COGKNIFE: FOOD RECOGNITION FROM THEIR CUTTING SOUNDS

Takamichi Kojima, Takashi Ijiri, Jeremy White, Hidetomo Kataoka, Akira Hirabayashi
College of Information Science and Technology, Ritsumeikan University

ABSTRACT

In this study, we present "CogKnife", a knife device which can identify food. For this, a small microphone is attached to a knife, which records the cutting sound of food. We extract spectrograms from the cutting sounds and use them as feature vectors to train a classifier. This study used the k-Nearest Neighbor method (k-NN), the support vector machine (SVM) and the convolutional neural network (CNN) to verify differences of the classification methods. To evaluate the accuracy of our technique, we performed classification experiments with six kinds of foods (apples, bananas, cabbages, leeks and peppers) in a laboratory environment. From 20-fold cross validation, we confirmed high recognition accuracies, such as 83% with k-NN, 95% with SVM and 89% with CNN.

Index Terms— food recognition, sound recognition, cooking support, machine learning, pattern recognition

1. INTRODUCTION

Monitoring and recording food consumption at home has a potential to solve various social problems, such as food wastage and lifestyle diseases. For instance, overbuying could be prevented if people accurately knew the remaining amount and quality of household food. If food nutritional value can be accurately recorded, we may be able to reduce risk of disease caused by our lifestyle. Monitoring food at home is a fundamental technology to develop an application to support cooking. If the food being cooked can be recognized, detailed information in relation to the effective use of these products, such as the amount of seasoning needed, or the most efficient way to cut the product could be found.

Many studies have been performed on food detection and their applications [1]-[3]. Existing studies detect food in kitchen environments by using multiple sensors, such as cameras, microphones, acceleration sensors, and pressure sensors. However, existing studies require a smart kitchen to be built or placing many sensors within in a kitchen. It is difficult to adopt food recognition systems in a standard house, because it is not practical and the cost of installation is too high.

Takamichi Kojima, Takashi Ijiri, Hidetomo Kataoka, Jeremy White, Akira Hirabayashi. CogKnife: Food Recognition From Their Cutting Sounds, Proc. CEA2016: 8th Workshop on Multimediafor Cooking and Eating Activities, 2016/7/15.

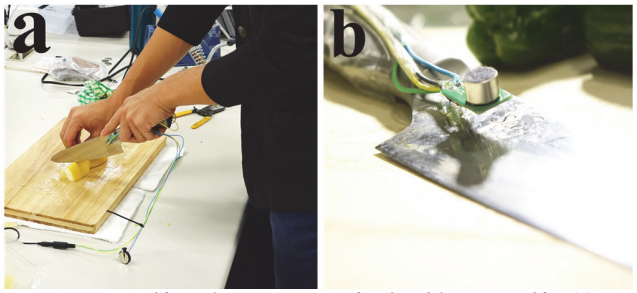

Fig. 1. CogKnife. The user cuts food with CogKnife (a). A microphone sensor is attached to a knife (b).

In this study, we present the CogKnife (Cognitive Knife) system. CogKnife identifies food by using a cheap and small microphone attached to a knife (Fig. 1). The key idea is to utilize the cutting sounds of food. We record the cutting sounds that correspond to single strokes, and extract the spectrograms of them to train a classifier. We adapt three different classifiers: k-nearest neighbors (k-NN), support vector machine (SVM), and convolution neural network (CNN).

To verify the accuracy of our technique, we performed cross validations by collecting cutting sounds of six fruits and vegetables: apples, bananas, carrots, cabbages, leeks and peppers. We confirmed that our technique classifies the six vegetables with 95% accuracy using SVM, 83% with k-NN, 89% with CNN. We also implemented a prototype, which classifies food and provides feedback immediately. Our contributions are listed as follows;

(1) Provide a framework for recognizing food based solely on their cutting sounds.
(2) Achieve accurate classification results by using spectrograms of the cutting sounds.
(3) Implement a prototype of CogKnife.

Since CogKnife requires a small and cheap microphone, it significantly reduces installation cost and provides an easy to use monitoring tool.

2. RELATED WORK

Food consumption monitoring and their applications are important topics within the human computer interaction field. We surveyed studies on food monitoring techniques for cooking support, household management, and health control.

Cooking Support. Food monitoring has the potential to enhance cooking navigation tools. Uriu et. al. present a cooking support system, named panavi [1]. They capture the temperature and motion of a fry pan with thermocouple and acceleration sensors. The captured information is used to navigate the user to manage correct cooking processes. Yamakawa et. al. install various sensors, such as infrared cameras, visible light cameras, and microphones in to a kitchen environment to estimate cooking activities for navigation [3].

Household Foods Management. Monitoring household foods helps to prevent overbuying and food wastage at home [4]. Fan et. al. present a method to measure the remaining contents in packaged items, such as beverages and snacks [5]. They probe a sweep sound to stimulate a target and estimate its container by using a response impulse. Diezma et. al. probe a sound to check the existence of hollow heart in a water melon by using impulse response [6].

Health Control. In general, diet has a significant influence on ones' health. To monitor and record eating is an important topic in the field of food-media. Chi et. al. estimate the calories of dishes [7] by using a camera mounted above the kitchen counter and two weighing sensors installed under a kitchen counter and under a stove. Amift et. al. classify types of food by using the chewing sound [8]. Hapifork system measures eating speed, the amount of fork-servings per minute, so as to prevent eating too fast [9]. Kadomura et. al. presented Sensing Fork to detect and modify eating behaviors of children [10].

The existing food monitoring systems require building a whole specialized system [1] [7] or multiple sensors installed in sufficient places [2] [3], resulting in high installation costs. In contrast, the CogKnife system only requires a small and cheap microphone on a knife. Its installation and maintenance cost is smaller than that of the existing systems.

The applications of the CogKnife system expand beyond the above three fields. Since the CogKnife classifies food under cooking, the obtained food information would be used for detailed cooking navigation as in [1]. If we record food consumption at home using the CogKnife, the household stock could be managed to prevent overbuying, and nutrition intake history could be visualized to control health. The future work of the research team is to develop these applications.

3. OVERVIEW

3.1. Fundamentals of CogKnife

CogKnife uses the differences of the cutting sounds. Since different fruits and vegetables have different internal structures, their sounds when cut by a knife are different from one another. For example, Fig. 2 shows audio waves and spectrograms of the cutting sounds of an apple and a cabbage. They clearly depict the difference of cutting sounds between the two. To some extent, we can also recognize the differences of their cutting sounds by listening.

There are various ways to cut food (e.g., chopping on a board and peeling). In this study, we focus on chopping on a board where a knife moves down to the board through a target, because it is the most basic cutting method. We deal with a sound which corresponds to a single chopping stroke as audio data. Notice that each audio data commonly contains a single peek when the knife hits the board. Figs. 2a and 2b show sounds correspond to single strokes as indicated by the arrows at the peak.

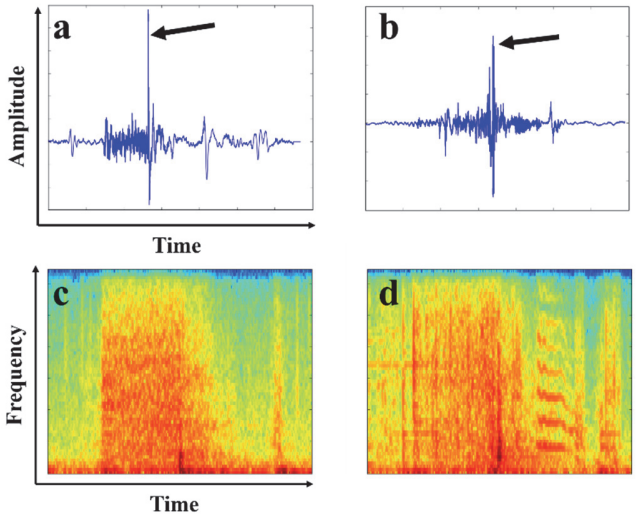

Fig. 2. Cutting sound waves and spectrograms of apples (a) (c) and cabbages (b) (d).

3.2. Microphone Attachment

In this study, we attached a small microphone sensor to a knife to collect cutting sounds while cooking. This study uses a 20mm×16mm sized condenser microphone equipped with an op amp with gain of 100 (Fig. 3). It requires 3V external power supply.

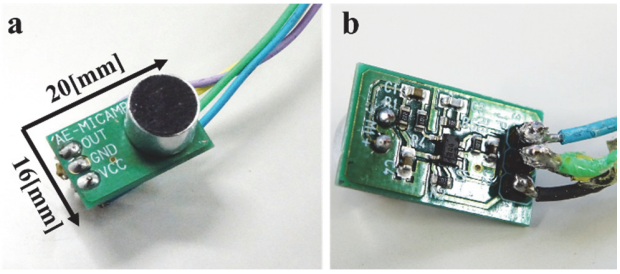

Fig. 3. A microphone sensor. Surface of sound collection (a) and the opposite side (b). It costs less than US $5.

After testing several ways of attaching of the microphone, we decided to attach it to a knife so that its sound correction surface is oriented to the opposite direction from the knife plane as in Fig. 1b. We attached the microphone on the opposite side from the user's dominant hand; the right-side knife for left-handed user. We placed the microphone floating in the air without contact to the knife in order to capture sounds propagated in the air. Notice that CogKnife is based on the fact that humans can identify food from the differences of their cutting sounds. This study uses the sound propagated in the air.

3.3. Workflow

The workflow of CogKnife consists of three steps. i) We collect cutting sounds by using a microphone attached to a knife (Fig. 3a). ii) Spectrograms are extracted from all the sounds as feature vectors to train a classifier (Fig. 3b). iii) We then classify unknown cutting sounds by using the trained classifier. We adapt three different methods, k-NN, SVM and CNN, as classifiers to compare their performances.

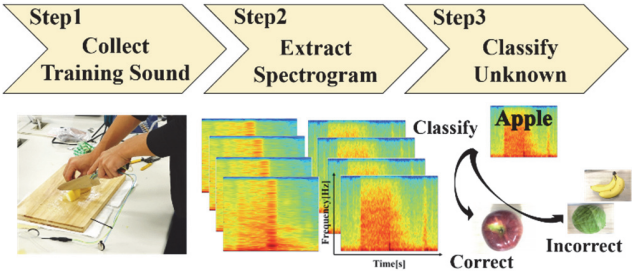

Fig. 4. The workflow of training.

4. CLASSIFICATION BY CUTTING SOUNDS

4.1. Collecting Training Sounds

To train our classifiers, we collect cutting sounds of fruits and vegetables that are labeled with their names. We record multiple strokes (e.g., five strokes) for our target food into an audio file. We stroke a knife with an interval of 0.5-1.0 seconds (Fig. 5a). We then manually divide the audio file into multiple cutting sounds each of which is associated with a single stroke (Fig. 5b).

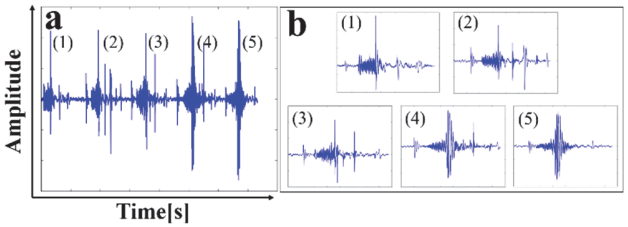

Fig. 5. An audio file for multiple strokes (a) and cutting sounds associated with single strokes (b).

4.2. Extracting Feature Vector

To train the classifiers, we use a spectrogram because it is able to encode both the spectral characteristics and temporal variation of input sounds. In our experience, the temporal variation of amplitude is especially important for accurate classification. We extract spectrograms from the obtained cutting sounds using the following three steps.

Downsampling. CogKnife is motivated by the fact that we could recognize the difference of cutting sounds of different food from listening. This suggests that cutting sounds contain enough information within the audible band to classify food. We then downsample the cutting sound recorded with 44.1 kHz to S kHz. We test different sampling rates, S = 32, 16, or 8. Since this downsampling removes unnecessary higher frequency components, it not only improves accuracy of classification but also accelerates the following processes. We used SoX [11] software for this process.

Alignment and trimming. For accurate classification, each cutting sound should be temporally aligned. In other words, its central point of the time window coincides to the time when the knife hits the board. To do this, we detect the time by smoothing the sound signal (Fig. 6c) and searching the highest volume point (Fig. 6d). We then clip a part of the sound 0.25 seconds before and 0.25 seconds after the detected time (Fig. 6de). The cutting sound is abandoned if the clipping range protrudes out of its original time window.

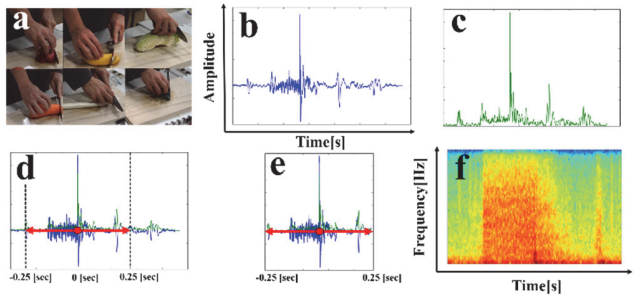

Fig. 6. Extracting feature vector. Given input audio file (a), we dowmsample it (b). We detect the time when the knife contacts to the chopping board by smoothing the audio signal (c) and searching the highest volume point (Red Dot in d). We clip 0.5 seconds from the input audio to obtain the aligned cutting sound (e) and convert it to a spectrogram (f).

Extracting spectrogram. We finally extract spectrograms from the aligned and trimmed cutting sounds with DFT window size, W, and 5% overlap. This study tested different window size, W = 2048, 1024, 512, or 256, to conform differences caused by it. Spectrograms become a two dimensional image as in Fig. 6f.

4.3. Training Classifier

We train k-NN, SVM, and CNN by using the obtained spectrograms. For k-NN, we use scikit-learn [12], a commonly used library for python, with the uniform weights and the ball tree algorithm. We vary k value to examine its effects. For SVM, we also use scikit-learn with the linear kernel. For CNN, we use digits 2.0, developed by NVIDIA Corporation [13], with AlexNet model [14]. We convert the spectrograms to 256×256 color images and used them to train CNN.

5. RESULTS AND DISUSSION

5.1 Validation

To evaluate the accuracy of our technique and to confirm the best set of classifiers and parameters, we performed cross-validation. We prepared six types of fruits and vegetables, such as apples, bananas, cabbages, carrots, leeks, and peppers. These foods are selected based on previous studies [2]. We recorded 50 strokes for each food and obtained 6×50 audio files in total, and then extracted spectrograms from them. In the spectrogram computation, 7 strokes were abandoned because of alignment errors. All of them were sounds for pepper.

Leave-one-out validation. We performed leave-one-out validation on the given 293 spectrograms of six types of foods. We varied the sampling rate S kHz as (32, 16, 8) and the window size W as (2048, 1024, 512, 256) so that we considered 12 combination patterns in total for each classifier. As results, SVM showed the best accuracy, 95%, with S=16 and W=1024, and k-NN showed its best accuracy, 85%, with k=6, S=16, and W=1024 (see Tables 1 and 2). We did not perform leave-one-out variation for CNN because of time and computational resource limitation.

Our SVM-based classification showed clearly higher accuracy than Kranz et. al., [2]. Notice that Krantz et. al., [2] performed experiments with apples, bananas, kohlrabis, carrots, leeks, and peppers; we used cabbage instead of kohlrabi, since it was difficult to source.

In our experiments, differences of parameters did not have strong influence on classification accuracies; In SVM-based classification, the difference of accuracy between the best and worst parameter sets was about 4%.

In Table 1, the results of SVM for bananas and carrots were almost perfectly classified. However, peppers and leeks showed slightly higher classification errors. To investigate this reason, we visualize representative spectrograms of the six foods in Fig. 7 and average spectrograms in Fig. 8. Peppers and leeks often provided similar spectrograms and thus tended to be miss-classified. Carrots have a higher volume of cutting sound than other fruit and vegetables, so

Table 1. Confusion matrix of the best classification accuracy with SVM with S=16, W=1024.

	apple	banana	cabbage	carrot	leek	pepper	accuracy
apple	48	0	0	2	0	0	**96%**
banana	0	49	0	0	0	1	**98%**
cabbage	2	0	47	0	1	0	**94%**
carrot	1	0	0	49	0	0	**98%**
leek	0	1	0	0	46	3	**92%**
pepper	0	1	0	0	3	39	**91%**

Table2. Confusion matrix of the best classification accuracy with k-NN with S=16, W=1024, k=6.

	apple	banana	cabbage	carrot	leek	pepper	accuracy
apple	43	3	0	3	0	1	**86%**
banana	2	46	0	1	0	1	**92%**
cabbage	2	4	38	1	3	2	**76%**
carrot	2	1	0	46	0	1	**92%**
leek	1	5	0	0	39	5	**78%**
pepper	0	5	0	0	2	36	**84%**

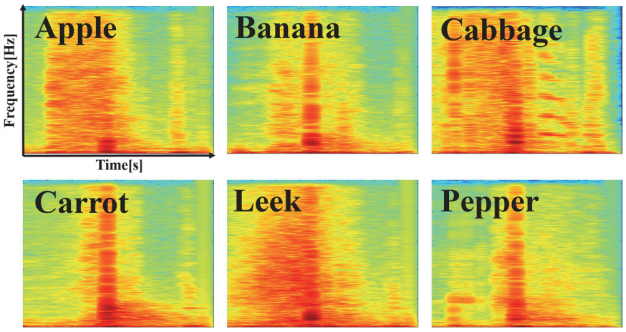

Fig. 7. Spectrograms of representative examples each food. They are extracted by parameters S=16, W=1024.

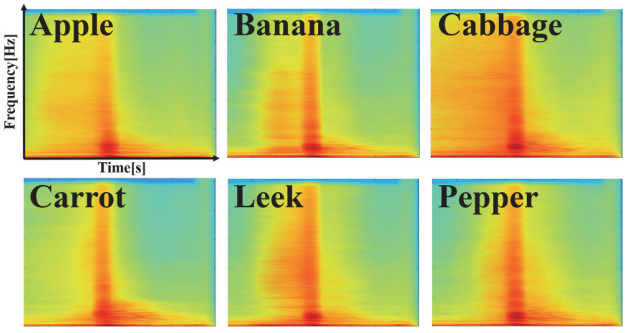

Fig. 8. Average spectrograms that are made from 50 spectrogram images per food.

that it allowed for high accuracy. Bananas also showed high accuracy, because they have low volume space before the knife contacts to the board (Fig. 9).

Table 2 provides the results of k-NN. Bananas or carrots were still almost perfectly classified even though accuracy is lower than the SVM result. Other fruits and vegetables accuracy was clearly less than the SVM result because k-NN has difficulty in sensing small differences of each spectrogram. Leeks and peppers have similar spectrograms, and they are slightly similar to the spectrogram of bananas. k-NN could not detect the difference of them compared to SVM. Cabbages classification accuracy greatly showed down than SVM. Referring to table 2, wrong cabbage classifications are classified to every other food because cabbage contains others small feature which was difficult for k-NN to classify correctly.

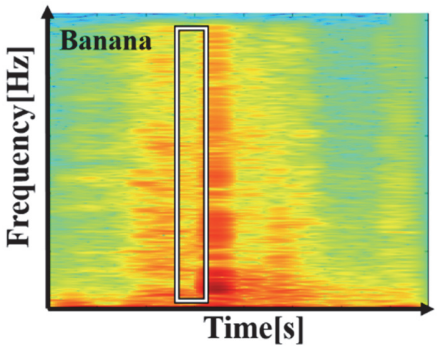

Fig. 9. A spectrogram of a banana with *S*=16 and *W*=1024. We often observed low volume region before the knife contacts the board.

20-Foldout validation. In recent years, CNN has been attracting attention and adapted to the food media applications. For instance, Yanai et. al. presented a method that can classify dish images that are collected from twitter stream by using a kind of CNN [15]. In this paper, we also adapt CNN for identifying food from cutting sounds.

To validate the performance of CNN, SVM and k-NN, we conducted a 20-foldout validation with them. We also varied the parameters *S* and *W* similarly to the leave-one-out validation. As a result, SVM showed best accuracy 95% with *S*=16, *W*=1024, k-NN's best accuracy was 83% with *k*=6, *S*=16, *W*=1024, and CNN showed best accuracy 89% with *S*=16, *W*=1024. In our experiments, SVM provided slightly higher classification accuracy than CNN. Since CNN is not specialized to sound classification, we would like to adopt deep learning technique specialized to sound classification in the future.

Notice that we evaluated our classification technique by using "same or similar" fruits and vegetables. For instance, we bought a package of carrots and recorded cutting sound of them. Therefore, our evaluation did not consider the differences of ripe stage or breeds. Detailed evaluation with foods with various conditions remains as our future work.

5.2. COGKNIFE IMPLEMENTATION

Based on the presented technique, we implemented a simple prototype of CogKnife that recognizes foods from cutting sounds. The system loads cutting sounds and trains SVM. We used the 293 sounds of six types of foods for this (Fig. 10a). After the training, the system waits cutting sound input. If the user cut vegetables, the system loads the sound and converts it to a spectrogram (Fig. 10b). The system then classifies the obtained spectrogram and provides the results to the user (Fig. 10c). Our prototype system is implemented in Python and uses *W*=16, *S*=1024, and SVM. We tested it on a Microsoft Surface Pro 3 with Intel Core i5-4300U CPU and 8.0 GB DDR3 RAM and it took about 10 seconds to classify a cutting sound.

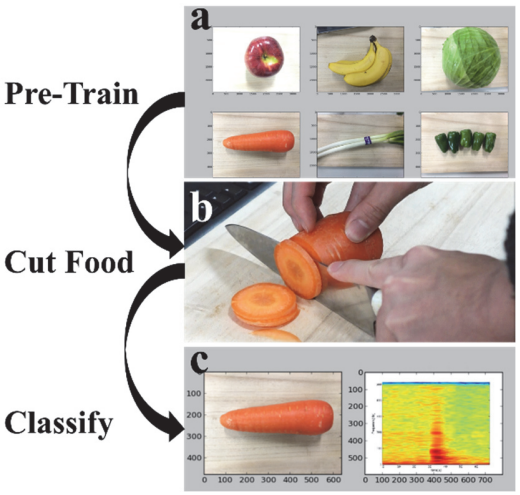

Fig. 10. A prototype system of CogKnife.

6. CONCLUSIONS

In this study, we have presented the CogKnife system that classifies food in cooking by using cutting sounds. To evaluate the accuracy of our technique and verify the best set of parameters, we performed leave-one-out and 20-foldout validation by using six types of foods, such as apple, banana, cabbages, carrots, leeks, and peppers. We found the highest classification accuracy, 95%, with SVM with *W*=16, *S*=1024 for the leave-one-out validation, and the highest accuracy, 95%, with SVM with *W*=16, *S*=1024 for the 20-foldout validation. We also implemented our prototype system, which classifies foods and provides classification results immediately.

CogKnife requires only a cheap and small microphone and thus its installation cost is lower than current food recognition systems. CogKnife achieves high classification accuracy relying only on cutting sounds. To the best of our knowledge, CogKnife system provides the first attempt to adapt a deep learning, i.e., CNN, for sound-based food classification.

Limitations and Future work. CogKnife is limited to food with a clear cutting sound; it is difficult to identify soft food such as tofu and jelly-like products. It is also difficult to identify food with a similar cutting sound caused by similar internal structures, such as bell pepper and paprika. For such food, it is important to corporate with other food recognition system relying on different modality, such as a camera-based system. Our evaluation is limited to six types of food; we would like to evaluate our technique by using a greater variety of food.

In this study, we recorded the cutting sounds in a quiet laboratory setting. It is difficult to adopt the current CogKnife to cutting sounds recorded in a natural environment with a lot of noise. In the future, we would like to improve both the hardware and software to be robust to noises. In our experiments, all sounds were provided by one person. In the future, we would like to perform larger scale experiments considering inter-user differences. Other future work includes developing an application to maintain household foods and to visualize the history of food consumption for health control by using CogKnife.

Acknowledgements. We appreciate anonymous reviewers for their useful comments. This work was supported in part by Grant-in-Aid for Scientific Research on Innovative Areas 15H05924.

7. REFERENCES

[1] Uriu, D., Namai, M., Tokuhisa, S., Kashiwagi, R., Inami, M. and Okude, N. "Panavi: recipe medium with a sensors-embedded pan for domestic users to master professional culinary arts," in *Proc. CHI 2012*, pp.129-138, 2012.

[2] Kranz, M., Schmidt, A, Rusu, R. B. and Rigoll, G. "Sensing Technologies and the Player-Middleware for Context-Awareness in Kitchen Environments," in Proc. *Fourth International Conference on Networked Sensing System*, pp. 179-186, 2007.

[3] Yamakawa, Y., Shoji, T., Kakusho, K. and Minoh, M. "Automatic Cooking Archiving with Spoken Dialogue with Assistant Agent," in *Technical report of IEICE*, 105, pp.55-60, 2005.

[4] Ganglbauer, E., Fitzpatrick, G. and Molzer, G. "Creating visibility: understanding the design space for food waste," in *Proc. MUM '12*, 2012.

[5] Fan, M. and Troung, K. N. "SoQr: sonically quantifying the content level inside containers," in *Proc. UbiComp '15*, pp.3-14, 2015.

[6] Diezma-Iglesias, B., Ruiz-Altisent, M., and Barreiro, P. "Detection of internal quality in seedless watermelon by acoustic impulse response," *Biosystems engineering, 88(2),* pp. 221-230, 2004.

[7] Chi, P. Y, Chen, J. H, Chu, H. H. and Lo, J. L. "Enabling Calorie-Aware Cooking in a Smart Kitchen," in Proc. *3rd international conference on Persuasive Technology*, pp.116-127, 2008.

[8] Amift, O., Stäger, M., Lukowicz, P. and Tröster, G. "Analysis of Chewing Sounds for Dietary Monitoring," in *Proc UbiComp '05*, pp.56-72, 2005.

[9] HAPILABS, Inc, "hapifork". www.hapi.com/product/ hapifork, 2016.

[10] Kadomura, A., Li, C.-Y., Tsukada K., Chu H.-H., and Siio, I., "Persuasive Technology to Improve Eating Behavior using a Sensor-Embedded Fork," in *Proc. UbiComp '14*, pp.319-329, 2014.

[11] "Sox : Sound eXcahnge", sox.sourceforge.net/, 2016.

[12] Machine Learning in Python, "scikit-learn," https://scikit-learn.org/stable/, 2016.

[13] NVIDIA, "DIGITS," https://developer.nvidia.com/digits, 2016.

[14] Krizhevsky, A., Sutskever, I., Hinton, E. G., "ImageNet Classification with Deep Convolutional Neural Networks," *NIPS*, 2012.

[15] Yanai, K., and Kawano, Y. "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. IEEE ICMEW 2015*, pp.1-6, 2015.