INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multimodal Response Obligation Detection with Unsupervised Online Domain Adaptation

Shota Horiguchi, Naoyuki Kanda, Kenji Nagamatsu

Hitachi, Ltd.

{shota.horiguchi.wk, naoyuki.kanda.kn, kenji.nagamatsu.dm}@hitachi.com

Abstract

Response obligation detection, which determines whether a dialogue robot has to respond to a detected utterance, is an important function for intelligent dialogue robots. Some studies have tackled this problem; however, they narrow their applicability by impractical assumptions or use of scenario-specific features. Some attempts have been made to widen the applicability by avoiding the use of text modality, which is said to be highly domain dependent, but it decreases the detection accuracy. In this paper, we propose a novel multimodal response obligation detector, which uses visual, audio, and text information for highly-accurate detection, with its unsupervised online domain adaptation to solve the domain dependency problem. Our domain adaptation consists of the weights adaptation of the logistic regression for every modality and an embedding assignment for new words to cope with the high domain dependency of text modality. Experimental results on the dataset collected at a station and commercial building showed that our method achieved high response obligation detection accuracy and was able to handle domain change automatically.

Index Terms: response obligation detection, addressee estimation, domain adaptation

1. Introduction

Current advances in various key technologies, e.g. automatic speech recognition (ASR), language understanding, and dialogue management systems, have improved the practicality of dialogue robots. Dialogue robots are typically designed to respond to each detected utterance. However, the utterances are not necessarily directed to the robots; they might be human-human conversations, monologues, or public announcements. These utterances potentially cause unnecessary and wrong responses from the robots. Therefore, it is very important for the robots to have the ability to judge whether they have to respond to the utterance. This task is called response obligation detection (ROD).

Our research target in this paper is ROD that does not restrict the operating environment. Two main factors restrict the operating environment, i.e. narrow the applicability, of ROD. The first one is the strategy of ROD. For example, use of an explicit key to notify robots that a following utterance is directed to them, such as a wake-up-word [1, 2, 3], is proven to be effective for highly accurate ROD. However, it restricts users to only ones who know the key phrase in advance, so it is not fit for robots that operate in various environments. Therefore, ROD should at least be conducted by only implicit information such as gaze and voice tone, which could limit the accuracy of ROD. The second one is the datasets used to train a response obligation detector. The robot operation environment, i.e. target domain, may be very different from the training data collection environment, i.e. source domain; therefore, the detector may not work accurately when in operation. It is necessary for the robots to resolve this domain conflict.

Conventional studies mainly focused on how to achieve highly-accurate ROD using humans' behavioral cues of multimodality [4, 5, 6, 7, 8, 9, 10, 11, 12]. Although they do not use an explicit key, they still contain factors that narrow the applicability of ROD, such as use of scenario-specific features that cannot be used in other scenarios [7, 5], assumptions that strictly hinder the availability of the applications [5, 4], and the use of external sensors installed in the environment [6, 7].

On the other hand, a few attempts have been made to resolve the conflict between source and target domains [6, 12]. They noted that text modality severely depends on the scenario; thus, they avoided using the text information to improve the domain independency. However, because text information contributes highly to performance improvement, and text itself can be obtained regardless of the kind of tasks, it should be manageable for ROD to utilize the text information. Domain adaptation (DA) could be a solution for this, but as far as we know, there is no study that resolves the domain dependency of ROD by DA.

For not only highly-accurate but also highly-applicable ROD, we propose a novel multimodal response obligation detector and its unsupervised online DA method. Our detector combines the result from visual, audio, and text-based detectors, and no scenario-specific feature is used. The scenario dependency of the detector, especially the text-based detector, is resolved by DA that consists of (i) an update of the logistic regression part of every modality and (ii) an embedding assignment for new words to resolve the heavy scenario dependency of text modality. To evaluate our method in a real environment, we built a spoken dialogue corpus by operating a humanoid robot at a station and a commercial building. The experimental results showed that our proposed method attains high ROD accuracy by using multimodality and automatically fits other domains by using our DA.

2. Related work

2.1. Response obligation detection

A common approach of ROD is to use multimodal inputs to attain high accuracy, e.g. text and image features [4], utterance length, dialogue status, key word existence [5], and motion and acoustic features [6]. This is because those modalities complementarily contribute to accuracy improvement. It is also studied in the context of addressee estimation [7, 8, 9]. Deep learning-based methods have also been proposed [10, 11, 12].

However, most of the studies have components that narrow the applicability of the detectors. One example is that some methods are based on assumptions that strictly hinder the availability of the applications, e.g. fixing the number of people [5] or accepting only command-like inputs [4]. Using scenario-

http://dx.doi.org/10.21437/Interspeech.2019-1313

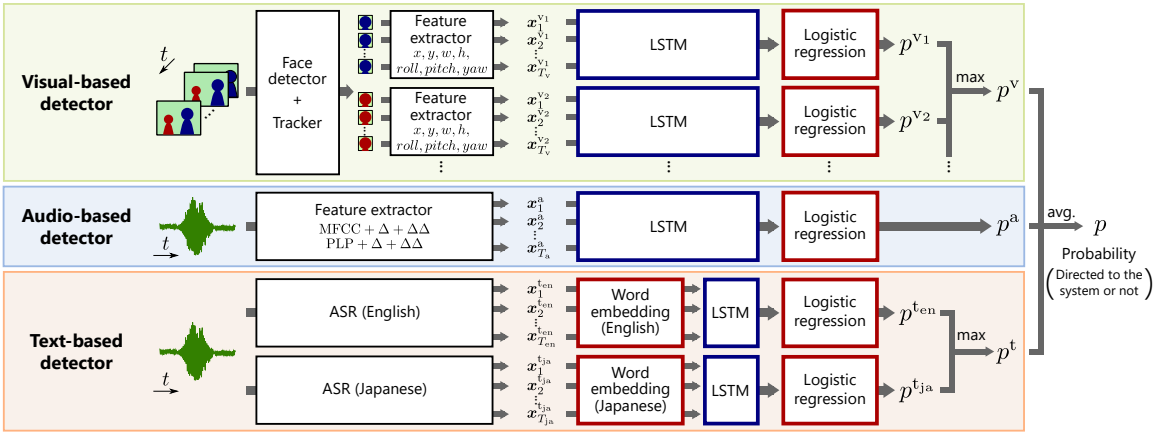


Figure 1: *Overview of our multimodal response obligation detector. Trainable parameters are in word embedding, LSTM, and logistic regression parts of the detector. When applying DA, we only update the bold red boxes, namely, the word embedding and logistic regression part, while fixing LSTM parameters depicted by the bold blue boxes.*

specific features such as the difficulty of a question that a robot sets in a quiz game scenario [7] or the existence of a limited number of keywords [5] also narrows the applicability. Moreover, the use of external sensors installed in the environment such as Kinect® [6] or Vicon® [7] becomes an obstacle for broad use of the robot. Because of these reasons, they are not fit for a dialogue robot that aims to operate in various environments.

Some attempts to widen the applicability of the detectors have been achieved by avoiding the use of text modality, which is said to be heavily domain dependent [6, 12]. However, this approach has a negative effect on detection accuracy. In this way, in general, there is a trade-off between the accuracy and applicability of ROD. In this paper, we utilize the text modality with DA to manage both accuracy and applicability.

2.2. Unsupervised online domain adaptation

DA is a technique for transferring knowledge from a source domain to a target domain. In particular, unsupervised online DA, which adapts an estimator using unlabeled target domain samples on-the-fly, has been used in tasks that are necessary for dealing with environments that vary from moment to moment, e.g. traffic camera analysis [13, 14, 15] or image recognition for kitting robots [16]. We believe that this approach is quite helpful for dialogue robots operating in various environments. In this paper, we use text modality to attain high accuracy while dealing with its scenario dependency by DA.

3. Method

3.1. Multimodal response obligation detection

For more accurate ROD, using multimodal information is the key. In this section, we introduce a multimodal ROD method that utilizes visual, audio, and text information.

Our multimodal response obligation detector is depicted in Fig. 1. We first detect the start and the end of an utterance by thresholding power of the input audio. For each detected utterance, we conduct ROD using images and audio captured during the interval between the start and end timing. The images and audio are processed by the following three detectors.

Visual-based detector: This detector decides the response obligation from the facial information of people in front of the robot. We first detect and track faces in the images using dlib [17]. For each detected face, we extract location (x, y), size $(width, height)$, and direction features $(roll, pitch, yaw)$. As a result, we obtain a seven-dimensional feature for each person per image. For each of N persons, we process the extracted features by 2-layer long short-term memory (LSTM) with 128 cells and obtain ROD probabilities $\{p_i^{\mathrm{v}}\}_{i=1}^{N}$. The final probability is obtained by choosing the maximum values, i.e. $p^{\mathrm{v}} = \max_i p_i^{\mathrm{v}}$.

Audio-based detector: This detector estimates the response obligation from acoustic features. We first extract MFCC + Δ + $\Delta\Delta$ and PLP + Δ + $\Delta\Delta$ features from input audio with 25 ms frame length and 10 ms frame shift, which results in a 78-dimensional feature per frame. We process them by 2-layer LSTM with 128 units, and obtain probability p^{a} by logistic regression.

Text-based detector: This detector estimates the response obligation on the basis of textual information. We first apply ASR to the input audio. Because our evaluation data contain English and Japanese, we use ASR for each language, and obtain $\{x_i^{\mathrm{t_{en}}}\}$ and $\{x_i^{\mathrm{t_{ja}}}\}$, respectively. The recognition results are processed by word embedding modules and are then input to 1-layer LSTM, obtaining probabilities for each language by logistic regression: $p^{\mathrm{t_{en}}}$ and $p^{\mathrm{t_{ja}}}$. Finally, we choose $p^{\mathrm{t}} = \max\left\{p^{\mathrm{t_{en}}}, p^{\mathrm{t_{ja}}}\right\}$.

The final detection probability is calculated by integrating outputs from each detector. In this paper, we use a simple average of them as follows:

$$p = \frac{1}{3}\left(p^{\mathrm{v}} + p^{\mathrm{a}} + p^{\mathrm{t}}\right). \tag{1}$$

3.2. Online domain adaptation for the multimodal detector

To guarantee the applicability of ROD, solving domain dependency of the model is necessary. In this section, we propose a DA method for the multimodal response obligation detector.

Our domain adaptation method is characterized by the following two parts: an update of the logistic regression part and an assignment of embeddings to new words, which do not appear in source domain datasets. Note that we fixed the parameters of the LSTMs to reduce the calculation cost for DA.

The algorithm of our domain adaptation is described in Algorithm 1. For preparation, we trained the multimodal ROD model described in Sec. 3.1 using a source domain dataset. We notated the LSTM and logistic regression parts of the model as $\mathcal{M}$, and the word embedding part as $\mathcal{W}_{src}$. Note that $\mathcal{W}_{src}$ includes a special embedding assignment for *unknown* word as in

Algorithm 1: Unsupervised online domain adaptation for multimodal response obligation detection.

Input: $\mathcal{D}_{tgt} = \left\{\left(\mathcal{X}_t^{\mathrm{v}}, \mathcal{X}_t^{\mathrm{a}}, \mathcal{X}_t^{\mathrm{t}}\right)\right\}_{t=1}^{T}$ // Target data
$\mathcal{M}$ // LSTM & regression part
$\mathcal{W}_{src} = \{(w_j, \boldsymbol{e}_j)\}_j$ // Embedding part
$K \in \mathbb{N}$ // # of cluster centers

1 $\{\boldsymbol{c}_k\}_{k=1}^{K} \leftarrow \mathsf{KMeansClustering}\left(\mathcal{W}_{src}, K\right)$
2 $\mathcal{W}_{tgt} \leftarrow \mathcal{W}_{src}$ // Word embedding for target data
3 $\mathcal{H} \leftarrow \emptyset$ // Array of tuples (word, embedding)
4 **for** $t = 1$ to T **do**
5 Predict label $\hat{y}$ of $\left(\mathcal{X}_t^{\mathrm{v}}, \mathcal{X}_t^{\mathrm{a}}, \mathcal{X}_t^{\mathrm{t}}\right)$ by $\mathcal{W}_{tgt}$ and $\mathcal{M}$
6 **for** $l = L_t$ to 1 **do** // $\mathcal{X}_t^{\mathrm{t}} = \left[x_{t,1}^{\mathrm{t}}, \ldots, x_{t,L}^{\mathrm{t}}\right]$
7 **if** $x_{t,l}^{\mathrm{t}}$ is not in $\mathcal{W}_{src}$ **then**
8 **if** $\hat{y}_t = 1$ **then** // positive label
9 Find k that maximizes the probability when $x_{t,l}^{\mathrm{t}}$ is embedded to $\boldsymbol{c}_k$
10 **else** // negative label
11 Find k that minimizes the probability when $x_{t,l}^{\mathrm{t}}$ is embedded to $\boldsymbol{c}_k$
12 Add $\left(x_{t,l}^{\mathrm{t}}, \boldsymbol{c}_k\right)$ to $\mathcal{H}$
13 Add/update embedding of $x_{t,l}^{\mathrm{t}}$ in $\mathcal{W}_{tgt}$ to the most assigned vector by referring to $\mathcal{H}$
14 Update regression part of $\mathcal{M}$ using $\mathcal{W}_{tgt}$ and $\hat{y}$

Table 1: *Details of the Station dataset and the Building dataset.*

Dataset	Duration	#Utterances	#Positives	#Negatives
Station	28:15:04	5306	2258	3048
Building	17:48:32	2238	924	1314

[18] so that we can calculate probability even when the input text contains a word that is not in $\mathcal{W}_{src}$. Embedding candidates for new words are calculated by applying k-means clustering to all the embedding vectors in $\mathcal{W}_{src}$ (L1). The word embedding on target domain datasets is initialized by that on source domain $\mathcal{W}_{src}$ (L2), and an array $\mathcal{H}$ for storing the embedding assignment history is also initialized by the empty set (L3). For each resulting target sample, we first calculate pseudo-label $\hat{y}$ by using current detector $\mathcal{W}_{tgt}$ and $\mathcal{M}$ (L4–5). $\hat{y}$ is calculated by thresholding the probability p evaluated by Eq. 1 as follows:

$$\hat{y} = \begin{cases} 1 & (\text{if } p > 0.5) \\ 0 & (\text{otherwise}) \end{cases}. \tag{2}$$

If the input text $\mathcal{X}_t$ contains a word $x_{t,l}^{\mathrm{t}}$ not in the $\mathcal{W}_{src}$ (L6–7), we first find the most appropriate embedding from the candidates; if the pseudo-label is positive, find $\boldsymbol{c}_k \in \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$ that maximizes the estimated probability when $x_{t,l}^{\mathrm{t}}$ is embedded to $\boldsymbol{c}_k$ (L8–9), and if the label is negative, find one that minimizes the probability (L10–11). Once the vector is found, we add a word-embedding pair $\left(x_{t,l}^{\mathrm{t}}, \boldsymbol{c}_k\right)$ to the embedding assignment history $\mathcal{H}$ (L12). Then the embedding for the word $x_{t,l}^{\mathrm{t}}$ is updated by the most frequent assignment by referring to $\mathcal{H}$ (L13). When all the words are processed, update the logistic regression parameters by stochastic gradient descent (L14).

4. Experiments

4.1. Dataset

To evaluate our method on real world datasets, we collected a spoken dialogue corpus by operating a humanoid robot shown in Fig. 2 at a station and a commercial building. Our robot conducted guidance operations for surrounding facilities and services. Interactions between humans and the robot are recorded by a microphone and RGB camera installed in the head of the robot as shown in Fig. 2. As a result of this operation, we obtained 16 bit / 16 kHz audio and 10 fps images, which are roughly synchronized. We detected utterances from the recorded audio by thresholding the audio power and annotated all the utterances with labels that indicate whether the robot has to respond to the utterances. The dataset details are shown in Table 1. In this paper, we refer to the dataset collected at the station as "Station," and that collected at the commercial building as "Building." For evaluation purposes, both datasets are divided into five subsets by recording dates, respectively.

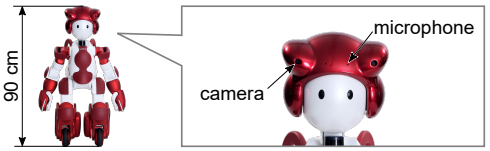


Figure 2: *A humanoid robot that we used for data collection.*

4.2. Evaluation protocol

In this paper, we conducted two kinds of experiments. One is in-domain evaluation (Sec. 5.1). In this case, we conducted offline 5-fold cross validation for each dataset using five subsets described in Fig. 4.1. Another is out-of-domain evaluation for calculating DA performance (Sec. 5.2). In this case, we first prepare two datasets: a source domain dataset and a target domain dataset. First, we trained the detector using the entire source domain dataset in an offline manner. Then we used four subsets of the target domain dataset for online DA of the detector and the remaining one for evaluation. We repeat this procedure five times to conduct 5-fold cross-validation.

We used two metrics for evaluation. One is the accuracy for evaluating the standard 2-way classification performance. However, in real operation, we will set the threshold to decide whether robots respond to an input or ignore it. To calculate the performance over various thresholds, we also evaluated the area under the receiver operating characteristic curve (ROC-AUC).

4.3. Implementation

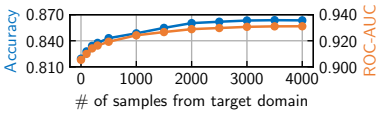
In the visual-based detection part, we used dlib [17] for face detection and face landmark detection. We calculated face direction by solving a perspective-n-point problem [19] on a reference 3D facial model and the detected face landmarks.

In the audio-based detection part, we used the Kaldi toolkit [20] for extraction of $\mathrm{MFCC} + \Delta + \Delta\Delta$ and $\mathrm{PLP} + \Delta + \Delta\Delta$, both of which are 39-dimensional features. When training the detector in in-domain evaluation or training the detector using the source domain in out-of-domain evaluation, we used 0.8x and 1.2x speed perturbation in addition to normal speed audio.

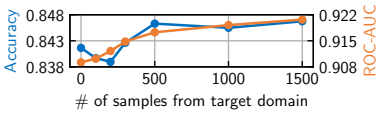
In the text-based detection part, we used our in-house speech recognition system specially tuned for the robot. When training the detector in in-domain evaluation or training the detector using the source domain in out-of-domain evaluation, we treat the words that emerged only once in an offline training set as the same word *unknown* to deal with new words following [18]. When evaluating performance in in-domain evaluation or training the detector using the source domain in out-of-domain evaluation, we used transcriptions of utterances in addition to ASR results to train the detector.

Table 3: *Out-of-domain ROD. We show the results with and without our unsupervised online domain adaptation. We also recapped the results from Table 2 when the detector was trained on the target domain in "Ours (Target only)" rows. "All" denotes that all the modalities, i.e. visual, audio, and text, are used for detection.*

Method	Source → Target	Accuracy				ROC-AUC			
		Visual	Audio	Text	All	Visual	Audio	Text	All
Ours (w/o DA)	Building→ Station	0.775	0.738	0.743	0.819	0.843	0.795	0.829	0.905
Ours (w/ DA)	Building→ Station	0.805	0.750	0.817	0.862	0.869	0.817	0.887	0.930
Ours (Target only)	Station → Station	0.807	0.810	0.849	0.894	0.879	0.876	0.908	0.951
Ours (w/o DA)	Station →Building	0.697	0.788	0.805	0.841	0.729	0.847	0.859	0.909
Ours (w/ DA)	Station →Building	0.729	0.786	0.782	0.849	0.776	0.853	0.876	0.920
Ours (Target only)	Building→Building	0.746	0.775	0.809	0.848	0.781	0.835	0.880	0.922



(a) *Building→Station*



(b) *Station→Building*

Figure 3: *Number of samples vs. detection performance of DA.*

Table 2: *In-domain ROD using various modalities.*

Dataset	#Modality	Modality			Accuracy	ROC-AUC
		Visual	Audio	Text		
Station	1 modal	✓			0.807	0.879
			✓		0.810	0.876
				✓	0.849	0.908
	2 modals	✓	✓		0.860	0.926
		✓		✓	0.863	0.936
			✓	✓	0.870	0.935
	3 modals	✓	✓	✓	**0.894**	**0.951**
Building	1 modal	✓			0.746	0.781
			✓		0.775	0.835
				✓	0.809	0.880
	2 modals	✓	✓		0.785	0.876
		✓		✓	0.803	0.895
			✓	✓	0.839	0.912
	3 modals	✓	✓	✓	**0.848**	**0.922**

5. Results

5.1. In-domain multimodal response obligation detection

Table 2 show the results of in-domain ROD on the Station and the Building datasets. Both accuracy and ROC-AUC were higher when a greater number of modalities were used for ROD.

Here we consider which modality contributed to the detection performance best. When only one modality was used, text modality performed better than visual or audio-based detection. When two modalities were used, visual and audio modalities, i.e. only text was not used, showed the worst performance. From these results, avoiding the use of the text modality [6, 12] is inappropriate for achieving highly-accurate ROD.

5.2. Unsupervised online domain adaptation for multimodal response obligation detection

We evaluated DA performance on the following two settings: Building→Station and Station→Building. The results are shown in Fig. 3. We adapted both the logistic regression part and word embedding part in this case. The number of clusters K was set to 10. We first observe that our DA successfully improved the detection performance. In the Building→Station setting, the accuracy on the text modality was the most degraded (-0.106) because of domain mismatch, but our DA successfully recovered the accuracy by 0.074. In the Station→Airport setting, not the text but the visual modality shows the worst performance degradation by domain difference. Our method also improved the detection performance on this setting.

Evaluation of embedding assignment for new words: To evaluate the effect of our word embedding for new words in DA, we compared our method with two conventional approaches. The first one is that all the words that are not in a source domain dataset are treated as the special *unknown* symbol in DA and the testing phases. This is the case in which the word embedding parts are trained on a source domain dataset and are not updated during DA. The second one is a subword-based embedding, which is another approach for dealing with the appearance of a new word. In this paper, we use SentencePiece [21] trained as a part of the language model for ASR [22]. The results are shown in Table 4. Our method with the embedding assignment structure for new words stably outperformed the two conventional approaches when $K \geq 2$.

Table 4: *Effects of embedding assignment for new words in DA of ROD. K is the number of clusters for calculating the embedding candidates for new words. "Source only" denotes that the word embedding parts are not updated during DA.*

Word embedding method	Source→Target	Accuracy	ROC-AUC
Source only	Building→Station	0.857	0.927
SentencePiece [21]	Building→Station	0.849	0.927
W/ embed. assign. ($K = 1$)	Building→Station	0.856	0.927
W/ embed. assign. ($K = 2$)	Building→Station	0.866	0.932
W/ embed. assign. ($K = 5$)	Building→Station	0.862	0.931
W/ embed. assign. ($K = 10$)	Building→Station	0.862	0.930
W/ embed. assign. ($K = 20$)	Building→Station	0.863	0.930
W/ embed. assign. ($K = 50$)	Building→Station	0.862	0.930
W/ embed. assign. ($K = 100$)	Building→Station	0.861	0.930
Source only	Station→Building	0.843	0.917
SentencePiece [21]	Station→Building	0.828	0.917
W/ embed. assign. ($K = 1$)	Station→Building	0.843	0.916
W/ embed. assign. ($K = 2$)	Station→Building	0.845	0.919
W/ embed. assign. ($K = 5$)	Station→Building	0.847	0.920
W/ embed. assign. ($K = 10$)	Station→Building	0.849	0.920
W/ embed. assign. ($K = 20$)	Station→Building	0.849	0.920
W/ embed. assign. ($K = 50$)	Station→Building	0.850	0.921
W/ embed. assign. ($K = 100$)	Station→Building	0.850	0.920

Effect of the number of samples: We show the effect of the number of samples on the detection performance in Fig. 3. As the number of target samples increase, accuracy and ROC-AUC become higher. We conclude that our method was able to make the detector adapt to new environments automatically.

6. Conclusion

In this paper, we proposed a multimodal response obligation detector, which consists of visual, audio, and text-based detection modules. Assuming that the detector is used in various environments, we also proposed an unsupervised online domain adaptation method for the detector, which consists of the logistic regression update and the word embedding assignment. To evaluate our method on real world datasets, we constructed datasets by operating a robot in a station and a commercial building. Evaluation on the datasets showed that our detector achieved highly-accurate ROD on in-domain evaluations and was able to adapt to new environments automatically.

7. References

[1] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *ICASSP*, 2014, pp. 4087–4091.

[2] G. Tucker, M. Wu, M. Sun, S. Panchapagesan, G. Fu, and S. Vitaladevuni, "Model compression applied to small-footprint keyword spotting," in *INTERSPEECH*, 2016, pp. 1878–1882.

[3] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *ICASSP*, 2017, pp. 5670–5674.

[4] X. Zuo, N. Iwahashi, R. Taguchi, S. Matsuda, K. Sugiura, K. Funakoshi, M. Nakano, and N. Oka, "Robot-directed speech detection using multimodal semantic confidence based on speech, image, and motion," in *ICASSP*, 2010, pp. 2458–2461.

[5] K. Komatani, A. Hirano, and M. Nakano, "Detecting system-directed utterances using dialogue-level features," in *INTERSPEECH*, 2012, pp. 230–233.

[6] T. Sugiyama, K. Funakoshi, M. Nakano, and K. Komatani, "Estimating response obligation in multi-party human-robot dialogues," in *Humanoids*, 2015, pp. 166–172.

[7] S. Sheikhi, D. Babu Jayagopi, V. Khalidov, and J.-M. Odobez, "Context aware addressee estimation for human robot interaction," in *GazeIn*, 2013, pp. 1–6.

[8] Y. I. Nakano, N. Baba, H.-H. Huang, and Y. Hayashi, "Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems," in *ICMI*, 2013, pp. 35–42.

[9] O. Akhtiamov, M. Sidorov, A. Karpov, and W. Minker, "Speech and text analysis for multimodal addressee detection in human-human-computer interaction," in *INTERSPEECH*, 2017, pp. 2521–2525.

[10] H. Ouchi and Y. Tsuboi, "Addressee and response selection for multi-party conversation," in *EMNLP*, 2016, pp. 2133–2143.

[11] T. L. Minh, N. Shimizu, T. Miyazaki, and K. Shinoda, "Deep learning based multi-modal addressee recognition in visual scenes with utterances," in *IJCAI*, 2018, pp. 1546–1553.

[12] A. Pugachev, O. Akhtiamov, A. Karpov, and W. Minker, "Deep learning for acoustic addressee detection in spoken dialogue systems," in *AINL*, 2017, pp. 45–53.

[13] J. Hoffman, T. Darrell, and K. Saenko, "Continuous manifold based adaptation for evolving visual domains," in *CVPR*, 2014, pp. 867–874.

[14] A. Bitarafan, M. S. Baghshah, and M. Gheisari, "Incremental evolving domain adaptation," *TKDE*, vol. 28, no. 8, pp. 2128–2141, 2016.

[15] M. Wulfmeier, A. Bewley, and I. Posner, "Incremental adversarial domain adaptation for continually changing environments," in *ICRA*, 2018, pp. 4489–4495.

[16] M. Mancini, H. Karaoguz, E. Ruccu, P. Jensfelt, and B. Caputo, "Kitting in the wild through online domain adaptation," in *IROS*, 2018, pp. 1103–1109.

[17] D. E. King, "Dlib-ml: A machine learning toolkit," *JMLR*, vol. 10, pp. 1755–1758, 2009.

[18] J. Park, X. Liu, M. J. Gales, and P. C. Woodland, "Improved neural network based language modelling and adaptation," in *INTERSPEECH*, 2010, pp. 1041–1044.

[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communication of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmerm, and K. Veselý, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[21] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *EMNLP*, 2018, pp. 66–71.

[22] N. Kanda, Y. Fujita, and K. Nagamatsu, "Lattice-free state-level minimum Bayes risk traning of acoustic models," in *INTERSPEECH*, 2018, pp. 2923–2927.