SUMMARY

In this paper we introduce a new framework of audio processing, which is essential to achieve a trigger-free speech interface for home appliances. If the speech interface works continually in real environments, it must extract occasional voice commands and reject everything else. It is extremely important to reduce the number of false alarms because the number of irrelevant inputs is much larger than the number of voice commands even for heavy users of appliances. The framework, called Intentional Voice Command Detection, is based on voice activity detection, but enhanced by various speech/audio processing techniques such as emotion recognition. The effectiveness of the proposed framework is evaluated using a newly-collected large-scale corpus. The advantages of combining various features were tested and confirmed, and the simple LDA-based classifier demonstrated acceptable performance. The effectiveness of various methods of user adaptation is also discussed.

key words: speech recognition, speech/non-speech discrimination, VAD, utterance verification, emotion recognition, hands-free, trigger-free, IVCD

1. Introduction

Speech interfaces exhibit their greatest advantages when they are implemented completely hands-free. In this paper, we aimed at developing a hands-free speech interface for home appliances, such as TVs and air conditioners. Voice is the most natural and convenient modality if user hands are occupied or unclean. Voice only interfaces, which are achieved by continuous sound acquisition without a trigger button, would particularly be the most convenient. Moreover, if users were reluctant to lean toward a microphone, distant-talk style interfaces would be favored. Obviously, the most convenient situation would be where users could ask a human servant through voice as “turn on the TV, please.” Therefore, why don’t we do the same with computers?

The hardest problem to solve in attaining such an interface is the frequent false alarms caused by continuous trigger-free sound acquisition. Since a real home environment is filled with all kinds of sounds, the system always receives various inputs. A heavy TV viewer may control the TV several hundred times a day, but the system receives more than a thousand other signals. Therefore, the false-alarm problem is more serious than the misrecognition problem. Although we already have many systems that recognize valid voice commands accurately, hands-free speech interfaces for home appliances cannot generally be accepted unless we succeed in developing accurate classifiers for valid and invalid sound inputs. Our target system must particularly reject most of the invalid inputs observed in real home environments.

We bring together a variety of speech and audio processing techniques to accomplish such a system. Automatic activation of the speech interface is basically defined as voice activity detection (VAD) or speech/non-speech discrimination. However, we have to eliminate not only non-speech inputs, but also irrelevant speech such as chats. The task of verifying utterances against given (recognized) keywords is called utterance verification, which is also important to attain a trigger-free system. Moreover, two identical words or phrases can occasionally be uttered with different intentions; the user may speak to the machine as a voice command or refer to the same word/phrase as a part of a human-to-human conversation. Hence, it is also necessary to recognize if the user uttered a voice commands with a clear intention to control the TV, or the utterance matched one of the voice commands by accident.

In this paper, we propose a new framework of speech and audio processing, referred to as Intentional Voice Command Detection (IVCD). The goal of IVCD is to determine whether or not the input sound is an intentional voice command to the speech interface. From a technical point of view, IVCD can be realized as a combination of various speech processing techniques such as VAD, utterance verification, and emotion recognition. From a practical point of view, it is important for IVCD to work effectively and robustly in real environments, whereas IVCD is as easy as simple VAD in a controlled environment. Therefore, our research was initiated by creating a large-scale corpus including various irrelevant sounds in the real world. The effectiveness of the proposed IVCD system was evaluated using this corpus.

The remainder of this paper is organized as follows. In the next section, we analyze our task phenomenologically, and describe how these phenomena can be interpreted as feature parameters. In Sect. 3, the details on our corpus are described. In Sect. 4, we introduce the concept of IVCD and describe our implementation. Section 5 presents the experimental results, and the last section gives the conclusions and future work.
2. Phenomenological Analysis of Speech and Non-speech Signals

Our target system focuses on the difference between valid commands and other sounds. A voice command is generated through a very complicated process from the human brain to the lips, and each subprocess adds a characteristics to the voice. Therefore, it is necessary to phenomenologically analyze the generation of voice commands to define the features for IVCD. Further, the generation of other sounds, whether by a non-command human voice or not, includes various human activities and other phenomena. In this section, we describe how those phenomena can be observed in the sound signal, in view of previous work on various speech and audio processing applications.

2.1 Voice Activity

Detection of the presence of voice activity is the basis of trigger-free speech interfaces. From the viewpoint of human speech generation, the most important process is the vibration of vocal chords caused by air-flow from the lungs. This can be detected by measuring the sound wave power, which is the most popular feature of voice activity detection (VAD).

VAD has a long history of research. It was originally a part of speech communication systems [1] and used to reduce the bandwidth utilization and computational cost by removing silent parts from the input signal. VAD was also later applied to speech recognition systems, in which the precise endpoints of the utterance were estimated. In our IVCD system, a simple VAD precedes the more sophisticated classification parts to remove obvious non-speech part. In quiet situations, the presence/absence of voice activity can be observed as the presence/absence of sound wave power.

Various improvements have been added to the simple power-based VAD. The zero crossing rate was also used in some previous work [2] to detect unvoiced sounds near the boundary of the utterance, which is difficult to detect by only using the power. In a more sophisticated manner, only the periodic components were extracted to calculate the power [3], so that the VAD method was robust in the presence of aperiodic noise.

2.2 Phonetic Characteristics

If there is background noise, more detailed features are required to discriminate speech from non-speech. Such features can be obtained if we shift our focus from the vocal chord to the vocal tract. Within a time-scale of 10 to 20 ms, these features mostly represent the characteristics of phonemes. Since the recognition of phonemes is the main part of automatic speech recognition, we can use the same features for our task, typically Mel-frequency cepstral coefficients (MFCCs). It has been reported that simple cepstrum-based VAD exhibits acceptable performance [4], and it can be improved by introducing a speech model created from the training data. Such a model-based approach can also be extended to audio classification [5], in which speech is one of the classes as well as music and noise. Gaussian mixture model (GMM) is often used to model speech [6], where a non-speech model is created in the same manner for comparison. It is also possible to model the dynamics of speech and non-speech generation using a switching Kalman filter [7].

In addition to the discrimination of phonemes, it is also useful for IVCD to analyze how each phoneme is pronounced. It has been reported that cepstral features convey a certain amount of information about the emotional state of the person speaking [8]. Since the user’s intentions may influence his/her attitudes, cepstral features can be used to discriminate between the intentional voice commands and other utterances.

2.3 Linguistic Characteristics

Within longer time-scales, the role of our brains becomes more important, which reflects the linguistic nature of human speech generation. Phonemes in human speech are interrelated in various ways, and the likelihood that they belong to a specific class of speech can be estimated using various measures in their interrelation.

Although our target is not equal to utterance verification, estimating the likelihood to be an in-vocabulary word is helpful to our task from a Bayesian perspective, because an in-vocabulary utterance is more likely to be a voice command. Therefore, the features used in utterance verification and out-of-vocabulary (OOV) detection are promising candidates of the IVCD feature. Utterance verification and OOV detection can be realized by applying a threshold to the normalized likelihood score, also as known as the confidence measure. The raw likelihood score of the best hypothesis is normalized using the likelihood scores of other words, unconstrained phoneme or HMM state sequences, etc [9]–[11].

Even without lexical knowledge, humans can classify speech into various classes. For instance, we can identify the language by listening to speech, without explicitly recognizing it. That task, known as language identification, provides some insight into IVCD. Even when the speaking style, background noise, and lack of lexical information on the OOV utterances make it difficult to recognize each phoneme accurately, imperfect phoneme recognition can help the classification task with assistance from class-labeled training data. Such an approach is called phone recognition followed by language modeling (PRLM), and has proved to be more effective than the GMM-based approach for the task of language identification [12].

2.4 Prosodic Characteristics

Prosody conveys various kinds of information attached to
speech [13], including linguistic information in tonal languages and paralinguistic information, such as emotion, in almost all languages. In most of the previous work, prosody has played an important role in enabling the speaker’s emotional state to be recognized [14], [15]. Moreover, prosodic features can be used to discriminate the positive and negative attitudes of the speaker from the same phrase [16]. Based on the previous work, we expect that the difference in user’s attitudes (asking a machine or talking to other people) can also be recognized in a similar manner.

2.5 Spectral Characteristics

Since some sounds in the real world are narrow-banded, spectral features can effectively be used for audio classification. Various frequency-domain features have been proposed for audio classification [17], [18], including subband powers, bandwidth, and spectral centroid. It was reported that reasonable results were obtained by using both cepstral and frequency-domain features in the experiments to classify instrumental sounds, mechanical noises, human voices and other sounds [18]. Spectral entropy, which is another simple form of spectral features, has also been proved to be effective for VAD [19].

3. HITHOME07/08 - Data Collection in Real Home Environments

3.1 Data Collection

The success of speech recognition systems has been heavily dependent on the variety of speech corpora. If we only focus on controlled speech, clean speech corpora such as JNAS [20] and CSJ [21] would suffice. If we focus on a specific type of noise, structured noise databases such as JEIDA Noise Database and RWCP Sound Source Database [22] would be helpful. However, we have to deal with all kinds of speech and noise that can be detected in the target environment. We have to know what kinds of speech and noise exist, how often each of them is detected, and how harmful each noise is for the speech interface. Since there are no publicly-available corpora which provide these types of information, we created a proprietary database for a specific task [23]. The requirements for the database were as follows.

(a) The environment had to look like a typical home, in respect to the room size, wall material, and distribution of furniture.
(b) The subjects had to behave as if they were in their own home.
(c) Voice command utterances had to be included.
(d) Every sound event had to be recorded.
(e) A distant microphone had to be used for recording.

where (a) to (d) are essential for a trigger-free speech interface in home environments, and (e) is an additional requirement for a user who is reluctant to lean toward a microphone. To satisfy these requirements, we rented an apartment, arranged it like a typical living room, hired some “inhabitants” who were from the same family or friends and asked them to act just as they would be in their own home (for requirements (a) and (b)), except that they had to use the speech interface to control the TV (for (c)). All the sounds were recorded from morning to evening (for (d)).

Data collection was conducted in two separate rounds. The dataset collected in the first round is called HITHOME07, and the other is called HITHOME08. Figures 1 and 2 outline the setups for each round. The greatest difference between the two rounds is in the positions of the microphones. In HITHOME07, the user held a remote controller with microphones. In HITHOME08, the microphones were attached to the ceiling. Therefore, HITHOME07 only satisfies the requirements (a) to (d), whereas HITHOME08 satisfies them and additional requirement (e). Figure 3 has a photograph and a floor plan of one of the rooms used for HITHOME08 data collection. Here, the microphone was located above the table between the TV and sofa, where the largest portion of the IVC utterances would be made. If so, the distance between the mouth and microphone was about 1.5 m. For data collection purpose, all sound signals were sent to a multi-track recorder (MTR) and recorded ceaselessly.

The input signal was also used to control the TV, so that a natural user attitude for the speech interface could be maintained. Although our target was a trigger-free input.
Fig. 3 Photograph and floor plan of one of the rooms used for HITHOME08 data collection.

Table 1 Age and gender distributions of HITHOME07/08 inhabitants.

<table>
<thead>
<tr>
<th>Age</th>
<th>HITHOME07</th>
<th>HITHOME08</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>10-19</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20-29</td>
<td>17</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>30-39</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>40-49</td>
<td>4</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>50-59</td>
<td>5</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>60-69</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>45</td>
<td>43</td>
</tr>
</tbody>
</table>

The data collection of HITHOME07 was conducted in two locations for 36 days in total. HITHOME08 was collected in two other locations for 49 days in total. Each daily session was approximately 7.5 hours long from morning until evening. One, two or three new inhabitants came each day, and they were asked to cook their lunch, and to use the vacuum cleaner and washing machine at least once a day. Table 1 lists the age and gender distributions of inhabitants, which indicates a good coverage of males and females from their 20’s to 60’s. It should also be noted that none of the inhabitants were engaged in speech research or speech-related industries.

3.2 Preliminary Analysis

After the recording sessions were finished, we had 278 hours (HITHOME07) and 375 hours (HITHOME08) of unsegmented data and timestamps of the push button. These data were segmented using a simple power-based VAD program with a rather conservative threshold, resulting in huge amount of short segments. When VAD was applied, the duration of a segment was limited up to 5 seconds (HITHOME07) or 30 seconds (HITHOME08). Assuming that a conservative threshold achieved coverage of all IVCs, we labeled all the short segments as either IVC or Non-IVC. In addition, IVCs were subcategorized into command categories and OOV, and Non-IVCs were subcategorized into laughs/chats††, coughs, and noises. Tables 2 and 3 show the details of the segment classification.

As the labeling results had shown, the most significant counterpart of IVC in typical home environments is not mechanical noise, but human voices such as laughing voices and chats. Approximately 87-89% of the segments belong to the laughs/chats class. This reveals that IVCD is not equivalent to VAD, and various techniques must be added to VAD to achieve high IVCD performance.

4. Proposed IVCD Framework

4.1 Task Analysis of IVCD

Basically, IVCD is defined as a simple two-class classifi-
cation task. However, a simple classification rate is not a good indicator of its performance. Since about 97% of the segments are Non-IVCs, 97% classification rate can be achieved, simply by always giving a Non-IVC label. A more reasonable criterion is the average classification rate (ACR), which is the average of IVC acceptance rate (IAR) and Non-IVC rejection rate (NIRR). Assuming that \( n_1 \) IVCs were accepted, \( n_2 \) IVCs were rejected, \( n_3 \) Non-IVCs were accepted, and \( n_4 \) Non-IVCs were rejected, the following definitions are obtained.

\[
IAR = \frac{n_1}{n_1 + n_2} \times 100\% \quad (1)
\]

\[
NIRR = \frac{n_4}{n_3 + n_4} \times 100\% \quad (2)
\]

In addition, from the viewpoint of information retrieval, precision (\( P \)), recall (\( R \)), and F-measure (\( F \)) are defined as follows.

\[
P = \frac{n_1}{n_1 + n_3} \times 100\% \quad (3)
\]

\[
R = \frac{n_1}{n_1 + n_2} \times 100\% (= IAR) \quad (4)
\]

\[
F = \frac{2PR}{P + R} = \frac{2n_1}{2n_1 + n_2 + n_3} \times 100\% \quad (5)
\]

Basically, ACR is a good performance indicator of an IVCD system because it deals with IVCs and Non-IVCs equally. However, frequent false alarms are more annoying for the user than an occasional lack of responses to IVCs, so we have to pay attention to the F-measure when we discuss practical issues with home appliances.

4.2 Features and Classifiers

Our IVCD framework is comprised of a feature extractor and a classifier[24], just like conventional machine learning approaches. Since IVCD has common ground with other audio and speech processing tasks, the various features used in these tasks can be imported.

Table 4 is the list of features used in our IVCD framework. The input data were sampled with 16 kHz frequency, and frame-based features were calculated every 10 ms with frame sizes of 26 ms (MFCC) and 40 ms (F0). MFCC is the most popular feature used in speech recognition. It is also used for VAD, emotion recognition, audio classification, etc. by some researchers. In our system, thirteen dimensional MFCCs (including 0-th) were calculated for each frame and then abstracted by their average and standard deviation over the segment. Including AMFCC standard deviations, there were 39 cepstral features.

GMM scores are often used for VAD. We created IVC GMM and Non-IVC GMM using the 13 dimensional MFCCs, and calculated corresponding scores for each segment. The confidence measure generated by an automatic speech recognition (ASR) system was also added, which is the principal feature in utterance verification.

Prosodic features were borrowed from emotion recognition. F0-based and log-power based features were calculated from corresponding frame features. The YIN algorithm[25] with minor modifications was used to estimate F0. These features were abstracted by using the maximum, minimum, position of the maximum and minimum, regression coefficient and mean square error, average and standard deviation. The F0 of the first and last voiced frames were also used. The voiced/unvoiced features consist of the number of voiced/unvoiced frames, the number of voiced/unvoiced regions, and the length of the longest voiced/unvoiced regions. The average and standard deviation of jitter (difference in the pitch period between adjacent frames) were also used.

The spectral features were borrowed from audio classification. For each of eight subbands, the average, standard deviation, and entropy were calculated over the segment using the frame-wise subband log-power. In addition, four more features related to fullband power and spectral entropy were added.

PRLM (phone recognition followed by language modeling) features were borrowed from language identification. Unconstrained monophone recognition was conducted using MFCCs and the standard acoustic model. Two language models were trained using IVC data and Non-IVC data, and they add two language model scores to the acoustic model score. The same procedure was repeated by using syllable recognition (SRLM), and we have six features in total.

After that, we added two more ad-hoc categories of features. The inter-segmental features were calculated using information outside the segment, such as the observation frequency of recent segments, power and cepstrum of the previous segment, etc. The final class of features was calculated by changing the preceding segmentation threshold.
We counted the number of segments and fragments included in the original segment, where the fragment means the active frame sequence divided by short pauses. The length of the largest segment was also added.

As for the classifier, we evaluated linear discriminant analysis (LDA), decision tree (DT), and support vector machine (SVM). These classifiers are efficient and easy to implement. LDA is simple enough, but its classification ability is heavily dependent on the design of the feature set. DT is advantageous if we want to interpret the trained model and reuse it, but it may suffer from overfitting problems. SVM is a strong classifier that can even model nonlinearly distributed objects. However, its training process is relatively slow.

5. Experimental Results

5.1 Comparison of Features

We carried out a set of evaluation experiments on our IVCD framework using HITHOME07/08. Prior to the experiments, each dataset was divided into two subsets according to the data collection location. GMM, monophone LM, and syllable LM were made for IVC and Non-IVC of each subset of each dataset (in short, 8 GMMs and 16 LMs in total). The test data of Location A of HITHOME07 were scored using models trained by the Location B data of HITHOME07 and vice versa. The same procedure was applied to HITHOME08. In addition, all segments were recognized by Julius [26] using the same command lexicon as in online recognition. For this purpose, we used the acoustic model trained by a clean speech corpus, which was attached to the Julius distribution. The recognition results were discarded, and only the confidence scores were used.

All the experiments were conducted by leave-one-out cross validation within each dataset. When one daily session of HITHOME07 was used for test, the other 35 daily sessions were used for training of the classifier, and this was repeated for all daily sessions. HITHOME08 was evaluated in the same manner. The final performance measures such as IAR, NIRR, and F-measure were calculated by counting the classification results for all segments.

Figure 4 shows the ROC (receiver operating characteristic) curves of HITHOME07 obtained by various features with LDA. The ROC curve represents the relationship between IAR and NIRR. In these experiments, all IVC segments were given larger weight (≈ 26.6) so that the separating hyperplane put equal weights on IVCs and Non-IVCs. After LDA coefficients were calculated, a variable threshold was applied to obtain the ROC curve. In addition to the six categories listed in Table 4, simple thresholding of the average power was plotted. Finally, the combination of all features in Table 4 was evaluated. Although all curves have similar shapes, it is obvious that the best performance was obtained by using the combined features. The same results were replotted from the viewpoint of the F-measure, as shown in Fig. 5. If we use the combined feature and threshold obtained by LDA to optimize the F-measure, we would have a false alarm every 28 minutes, whereas 79.7% of IVCs are correctly accepted.

The results for the same experiments using HITHOME08 are shown in Figs. 6 and 7. The weight for the IVC segments was about 29.0. In the best case using the combined feature, we have a false alarm every 26 minutes, whereas 67.1% of IVCs are correctly accepted.

In the HITHOME07 experiments, cepstral, spectral, and prosodic features are highly ranked. Taking into account that cepstral features include $C_0$, these results can be attributed to the power. This is reasonable because many users held the remote controller when they talked to the speech interface. In contrast, Power-independent features such as PRLM/SRLM and GMM score were ranked higher.
in the HITHOME08 experiments. Although power itself was ranked as the second worst in both cases, it was rather competitive against other features in the former.

It might have been unexpected that the ASR confidence measure was the worst feature. This can partly be attributed to the dimensionality of the feature. It must also be noted that the acoustic model was not matched to the environment. Introducing various robust speech recognition techniques would improve IVCD accuracy using the ASR confidence measure, although this is beyond the scope of this paper.

Figure 8 plots the difference in ACR for daily sessions with various features. Although we cannot distinguish inhabitants in the same daily sessions, these results imply variations in individuals. Power-based IVCD, which is the second worst of the various features, also has the largest standard deviation. In the case of the combined feature, the standard deviation of ACR is 4.0% for HITHOME07 and 7.0% for HITHOME08. There is an overall tendency for the feature with lower ACR to have the larger standard deviation, but the only exception is the ASR confidence measure, whose standard deviation is 6.7% for HITHOME07 and 6.6% for HITHOME08.

Figure 9 plots the trend in ACR during 7.5 hours from morning until evening. The leftmost plotting point represents the ACR of the segments observed in the first 30 minutes for all sessions, and other points represent corresponding 30 minute ACRs. Since no on-line IVCD was used during data collection, the only feedback which the inhabitants received was whether each IVC utterance was correctly recognized or not. Consequently, no signs of adaptation of user attitudes can be observed in Fig. 9. Observing changes in user attitudes through on-line IVCD experiments is important work we intend to do in the future.

More detailed comparison of features was done by feature selection experiments using Backward Stepwise Selection (BSS), where we started by using all the available features, and found a feature whose removal provided the most accurate result of all possible removals. One feature was removed by using this single step, and the same procedure was repeated until we only had one feature. The results for the HITHOME07 dataset are shown in Fig. 10. By using only 20 features, both the ACR and F-measure reached almost the same value as the full 115 features.

Moreover, we conducted random feature selection experiments. Ten of 115 features were randomly sampled. After 100,000 trials, we picked up 1% best trials and counted the appearance of features in these trials. Table 5 shows the 10 best features obtained through these experiments. In the column of BSS, the best feature set of BSS experiment is shown. In the column of Random, the 10 most frequently-appearing features in the top 1% trials are shown. In both cases, best features vary between ACR-based and F-measure-based selections. Of the 115 features, E(C₃) and ASR-CM seem to be very important, as well as $\sigma(\log(P))$
Table 5  Ten best features for HITHOME07, obtained by BSS and random selection.

<table>
<thead>
<tr>
<th>Feature</th>
<th>ACR</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BSS</td>
<td>Random</td>
</tr>
<tr>
<td>E(C_i)</td>
<td>E(C_i)</td>
<td>E(C_i)</td>
</tr>
<tr>
<td>E(\sigma(C_i))</td>
<td>E(\sigma(C_i))</td>
<td>E(\sigma(C_i))</td>
</tr>
<tr>
<td>ASR-CM</td>
<td>ASR-CM</td>
<td>ASR-CM</td>
</tr>
<tr>
<td>ASR-CM</td>
<td>ASR-CM</td>
<td>ASR-CM</td>
</tr>
<tr>
<td>max(L_v)</td>
<td>max(log(P))</td>
<td>max(log(P))</td>
</tr>
<tr>
<td>\sigma(B_i)</td>
<td>\sigma(B_i)</td>
<td>\sigma(B_i)</td>
</tr>
<tr>
<td>\sigma(B_5)</td>
<td>\sigma(B_5)</td>
<td>\sigma(B_5)</td>
</tr>
<tr>
<td>\sigma(\epsilon(B_7))</td>
<td>\sigma(\epsilon(B_7))</td>
<td>\sigma(\epsilon(B_7))</td>
</tr>
<tr>
<td>\epsilon(B_7)</td>
<td>\epsilon(B_7)</td>
<td>\epsilon(B_7)</td>
</tr>
<tr>
<td>BSS</td>
<td>BSS</td>
<td></td>
</tr>
<tr>
<td>\epsilon(\epsilon(B_7))</td>
<td>\epsilon(\epsilon(B_7))</td>
<td>\epsilon(\epsilon(B_7))</td>
</tr>
<tr>
<td>L_{SIVC}</td>
<td>L_{SIVC}</td>
<td>L_{SIVC}</td>
</tr>
<tr>
<td>L_{NonIVC}</td>
<td>L_{NonIVC}</td>
<td>L_{NonIVC}</td>
</tr>
</tbody>
</table>

Fig. 11  Comparison of classifiers (HITHOME07). Numbers in the parentheses are the highest ACR for each classifier.

Fig. 12  Comparison of classifiers (HITHOME07). Numbers in the parentheses are the highest F-measure for each classifier.

Table 6  Comparison of proposed method and previous work. GMM and PercCeps8 were implemented according to [6] and [18], respectively. Proposed method was evaluated using combined feature and SVM.

<table>
<thead>
<tr>
<th></th>
<th>HITHOME07</th>
<th>HITHOME08</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACR (%)</td>
<td>F-meas. (%)</td>
</tr>
<tr>
<td>GMM</td>
<td>88.5</td>
<td>39.7</td>
</tr>
<tr>
<td>PercCeps8</td>
<td>94.2</td>
<td>61.7</td>
</tr>
<tr>
<td>Proposed</td>
<td>94.4</td>
<td>62.1</td>
</tr>
</tbody>
</table>

However, even if we eliminate these five features, 91.7% of ACR and 75.9% of the F-measure were obtained in the best case, which indicates the redundancy of our feature set.

5.2 Comparison of Classifiers

Next, various classifiers such as LDA, DT, and SVM were compared using the combined feature and the HITHOME07 database. For the DT and SVM experiments, we used an open-source data mining software called WEKA [27]. A two-dimensional polynomial kernel was used for SVM. Since the SVM training process using a large amount of data does not converge within a reasonable time, only 1/20 of Non-IVC segments were randomly sampled for training, while the IVC segments were all used. A few pairs of IAR and NIRR (or precision and recall) were obtained by giving various weights to IVC segments.

The comparison results are shown in Figs. 11 and 12. The accuracy of DT is clearly worse than those of LDA and SVM for both datasets. Since the feature vector has very large dimensions and there are only small number of IVC samples, though we have many Non-IVC samples, such degradation may be attributed to overfitting of the decision tree. The accuracy of SVM is almost the same as that of LDA, which is inconsistent with our previous work [24]. An important modification from [24] is that all power-related features have been replaced with their logarithms. The trends in Figs. 11 and 12 suggest that the shape of feature space was smoothed by the sophistication of the features.

5.3 Comparison with Previous Work

Next, we compared our results with two typical approaches [6], [18]. GMM-based rejection, proposed by Lee et al. [6], is widely used as a front-end of speech recognition/dialog systems. They used five classes of child, adult, laughter, coughing, and noise to create GMMs. The label definition in our experiment was not the same as theirs, and we created GMMs for in-vocabulary, out-of-vocabulary, chat/laugh, coughing, and noise. We conducted cross validation experiments between two locations, and the exact class matching rates were 84.8% for HITHOME07 and 75.8% for HITHOME08. After merging the first two classes as IVC and the other three as Non-IVC, the ACR and F-measure were those listed in Table 6. Obviously, the ACR and F-measure of the GMM-based method were much lower than those of the proposed method.\footnote{Difficulty with threshold optimization affects the F-measure more due to many false acceptances of Non-IVCs. Therefore, the F-measures of all three methods were much lower than those in Figs. 5 and 7.}

The audio classification method proposed by Guo and Li [18] has a lot in common with ours, and is a suitable target for comparison. Of the various features they tested, we...
chose the so-called PercCep8, which they found provided the best performance. The PercCep8 features consist of full-band and subband powers, brightness, bandwidth, pitch, and 1st to 8th MFCCs. Their average and standard deviation over the segment were calculated, and the pitched ratio was added. The silence ratio was removed in our implementation because our segments did not have any explicit boundary outside of the active period. Instead, subband power from 4 to 8 kHz was added, since our sampling frequency was twice as large as theirs. By applying SVM, the results in Table 6 were obtained. Their method performed very closely to ours for HITHOME07, but the method we propose still had tremendous advantages for HITHOME08.

5.4 User Adaptation

More detailed analysis of the experimental results have shown that the ACR and F-measure of IVCD varied considerably from day to day. For HITHOME07, the daily ACR ranged from 83.3% to 99.6%. The daily F-measure ranged from 46.5% to 96.5%. Therefore, it would be an effective approach to focus on some specific days when there are more errors.

Adaptation for specific conditions can be categorized into two cases. Supervised adaptation assumes labeled adaptation data. Unsupervised adaptation does not require pre-labeled data, and usually the adaptation data are given by automatic labeling.

In Table 7, some adaptation methods are listed with the corresponding ACR and F-measure in the case of HITHOME07. Method 1 is the baseline shown in Figs. 4 and 5 as “Combined.” First, we assumed that it would be reasonable to ask the user to only adjust one parameter of the threshold (this is like a TV having microphone gain control). In methods 2 and 3, only the threshold was adapted. In 2, the daily data were divided into five subsets, four of these were used for adaptation, and the remaining subset was evaluated. In method 3, all the daily data were used for adaptation and evaluation. They can be interpreted as occasional (method 2) and precise (method 3) threshold tuning. Even method 2 made reasonable improvements to the F-measure, while method 3 was very effective as expected. In method 4, only the training data from the same day were used to obtain the LDA coefficients and threshold. The poor results are due to the lack of data. In method 5, the adaptation data were mixed with the open training data, and the LDA coefficients and threshold were both re-calculated. Despite its heavy computational cost, it could not exceed threshold-only adaptation.

Figure 13 shows the session-wise ACRs of three methods: power only, combined, and combined with threshold adaptation. As discussed with Fig. 8, power-only IVCD has a large variation of the session-wise ACR, and the results include as bad a session as 65.3% of ACR. The variation becomes smaller if the combined feature was used, and even smaller by threshold adaptation.

Finally, we tried unsupervised adaptation. Since the numbers of IVC and Non-IVC were ill-balanced, automatic labeling would create more confusion. Instead, we applied simple normalization to each of the 115 features by removing the average and dividing by the standard deviation. In method 6, all the daily data were used to calculate the average and standard deviation. In method 7, only the data up to the current sample were used with MAP estimation (the MAP weight was set at 100). However, we only achieved minor improvements in both cases.

6. Conclusions

In this paper, we introduced a new framework for audio processing, referred to as Intentional Voice Command Detection (IVCD), which is one of the key functions to realize trigger-free speech interface, especially in home environments. This framework was constructed by combining a classifier with various features representing various sound-producing phenomena. Utilization of these features was inspired by various speech and audio processing tasks, such as VAD, utterance verification, emotion recognition, language identification, and audio classification.

We evaluated the proposed framework using a large scale corpus which we collected in real home environments. Preliminary analysis of the corpus suggested that the key issue is how to distinguish between intentional commands and other voices. From the viewpoint of the average classification rate, we achieved 94.5% for the remote-controller microphone and 92.0% for the ceiling microphone. From the viewpoint of the F-measure, which would be more realistic
in terms of user feelings, our results were 82.4% and 72.5% respectively, which are still acceptable. A more encouraging finding is that the accuracy could be greatly improved by optimizing only one parameter of the classification threshold.

In addition to the contribution made by this paper, which focused on single-microphone stand-alone implementation, two important things are left for future work. First, some of the false alarms caused by TV sounds can be removed if the system is equipped with an echo canceler. We did not consider this option because our targets included various unnetworked appliances, but it may be possible in the future that all appliances exchange reference signals through a network. Second, the direction-of-arrival information would be helpful for IVCD. Most people prefer staying at a few specific positions in the room, and many noise sources do not move. Introducing recent achievements in microphone array research would result in additional improvements to IVCD accuracy.

Acknowledgements

The authors are thankful to Prof. Sadaoki Furui of Tokyo Institute of Technology and Prof. Tetsunori Kobayashi of Waseda University for their valuable comments. This work was supported by the Ministry of Economy, Trade and Industry, Japan.

References

[27] The University of Waikato, Weka 3 - Data mining with open source machine learning software in java (online), <http://www.cs.waikato.ac.nz/ml/weka/> (confirmed on 2009-10-28)
Yasunari Obuchi received the B.S. degree and the M.S. degree in physics in 1988 and 1990 respectively, and the Ph.D. in information science and technology in 2006, all from the University of Tokyo. Since 1992, he has been working for Central Research Laboratory and Advanced Research Laboratory, Hitachi Ltd. He worked for Carnegie Mellon University as a Visiting Researcher from 2002 to 2003. He has also been a Visiting Researcher at Waseda University from 2005 to 2010. Currently he is a Senior Researcher of Central Research Laboratory, Hitachi Ltd. His research interests include robust speech recognition, spoken dialog systems, speech recognition in small devices, speech-to-speech translation, language identification, emotion recognition, and microphone arrays. Dr. Obuchi was a co-recipient of the Technology Development Award of the Acoustical Society of Japan in 2000. He is a member of IEEE, IPSJ, and ASJ.

Takashi Sumiyoshi received the B.S. degree and the M.S. degree in informatics in 2001 and 2003 respectively from Kyoto University, Japan. Since 2003, he has been working for Central Research Laboratory, Hitachi Ltd. His research interests are speech recognition, speaker diarization, spoken document retrieval, etc. He is a member of IPSJ and ASJ.