Cooking state recognition based on Acoustic event detection

Anonymous Author(s)*

ABSTRACT

In this research, the cooking sound analysis for understanding cooking activities was conducted toward the cooking support system. Although there have been attempts to use images and signals from motion sensors and temperature sensors to understand cooking behavior, only limited studies have been conducted using acoustic signals. The data set was newly constructed by actually cooking and recording. When humans cook, different cooking sounds are generated depending on the type of cooking behavior. By learning each cooking sound and restricting the action sequence from the recipe structure, it was achieved to estimate the cooking action sequence effectively.

CCS CONCEPTS

• **Information systems**;

KEYWORDS

Cooking activities recognition, Acoustic event detection, Signal processing, Nonnegative matrix factorization

ACM Reference Format:

Anonymous Author(s). 2018. Cooking state recognition based on Acoustic event detection. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Cooking activities are fundamental to human life and supporting it with information processing technology will become more important in order to improve the quality of life[9].

For speech information processing, support by a spoken dialogue system is considered, but in order to support cooking activities, it is necessary to recognize the state in a way that does not bother the cooker, and the question answering system is unsuitable. Therefore, it is necessary to recognize the states of cooking and use them for cooking activities understanding.

Studies using images or acceleration sensors for cooking activities recognition have been performed[4][7][3]. However, studies using sounds (cooking sounds) generated at the time of cooking are performed only limitedly[2], and research to tackle the recognition of the entire cooking process has not been conducted.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

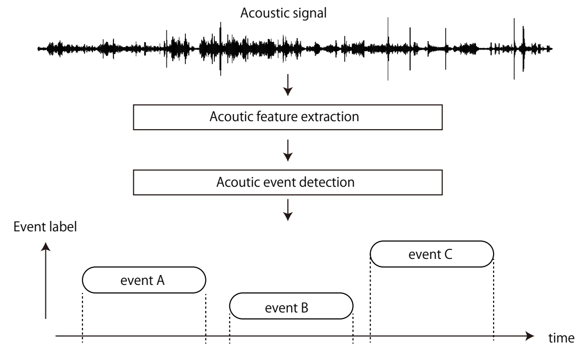

Figure 1: Probrem of Acoustic event detection

While image recognition is good at recognizing static objects such as food materials, cooking sounds are superior in recognition of cooking behavior such as "cutting" and "stir-frying", and it is suitable for this task.

Identification of cooking sound can be considered as a task of Acoustic Event Detection and Classification (AEDC) (Fig. 1), but it is difficult to define the target to be recognized. What can be thought of as a recognition unit is cooking behavior such as "cutting" or "stir-frying" as described in the recipe. However, in order to perform recognition in cooking behavior units, it is necessary to model cooking sounds labeled for each cooking behavior appropriately, and it is necessary to appropriately model cooking sounds labeled for each cooking behavior, A method that directly associates features is not suitable[6].

In this research, assuming that the cooking sound when doing cooking behavior can be described by the overlapping pattern of the basis of plural sound events, by extracting the sound event feature amount by nonnegative matrix factorization and modeling it, cooking We examined whether it is possible to identify behavior.

2 RELATED WORK

2.1 Cooking Activities recognition

Research on cooking behavior recognition using Egocentric videos and research on cooking behavior recognition using image data during cooking are being conducted. [8] [3]. In these researches, data during cooking and the corresponding label data are collected to conduct experiments and evaluations. Similarly in this paper, data sets are constructed by collecting acoustic data during cooking and manually assigning the corresponding label data, and experiments and evaluations are performed.

2.2 Acoustic event detection

Acoustic event detection is a theme that detects perceived events from acoustic signals such as bird calls and gunshots. Although research has been conducted by applying speech recognition technology, it is also addressing difficult problems by technological development of deep learning. There are many types of acoustic events, so research for using low-quality data sets and research for using small-scale data sets are issues[1]. Since the data set including environmental sound during cooking is not released, in this paper, a new data set was constructed, and experiments and evaluations were performed.

3 PROPOSED METHOD

3.1 Feature extraction

In this paper, two feature quantities are used. The first is the Mel frequency cepstrum coefficient (MFCC), and the other is the activation matrix using Nonnegative Matrix Factorization (NMF).

MFCC is a feature that is widely used in the speech recognition field, but it is also widely used as an effective feature in the field of AED.

NMF is one of multivariate analysis methods aimed at decomposing nonnegative value data into additive constituents, and attracts attention in various fields in recent years There is technology to be [5].

NMF is applied to the analysis of acoustic signals. The input spectrogram V as a two-dimensional matrix of time-frequency and approximate it to the product of two matrices H, U ($V \simeq HU$). One is called the basis matrix, and multiple characteristic spectral distributions that make up the original spectrogram are obtained. The other is called a weighting matrix, which indicates how much weight each basis vector is added at each time. Define the distance between V and HU and perform minimization to find H, U. Using the generalized Kullback-Leibler divergence as a distance function, find the stopping point by the auxiliary function method.

3.2 Cooking state modeling

The spectral basis is expected to be the spectrum of sound events that occur frequently in cooking sounds. However, the recognition unit of cooking sound has a longer duration than these sound events. Therefore, we propose to model cooking behavior using the frequency of each sound event appearing in a long time interval corresponding to cooking behavior.

Divide the weight vector sequence U_t by the window length N to obtain the weight matrix sequence $X_0, \ldots, X_L$. The NMF basis histogram $\boldsymbol{x}_i$ is obtained by summing the weight vectors for each weight matrix.

$$\boldsymbol{x}_i = \sum_{n=0}^{N-1} X_{i,n} \tag{1}$$

, where $X_{i,n}$ is the nth row of the weight matrix X_i. The histogram $\boldsymbol{x}_i$ of the cooking behavior A is modeled by the output probability distribution $P(x|A)$.

(a) the ingredients **(b) the dish**

Figure 2: the ingredients and dish

4 EXPERIMENT

4.1 Dataset

The cooking sound data set was constructed. Different cooks cook in different environments to avoid over-learning by recording in a specific environment, and the recipe made for the sake of simplicity is the same as the recipe specified. The specified food is "Yakisoba", which is Fried noodles. The recipe specified a simple one. For recording, we used an iPhone 8 built-in microphone, and recorded using 48 kHz sampling, and in the experiment we used downsampling to 16 kHz. The recorded data is 6 times, about 130 minutes in total. The labeling was done by the cook himself, and I asked them to fill in each section of the procedure described in the recipe.

4.2 Frame by frame cooking state identification

Four classification experiments were conducted to classify cooking noise into three classes: "Cut", "Fish", and "Other". A six-fold cross validation was performed using the six recorded data recorded. The probability density function for all experiments was Gaussian mixture model (GMM). First, as a baseline, the cooking sound was converted to a feature sequence, and a discrimination experiment was performed assuming that each time frame always follows GMM (MFCC). The feature is a 36-dimensional vector combining 12-dimensional MFCC with 1st-order and 2nd-order dynamic feature. Both frame length and frame shift are 16 ms. Frames are grouped into segments so that the identification is performed for each segment.

Next, we tried soft clustering using GMM for all cooking sound data, and tried to generate a histogram using the class posterior probability for each frame (GMM). The mixture number of GMM was 20, and the histogram was set to be generated every second.

Third, we used the NMF basis histogram, which is the proposed method (NMF). The specified number of NMFs was 20, and the histogram was set to be generated every second.

Finally, in addition to the proposed method, we also examined a method for decomposing a spectrogram into two spectrograms before base decomposition with NMF (HPSS).

The decomposition method is intended to preserve the continuity in the time direction and the frequency direction respectively by median filter, and it is expected that the spectral basis will be learned

Table 1: MFCC results

	mix-1	mix-2	mix-4	mix-8	mix-16	mix-32
val 1	61.84	63.70	65.79	72.60	70.35	74.50
val 2	48.55	68.33	60.70	65.99	64.69	62.11
val 3	59.67	60.30	61.56	63.23	54.67	56.10
val 4	63.52	69.48	75.28	54.64	55.19	55.01
val 5	77.15	78.60	79.70	79.62	81.07	80.80
val 6	65.72	69.61	80.12	71.90	74.64	77.15
average	62.74	68.34	70.53	68.00	66.77	67.61

Table 2: NMF results

	mix-1	mix-2	mix-4	mix-8	mix-16	mix-32
val 1	57.98	65.12	67.29	66.49	66.68	70.21
val 2	50.41	65.92	66.20	66.71	67.84	72.39
val 3	56.49	58.04	51.86	48.26	52.40	54.49
val 4	49.91	61.18	62.85	47.88	32.88	32.98
val 5	56.15	72.58	74.98	71.07	75.32	70.80
val 6	51.29	57.46	52.44	51.83	60.00	55.86
average	53.71	63.38	62.60	58.70	59.18	59.45

appropriately by this. The basis number of NMF was set to be 10 for each of two resolved spectrograms, and the histogram was generated every one second. The mixing number of GMM was only 2 mixed with MFCC, and the other 3 methods were 32 mixed.

The table 1 2 shows the F value of each class and the whole for each method.

The MFCC method directly associates the acoustic features with the labels, and it is considered that the difference between various cooking environments was not successfully absorbed. Moreover, the GMM method is characterized by the class posterior probability and is expected to be robust to differences in environment, but it has the lowest F value. This was the worstGMM learns that each time frame belongs to only one cluster, so unlike NMF where multiple bases are allowed to be included at the same time, cooking behavior could not be modeled well It is thought that there is no. On the other hand, the proposed NMF method can model cooking behavior by the superposition pattern of sound events, and it is thought that good results have been obtained. When HPSS is used in addition to the proposed method, it was expected that the basis could be learned more properly by separating the pulsing sound and the steady sound in advance. It was suggested that it did not fit the purpose of expressing the probability distribution of each action.

4.3 Post processing

In frame-by-frame state estimation, the problem is that the state changes frequently. As a method to avoid this, it is generally used to constrain the state transition using HMM. In this paper, we use the method of simply making the number of state transitions known and applying a smoothing filter until the number of state transitions is correct.

5 CONCLUSION

In this research, we presented a method of utilizing acoustic signals for cooking activities recognition for cooking support systems, and examined cooking state identification using cooking sound with real environment data.

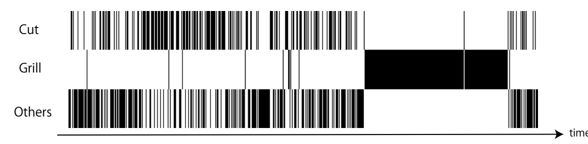

Figure 3

Figure 4: result

Using the MFCC and GMM widely used in the audio field, we could obtain over 70% identification results. The method of modeling the cooking state by the appearance distribution of the sound event resulted in low accuracy. Further research is required in the future.

REFERENCES

[1] Keisuke Imoto. 2018. "Introduction to acoustic event and scene analysis". *Acoustical Science and Technology* 39, 3 (2018), 182–188. https://doi.org/10.1250/ast.39.182

[2] T. Kojima, T. Ijiri, J. White, H. Kataoka, and A. Hirabayashi. 2016. CogKnife: Food recognition from their cutting sounds. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*. 1–6. https://doi.org/10.1109/ICMEW.2016.7574741

[3] H. Kuehne, A. B. Arslan, and T. Serre. 2014. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*. 780–787.

[4] Fernando De la Torre et al. 2008. *"Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database"*. Technical Report. CMU.

[5] Daniel Lee and H Sebastian Seung. 1999. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* 401 (11 1999), 788–791. https://doi.org/10.1038/44565

[6] Panu Maijala, Shuyang Zhao, Toni Heittola, and Tuomas Virtanen. 2018. Environmental noise monitoring using source classification in sensors. *Applied Acoustics* 129 (01 2018), 258–267. https://doi.org/10.1016/j.apacoust.2017.08.006

[7] Atsushi Shimada, Kazuaki Kondo, Daisuke Deguchi, G¥'eraldine Morin, and Helman Stern. 2013. Kitchen Scene Context Based Gesture Recognition: A Contest in ICPR2012. In *Advances in Depth Image Analysis and Applications*, Xiaoyi Jiang, Olga Regina Pereira Bellon, Dmitry Goldgof, and Takeshi Oishi (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 168–185.

[8] Shuichi Urabe, Katsufumi Inoue, and Michifumi Yoshioka. 2018. Cooking Activities Recognition in Egocentric Videos Using Combining 2DCNN and 3DCNN. In *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management (CEA/MADiMa '18)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3230519.3230584

[9] Daisuke Uriu, Mizuki Namai, Satoru Tokuhisa, Ryo Kashiwagi, Masahiko Inami, and Naohito Okude. 2012. Panavi: Recipe Medium with a Sensors-embedded Pan for Domestic Users to Master Professional Culinary Arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 129–138. https://doi.org/10.1145/2207676.2207695