# IMPROVEMENT OF UTTERANCE CLUSTERING BY USING EMPLOYEES' SOUND AND AREA DATA

*Tetsuya Kawase*[†] *Masanori Takehara*[†] *Satoshi Tamura*[†] *Satoru Hayamizu*[†]
*Ryuhei Tenmoku*[‡] *Takeshi Kurata*[‡]

[†] Department of Engineering, Gifu University
1-1 Yanagido, Gifu, Gifu, 501-1193 Japan
[‡] Center for Service Research, National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki, 305-8568 Japan

## ABSTRACT

In this paper, we propose to use staying area data toward the estimation of serving time for customers. To classify utterances enables us to estimate conversation types between speakers. However, its performance becomes lower in real environments. We propose a method using area data with sound data to solve this problem. We also propose a method to estimate the conversation types using the decision trees. They were tested with the data recorded in a Japanese restaurant. In the experiment to classify utterances, the proposed method performed better than the method using only sound data. In the experiment to estimate the conversation types, we succeeded to recover 70% of the mis-classified conversations using both of sound and area data.

***Index Terms***— Utterance classification, Area data, Serving time for customers , Conversation type

## 1. INTRODUCTION

In recent years, service engineering becomes one of the attractive themes related with signal processing and pattern recognition fields. In the service engineering, usage of various sensor data has been investigated so as to improve work efficiency and service quality; visualization and data mining are conducted as comprehensive analysis. We focus on activities for facilities e.g. restaurants and hotels [1]. To estimate employee's operations, sound signal is an important cue. If conversations with customers or other employees could be classified, we can clarify the situation around the employee.

However, there are two problems when we conduct the utterance classification in real environments. The first one is about noise. In real environments, there are many types of noises, and they often make the classification performance decrease. In addition, when focusing on conversations with a certain person, those with the others are no longer useful. The second is about speaker anonymity. Building speaker-dependent models for the employees needs numerous costs. Furthermore, there are many customers in the facilities as anonymous speakers, which makes the issue more difficult.

As the related works, there are researches about speaker turns [2], clustering of broadcast news audio [3], clustering for speech synthesis [4], and speaker diarization [5]. In these works, the research about speaker diarization is most similar to our work. The speaker diarization is often employed for conference recording or retrieval system. The research [5] uses relative positions between speakers as well as a microphone array and a camera in fixed positions. However, in our case speakers frequently move in the facility, in addition as mentioned, unknown speakers often appear and disappear. Our work thus has serious issues that are difficult to solve.

To overcome the above problems, in this paper we record and use not only sound data by bone conductive microphone but also staying area data of employees by motion sensor simultaneously. We also propose bimodal utterance classification to label conversations with employees or customers, i.e. conversation type. The results are quite useful to analyze employee's operation, especially serving time for customers.

This paper is organized as follows: Section 2 describes utterance classification using sound data. The detail of our bimodal utterance classification is explained in Section 3. Experiment to estimate conversation type is described in Section 4. Finally Section 5 concludes this paper.

## 2. SOUND-ONLY UTTERANCE CLASSIFICATION

### 2.1. Overview

In the conventional works for speaker identification and clustering, statistical models e.g. GMM (Gaussian Mixture Model) or classifiers e.g. SVM (Support Vector Machine) were often employed [6, 7, 8]. And most works chose MFCCs and a power coefficient as acoustic features.

The utterance classification is defined to classify sound data into several groups, each corresponding to employee's own speech, utterances by other employees and customers. The utterance classification is thus similar to speaker iden-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tification. Therefore, we also adopted MFCCs and a power coefficient as well as SVM classifiers in our work.

### 2.2. Binary classification for employee's speech

The goal of our work is to estimate conversation type and serving time for customers. To estimate those, it is necessary to classify speech data into the following three classes: "Own" (employee's speech on a microphone), "Others" (other employees' speech), and "Customers" (customers' speech). Before three-class classification, we did a preliminary experiment to extract "Own". Because sound data of "Own" are different from the others in the point of speaker specificity and energy coefficient, this binary classification is much easier. The result is useful to estimate speech amount of employee, who attaches the microphone.[1]

Fig.1 shows the flow of the binary classification. In an utterance, MFCCs are extracted for each frame. SVM is performed on each frame. A majority voting scheme is applied to obtain the result for the utterance. Table 1 shows experimental conditions. We used SVM-Light [9] to apply SVM. We used linear kernel and manually set a margin trade-off parameter. In our experiments, two data sets (speakerA and speakerB) in real environments, each including one-hour sound data were used. We conducted the experiment in open condition, and utterances were manually extracted and labeled. Details of the sound data are described in our previous research.

We evaluated how audio frames or utterances were correctly classified. For speakerA and speakerB, 79.2% and 71.7% frame-level accuracies were obtained, respectively. Fig.2 shows histograms for both data sets, according to frame-level rate in an utterance. SpeakerA includes 361 utterances and speakerB includes 394 utterances. From Fig.2, we can classify about 90% utterances correctly if the simple majority rule is employed. The result suggests that "Own" can be extracted easily even in real environments.

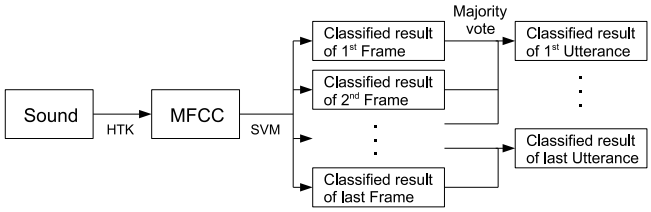



**Fig. 1**. A flow of binary classification for "Own".

**Table 1**. Experimental conditions for utterance classification.

| | | |
|---|---|---|
| train data | positive | 3000 frames of "Own" |
| | negative | 1500 frames of "Others" and "Customers" |
| test data | positive | utterances of "Own" |
| | negative | utterance of "Others" and "Customers" |
| frame size | 25msec | |
| frame shift | 10msec | |
| features | 39-dimension<br>MFCC(12), Δ MFCC(12), ΔΔ MFCC(12)<br>Power(1), Δ Power(1), ΔΔ Power(1) | |

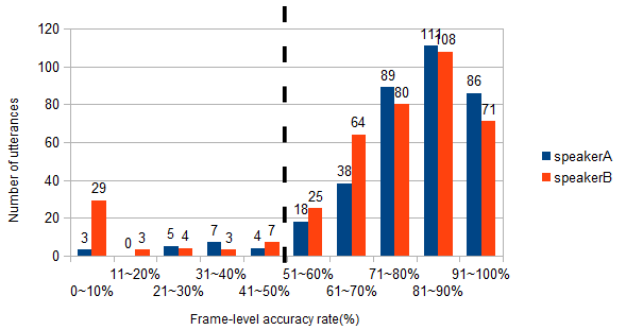



**Fig. 2**. The histograms of rates about utterances.

### 2.3. Ternary utterance classification

#### *2.3.1. Toward ternary classification*

There are two approaches to classify sound data into "Own", "Others" and "Customers": to conduct ternary (three-class) classification, or to apply the binary classifier in the last subsection and subsequently another classifier which distinguishes "Others" and "Customers". If we choose the latter, we must use results by the first classifier which may include mis-classification, and the error causes further performance degradation in the second classification. Thus we employ the former strategy in this paper.

#### *2.3.2. Experimental conditions*

The training data consists of 3,000 frames in "Own", "Others" and "Customers". The test data also consists of frame data derived from the three classes. And we used SVM-Multiclass as a classifier in this experiment. The other conditions are the same as the previous experiment.

#### *2.3.3. Experimental results and discussions*

Table 2 shows utterance-level accuracies for the three classes. Sound data for "Own" were classified almost correctly similar to the previous experiment. On the other hand, the accuracy for "Customers" is low. This occurred because "Others" and "Customers" were for unspecified or anonymous speakers. The another reason is that the power coefficient for these classes was unstable, while the coefficient for "Own" had usually a large value and was stable. To solve this problem, the scheme to make individual models for all employees and apply multi-class classification is considered. But the scheme requires a lot of costs, since further annotation (which employee spoke) is essential. So alternatively, we propose another approach where the area data recorded by motion sensor are utilized. Details of the area data are described in the next section.

**Table 2**. Experimental results for ternary classification.

| data set | Own | Others | Customers |
|---|---|---|---|
| speakerA | 96.4% | 68.8% | 44.5% |
| speakerB | 88.6% | 53.6% | 20.5% |

## 3. BIMODAL UTTERANCE CLASSIFICATION WITH AREA DATA

### 3.1. About area data

We have recorded area data that represent where the employee was. Area data enable us to limit the target whom the employee talks to. For example in the guest room, the employee probably talks to customers.

The area data in our work indicate the place in which the employee stayed, using the 3-dimensional coordinates recorded by the sensor. Fig.3 shows a part of layouts of the restaurant where we recorded data. Area data types are guest room(G), kitchen(K), passage(P), cash desk(C), and stairs(S).

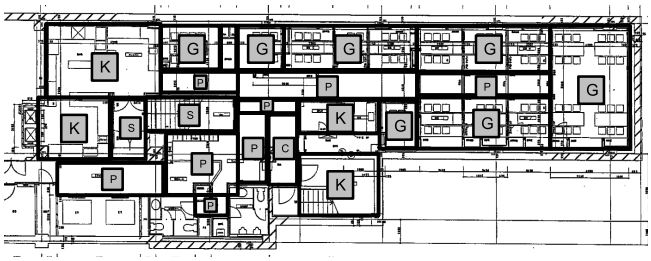


**Fig. 3**. Layout of the restaurant for this work.

### 3.2. Bimodal utterance classification

In this paper we propose a bimodal utterance classification with sound data and area data. Fig.4 illustrates the flow of the bimodal utterance classification. The classification is performed based on the following formulas:

$$D_1 = \lambda_s R_s(others) + \lambda_a R_a(kit) \quad (1)$$

$$D_2 = \lambda_s R_s(customers) + \lambda_a R_a(gr, ca) \quad (2)$$

$$D_3 = R_s(others) - R_s(own) \quad (3)$$

$$D_4 = R_s(customers) - R_s(own) \quad (4)$$

$$D_5 = R_s(others) - threshold \quad (5)$$

In the formulas (1)(2)(3)(4)(5), each variable represents:

- $\lambda_s, \lambda_a$ : weight factors for sound and area data,
- $R_s(x)$ : a ratio of the class $x$ in the sound-only ternary classification result,
- $R_a(kit)$ : a ratio of the time the employee with a microphone stays in kitchen,
- $R_a(gr, ca)$ : the ratio of the time the employee with a microphone stays in guest room or cash desk.

The utterances estimated as "Own" using only sound data are kept because they were classified clearly in the previous experiments. Sound-only results for "Others" and "Customers" are fixed using (1)-(4). The comparison between $D_1$ and $D_2$ determines which the employee talks to "Others" or "Customers". The sign of $D_3$ classifies "Own" or "Others", while that of $D_4$ classifies "Own" or "Customers". The formula (5) is for "Others". The sign of $D_5$ can fix the mis-classification for the other classes.

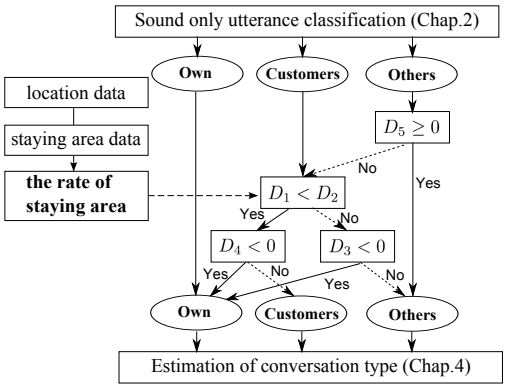



**Fig. 4**. A flow of bimodal utterance classification.

### 3.3. Experiment using area data

#### *3.3.1. Experimental conditions*

To evaluate the effectiveness of our proposed method, we conducted bimodal utterance classification. Table 3 shows the experimental conditions.

The sound data were the same as the previous experiment. The threshold of "Others" was determined in 0.75 empirically. To find the appropriate ratio of sound data and area data, the weight factors were changed by 0.1 satisfying the formula (6).

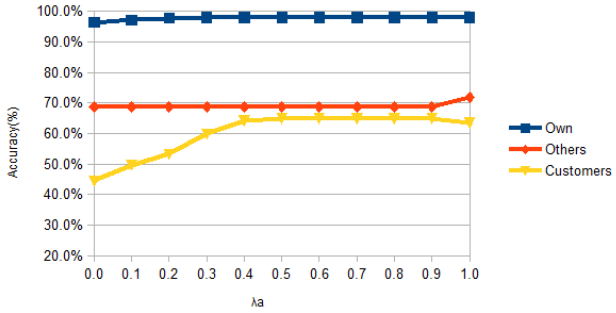
$$\lambda_s + \lambda_a = 1 \quad (6)$$

**Table 3**. Experimental conditions of bimodal utterance classification.

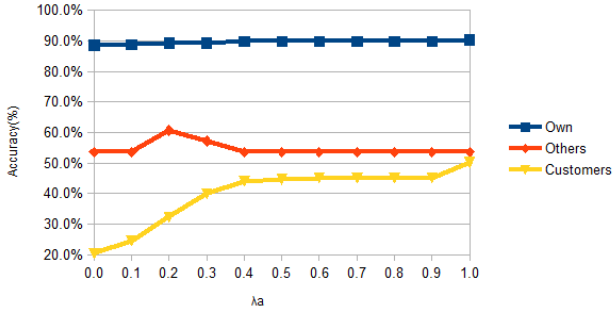
| | | |
|---|---|---|
| Sound | train | each 3000 frames ("Own", "Others", "Customers") |
| | test | each utterance ("Own", "Others", "Customers") |
| | frame size | 25msec |
| | frame shift | 10msec |
| | features | 39-dimension<br>MFCC(12), Δ MFCC(12), ΔΔ MFCC(12)<br>Power(1), Δ Power(1), ΔΔ Power(1) |
| Area | interval | 1sec |
| | area type | guest room, cash desk, kitchen, the others |

#### *3.3.2. Experimental result and discussions*

Fig.5 shows the accuracy rates changing weight $\lambda_a$. From Fig.5, we can know that the bigger the weight of area data became, the better utterances of customers were estimated. Furthermore, from Fig.5, we could decrease absolutely 20.4% mis-classification of customers in speakerA and 29.7% in speakerB.




(a) speakerA




(b) speakerB

**Fig. 5**. Utterance-level classification results of our proposed method.

## 4. ESTIMATING CONVERSATION TYPE

### 4.1. Definition of conversation types and experimental condition

In the experiment in Section 3, the performance of utterance classification was improved by using area data. In this section, we experimented to estimate the conversation types by combining several utterances into one conversation according to utterance classification results, toward the estimation of serving time for customers. Conversations are classified into three types shown in Table 4. Note that the conversations between "Others" and "Customers" were not observed.

Because type2 and type3 are important for analyzing operation of employees and estimating serving time for customers, we aimed at these two types. The conversation labels are determined based on the utterance time, "Others" or "Customers". Utterances were grouped into conversations by silence interval.

**Table 4**. Conversation types.

| | Own | Others | Customers | Example |
|---|---|---|---|---|
| type1 | ○ | | | greet, monologue |
| | ○ | ○ | ○ | large voice |
| type2 | ○ | ○ | | business contact |
| type3 | ○ | | ○ | order, check |

### 4.2. Experimental result and discussion

We evaluated three methods with different silence intervals, based on recovery rate which means how mis-recognized conversations are recovered using sound and area data, in the whole mis-recognized conversations caused by sound-only classification. The experimental results are shown in Fig.6. From Fig.6, we can know that it is possible to recover 70% of the mis-estimated frames in sound-only classification results by using the bimodal method. In this experiment we tested three intervals: 7sec, 5sec and 3sec. Among the three intervals, no significant differences were observed. So it is found that the method can estimate conversation types without strongly depending on silence intervals. And this experimental results suggest that using area data is useful to estimate more accurate serving time for customers.

Finally, discussing results in Section 3 and 4, bimodal utterance classification is much useful to recover the performance. And the effect is more strongly observed in conversations than utterances. It thus turns out that we can benefit from bimodal techniques much more when we employ larger recognition unit. The fact may be useful for other bimodal signal processing fields.

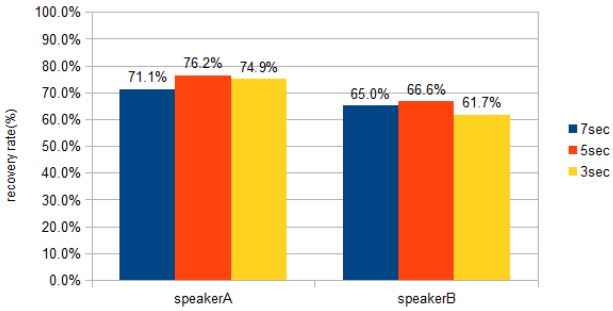



**Fig. 6**. Recovery rates of mis-estimated frames in sound-only results by the bimodal method.

## 5. CONCLUSION

In this paper, we showed the efficiency of bimodal classification with area data when classifying utterances and conversations. In the experiment for utterances, the method using area data performed better than the method using sound-only data. In the experiment for conversations, about 70% of misclassified conversation labels were correctly recovered by using utterance labels with area data.

Our future work includes use of other data such as fiscal data and making classifiers using both sound and area data.

## 6. ACKNOWLEDGMENT

The authors thank Ganko Food Service Co., Ltd., and staff in the Ganko Ginza 4-chome restaurant for their cooperation with our experiment.

## 7. REFERENCES


[1] Takehara, M. and Tamura, S. and Tenmoku, R. and Kurata, T. and Hayamizu, S., "The role of speech technology in service-operation estimation," Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on, pp.116-119, 2011

[2] Johnson, S.E., "Who Spoke When? - Automatic Segmentation and Clustering for Determining Speaker Turns," Proc. Eurospeech, Vol. 5, pp. 2211-2214, 1999.

[3] M. Siegler, U. Jain, B. Ray and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio", Proceedings of the Speech Recognition Workshop, pp 97-99, 1997.

[4] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in Eurospeech, 1997, pp. 601-604.

[5] Araki, S.; Hori, T.; Yoshioka, T.; Fujimoto, M.; Watanabe, S.; Oba, T.; Ogawa, A.; Otsuka, K.; Mikami, D.; Delcroix, M.; Kinoshita, K.; Nakatani, T.; Nakamura, A.; Yamato, J., "Low-latency meeting recognition and understanding using distant microphones," Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on, pp.151,152, 2011

[6] Matsui Tomoko, and Tanabe Kunio, "Comparative study of speaker identification methods: dPLRM, SVM and GMM.," IEICE transactions on information and systems 2006, pp.1066-1073, 2006

[7] Reynolds, D.A.; Rose, R.C., "Robust text-independent speaker identification using Gaussian mixture speaker models," Speech and Audio Processing, IEEE Transactions on , vol.3, no.1, pp.72,83, Jan 1995

[8] Chen, Shi-Huang, and Yu-Ren Luo., "Speaker verification using MFCC and support vector machine.," Proceedings of the International MultiConference of Engineers and Computer Scientists. ol.1. 2009

[9] SVM-Light,
http://svmlight.joachims.org/