DOI: 10.61782/fa.2025.0803

FORUM ACUSTICUM
EURONOISE 2025

SOUNDSCAPE SYSTEM BASED ON PASSENGER EMOTION AND CONVERSATION PERCEPTION OF AUTONOMOUS VEHICLES

Kyoung-Jin Chang[1*] **Jeonggoo Kang**[1] **Dong Chul Park**[1]
Gyumin Cho[2] **Chang Wook Ahn**[2]
[1] Hyundai Motor Company, Hyundai-ro 150, Namyang-eup, Hwaseong-si, Gyeonggi-do, 18280, Korea
[2] Graduate School of AI, Gwangju Institute of Science and Technology, Korea

ABSTRACT

When passengers of an autonomous vehicle have conversations or engage in various activities, the soundscape of the vehicle that suits the atmosphere can play an important role. This study introduces the results of developing a system that recognizes emotions and conversation states from passengers' face images in autonomous vehicles, selects music or the sounds of nature that match this atmosphere, and automatically adjusts the volume to not interfere with passengers' conversation. This study consists of the following process. First, algorithms for emotion classification and conversational state recognition using facial expressions are developed, and the program is lightened and implemented in the vehicle's deep learning control system. Second, a soundscape system is developed that matches and plays sound sources suitable for emotions and conversational states estimated in real time in vehicles. Third, the control and soundscape system are installed in an autonomous vehicle and the accuracy and satisfaction of the systems are evaluated by users. This technology is expected to provide new auditory experiences and increase satisfaction, especially for passengers of autonomous vehicles such as robotaxi.

Keywords: *artificial intelligence, soundscape, emotion recognition*

1. INTRODUCTION

In recent years, automobiles are expanding from simple means of transportation to spaces that improve the quality of life, and at the same time, the technology to provide new experiences in accordance with the emotions of passengers in the vehicle is becoming more important. Particularly in autonomous vehicles, playing music that best suits the mood and emotions without disturbing the conversation of several passengers is expected to give high satisfaction.

Research on recognizing emotions from passengers' facial expressions using cameras in vehicles is an important technology for personalization, and many studies have been conducted so far [1-3]. For example, W. Li et al. [1] had a camera installed in a vehicle simulator and analyzed the driver's emotions into five categories: anger, happiness, sadness, fear, and surprise. Z. Chen et al. [2] developed a face recognition and emotion classification system for autonomous vehicles with a low latency of 2ms and high accuracy (F-score 96%) to enable real-time implementation in Internet of Things (IoT) systems. K.-J. Chang et al. [3] developed a new soundscape provision system that classifies emotions and provides customized driving sound using a multimodal analysis method that combines facial expressions, heart rate, and skin conductivity of vehicle drivers. These studies show that there is a growing interest in in-cabin camera utilization and emotion recognition technology to provide personalized entertainment services. However, when implemented in a vehicle controller, it is very necessary to design an efficient algorithm that can achieve high performance at a low capacity so as not to interfere with other control performance of a vehicle.

On the other hand, in terms of technology for recognizing the conversation state, only methods for measuring and analyzing users' voices have been mainly studied [4-5].

[*] ***Corresponding author***: *changkj@hyundai.com.*

In contrast, this study introduces a technology that uses only an in-cabin camera to identify the passenger's emotions and conversation state from the passenger's facial image information and provides the most appropriate soundscape in the vehicle. In addition, an algorithm that operates with minimal capacity and calculation is produced and implemented in the AI controller of an autonomous vehicle for robotaxi.

2. AI ALGORITHMS

2.1 Emotion recognition

First, the face recognition model of this study uses Multi-task Cascaded Convolutional Neural Networks (MTCNNs) [6]. This is a cascade structure that sequentially passes through three lightweight convolutional neural networks (CNNs) called proposal network (P-Net), refinement network (R-Net), and output network (O-Net), so it is fast to process and accurate even in low-quality images. Fig. 1 shows the architecture of MTCNN. The face recognition model of this study has been designed with a small capacity of 600 KB and has achieved a high recognition rate of 98.02% for 24,111 images on the FER-2013 dataset. Next, the emotion classification model uses the Mini-Xception algorithm [7], which is a lightweight and efficient convolutional neural network (CNN) architecture. This is known for its depthwise separable convolutions that reduce computational cost while maintaining performance. In this study, the emotion classification model has been designed with a small capacity of 853 KB and has showed 74% accuracy for 10,767 images on the FER-2013 datasets. Table 1 shows the results of testing the accuracy of the emotion classification model used in this study. For use in vehicles, confidence for emotions such as Happy, Sad, and Neutral is extracted for each basic frame according to the passenger's facial expression, and the emotion with the greatest value is classified as a cumulative emotion by accumulating the confidence for 30 seconds. Here, 30 seconds is a cycle in which emotions are updated to change the soundscape, and as the emotion update time approaches, confidence is given a high weight so that the recent emotional state can be reflected more. When several passengers board, the above analysis is performed for each passenger, and the most common emotion is selected as the representative emotion.

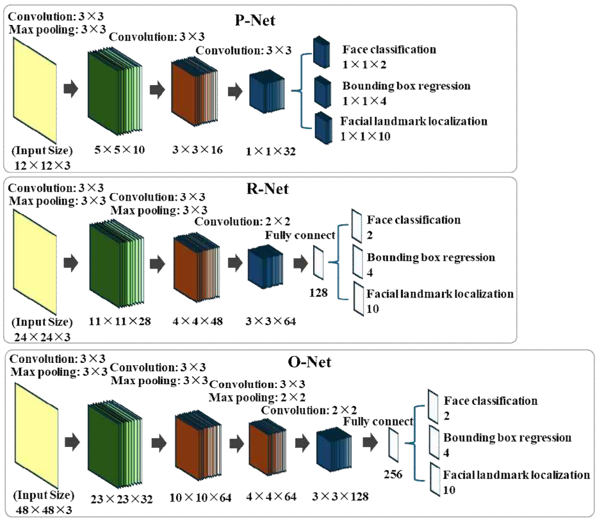

Figure 1. Architecture of MTCNN.

Table 1. Emotion classification test results using the Mini-Xception.

Predicted Actual	Neutral	Happy	Sad
Neutral	5198	553	224
Happy	857	2257	75
Sad	912	209	482

2.2 Conversation state recognition

When several passengers talk to each other, it is important to select music suitable for the atmosphere and provide it at an appropriate volume so as not to interfere with the conversation. In this study, the conversational state of passengers is determined through CNN algorithm and calculation process from the passenger's facial image data, not the passenger's voice. This method has the advantage that it is not necessary to add a microphone and is not affected by various vehicle noises because it can be performed with an in-cabin camera. A model with 68 face landmarks from the passenger's face image is constructed, and the passenger's conversation state is determined by calculating the number of openings and the opening distance of the landmark around the mouth for a certain time. The above process is repeated for all passengers' mouths, and the conversational state is divided into three categories: No talk, Weak talk, and Strong talk from the calculation results. Fig. 2 shows an example of landmarks around the mouth estimated by facial images.

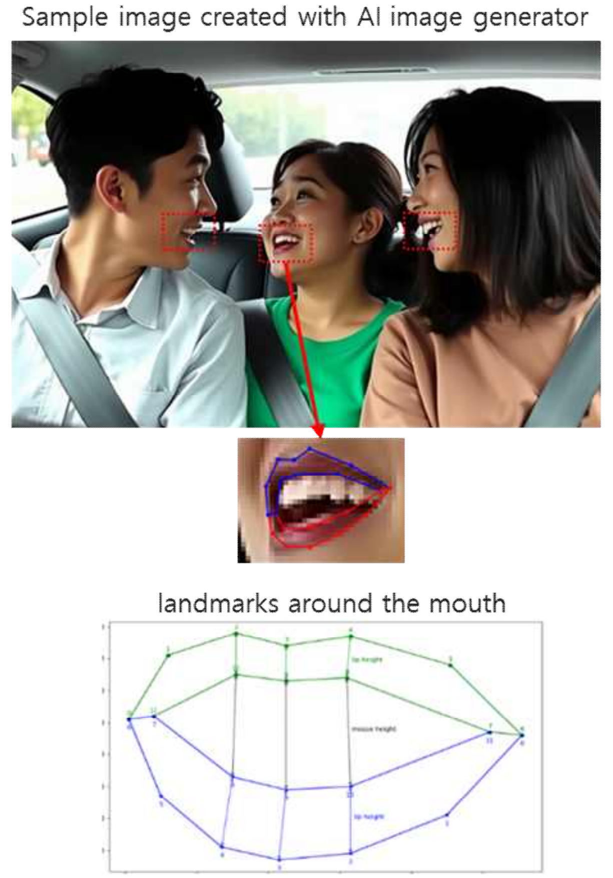
Sample image created with AI image generator
landmarks around the mouth
lip height
mouse height
lip height

Figure 2. Examples of landmarks around the mouth for conversation state recognition.

3. SOUNDSCAPE

3.1 Soundscape library

Sound sources collected in the library of four concepts (i.e., Neutral, Happy, Sad, Talkative) are used to provide soundscape suitable for passengers' emotions and conversational states. These sound sources have been produced using an AI composition program and improved through correction by music experts. The Happy library contains bright and fast beat music, and the Sad library contains blue and melancholic music. The Neutral library contains comfortable and calm music, and the Talkative library contains simple music with the beats removed so as not to disturb the conversation.

3.2 Soundscape control logic

The Soundscape library selects the sound source that best matches the results of the passengers' representative emotions and conversational states analyzed in real-time in the vehicle. If multiple passengers show various emotions, four complex emotions can be selected as follows: Neutral + Happy (NH), Neutral + Sad (NS), Happy + Sad (HS), and Neutral + Happy + Sad (NHS). In the case of complex emotions, music that satisfies both or more emotions is matched, and if there is a dominant emotion (e.g., cumulative confidence of 80% or more), music that considers that emotion first is selected. Fig. 3 shows the distribution of sound sources matched according to the four complex emotions in this study. On the other hand, when the conversation of passengers is detected, music suitable for the conversation state is played. In addition, the automatic volume logic is applied to reduce the volume of music if it corresponds to Strong talk and increase the volume of music again if it corresponds to Weak talk by determining the stage of conversation with the method introduced in Section 2.2.

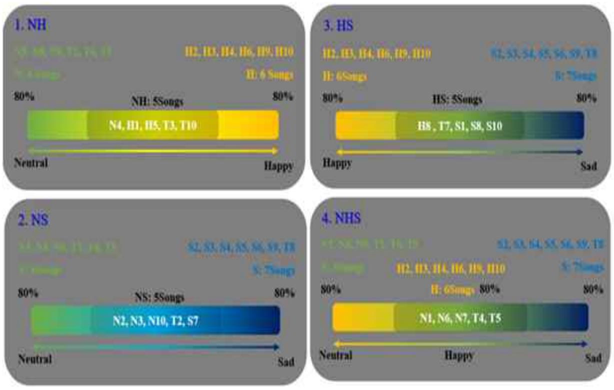

Figure 3. Distribution of sound sources according to the four complex emotions.

3.3 Jury test

Jury evaluation has been performed on 20 people with the camera, AI controller, and soundscape provision system developed in this study. The average age of the subjects is 26.3 years old, and they are organized to have various emotions while looking at various images in the listening room. At the same time, music suitable for the emotions recognized from the subjects' facial expressions is played, and the accuracy of emotion recognition, the accuracy of matching music and emotion, and auditory satisfaction are evaluated. The results of evaluating each item with a 10-

point Likert scale are shown in Fig. 4, and the average value is 7 points or more, showing relatively good results.

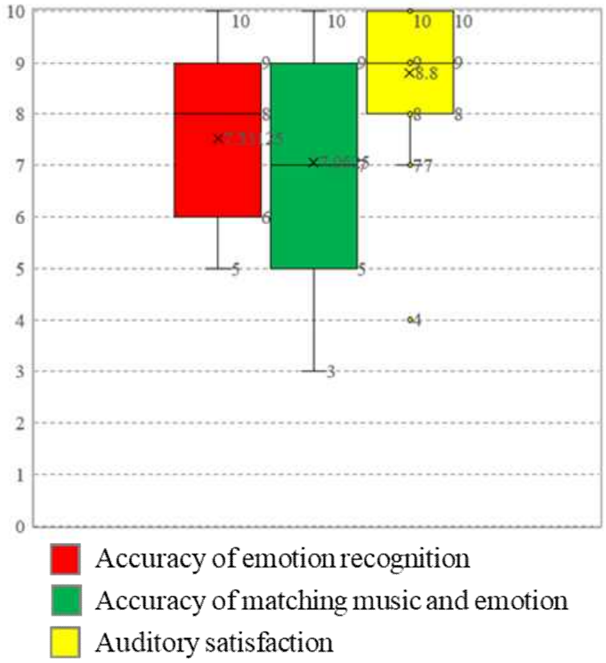

Figure 4. Jury test results for AI algorithms and soundscape.

4. IMPLEMEMTATION IN AUTONOMOUS VEHICLES

4.1 AI Controller

The AI model of this study has been programmed with TensorFlow and embedded in an Ambarella EVK board capable of AI image processing to operate in an autonomous vehicle. Fig. 5 shows the Ambarella EVK board used in this study, which is equipped with CV25, a system-on-chip that supports EasyAI libraries and provides a vector processor with instruction sets that handles a lot of data. The camera installed in the front center of the vehicle interior to capture images of passengers is the ON Semiconductor's AR0239, which is a 1/2.7-inch CMOS digital image sensor with an active-pixel array of 1936 × 1188.

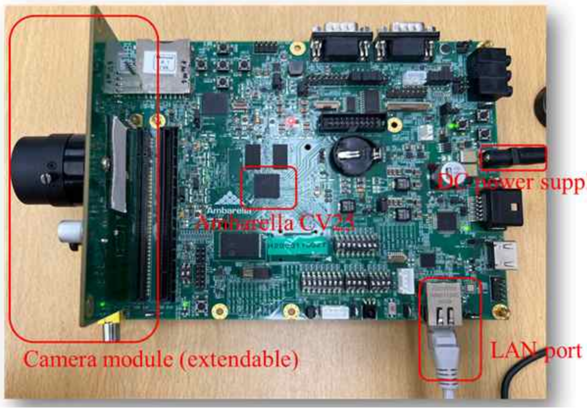

Figure 5. Ambarella EVK board.

4.2 User Interface system

An application for Microsoft Surface Pro tablets has been created to allow passengers to check the emotional and conversational status analyzed in real time, watch videos that fit well with the music determined by the proposed algorithm, and manually select an interface menu. Fig. 6 shows the screen image of this application, including the function of automatically adjusting the volume according to the conversation state. The determined music is played by the vehicle audio amplifier and speaker through the added external sound card. Fig. 7 shows a camera and tablet mounted inside the vehicle, an Ambarella controller mounted in the trunk room, and a sound card.

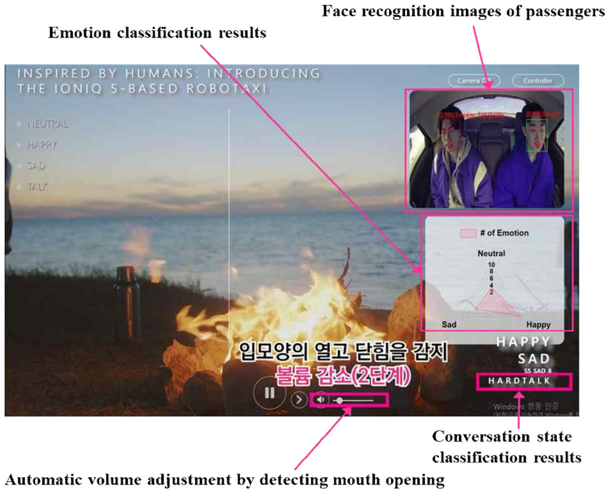
Face recognition images of passengers
Emotion classification results
INSPIRED BY HUMANS: INTRODUCING THE IONIQ 5-BASED ROBOTAXI
Camera Off
Controller
NEUTRAL
HAPPY
SAD
TALK
# of Emotion
Neutral
10
8
6
4
2
Sad
Happy
HAPPY
SAD
HARDTALK
입모양의 열고 닫힘을 감지
볼륨 감소(2단계)
Conversation state classification results
Automatic volume adjustment by detecting mouth opening

Figure 6. Screen image of an app running on a tablet.

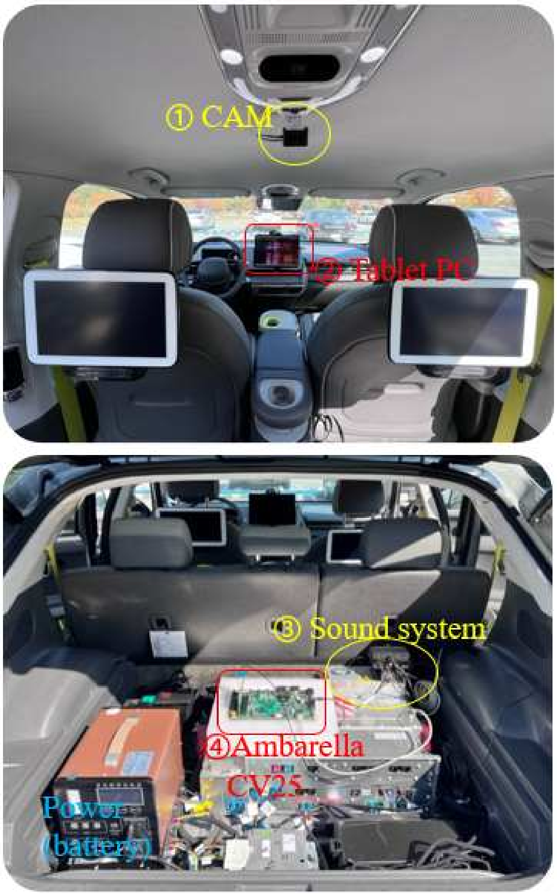

Figure 7. AI controller and camera mounted on an autonomous vehicle.

4.3 Driving Demonstration

The control and soundscape system of this study has been mounted on an autonomous test vehicle and demonstrated while driving with several passengers on board. In the demonstration and evaluation involving a total of 16 people, the overall satisfaction score is 8.4 out of 10. Compared to Section 3.3 evaluated in the listening room, the vehicle evaluation result shows higher satisfaction. Many users are found to be satisfied with providing a soundscape suitable for customers' emotions and moods in the vehicle.

5. CONCLUSIONS

In this study, a lightweight emotion/conversation recognition and soundscape matching algorithm using face images that can be embedded in a deep learning controller of an autonomous vehicle has been proposed and implemented in a test vehicle. As a result of user evaluation, it has been confirmed that the accuracy of emotion recognition, the accuracy of matching music and emotions, auditory satisfaction, and user satisfaction in vehicles are all better than 70%. The results of this study can be used as important guidelines for designing autonomous vehicle services. That is, it is possible to provide a more personalized and immersive auditory experience to passengers through a soundscape suitable for the recognized emotion and conversation state. In addition, if the results of this study are used for self-driving taxis, it is expected to improve the service quality of self-driving taxis and secure a competitive advantage over other transportation methods by providing a personalized sound environment experience in the vehicle.

6. REFERENCES

[1] W. Li, Y. Cui, and et al: “A Spontaneous Driver Emotion Facial Expression Dataset for Intelligent Vehicles: Emotions Triggered by Video-Audio Clips in Driving Scenarios,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 747-760, 2023.

[2] Z. Chen, X. Feng, and S. Zhang: “Emotion detection and face recognition of drivers in autonomous vehicles in IoT platform,” *Image and Vision Computing*, vol. 128, no. 104569, pp. 1-11, 2022.

[3] K.-J. Chang, G. Cho, and et al: “Personalized EV Driving Sound Design Based on the Driver's Total Emotion Recognition,” *SAE Int. Journal of Advances & Current Practices in Mobility*, vol. 5, no. 2, pp. 921-929, 2023.

[4] S. Chamishka, I. Madhavi, and et al: “A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling,” *Multimedia Tools and Applications*, vol. 81, pp. 35173–35194, 2022.

[5] C. Gobl and A. Chasaide: “The role of voice quality in communicating emotion, mood and attitude,” *Speech Communication*, vol. 40, no. 1–2, pp. 189-212, 2003.

[6] K. Zhang, Z. Zhang, and Z. Li: “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” *IEEE Signal Proc. Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.

[7] C. Dongyan and H. Shen: “Facial Expression Recognition Based on Mini-Xception Network,” *Int. Journal of Computational Intelligence Systems and Applications*, vol. 11, no. 1, pp. 50-57, 2024.