MULTI-MODAL SERVICE OPERATION ESTIMATION USING DNN-BASED ACOUSTIC BAG-OF-FEATURES

Satoshi Tamura[*], *Takuya Uno*[*], *Masanori Takehara*[*], *Satoru Hayamizu*[*], *Takeshi Kurata*[†]

[*] Department of Information Science
Gifu University, Japan

[†] Center for Service Research
AIST, Japan

ABSTRACT

In service engineering it is important to estimate when and what a worker did, because they include crucial evidences to improve service quality and working environments. For Service Operation Estimation (SOE), acoustic information is one of useful and key modalities; particularly environmental or background sounds include effective cues. This paper focuses on two aspects: (1) extracting powerful and robust acoustic features by using stacked-denoising-autoencoder and bag-of-feature techniques, and (2) investigating a multi-modal SOE scheme by combining the audio features and the other sensor data as well as non-sensor information. We conducted evaluation experiments using multi-modal data recorded in a restaurant. We improved SOE performance in comparison to conventional acoustic features, and effectiveness of our multi-modal SOE scheme is also clarified.

Index Terms— Service operation estimation, multi-modal signal processing, stacked denoising autoencoder, bag of features, environmental sounds.

1. INTRODUCTION

In recent years, service engineering becomes one of attractive themes related to signal processing and pattern recognition fields. The purpose of service engineering is to improve work efficiency and service quality in shops, restaurants, hotels, factories, warehouses, and any other applicable places. In service engineering, at first many kinds of sensor data (e.g. speech data, camera images, location information of workers) are obtained, and non-sensor data (e.g. Points-Of-Sales (POS) data) are also collected. These data are subsequently analyzed to extract and find any knowledge for improving the efficiency and quality, not only using signal processing methods but also employing data mining and visualization approaches. According to the results, finally work routines are reconsidered and improved so as to provide better service qualities.

Estimating worker's operations, called Service Operation Estimation (SOE), is an essential process in service engineering; features extracted from observed data are recognized as service operations, e.g. greeting, taking a order, serving a meal, and making an account in a restaurant. Among the data available for SOE, particularly speech information in an audio channel plays an important role; if we could know when or how often a worker utters, what a worker speaks, and whom a worker talks to, we can understand situation around the worker. Moreover, such results enable us to estimate worker's skills and service quality. From this point of view, we have investigated how to utilize speech information for SOE [1–3]; for example, we proposed to use Voice Activity Detection (VAD) to estimate how long and how often an employee speaks, which indicates serving time for customers [1]. On the other hand, background sounds in the audio signal have not been well investigated. Background sounds have a lot of information about environments, situations, and activities of workers [4]; if in a restaurant a glass crash sound is observed on employee's microphone, then employee must stay in a customers' room or pantry. Utilizing environmental sounds thus has a great potential to clarify worker's operations and service level.

Since not only an audio channel but also the other sensor data such as location data of workers are available in most cases, employing these data together is obviously useful to obtain more accurate information and various evidences for SOE. In addition, we can often employ non-sensor data like POS data. Multi-modal signal processing enables us to combine these useful modalities. The effectiveness of multi-modal signal processing have been shown in many fields; for instance, using visual information can enhance robustness of speech recognizer against acoustic noise, achieving higher recognition performance [5]. In service engineering, multi-modal processing also greatly helps us as well [2, 6].

This paper investigates (1) effective acoustic features of background sounds for SOE, as well as (2) a multi-modal service operation estimation scheme incorporating sensor data including proposed audio information and non-sensor data. To extract audio features, Stacked Denoising Autoencoder (SDA) [7] and Bag Of Feature (BOF) [8] are employed. The audio feature and the other ones are subsequently integrated. Support Vector Machine (SVM) is adopted in our work, to recognize the combined features as service operations. We conducted experiments using real data obtained in a restaurant, and tested our proposed acoustic feature and SOE scheme. SOE was finally evaluated by measuring performance of operation recognition.

The rest of this paper is organized as follows: Section 2 explains acoustic feature extraction. Our multi-modal SOE method is introduced in Section 3. Section 4 describes data

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

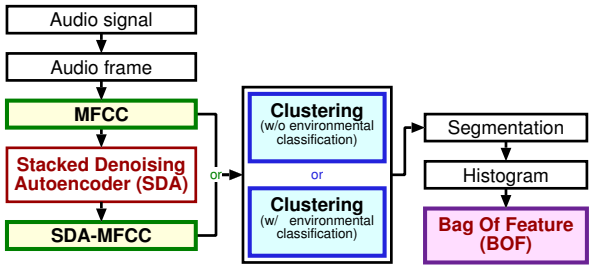


Fig. 1. Acoustic feature extraction.

specifications, experimental setup, results and discussions. Finally Section 5 concludes this paper.

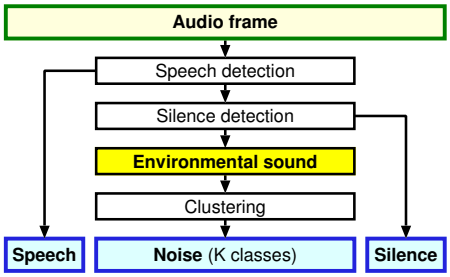
2. ACOUSTIC FEATURE

Figure 1 introduces our acoustic feature extraction strategy. An audio signal is firstly divided into audio frames with certain frame length and frame shift. Next, Mel-Frequency Cepstral Coeffcient (MFCC) parameters, that is often used in audio signal processing, are computed in each frame. SDA is then applied to MFCC vectors, generating new feature vectors: SDA-MFCC. Using either MFCC or SDA-MFCC vectors, clustering is subsequently carried out. Clustering is based on Gaussian Mixture Model (GMM) with/without environmental classification, in an unsupervised manner. Several frames are grouped into one segment to obtain a histogram of codebook elements, applying BOF techniques. An acoustic feature vector for SOE is finally obtained.

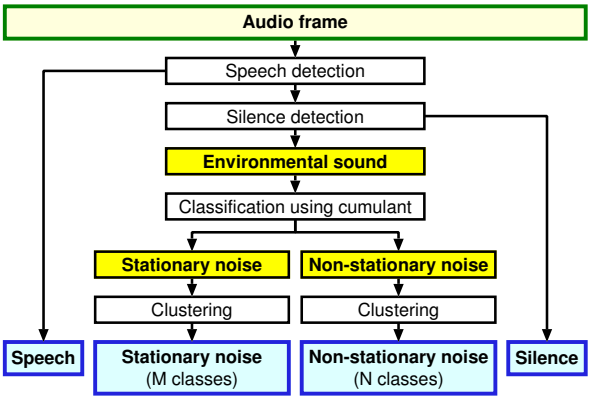
2.1. Stacked Denoising Autoencoder

In order to acquire high-performance acoustic features that is robustly applicable in real environments, this paper employs SDA [7]. SDA is one of Deep Neural Network (DNN) implementations, in which an input layer corresponds to a noisy feature vector and an output layer corresponds to a clean one. SDA is usually expected to remove noisy influence and distortion. In service engineering, audio data are recorded in real conditions where many kinds of background sounds simultaneously exist. The variation of audio features becomes so large, making the SOE performance significantly decrease. We believe SDA has a possibility to deal with the distortion.

When training an SDA, noisy data are often artificially generated from clean training data in order to implement noise reduction abilities. In our case, however, it is unnecessary to reduce all background noises, and rather, SDA is expected to flexibly deal with the variations. Therefore, we used feature vectors converted from real-environment training samples as "clean" data. Note that only pre-training is conducted in this paper, because fine-tuning is a supervised training method and it is quite difficult to manually make background noise transcriptions for numerous training data. In practice, we use pylearn2 [9] to perform SDA. Details about SDA should be also referred to tool documents.



(a) Without environmental classification



(b) With environmental classification

Fig. 2. Clustering of audio frames.

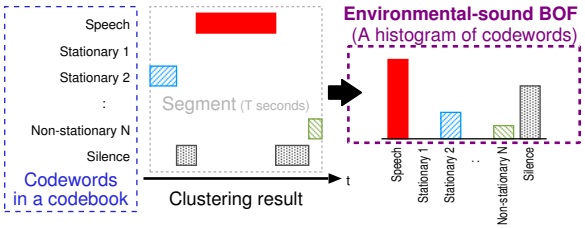


Fig. 3. Bag of features for an audio segment.

In this work, we obtain new features (SDA-MFCC) from conventional features (MFCC) by using SDA. In the following processing (Section 2.2), either MFCC or SDA-MFCC is exclusively chosen.

2.2. Clustering

Figure 2 (a) is a clustering method without environmental classification [4], and (b) is a proposed approach in this paper with environmental classification. In both schemes, at first, a conventional VAD technique is applied causing three classes: speech, silence, and environmental noise. Separating speech turns are described in [1], and detecting silence turns is done using a simple power threshold scheme.

Next, environmental classification is carried out in (b); the environmental-noise class is further divided into stationary and non-stationary noise classes using a quartic cumulant

Table 1. Basic features used in our multi-modal SOE.

Modality	Feature [dimension]
Acoustic	Environmental-sound BOF [10]
Location	Staying-area, orientation and action [18]
POS	Work-related [17]

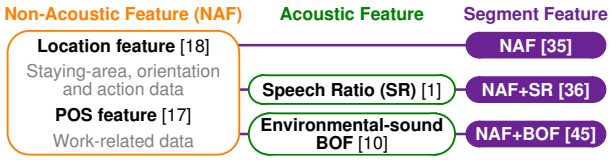


Fig. 4. Segment features from acoustic and non-acoustic ones.

coefficient C_4^s:

$$C_4^s = \frac{E[s(t)^4]}{E[s(t)^2]^2} \tag{1}$$

where $s(t)$ indicates an amplitude at time t in a frame. If this cumulant value is lower than a pre-defined threshold, the corresponding frame is classified into the stationary noise class. Otherwise, the frame is classified into the non-stationary noise class.

In either (a) or (b), non-hierarchical unsupervised clustering is further conducted for each noise class. In this work, a GMM-based approach is adopted. A GMM represents a distribution of training data as:

$$G(\boldsymbol{x}) = \sum_{k=1}^{K} c_k N(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k) \tag{2}$$

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ is a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, K is the number of Gaussian components, and c_k is a mixture weight factor subject to:

$$\sum_{k=1}^{K} c_k = 1 \ \ (0 \leq c_k \leq 1). \tag{3}$$

When training, we estimating the above parameters c_k, $\boldsymbol{\mu}_k$, and Σ_k. A trained GMM is then used for clustering, by estimating the most relevant Gaussian component for given test data; a class index $\hat{k}$ for an input vector $\boldsymbol{x}_i$ in a test set can be estimated as:

$$\hat{k} = \underset{k}{\operatorname{argmax}} \left\{ c_k N(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k) \right\}. \tag{4}$$

Note that in the following, let us denote the numbers of stationary and non-stationary classes in Figure 2 (b), by M and N, respectively.

2.3. Bag-of-features

BOF is originally proposed in a natural language processing domain as "Bag of words [8]," but currently it has been widely employed in the other fields; BOF is also effective in acoustic classification, e.g. [10, 11]. We utilize a BOF approach to finally obtain acoustic features.

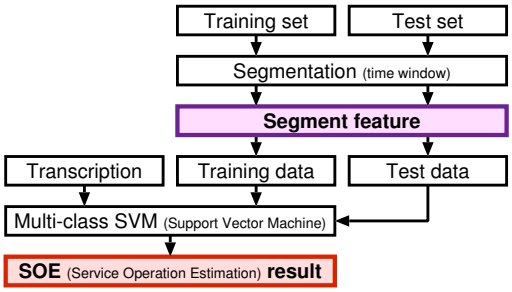


Fig. 5. Our service operation estimation method.

Table 2. Data specification and experimental condition.

Data	# of subjects: 2, # of segments: 1,683 (2h 20m)
Audio	Bitrate: 256kbps, Segment length: $T = 5$sec, Frame length: 25ms, Frame shift: 10ms
SDA	# of training data: 64,000 frames, # of hidden layers: 5, # of perceptrons: 1024–512–256–128–39, # of iterations of training: 10, Input & output layers: 39 dim feature vector (12 dim MFCCs and powe, their Δ and $\Delta\Delta$)
Clustering	W/o environmental classification: $K = 128$, W/ environmental classification: $M = 128$, $N = 32$
SVM	Kernel: RBF

Figure 3 depicts a flow. A codebook consisting of codewords, equivalent to classes in Section 2.2, is generated beforehand. Clustering results in consecutive several frames are collected into one segment. For each class, the number of frames classified is counted to generate a histogram of the codewords. Then the histogram is simply employed as an acoustic feature vector of the segment: Environmental-sound BOF. The dimension thus corresponds to $N+M+2$.

3. MULTI-MODAL SERVICE OPERATION ESTIMATION

3.1. Features

In addition to the acoustic feature described in Section 2, several types of features are prepared. Table 1 summarizes all the features. Location and POS features are extracted based on our previous works [6, 12]. From location, orientation and acceleration sensors, we extract following time-rate features: 10-dimensional staying-area features indicating where and how long a worker stayed, 6-dimensional orientation features indicating head orientation of worker, and 2-dimensional action features. From POS data, we compute 17-dimensional work-related features consisting of the numbers of ordering and customers, and so on.

Before integration, according to preliminary experiments, Principal Component Analysis (PCA) is additionally per-

Table 3. Service operations.

SO No.	Service operation
1	Greeting and offering customers to tables
2	Moving and carrying foods and drinks
3	Accounting
4	Taking orders
5	Serving foods and drinks
6	Cleaning up and setting tables

formed to the acoustic feature in order to reduce the dimension. As a result, a 10-dimensional acoustic feature is extracted. Finally in each segment, we simply concatenate the acoustic feature and the non-acoustic features into one supervector (segment feature). Figure 4 illustrates segment features used in this paper.

3.2. Estimation method

Figure 5 shows our SOE method. Feature supervectors are computed segment by segment. Multi-class SVM is built using training data and corresponding transcription labels. Test data are recognized by using the SVM.

4. EXPERIMENTS

We conducted two experiments: audio-only SOE comparing MFCC and SDA-MFCC (Section 4.3.1), and multi-modal SOE using acoustic features in addition to features derived from the other modalities (Section 4.3.2).

4.1. Data specification

Table 2 shows data specification used in the experiments. We collected data in a Japanese-style restaurant. Every workers equipped their own microphone and recorder. SNR was roughly -5 to 5dB in busy time, or 10 to 15dB otherwise. Note that VAD performance was 5.8% False Acceptance Rate (FAR) and 11.5% False Rejection Rate (FRR).

4.2. Experimental setup

Experimental condition is also summarized in Table 2. Hyper-parameters appeared in Table 2 were determined empirically. Table 3 indicates six service operations that should be classified. We chose these operations based on our experiences, so that we could measure employees' activities and service quality by analyzing results. We evaluated each feature and method by F-measure. All the experiments were done in a leave-one-out manner. Note that in order to evaluate recognition ability of proposed features and method, we tested data segments only corresponding to the operations.

In the first experiment, **MFCC** and **SDA-MFCC** were compared. Because we would like to show the necessity of proposed environmental classification process shown in Figure 2 (b), acoustic features without the environmental classification (a) were also tested. The number of stationary and non-stationary inner classes, that is M and N, were experimentally chosen.

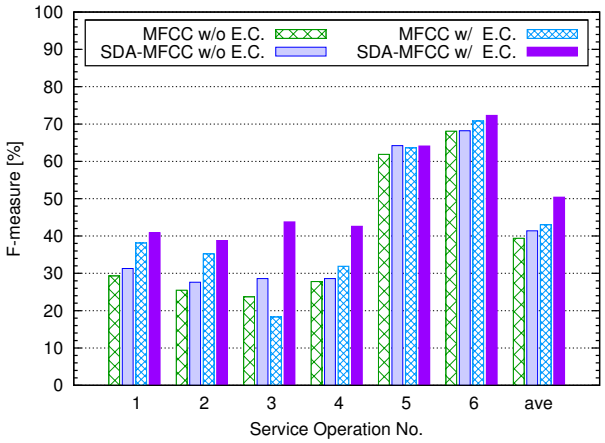


Fig. 6. F-measures for several acoustic features and E.C. (E.C.=Environmental classification)

Table 4. A confusion matrix using SDA-MFCC with environmental classification.

	SO No.	Estimated 1	2	3	4	5	6
Correct	1	**110**	7	20	33	29	70
	2	11	**33**	2	6	11	47
	3	21	2	**39**	7	12	19
	4	49	7	8	**106**	55	58
	5	40	4	4	39	**239**	35
	6	38	7	5	24	39	**447**

In the second experiment, we compared three features (see also Figure 4): Non-Acoustic Feature only (**NAF**), the location and POS features in addition to a conventional acoustic feature [1] (**NAF+SR**), and combination of the non-acoustic features and the proposed BOFs (**NAF+BOF**).

4.3. Experimental results and discussions

4.3.1. Audio-only SOE

We conducted audio-only SOE to clarify which feature is better (**MFCC** v.s. **SDA-MFCC**), and whether environmental classification is effective or not. Figure 6 shows experimental results for every service operations, and Table 4 indicates a confusion matrix when using **SDA-MFCC** with environmental classification. **SDA-MFCC** with environmental classification achieved the best performance among all the four features; especially No.3 and No.4 were significantly improved. For No.1 and No.2, environmental classification contributed to the improvements.

Since the numbers of GMM components were different ($K = 128$ v.s. $N + M = 160$), we did a supplemental experiment where $K = 160$. However, the result was slightly worse than $K = 128$. This indicates the improvement is not due to the difference of the number of Gaussian components but to the introduction of environmental classification.

The best performance (roughly 50%) seems to be insufficient. This comes from limitation of audio classification power; some operations are difficult to distinguish using au-

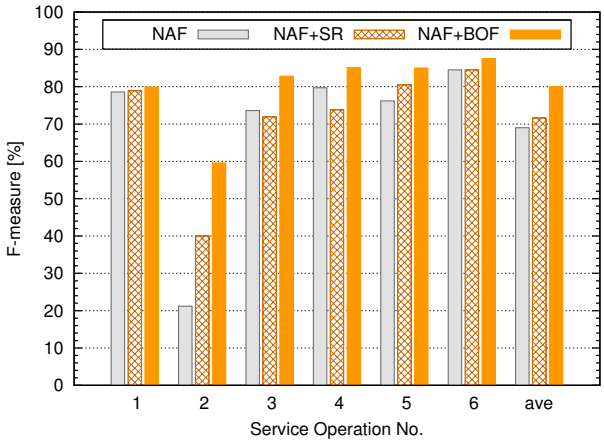


Fig. 7. F-measures of multi-modal SOE using non-acoustic features (**NAF**) in addition to conventional (**SR**) and proposed (**BOF**) acoustic features.

Table 5. A confusion matrix using non-acoustic features and environmental-sound BOF .

SO No.		Estimated 1	2	3	4	5	6
Correct	1	**245**	2	0	1	4	16
	2	18	**53**	3	5	5	26
	3	10	1	**82**	1	2	4
	4	22	4	6	**223**	16	12
	5	19	2	7	8	**307**	18
	6	30	5	0	3	27	**495**

dio cues only. Incorporating the other modalities is thus crucially expected.

4.3.2. Mulit-modal SOE

We performed multi-modal SOE as well as non-acoustic-only SOE for comparison. Figure 7 indicates experimental results. The best performance was observed when using our proposed multi-modal features (**NAF+BOF**); roughly 80% F-measure was obtained with more than 10% improvement from location-and-POS features (**NAF**). Especially, performance of No.2 was drastically improved. Table 5 shows a confusion matrix using **NAF+BOF**. It is also found that errors in all the cases significantly decreased in comparison with Table 4.

In conclusion, it is obvious that the performance of **NAF+BOF** is significantly higher not only than the non-acoustic-only SOE but also than the audio-only SOE. This indicates incorporating different modalities is successfully accomplished.

5. CONCLUSION

This paper investigates two aspects for SOE: (1) acoustic features based on environmental sounds by applying SDA and BOF methods, and (2) multi-modal SOE using the acoustic features and the other features. We recorded and annotated real data, and we also performed evaluation experiments using them. We found that our proposed acoustic feature could sufficiently improve the performance, and our multi-modal SOE successfully achieved obtaining almost 80% F-measure.

Our future works include investigation of using raw audio signals instead of MFCC in SDA, and evaluation of proposed features and SOE scheme in different environments. Because service operations are not always exclusive in some cases, that means features may belong to a couple of operations, a suitable evaluation scheme should be also explored. In this paper we combined non-acoustic features and PCA-applied acoustic BOFs to adjust their contributions in SVM. We will further investigate how to integrate both information more efficiently and to compact acoustic features using the other techniques e.g. Latent Semantic Indexing (LSI).

REFERENCES

[1] M.Takehara et al., "The role of speech technology in service-operation estimation," *Proc. Oriental COCOSDA 2011*, Oct 2011.

[2] T.Kawase et al., "Improvement of utterance clustering by using employees' sound and area data," *Proc. ICASSP 2014*, pp. 3071–3075, May 2014.

[3] M.Takehara et al., "Analysis of customer communication by employee in restaurant and lead time estimation," *Proc. APSIPA ASC 2014*, Dec 2014.

[4] K.Imoto, "User activity estimation method based on probabilistic generative model of acoustic event sequence with user activity and its subordinate categories," *Proc. INTERSPEECH 2013*, pp. 2609–2613, Aug 2013.

[5] S.Tamura et al., "A robust audio-visual speech recognition using audio-visual voice activity detection," *Proc. INTERSPEECH 2010*, pp. 2050–2053, Sep 2010.

[6] R.Tenmoku et al., "Service-operation estimation in a Japanese restaurant using multi-sensor and POS data," *Proc. APMS 2011*, Sep 2011.

[7] P.Vincent et al., "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, Jan 2010.

[8] T.Joachims, "Learning to classify text using support vector machines: Methods, theory, and algorithms," *Journal of Computational Linguistics*, vol. 29, no. 4, pp. 655–661, Dec 2003.

[9] "Pylearn2," http://deeplearning.net/software/pylearn2/.

[10] S.Pancoast et al., "Softening quantization in bag-of-audio-words," *Proc. ICASSP 2014*, pp. 1370–1374, May 2014.

[11] A.Plinge et al., "A bag-of-features approach to acoustic event detection," *Proc. ICASSP 2014*, pp. 3704–3708, May 2014.

[12] M.Takehara et al., "Improvement of service-operation estimation using voice activity detection of employee's speech," *IEICE Transaction on Information and Systems (Japanese Edition)*, vol. J97-D, no. 10, pp. 1563–1571, Oct 2014.