

Chapter 2

Driver Emotion Profiling from Speech

Norhaslinda Kamaruddin, Abdul Wahab, and Hüseyin Abut

Abstract Humans sense, perceive, and convey emotion differently from each other due to physical, psychological, environmental, cultural, and language differences. For example, as recognized and studied by psychologists more than a century, it is easier for someone of the same culture to judge and recognize emotion correctly compared to those from different culture. In this chapter, we attempt to study the speech emotion recognition problem by using two speech corpora from the Berlin dataset and the NAW datasets. We have investigated the universality as well as diversity of two different cultural speech datasets recorded by German and American speakers, respectively. Experiments were conducted for identifying three basic emotions, namely, angry, sad, and happy with neutral as emotionless state from these datasets. MFCC coefficients were used as feature sets in the experiments, and MLP was employed as classifiers to compare the performance of these datasets. In addition, real-time recorded speech from drivers was also tested to see the performance in a vehicular setting. Finally, speech emotion profiling approach was introduced to explore the universality and diversity of the speech emotion features.

N. Kamaruddin (✉)

Faculty of Computer Science and Mathematics, University Technology MARA (UiTM),
Shah Alam, Selangor, Malaysia
e-mail: norhaslinda@fskm.uitm.edu.my

A. Wahab

Faculty of Information and Communication Technology, International Islamic
University (IIUM), Kuala Lumpur, Selangor, Malaysia
e-mail: abdulwahab@pmail.ntu.edu.sg

H. Abut

ECE Department (Emeritus), San Diego State University, San Diego, CA, USA

EE Engineering Department, Boğaziçi University, Istanbul, Turkey
e-mail: abut@anadolu.sdsu.edu

Keywords Berlin dataset • Mel frequency cepstral coefficients (MFCC) • Multilayer perceptron (MLP) • NAW dataset • Speech emotion profiling • Speech emotion recognition

2.1 Introduction

During the last century, many researchers from different disciplines have tried to postulate a few basic emotions out of the entire range of emotions that are tinged and enlivened. One of the models suggests that every emotion is composed of different levels of certain basic components including arousal, intensity, aversion, self-directedness, and others. Among many models, the prevailing one conjectures that emotions arise much the same way as colors do – presenting a myriad of hues out of the basic few constituents [7]. To date, cognitive science does not possess a test to decide between various competing models of the basic emotion. However, researchers in various disciplines agree that some emotions are universally accepted as basic and many others as secondary. Cornelius has labeled six emotions as the “Big Six” [11], which are *angry*, *happy*, *sad*, *fear*, *surprise*, and *disgust*. These were chosen in this study. However, we focus only on *angry*, *sad*, and *happy* emotions with *neutral* as emotionless state in this chapter.

Emotion recognition from engineering perspective is a fairly new field of research compared to the psychologists’ community. With the understanding that human convey and perceive underlying emotion in the interaction, scientists and researchers are able to analyze massive amount of information transmitted from a speaker to the listener using the tools of signal processing today. Yet, we are struggling to understand emotion and, more critically, capture and/or process it in a form that is useful for technical purposes.

In 2001, Sherer et al. have conducted a study in nine different countries in Europe, United States, and Asia on vocal emotion portrayals using content-free sentences containing anger, sadness, fear, joy, and neutral voice [9]. They found that generally the accuracy decreased with increasing language dissimilarity in spite of the use of language-free speech samples. It is concluded that culture-and language-specific paralinguistic patterns may influence the emotion recognition process.

In this chapter, we address this issue by proposing Mel Frequency Cepstral Coefficients (MFCC) as our features for speech emotion recognition. Our feature extraction method based on Slaney’s [8] approach coupled with the WEKA multilayer perceptron (MLP) [12] classifier. These are adopted to identify the three basic emotions, namely, *angry*, *sad*, and *happy* emotional states. Initially, two different speech emotion datasets – using the NAW dataset (American actors) and Berlin dataset (German actors) – were employed to train and test the accuracy of the proposed system based on the K-fold validation technique. Next, we have extended our scope by using speech data recorded while driving in real time, to analyze and understand the driver behavior [6]. The driver was asked to interact with the passenger as well as

talking to a caller using a mobile phone equipped with a hands-free module as a safety precaution while driving. Three different scenarios were recorded based on:

- Driver under stress when talking on the mobile phone while driving
- Laughing while driving
- Driver feeling very sleepy

Data from these three driving scenarios were then compared with the two standard datasets, i.e., NAW and Berlin datasets.

In addition to the speech emotion recognition system under study, we also explore speech emotion profiling as an alternative tool to better understand speech emotion and in analyzing inter-and intra-cultural behavior. Such tool seems to provide a deeper insight to the hidden characteristics of speech emotion.

This paper is organized as follows: In Sect. 2, we present the theoretical and experimental framework for the proposed speech emotion recognition system based on a feature extraction method using Mel Frequency Cepstral Coefficients (MFCC) as features and MLP as classifier. In Sect. 3, experiments for the proposed speech emotion profiling system will be presented with some analysis of the driving dataset as compared to those using the NAW and Berlin datasets. Section 4 discusses the findings and conclusion of the study with some future work that can help extend the idea of profiling to its next level.

2.2 MFCC-MLP Speech Emotion

During the past couple of decades, MFCC feature has been successfully used for high-end speech recognition and speaker identification problems. However, there are many variations in applications in terms of number of filters, shape of filters, bandwidths, and the manner in which the spectrum is warped. In classification experiments, Slaney's approach [8] – founded on a study by Ganchev et al. [3] – gives a slightly better performance than many earlier works. Hence, we have adopted the approach described in Slaney's Auditory Toolbox for Matlab [8] in this study.

Once the MFCC features from the speech are extracted, the speech emotion is then classified/recognized using a multilayer perceptron (MLP) technique which is based on Bishop's work [1] wherein the preliminary experiments sought to determine the initial accuracy of the speech emotion recognition system. MLP uses a combination of several perceptron layers that are interconnected to each other and exhibit a high degree of connectivity, which is determined by the synapses of the network. It consists of three main layers which are the input layer, the hidden layer, and the output layer. In the input layer, the data is given to the network, thus the number of input neurons must be equivalent to the number of features for the data. Each data entry is given a weight by the network to be passed to the hidden layer where a nonlinear calculation will be carried out with the activation function. The output layer is the sum of the entire hidden layer outcomes. MLP uses the ubiquitous back-propagation algorithm as its learning procedure.

Fig. 2.1 Shows the proposed speech emotion recognition system

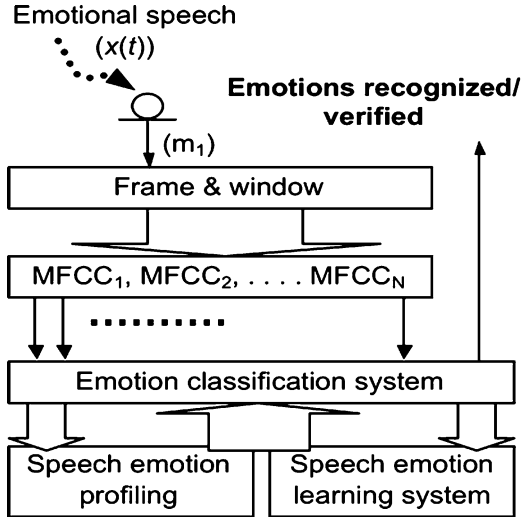


Figure 2.1 shows the proposed speech emotion recognition system where the emotional speech is first filtered and framed. Forty MFCC features were then extracted and later classified using the MLP. The other additional module on speech emotion profiling and learning system is meant to enhance the speech emotion recognition system to cater for inter-and intra-cultural differences. In this chapter, only the emotion classification and some preliminary work on emotion profiling are presented next.

2.2.1 Berlin Dataset

The Berlin Emotional Speech Database [2] contains ten sentences that have little emotional content textually. It is in German and covers seven emotion classes, namely, anger, fear, happy, sad, disgust, boredom, and neutral. The content of the spoken material is predefined and presented to five male and five female professional actors, respectively. The recording was done under studio conditions using a high-quality recording equipment and saved in mono wave format with an 8.0 kHz sampling rate. The complete database has been pre-evaluated through a manual perception test by 20 human subjects.

2.2.2 NAW Dataset

The NAW dataset [13] was collated using some video clips from movies and television sitcoms obtained from the Internet. The participants are native speaker of American English. The emotions portrayed by the speakers have been analyzed and

Table 2.1 Confusion matrix for human recognition performance for NAW dataset

	Happy	Angry	Disgust	Surprised	Sad	Neutral
Happy	76.5	0.0	1.5	12.0	0.0	10.0
Angry	0.0	90.0	5.0	0.0	4.0	1.0
Disgust	2.0	32.5	34.5	6.5	3.0	21.5
Surprised	9.0	2.0	8.0	64.5	1.5	15.0
Sad	0.0	0.0	0.5	0.0	98.0	1.5
Neutral	1.0	0.0	2.5	0.0	0.0	96.5

identified based on speech semantics, facial expression of the speaker, as well as basic understanding of the situations of the video clips occurrences. These video clips were converted to MP3 audio files at a sampling rate of 8.0 kHz, mono stream, and their amplitudes were scaled in the range $(-1, +1)$ V. A number of findings using this dataset have been reported earlier in [4, 5, 13].

2.2.3 Human Perception Test

In order to ensure that the video clips obtained for the NAW dataset were correctly perceived, manual perception test were subsequently carried out. In this test, a total of 40 human subjects – 11 from Nanyang Technological University, Singapore (nine males and two females), and 29 from International Islamic University, Malaysia (15 male and 14 female), with an age mean of 23 years – have volunteered to provide their perceived assessment of the speech emotion audio files presented.

The participating subjects have reported that they have experienced neutral emotion prior to the commencement of the human perception test. The survey was conducted in a laboratory environment where the judges can listen to the speech emotion audio files with minimal distraction. They sat in front of a computer and listened to the speech emotion audio files via a headphone to ensure that judges can hear audio files without interruption. For each speech emotion audio file, they indicated the perceived emotion on a six-force-choice format representing the emotion classes with neutral as shown in Table 2.1.

In order to avoid any misled perception, each speech emotion audio file's name was labeled using a file number that has no relation to the respective emotion. In addition, the file numbering was also randomized to avoid any prediction of the emotion pattern. The human judges were allowed to listen to any of the speech emotion audio files repeatedly prior to making an appropriate decision.

Table 2.1 shows the confusion matrix for human recognition performance for the NAW dataset. Here, it can be seen that most judges were able to identify *sad*, *angry*, *neutral*, and *happy* quite easily with at least 76% accuracy. This is followed by *surprised* with 64% accuracy and *disgust* with only 34% accuracy, respectively. Disgust yielded very low recognition, which shows that the judges were not clear with its definition that they might have perceived disgust as mild anger thus resulting in higher percentage of anger being perceived. Similarly, surprised

emotion also scored fairly low perception performance due to the judges' mixed perception that most of them categorized surprised as happy for positive surprised or disgust for negative surprised. Sad is the highest correctly perceived emotion with 98% recognition accuracy performance since it was observed from features that it has the most acoustically distinct features.

2.3 Speech Emotion Recognition and Profiling Experiments

2.3.1 Emotion Identification Experiments

Identification experiments were carried out to investigate the performance of our proposed system in determining the emotion for a given speech segment. As shown in Fig. 2.2, our proposed system can yield accuracy ranging from 47.9% to 75.4% for Berlin dataset and 61.4–71.2% for NAW dataset, respectively.

As it can be seen from Fig. 2.2, the maximum and minimum accuracy percentages for both datasets are consistent wherein *sad* emotion resulted in the highest accuracy and the *happy* emotion the lowest accuracy. Based on these results, we can see that NAW dataset result is comparable to the Berlin dataset, and the combination of MFCC as feature extraction coupled with MLP can achieve reasonable accuracy performance. This indicates that our proposed approach has potential to recognize emotion in speech.

2.3.2 Understanding Driver's Emotion

The approach was next applied to a pre-recorded driving data to identify emotional state of drivers while driving under varying scenarios. There were four scenarios for the driver emotional state, namely, *stress*, *laughing*, *neutral*, and *sleepy*, which were tested in these set of experiments. Stress data is taken while the driver was talking through a mobile phone while driving with the assumption that he/she

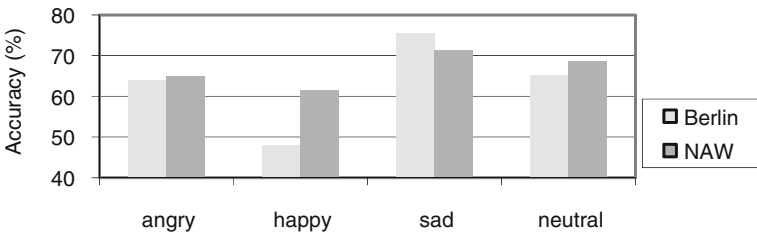


Fig. 2.2 Identification result for Berlin and NAW dataset

Table 2.2 Confusion matrix for identification results for real-time driving dataset

	Stress (%)	Laugh (%)	Neutral (%)	Sleepy (%)
Stress	55.6	19.9	13.3	11.2
Laugh	22.3	57.0	12.1	8.6
Neutral	5.4	6.9	74.5	13.2
Sleepy	8.5	7.4	17.6	66.5

Table 2.3 Confusion matrix for identification results from combination of Berlin, NAW, and real-time driving datasets

	Happy (%)	Sad (%)	Neutral (%)	Stress (%)	Angry (%)	Sleepy (%)
Happy	50.6	11.5	14.9	4.9	15.4	2.8
Sad	12.7	59.5	15.4	1.0	9.4	2.1
Neutral	13.5	10.5	62.3	1.8	6.5	5.5
Stress	22.0	6.1	22.4	39.3	2.5	7.7
Angry	22.4	9.2	9.2	1.0	56.1	2.2
Sleepy	12.9	4.8	21.4	2.8	1.2	56.9

Table 2.4 Confusion matrix for identification result of combination Berlin, NAW, and real-time driving dataset without neutral

	Happy (%)	Sad (%)	Stress (%)	Angry (%)	Sleepy (%)
Happy	59.4	13.5	5.8	18.1	3.2
Sad	15.0	70.3	1.2	11.1	2.5
Stress	28.4	7.9	50.6	3.2	9.9
Angry	24.6	10.1	1.0	61.8	2.4
Sleepy	16.4	6.1	3.6	1.5	72.4

needed to multitask between concentrating on his/her driving and at the same time providing appropriate responses to the caller prompts. The results are tabulated in Table 2.2.

From Table 2.1, we could see that the proposed system can identify at least 55.6% and can reach up to 74.5% of the driver emotional state using the same driving dataset. Neutral yielded the highest accuracy, and stress obtained the lowest accuracy.

In order to have better understanding of the proposed system performance, we have combined the three datasets consisting of Berlin, NAW, and the driving datasets, and have conducted identification experiments. Since *laughing* is a reaction when the driver is *happy*, we assumed that laughing is a subset of happy emotion. The identification results are provided in Table 2.3. It is clearly seen that the lowest accuracy yielded for stress data with only 39.3% accuracy while the maximum accuracy obtained by neutral with 62.3%.

The accuracy of such system can be improved if the neutral state is removed from the dataset. From the understanding of Schlosberg's affection space model [10], the neutral state is the speech emotion basis regardless of their emotion primitives' axes. Thus, pure emotion can be extracted from the processed speech if we can remove neutral from our findings. Confusion matrix of Table 2.4 shows a rather interesting result when the neutral is removed. The accuracy of the proposed system is increased by approximately 10%.

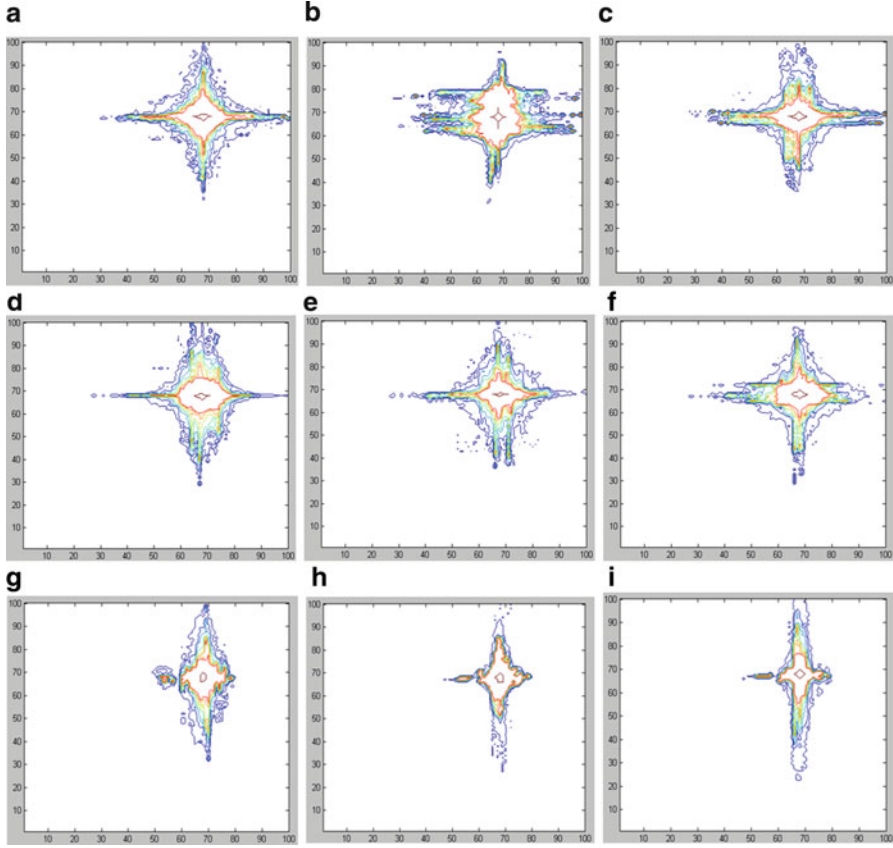


Fig. 2.3 Speech emotion profiling for Berlin, NAW, and driving datasets: (a) Berlin – angry; (b) Berlin – sad; (c) Berlin – happy; (d) NAW – angry; (e) NAW – sad; (f) NAW – happy; (g) Driving – stress; (h) Driving – sleepy; (i) Driving – laugh

2.3.3 Speech Emotion Profiling

Based on the results presented in Sect. 2, we apply speech emotion profiling method to the data in order to visualize the correlation between speech emotion signal and the neutral state. It is interesting to note from Fig. 2.3a–i that even though the data used is different, the pattern is similar for the same emotion across datasets, and yet the distinction is clearly observable for different emotion within a given dataset.

The most obvious example is the profile plot of *happy* emotion which has a cross pattern for all three dataset, although the data is completely different. Figure 2.3 also indicates that it is possible for us to visualize the inter- and intra-cultural variations of the speech emotion, which can lead us to better understand the effect of these cultural artifacts to improve speech emotion recognition globally.

2.4 Summary, Conclusion, and Future Work

Speech emotion profiling can be an effective tool for investigating intra- and inter-cultural variations from various perspectives. It enables one to visualize the interaction of the emotion that may give important information which is not observable using normal signal analysis tools, namely, the speech recognition and speaker identification. More work in understanding the profile especially in extracting the relevant features as well as the appropriate data processing are needed to benefit from such visualization tool. The speech emotion profile coupled with a three-dimensional affective-space model may be able to provide a better understanding of the dynamics of driver behavior. This work also illustrated that there are strong correlations between driver behavior and emotion which can be empirically measured using speech signals.

References

1. Bishop C (1997) Neural networks for pattern recognition. Clarendon, Oxford
2. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Proceedings of INTERSPEECH-ISCA, Lisbon, pp 1517–1520, 2005
3. Ganchev T, Fakotakis N, Kokkinakis G Comparative evaluation of various MFCC implementations on the speaker verification task. In: Proceedings of the 10th international conference on speech and computer (SPECOM 2005). Patras, vol 1, pp 191–194, 2005
4. Kamaruddin N, Wahab A (2008) Feature extraction for speech emotion. In: Proceedings of the 17th international conference on software engineering and data engineering (SEDE '08), Los Angeles, pp 120–125
5. Kamaruddin N, Wahab A (2008) Speech emotion verification system (SEVS) based on MFCC for real time applications. In: Proceedings of the 4th international conference on intelligent environments (IE '08), Seattle, pp 1–7
6. Khalid M, Wahab A, Kamaruddin N (2008) Real time driving data collection and driver verification using CMAC-MFCC. In: Proceedings of the 2008 international conference on artificial intelligence (ICAI '08), Las Vegas, pp 219–224
7. Plutchik R (2003) Emotions and life: perspective from psychology, biology and evolution, 1st edn. American Psychological Association, Washington, DC
8. Slaney M (1998) Auditory toolbox: Version 2. Technical Report #1998-010, Interval Research Corporation
9. Scherer KR, Banse R, Wallbott HG (2001) Emotion inferences from vocal expression correlate across languages and cultures. *J Cross-Cultural Psychol* 32(1):76–92
10. Schlosberg H (1954) Three dimensions of emotion. *Psychol Rev* 61:81–88
11. Cornelius, R. R. (1996). *The Science of Emotion: Research and Tradition in the Psychology of Emotion*, Upper Saddle River, NJ: Prentice-Hall
12. Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S. J. (1999) Weka: Practical Machine Learning Tools and Techniques with Java Implementations. In: N. Kasabov & K. Ko (Eds.). *Proceedings of the ICONIP/ANZIIS/ANNES'99 International Workshop on Emerging Knowledge in Engineering and Connectionist-Based Information Systems*. Dunedin, New Zealand, 192–196
13. Kamaruddin N. & Wahab A. (2009). Features Extraction for Speech Emotion. *Journal of Computational Methods in Science and Engineering (JCMSE)*, 9 (Supplement 1), S1–S12, 2009

Digital Signal Processing for In-Vehicle Systems and
Safety

Hansen, J.H.L.; Boyraz, P.; Takeda, K.; Abut, H. (Eds.)

2012, XXII, 326 p., Hardcover

ISBN: 978-1-4419-9606-0