講演番号	004

文献番号	20225004

Deep Learning Based Early Recognition of Emergency Vehicles using On-Broad Microphones

Chisato Takatsu[1)] Keisuke Yoneda [1)] Naoki Suganuma[1)]

1) Kanazawa University,Graduate School of Natural Science and Technology,Division of Frontier Engineering
Kakuma-machi, Kanazawa City, Ishikawa Prefecture, 920-1192, Japan

KEY WORDS: Safety, Intelligent vehicle, Road environment recognition, Speech recognition, FBANK, SVM [C1]

In recent years, research on self-driving vehicles has been conducted with the aim of reducing traffic accidents and assisting the elderly in driving. One of the technologies required for automated vehicles is the recognition of the surrounding environment. Peripheral environment recognition is a technology that recognizes objects that exist in the vicinity of the automated vehicle, such as lanes, stationary obstacles, and moving objects. Among them, with respect to moving objects, vehicles must take driving actions that do not interfere with the safe travel of emergency vehicles when they approach on public roads. When an emergency vehicle exists outside the range of the sensors of the automated vehicle in a situation where the vehicle must give way to an emergency vehicle, it is effective to recognize the emergency vehicle by the sound of its siren, because the emergency vehicle is traveling with its siren on. Considering the differences in sirens, ambulances, police cars, fire trucks, and motorcycle police cars are the four most common types of emergency vehicles. In this study, we aim to recognize the siren of an ambulance, which has a clear periodic feature, and the siren of a police car, which has many emergency vehicles with the same feature.

The flow of recognition is to extract sound features in accordance with the general sequence of speech recognition, and the features were trained by a support vector machine (SVM) to perform a two-class classification: siren or not siren. The features to be used are Log Mel filter bank features (FBANK). Soft-margin SVM (C-SVM) was used for SVM.

The sounds used in the training dataset were divided into 5-second segments. Three training datasets were created for ambulance recognition, and the dataset for police car siren recognition used the same configuration as the dataset that was most effective for ambulance recognition. In the following, we describe the training dataset for ambulance recognition.The first is a non-siren sound (environmental sound) plus only environmental sounds such as wind and horns (denoted as 'Only environmental sounds'), and the second is an environmental sound plus a police car siren (denoted as ' Add other siren'), the third one is the sound of 'Add other siren' with the sound of the wind superimposed on it and added to 'Add other siren' (denoted as 'Add mixture noise'). Among the three, the result of 'Add other siren' was the most valid, so we created a dataset for police car recognition with the siren as a police car siren and the siren in the ambient sound as an ambulance siren (denoted as 'Police car data'). Two evaluation datasets were used for ambulance recognition: one using sounds separated by 5 seconds as in the training dataset, and the other using 50 minutes of actual driving data. For police car recognition, only the one using sounds separated by 5 seconds was used for evaluation. Accuracy, Precision, and Recall were used as evaluation indices, and only Precision and Recall were used in the evaluation of actual driving data.

The evaluation results for 'add other sirens' are shown in Table 1 and for 'police car data' in Table 2. In the evaluation data set, more than 95% recognition was possible with 'Add other siren', which was the most effective, and more than 75% recognition was possible with the actual driving data. In addition, more than 90[%] of the data could be recognized in the 'Police car data'. The two types of misrecognition were the sirens with a lot of noise mixed in or the sirens themselves were low-pitched, as shown in Fig. 1(a), and the sirens with strong noise around the frequency of the sirens, as shown in Fig. 1(b). The misrecognition shown in Fig.1 (b) is thought to be due to the fact that features such as the frequency of the sirens changing over a certain period of time have not been successfully obtained.

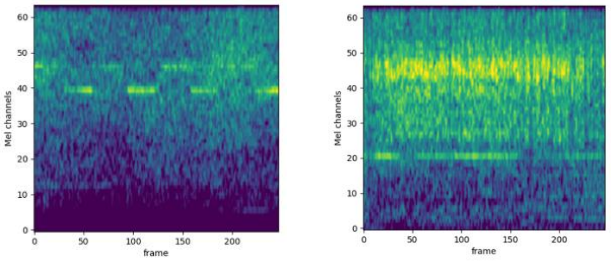
In the future, we intend to reduce false recognition by removing noise from the voice and extracting frequency features of sirens. In addition to recognizing ambulances and police cars, we also plan to make it possible to detect the presence of other sirens such as fire trucks and motorcycle sirens.

Table 1 Add other siren

	Evaluation dataset	Actual driving data
Accuracy	0.964	
Precision	0.966	0.938
Recall	0.960	0.789

Table 2 Police car data

	Evaluation dataset
Accuracy	0.945
Precision	0.930
Recall	0.926



(a) the siren with a lot of noise mixed in (b) wind noise

Fig.1 Misrecognition data