977.

[5] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003, doi: 10.1016/S0167-6393(03)00099-2.

[6] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

[7] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, no. June 2019, pp. 56–76, 2020, doi: 10.1016/j.specom.2019.12.001.

[8] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, 2019, doi: 10.1016/j.bspc.2018.08.035.

[9] B. H. Prasetio, W. Kurniawan, and M. H. H. Ichsan, "Pengenalan Emosi Berdasarkan Suara Menggunakan Algoritma HMM," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, pp. 168–172, 2017, doi: 10.25126/jtiik.201743339.

[10] A. Muppidi and M. Radfar, "Speech emotion recognition using quaternion convolutional neural networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, pp. 6309–6313, 2021, doi: 10.1109/ICASSP39728.2021.9414248.

[11] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," *Proc. - 9th Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2017*, vol. 2018-Febru, no. December, pp. 1613–1616, 2018, doi: 10.1109/APSIPA.2017.8282282.

[12] and J. P. . B. Eric J . Humphrey , Tae Min Cho, "LEARNING A ROBUST TONNETZ-SPACE TRANSFORM FOR AUTOMATIC CHORD RECOGNITION New York University , New York USA," *Ieeexplore.Ieee.Org*, pp. 1–4, 2012, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6287914/.

[13] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, no. D, pp. 3642–3646, 2017, doi: 10.21437/Interspeech.2017-1428.

[14] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Convolutional Neural Networks-based continuous speech recognition using raw speech signal," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015-Augus, pp. 4295–4299, 2015, doi: 10.1109/ICASSP.2015.7178781.

[15] F. M. Yajie Miao, Mohammad Gowayyed and Language, "EESEN : END-TO-END SPEECH RECOGNITION USING DEEP RNN MODELS AND WFST-BASED DECODING Yajie Miao , Mohammad Gowayyed , Florian Metze," *Proc. ASRU*, pp. 167–174, 2015.

[16] A. Graves, N. Jaitly, and A. Mohamed, "HYBRID SPEECH RECOGNITION WITH DEEP BIDIRECTIONAL LSTM Alex Graves , Navdeep Jaitly and Abdel-rahman Mohamed University of Toronto Department of Computer Science 6 King ' s College Rd . Toronto , M5S 3G4 , Canada," pp. 273–278, 2013.

[17] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 5, pp. 867–881, 2010, doi: 10.1109/JSTSP.2010.2057200.

[18] G. Bradski, "The OpenCV Library," *Dr. Dobb's J. Softw. Tools*, 2000.