Multimodal Activation: Awakening Dialog Robots without Wake Words

Liqiang Nie[†*], Mengzhao Jia[†], Xuemeng Song[†], Ganglu Wu[‡], Harry Cheng[†], Jian Gu[‡]
[†]Shandong University, Shandong, China, [‡]Alibaba Group, Zhejiang, China
{nieliqiang, jiamengzhao98, sxmustc, xaCheng1996}@gmail.com; {ganglu.wgl, gujian.gj}@alibaba-inc.com

ABSTRACT

When talking to the dialog robots, users have to activate the robot first from the standby mode with special wake words, such as "Hey Siri", which is apparently not user-friendly. The latest generation of dialog robots have been equipped with advanced sensors, like the camera, enabling multimodal activation. In this work, we work towards awaking the robot without wake words. To accomplish this task, we present a Multimodal Activation Scheme (MAS), consisting of two key components: *audio-visual consistency detection* and *semantic talking intention inference*. The first one is devised to measure the consistency between the audio and visual modalities in order to figure out weather the heard speech comes from the detected user in front of the camera. Towards this end, two heterogeneous CNN-based networks are introduced to convolutionalize the fine-grained facial landmark features and the MFCC audio features, respectively. The second one is to infer the semantic talking intention of the recorded speech, where the transcript of the speech is recognized and matrix factorization is utilized to uncover the latent human-robot talking topics. We ultimately devise different fusion strategies to unify these two components. To evaluate MAS, we construct a dataset containing 12,741 short videos recorded by 194 invited volunteers. Extensive experiments demonstrate the effectiveness of our scheme.

CCS CONCEPTS

• **Human-centered computing** → **Interaction paradigms**.

KEYWORDS

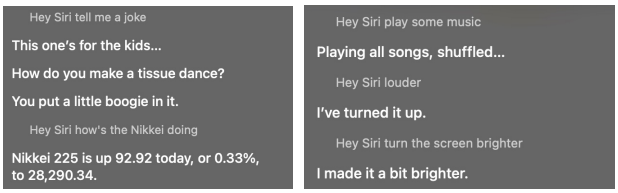
Multimodal Activation, Dialog Robots, Wake Words, Audio-Visual Consistency Detection, Semantic Talking Intention Inference

ACM Reference Format:
Liqiang Nie, Mengzhao Jia, Xuemeng Song, Ganglu Wu, Harry Cheng, Jian Gu. 2021. Multimodal Activation: Awakening Dialog Robots without Wake Words. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462964

*Liqiang Nie is the corresponding author.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462964



(a) (b)

Figure 1: Exemplars of robot activation with a wake word "Hey Siri".

1 INTRODUCTION

With the recent advances in speech recognition, considerable efforts have been dedicated to developing dialog robots in the AI era, due to the huge convenience they can bring to people' daily life, like setting reminders, playing music, and checking the weacher condition, simply based on the users' voice commands. Inevitably, dialog robots are quickly gaining popularity over the past few years around the world, and series of products flood the markets, such as Amazon's Echo and Alibaba's Tmall Genie. According to the statistics reported by Strategy Analytics[1], global sales of dialog robots hit a record high in 2019 with shipments of 146.9 million units, up 70% over 2018.

Before talking to the aforementioned dialog robots, users have to activate them with a special word or phrase, which is referred to a wake word, like "OK Google", and "XiaoduXiaodu". Wake words are vital in triggering dialog robots to recognize when to reply and execute commands, and when to merely listen and not to respond. Figure 1(a) and (b) demonstrate some examples of activating a dialog robot with the wake word "Hey Siri". As can be seen, each time the user wants some services from the dialog robot, he/she needs to speak the wake words first, even though the user is facing to the dialog robot. In a sense, although trickily linking the gateway between human and robots, wake word technology is less friendly or elegant as compared to our human-human interaction whereby a glance, a smile, or a short speech is sufficient to convey the talking intention instead of calling out a specifc name, especially during face-to-face talking.

Upgraded from voice interaction, the latest generation of dialog robots, like Ali's TMallGenie CC, have been equipped with smart touchscreens and enabled the robot sight with the front-facing camera. Therefore, users can awaken the robots in a more flexible way by either one or some of the tactile, acoustic and visual commands. In practice, there are three feasible ways for users to awaken the robot as shown in Figure 2. When a user is: 1) outside the view field of the robot, a wake word is compulsory; 2) in the field of view, the user can either activate the robot by touching

[1]https://tinyurl.com/y7m8p5mv.

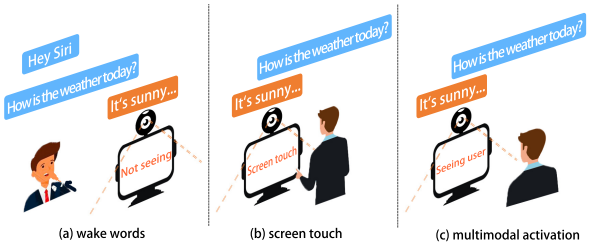


Figure 2: Three multimodal ways to awaken dialog robots. We work towards the multimodal activation one in this paper.

the screen with a special stylus and even fingers, or 3) visual cues together with certain speech semantics. In fact, the former two have been well studied thus far, while the last one, more user-friendly, remains largely untapped. In this paper, we hence define a new research task, i.e., multimodal activation, aiming to directly awaken the dialog robots by the audio-visual cues, which is complementary to wake words and screen touch.

Multimodal activation is indeed non-trivial considering the following facts. 1) Audio-visual consistency detection. As the dialog robot keeps open to the environment when it is in the standby mode, the heard speech by the robot does not have to be spoken by the detected user. As illustrated in Figure 3(a), a user A is eating with obvious mouse movement in front of the camera; while a user B is speaking loudly outside of the camera view that is heard by the given robot. As a result, it is prerequisite to figure out whether the heard speeh comes from the detected user. As reported in [11], the acoustic output and associated facial expression of our human are both rhythmic (in the 3 to 8 Hz range) and tightly correlated. Inspired by this, the consistency detection between the facial expression of the detected user and the heard speech is essential. However, the user facial expression can be rather sophisticated, including but not limited to lip movements, nose wrinkling, and eye gaze, and heterogenous to the speech modality. Thereby, how to access the audio-visual consistency is the first research challenge we are facing. And 2) semantic talking intention inference. The speech of the detected user is not necessarily meant to the dialog robot, although his/her facial expression is consistent with the speech. Considering the scenario in Figure 3(b) as an example, a user in the field of view is talking on the phone. Therefore, judging whether the detected user intends to talk to the given dialog robot is another tough challenge. Notably, in the context of multimodal activation, these two conditions must be satisfied at the same time like the example shown in Figure 3(c), otherwise the dialog robot should not be awakened.

To tackle these two challenges, we devise an end-to-end Multimodal Activation Scheme (MAS) as shown in Figure 4, which simultaneously leverages the visual and speech cues to automatically activate dialog robots from the standby mode. MAS comprises two key components. The first one is to judge whether the speech is from the detected user. To be more specific, the multimodal cues are regarded as two sequences of visual frames and speech segments, respectively. We first detect 68 facial landmarks for each frame to represent the visual cue in a fine-grained manner, and extract acoustic features, i.e, Mel Frequency Cepstral Coefficients (MFCC), to represent each speech segment. After that, we leverage a 3D and single channel CNN to characterize the

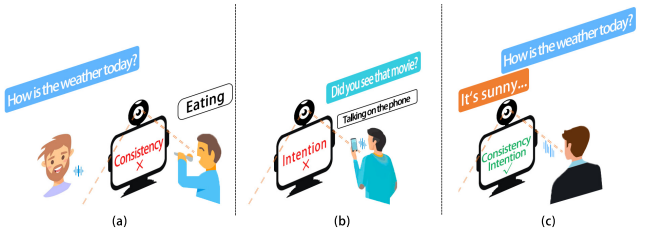


Figure 3: Illustration of one positive and two negative examples. a) The facial expression and speech is asynchronous. b) The user is on the phone without activation intention. And c) the user is talking to the robot.

spatial-temporal features of the video and the feature evolution of the speech sequence, respectively. Meanwhile, we project them into the same space whereby their consistency can be estimated directly. As to the second component, it analyzes the semantic speech content of the detected user and discriminate whether the user is intentionally talking to the dialog robot. To accomplish this, we perform automatic speech recognition (ASR) to collect text transcripts from the speeches and embed them with the XLNet model. We then stack the textual embedding of positive samples into a matrix and factorize the matrix to obtain the latent wake-up transcript patterns, which in a sense conveys the patterns of talking topics between the user and the dialog robot. Thereby, we infer the semantic talking intention based upon the similarity between the textual transcript and all the latent wake-up patterns. To jointly optimize these two components, we explored various fusion methods in this paper.

The contributions of this work can be summarized into threefold:

- To the best of our knowledge, this is the first work on multimodal activation that targets at awakening dialog robots directly. We clearly define the research scope of this task, and most importantly, we construct the first large-scale dataset. As a byproduct, we released the codes and parameter settings to facilitate other researchers in this community[2].
- We divide this complicated research task of multimodal activation into two key sub-problems, i.e., audio-visual consistency detection and semantic talking intention inference. Accordingly, we devise a novel end-to-end multimodal activation scheme, where the visual, audio, and textual (recognized from the speech) modalities are simultaneously explored.
- We introduce the fine-grained representation of the visual frames towards the audio-visual consistency detection, and adopt the matrix factorization to uncover the latent wake-up topic patterns for strengthening the talking intention inference.

The rest of the paper is organized as follows. In Section 2, we briefly review the related literature. Section 3 details our proposed scheme. We construct the dataset, conduct extensive experiments, and analyze the experimental results in Section 4, followed by conclusion and future work in Section 5.

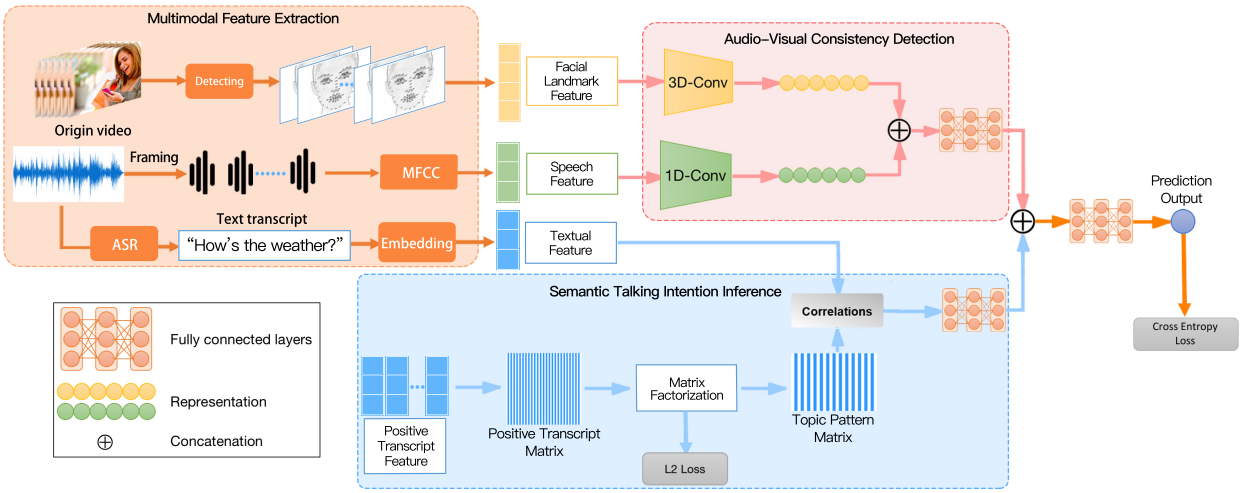


Figure 4: Schematic illustration of our proposed MAS scheme. Given a multimodal input, we first extract its fine-grained landmark features, MFCC features, and XLNet-based textual features. Then the fine-grained landmark features and MFCC features are fed into the audio-visual consistency detection component. Meanwhile, the textual features are passed into the sematic talking intention inference component, where matrix factorization is utilized to uncover the latent human-robot talking patterns.

2 RELATED WORK

Our work is related to activation methods for dialog robots, audio-visual consistency detection, and semantic talking intention inference.

2.1 Activation Methods

Smart speakers, a typical instance of dialog robots, have become almost ubiquitous as they can assist users in many aspects in their daily life, ranging from setting reminder to checking weather condition. Regarding this prosperity, the technique of wake word detection from the acoustic signal has been widely used to activate the smart speakers. In particular, the smart speaker always keeps in the standby mode waiting for the user's voice command and only goes into the state of higher power consumption to recognize the user's speech instructions when it hears the pre-defined wake words, like "Alexa" for Amazon's Echo.

Pioneers exploited keyword spotting methods to deal with the identification of wake words in utterances based on a large vocabulary word recognizer or templates [32]. However, keyword spotting is unable to discriminate whether the specific word is used in alerting or referential context. For instance, "Siri, what is the time now" exemplifies the wake word "Siri" in an alerting context; whereas "my friend Siri bought a T-shirt last night" shows the referential case. Although this problem can be somehow alleviated by using more sophisticated or uncommon words in practice like "Hey Siri", several studies have been dedicated to address this issue theoretically. For example, Kepuska et al. [17] introduced the Hidden Markov Model (HMM) triple scoring with SVM classifier based on multiple speech feature streams to discriminate the altering/referential context. Besides the HMM-based method, Wang et al. [38] presented a hybrid DNN/HMM wake word detection model with partially labeled training data, which makes the online detection possible. Beyond the above hard detection, Maekaku et al. [22] attempted to localize the wake words by simultaneously predicting its duration and endpoints with the multi-task learning. Most recently, Ahuja et al. [1] reformed wake word or command detection by inferring the direction of voice, namely, judging if the command is directed at the given robot.

Despite its significance and value, activating dialog robots with wake words has several limitations. 1) It is unnatural and stultifying for users to highlight the wake word before each instruction, especially when the user is facing the dialog robot. 2) It focuses on the unimodal acoustic input but overlooks the rich modality cues, like video, touch, gesture, gaze, as well as head and body movement, whose importance has been identified in [16]. And 3) the low recognition accuracy for identifying the misused wake words may cause accidental triggers. In light of this, Momeni et al. [26] proposed a novel key word spotting method, which can locate the word of interest simply based on the user's talking face, while the audio track can be used for performance improvement. Meanwhile, some work explored the feasibility of using gestures [31] and gazes [25] to trigger the smart speakers. Although at an infant stage, exploring innovative awakening methods is a big step forward as compared to the pure wake word detection.

2.2 Audio-Visual Consistency Detection

Due to its wide range of applications, such as spoof detection in biometrics and lip-syncing, audio-visual consistency detection that aims to judge whether the sequential facial expressions of a speaking person correspond to the accompanying audio track, has attracted a significant amount of research efforts.

Traditionally, solutions to this task largely rely upon pattern matching or shallow learning. For instance, in [20], a linear prediction method is adapted to recognize phoneme from a given audio, which is associated with mouth positions to provide keyframes for computer animation of speech. The methods presented in [19] and [24] model the relationship between the sound and mouth shape by having the speaker recording a standard set of sounds, typically, a set of vowel, and then correlate the face shape such as jaw position to the sound signal. Inspired by phonemes that are the

[2]https://mmacti2021.wixsite.com/mysite.

smallest linguistic sound units, Morishima et al. [27] classified the lip shapes into visemes as the phonemes' visual equivalents, and judged the synchrony via viseme-phoneme mapping. In addition to these intermediate analyses, some researchers performed canonical correlation analysis (CCA) [34] or co-inertia analysis (CoIA) [33] between speech and lip texture features, to identify their correspondence. Although the aforementioned methods achieve much success, they are still sub-optimal due to the fact that they seldom jointly consider the spatial and temporal information of audio and visual sequences, and only adopt the hand-crafted features.

Alongside with the advances of deep neural networks, researchers started to explore deep models to address this problem. Marcheret et al. [23] investigated the effectiveness of many deep neural models on two large audio-visual databases. In addition, Chung et al. [6] predicted the lip-sync error in a video by a two-stream ConvNet architecture, while Suwajanakorn et al. [36] adopted a recurrent neural network to learn the mapping from raw audio features to mouth shapes. Beyond existing methods, in addition to spatial and temporal cues, we consider a fine-grained facial expression in estimating the audio-visual synchrony.

2.3 Semantic Talking Intention Inference

The conventional solution to this task is typically a pipeline of an ASR of a speech utterance followed by intent inference via textual classification [39]. Considering the fact that textual features may carry different intentions depending on the manner of speech, Gu et al. [12] presented a novel multimodal deep learning structure that extracts features from textual-acoustic data for sentence-level speech intention classification. Specifically, textual and acoustic features are first extracted via two independent convolutional neural networks, then combined into a joint representation, and finally fed into a decision softmax layer. Ning et al. in [30] also noted the complementary information from users' speaking manners, and they defined Intention Prominence (IP) as the semantic combination of focus by text and emphasis by speech, and proposed a multi-task deep learning framework to predict IP.

Due to the fact that the errors accumulated at the ASR stage might affect the following intention classification, many researchers have resorted to the end-to-end solutions. For example, authors in [8] extended the ASR learning part to include an intention classification module and optimize the whole system. Later, Tian et al. [37] further improved the end-to-end speech-to-intent classification with Reptile [29], and tested its efficacy on four datasets of different languages and domains. Despite their great success, end-to-end solutions highly depend on large-scale training data, especially for complex intention classification. The data in new emerging domains are, however, difficult to acquire. To compensate this problem, we follow the two-stage scheme but model users' intention from the semantic topic level by learning a set of pattern bases.

3 METHODOLOGY

In this section, we first formulate our problem, and then introduce the feature extraction, followed by elaborating our proposed model.

3.1 Problem Formulation

For the ease of problem formulation, we first declare some notations. In particular, we use bold uppercase letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to denote matrices and vectors, respectively. We employ non-bold letters (e.g., N) to represent scalars, and Greek letters (e.g., λ) as parameters. If not clarified, all vectors are in the column form.

As aforementioned, the goal of this work is to devise a multimodal activation method for awakening dialog robots based on the multimodal cues. In a sense, the dialog robot should be activated only when the following two conditions are simultaneously met. 1) The speech modality perceived by the dialog robot should come from the detected user, whereas that of other users out of the camera view is not the focus of this work and will be treated as environmental noise. And 2) the speech semantic content delivers the user's talking intention.

Accordingly, suppose we have the training dataset $\mathcal{D} = \{d_i = (v_i, s_i, t_i, y_i^c, y_i^t)\}_{i=1}^N$. Thereinto, v_i, s_i and t_i refer to the visual, audio, and textual modalities of the i-th multimodal input, where t_i is recognized from s_i by the well-developed Alibaba ASR tool[3]. The symbol $y_i^c \in \{0, 1\}$ is the label indicating whether the i-th audio modality is consistent with the i-th visual one, i.e., the audio comes from the detected user. The other symbol $y_i^t \in \{0, 1\}$ is the label denoting whether the i-th speech delivers the talking intention. Apparently, samples that satisfy $y_i^c = 1$ and $y_i^t = 1$ at the same time are positive, denoted as $\mathcal{D}_{\{c+,t+\}}$, and the rest are negative. We have three types of negative samples: 1) samples with y_i^c=1 and y_i^c=0 constitute $\mathcal{D}_{\{c+,t-\}}$, referring to the cases of audio-visual consistency, but with no talking intention to the dialog robot; 2) Samples with y_i^c=0 and y_i^t=1, denoted as $\mathcal{D}_{\{c-,t+\}}$, refer to those whose speeches do carry the talking intention but from the other users outside of the camera view rather than the detected user; and 3) Samples in $\mathcal{D}_{\{c-,t-\}}$, with y_i^c=0 and y_i^t=0, correspond to the cases where the speeches neither belong to the detected users nor convey any talking intention. Based on these training samples, we are capable of learning a binary classifier composed of two key components: 1) the audio-visual consistency detection C based on the audio and visual modalities; and 2) the talking intention inference $\mathcal{T}$ relying on the semantic textual modality.

3.2 Multimodal Feature Extraction

In this subsection, we detail the process of multimodal feature extraction, as illustrated in Figure 4.

3.2.1 Visual Features. In our work, the visual features are mainly used for the audio-visual consistency detection, i.e., figuring out whether the audio modality belongs to the detected user. Beyond the coarse-grained visual information in the entire frame that may bring in the background noise, we focus on the more fine-grained and clean cues to characterize a person facial rhythmic movement. The underlying philosophy is that one's fine-grained facial movements, like whether the mouth is open or close, the eyes are looking up or blinking, is highly correlated to his/her acoustic output [15].

To accomplish this task, we first locate the face of a person within each frame with a bounding box, and then leverage the pre-trained facial landmark detector in the dlib library[4] to obtain the key landmarks of the detected face. Specifically, we select 68 representative landmark points corresponding to the human's key

[3] https://ai.aliyun.com/nls/asr.
[4] http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2.

facial structure, including the 17 jaw points, 5 right brow points, 5 left brow points, 9 nose points, 6 right eye points, 6 left eye points, 13 mouth points, and 7 lips points. To ensure that the visual feature has the rotation and translation invariance, and better captures the facial movement, we calculate the pairwise relative distance between each landmark point and the centroid one in the nose region along both the horizontal and vertical axes. In this way, each frame of the visual modality v_i can be represented by a 68×2 matrix. In this work, we fix the number of frames for all samples as 450, since the longest time duration of the video in our dataset is 15 seconds and each second contains 30 frames, where the padding zero operation is applied to the shorter samples. Accordingly, the i-th visual modality v_i can be represented as $\mathbf{V}_i \in \mathbb{R}^{68\times2\times450}$.

3.2.2 Speech Features. In order to encode our speech modality, we resort to the MFCC with 13 mel frequency bands. It describes the short-term power spectrum of a given audio input on the non-linear mel scale of frequency, which has been recognized as a powerful audio descriptor in many speech processing tasks [5, 28, 35]. Similar to the visual modality, to better capture the acoustic evolution that is highly correlated to the facial movements, we split the speech cue into a set of small segments, where the adjacent speech segments share certain overlaps due to the edge effect of the window function. In addition, to make each speech segment statistically stationary and support the reliable spectral estimate, we segment each speech signal into 25ms-segments with the step size of 10 ms. As all the speech signals are no longer than 15 seconds, we unify the length of all speech segment sequences to be 1,500, where the padding operation is used for the shorter signal. Ultimately, the speech modality s_i can be represented as $\mathbf{S}_i \in \mathbb{R}^{1500\times13}$, where the j-th row refers to the 13-D MFCC feature of the j-th speech segment.

3.2.3 Textual Features. As talking intention is one prerequisite to activate the dialog robot, it is natural to exploit the semantic content of the speech. Towards this end, we recognize the text modality t_i from the speech s_i with the help of the ASR API provided by Alibaba group, which has proven to be effective in many short speech recognition tasks [4]. To be more specific, we choose the pre-trained universal Chinese speech recognition model, due to the fact that this model can handle various speech scenarios, ranging from the voice input to social chatting. Based on the recognized text for each speech modality, we adopt the XLNet model [41], which has been widely used for text embedding learning [44]. Specifically, utilizing the XLNet model that is pre-trained on the cased Chinese simplified and traditional text[5], we represent each transcript of the i-th speech with the feature vector $\mathbf{t}_i \in \mathbb{R}^{D_t}$, whose dimension is empirically set to be $D_t = 768$.

3.3 Multimodal Activation

In this subsection, we detail our proposed MAS scheme and each of its component separately.

3.3.1 Audio-Visual Consistency Detection. In this component, we need to judge whether the speech is consistent with the given visual modality, in order to figure out whether the speech comes from the detected user. Towards this end, we propose to measure the correlation between the facial movements and the speech cue.

[5]https://huggingface.co/hfl/chinese-xlnet-base.

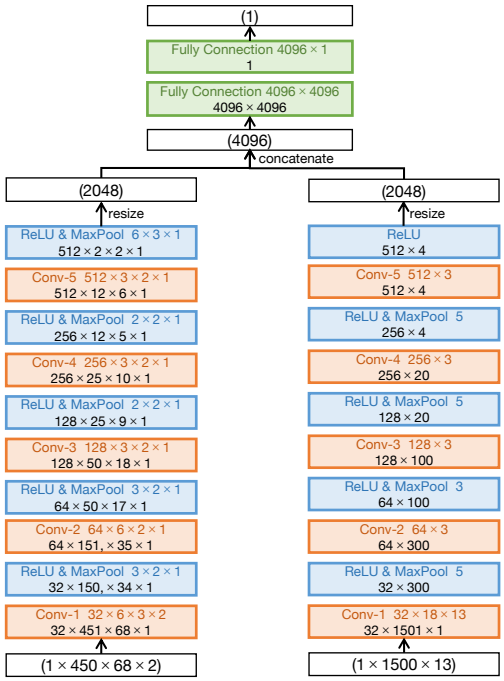


Figure 5: Parameter settings of audio-visual consistency detection.

To encode the visual facial movement, we resort to a 3D CNN model [14], which has achieved great success in many image or video understanding tasks. In a sense, the 3D CNN is able to capture the spatial correlation within the same face and temporal correlation resided in the facial landmark sequence, which benefits the facial movement understanding. To be more specific, as shown in Figure 4, by feeding the fine-grained features $\mathbf{V}_i$ of the i-th visual modality v_i into the 3D CNN, denoted by CNN_v, we can obtain the embedding $\hat{\mathbf{v}}_i \in \mathbb{R}^{D_v}$ of the visual modality v_i as follows,

$$\hat{\mathbf{v}}_i = CNN_v(\mathbf{V}_i|\mathbf{\Phi_v}), \tag{1}$$

where $\mathbf{\Phi_v}$ is the network parameter of CNN_v.

Towards the speech encoding, we also resort to the CNN, which has shown compelling performance in various speech processing tasks. As aforementioned that each speech modality s_i can be represented as a matrix $\mathbf{S}_i \in \mathbb{R}^{1500\times13}$, where each row corresponds to the 13-D MFCC feature of a speech segment. Inspired by TextCNN [18], we use a one-layer CNN to learn the feature evolution across segments. Formally, we have

$$\hat{\mathbf{s}}_i = CNN_s(\mathbf{S}_i|\mathbf{\Phi_s}), \tag{2}$$

where $\hat{\mathbf{s}}_i \in \mathbb{R}^{D_s}$ stands for the embedding of the speech modality s_i, and $\mathbf{\Phi_s}$ is the parameter of the CNN_s network.

Having obtained the speech and video embeddings, we then feed the concatenation of them to a multi-layer perceptron to learn the consistency between the facial expression and speech. In particular, we have

$$\hat{o}_i^c = MLP_c([\hat{\mathbf{v}}_i; \hat{\mathbf{s}}_i]|\mathbf{\Lambda_c}), \tag{3}$$

where $\mathbf{\Lambda_c}$ refers to the parameters of the multi-layer perceptron MLP_c. The parameter settings of MLP_c are shown in Figure 5.

3.3.2 Semantic Talking Intention Inference. After carefully going through the positive samples in our training dataset, we gain the insights that although the semantic topics users talk to the dialog robot are somehow diverse, they can be distinguished by certain talking patterns. Inspired by this, considering the remarkable performance in latent factor modeling [13], we adopt the matrix factorization to uncover the latent human-robot talking patterns. Specifically, we concatenate the textual embedding of all the positive samples with talking intention (i.e., $\mathcal{D}_{\{c+,t+\}} \bigcup \mathcal{D}_{\{c-,t+\}}$) into a big matrix $\mathbf{R} \in \mathbb{R}^{D_t \times N_p}$, where D_t refers to the representation dimension of each textual modality, and N_p denotes the total number of positive samples. Thereafter, we factorize the big matrix $\mathbf{R}$ into two matrices: topic pattern matrix $\mathbf{P} \in \mathbb{R}^{D_t \times K} = \{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_K}\}^T$ and latent representation matrix $\mathbf{H} \in \mathbb{R}^{K \times N_p}$, where K denotes the number of latent patterns. Each column of the topic pattern matrix $\mathbf{P}$ corresponds to a talking pattern basis, while that of $\mathbf{H}$ refers to the representation of each textual modality in the latent pattern space.

As to the matrix factorization, we reach the following objective function,

$$\mathcal{L}_{mf} = \min_{\mathbf{P},\mathbf{S}} \left\| \mathbf{R} - \mathbf{PS} \right\|^2. \tag{4}$$

Notably, the above latent pattern learning is not pre-trained but trained jointly with whole framework in an end-to-end manner.

Thereafter, given a textual modality, we can evaluate its talking intention by calculating its semantic similarity with each learned pattern, which can be formulated as follows,

$$m_{ij} = cos(\mathbf{t}_i, \mathbf{p}_j), \tag{5}$$

where m_{ij} denotes the semantic similarity between the given text modality t_i and the j-th chatting topic pattern. cos is the cosine similarity function. Accordingly, the similarity of the text modality t_i to all the patterns can be denoted with the embedding of $\mathbf{m} = [m_{i1}, m_{i2}, \cdots, m_{iN_p}]$. Similar to the audio-visual consistency component, we also pass the similarity embedding of the text t_i regarding chatting patterns into a multi-layer perceptron and get the talking intention as follows,

$$\hat{o}_i^t = MLP_t(\mathbf{m}_i | \mathbf{\Lambda_t}), \tag{6}$$

where $\mathbf{\Lambda_t}$ refers to the parameters of the multi-layer perceptron MLP_t, consisting of two fully connected layers. The first layer is activated by the Relu function, while the second is the linear transformation.

3.3.3 Activation Estimation. Based on the outputs of the aforementioned two key components, namely, the audio-visual consistency detection and the talking intention inference, we can predict whether the dialog robot should be activated for the given multimodal input. As a pioneer study on multimodal activation, we propose to explore multiple fusion methods to optimize the binary classification model for multimodal activation.

Early Fusion. As shown in Figure 4, we feed the concatenation of outputs of the two key components, i.e., the consistency score $\hat{o}_i^c$ and the talking intention score $\hat{o}_i^t$, into a fully-connected layer with the sigmoid activation to get the final predicted activation probability $\hat{y}_i$ as follows,

$$\hat{y}_i = fc([\hat{o}_i^c; \hat{o}_i^t] | \mathbf{\Lambda_e}), \tag{7}$$

Table 1: Statistics of four sample types in our dataset.

Data Type	Train	Val	Test	Total	Ave Duration(s)
$\mathcal{D}_{\{c+,t+\}}$	1,429	178	180	1,787	1.94
$\mathcal{D}_{\{c+,t-\}}$	631	80	79	790	2.10
$\mathcal{D}_{\{c-,t+\}}$	1,430	178	179	1,787	1.94
$\mathcal{D}_{\{c-,t-\}}$	6,702	838	837	8,377	5.15

where $\mathbf{\Lambda_e}$ is the network parameter.

Thereafter, using the cross-entropy loss for classification, we have the final objective function as follows,

$$\mathcal{L}_{early} = \min_{\mathbf{\Theta}} \frac{1}{N} \sum_i -[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)] + \mathcal{L}_{mf}, \tag{8}$$

where $y_i = y_i^c \times y_i^t \in \{0, 1\}$ is the ground truth for multimodal activation. $y_i = 1$ indicates that the dialog robot should be activated for the multimodal input (v_i, s_i, t_i), and $y = 0$ otherwise. $\mathcal{L}_{mf}$ refers to the matrix factorization loss, defined in Eqn. (4). $\mathbf{\Theta} = \{\mathbf{\Phi_v}, \mathbf{\Phi_s}, \mathbf{\Lambda_c}, \mathbf{P}, \mathbf{S}, \mathbf{\Omega_t}, \mathbf{\Lambda_t}, \mathbf{\Lambda_e}\}$ refer to the parameters of the whole framework. Once the framework gets well-trained, we can decide whether to activate the dialog robot by comparing the probability $\hat{y}_i$ with the threshold of γ.

Late Fusion. Distinct from the early fusion scheme, in this paper, we first introduce the cross-entropy loss for each component, and then jointly optimize them. The objective function is given as follows:

$$\begin{cases} \mathcal{L}_{late} = \mathcal{L}_c + \mathcal{L}_t, \\ \mathcal{L}_c = \min\limits_{\mathbf{\Theta_c}} \frac{1}{N} \sum\limits_i -[y_i^c log(\hat{o}_i^c) + (1-y_i^c) log(1-\hat{o}_i^c)], \\ \mathcal{L}_t = \min\limits_{\mathbf{\Theta_t}} \frac{1}{N} \sum\limits_i -[y_i^t log(\hat{o}_i^t) + (1-y_i^t) log(1-\hat{o}_i^t)] + L_{mf}, \end{cases} \tag{9}$$

where L_c and L_t are the losses for the audio-visual consistency detection and talking intention inference components, respectively. y_i^c and y_i^t are the ground truth labels indicating the audio-visual consistency and talking intention of the i-th multimodal input. Notably, different from the early fusion, once the networks get trained, the activation can be triggered if and only if $\hat{o}_i^c \geq \eta_1$ and $\hat{o}_i^t \geq \eta_2$, where η_1 and η_2 are the threshold parameters.

4 EXPERIMENTATION

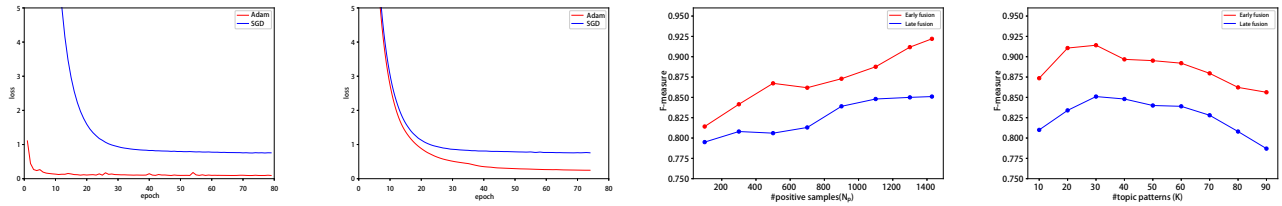
To evaluate the proposed multimoal activation scheme, i.e., MAS, we conducted extensive experiments on our newly created dataset by answering the following research questions:

- What is the effect of the key hyperparameters?
- Does the proposed MAS outperform the state-of-the-art methods?
- How is the performance of each component of MAS?

In this section, we first introduce the dataset as well as the experimental settings, and then provide the experimental results with detailed discussions over each above research question.

4.1 Dataset Construction

As this work is the first study on multimodal activation for dialog robots, there is no public benchmark dataset available. To address this problem, we created our own dataset with the 194



(a) Optimizer (early fusion) (b) Optimizer (late fusion) (c) The number of positive samples (d) The number of hidden talking patterns

Figure 6: Effects of the optimizer, number of positive samples for the talking pattern factorization, and number of hidden talking patterns.

invited volunteers interacting with dialog robots. Specifically, as the audio-visual consistency and the talking intention are the two key prerequisites for multimodal activation, we collected four types of video samples with two fine-grained labels as follows.

1) To collect the samples of $\mathcal{D}_{\{c+,t+\}}$, we invited 16 volunteers, and asked them to interact with dialog robots in a silent environment with scripts, like "How is the weather?" and "Can you recommend a movie?". All the interactions of each volunteer are recorded in a single video. Thereafter, we segmented the long video of each volunteer into short ones, where each contains a complete utterance. Notably, to ensure the quality of the samples, the segmented short video is manually checked to ensure it indeed carries the talking intention to the dialog robot.

2) Analogously, we constructed $\mathcal{D}_{\{c-,t-\}}$ by inviting 178 volunteers and recorded a single long video for each volunteer. It is worth noting that we invited much more volunteers here to generate purely negative samples from different aspects, making this dataset representative. Distinct from those in $\mathcal{D}_{\{c+,t+\}}$, volunteers during the video recording were only allowed to have facial expressions, like opening mouth, but refraining from producing any vocal sound, while the audio modality comes from the background noise in the environment, e.g., produced by a TV or other people's conversation.

3) As to the sample collection in $\mathcal{D}_{\{c+,t-\}}$, we resorted to the publicly accessible benchmark dataset, which is a Chinese speaker recognition corpus released by [10]. Samples of this corpus are mainly about Chinese celebrities and obviously all the speakers have no intention to activate a dialog robot.

4) Regarding $\mathcal{D}_{\{c-,t+\}}$, we synthesized its samples based on the other three datasets. To be more specific, for each positive sample in $\mathcal{D}_{\{c+,t+\}}$, we kept its audio part and replaced its visual part with that of a randomly selected sample from one of other three data types, i.e., $\mathcal{D}_{\{c+,t+\}}$, $\mathcal{D}_{\{c+,t-\}}$, and $\mathcal{D}_{\{c-,t-\}}$, whereby operations of padding or truncating frames were adopted for aligning the audio and visual modalities. In this way, the facial expression and the speech is inconsistent in the synthesized video, while the speech content still keeps the talking intention to activate the dialog robot.

To guarantee the quality of our dataset, we eliminated the videos with no clear facial signals, which may be caused by the long distance between the user and dialog robots. We finally obtained 12,741 short videos in our data collection, lasting for 51,904 seconds and containing 1, 557, 120 visual frames with facial landmarks in total. All the videos are in terms of 30 frames per second with 16kHz audio sample rate. The longest and shortest video lasts 15 and 0.63 seconds, respectively, while the average length of the short videos is 4.07 seconds. Regarding the dataset split, 80% of the available samples are randomly allocated for training, while the remaining 20% of samples are equally partitioned, composing the validation and testing datasets, respectively. Table 1 summarizes the statistics of our dataset with four data types.

4.2 Experimental Settings

Our task is indeed a binary classification problem. We hence employed the widely-used precision, recall and F-measure metrics to measure the model performance. Specifically, the precision for a class is the number of true positives divided by the total number of samples labelled as positive class. By contrast, recall is defined as the number of true positives divided by the total number of samples that actually belong to the positive class. As to the F-measure, it is the harmonic mean of the precision and recall.

We explored both stochastic gradient descent (SGD) and Adaptive Moment Estimation (Adam) to optimize the network. We adopted the grid search strategy to determine the optimal values of the number of samples for the pattern factorization (i.e., N_p) among values $[100, 300, 500, 700, 900, 1100, 1300, 1429]$, and the number of hidden talking patterns (i.e., K) within the range of $[10, 90]$ with the step of 10, respectively. In addition, γ, η_1, and η_2 were searched in the range of $[0, 1]$ with the step of 0.1. We empirically found that the proposed model achieves the optimal performance with Adam, $N_p = 1,429$, $K = 30$, $\gamma = 0.5$, $\eta_1 = 0.5$, and $\eta_2 = 0.5$.

All the experiments were implemented with Pytorch 1.6.0, and conducted on a server, equipped with GPU and Geforce RTX 2080Ti 12G graphics cards running over OS Linux.

4.3 Parameter Tuning (RQ1)

Our scheme involves some key parameters, including optimizer selection, the number of positive samples used for the latent topic pattern factorization, and the number of latent topic patterns.

It is well-known that a right optimizer is able to squeeze the last bit of accuracy out of the model. We thus compared the convergence speed and effectiveness between two widely-used optimizers: SGD and Adam, by fixing other parameters of our model. Figure 6(a) illustrates the loss curves of these two optimizers regarding the number of epochs. We observed that Adam converges remarkably faster than SGD, which may be owing to its momentum and adaptive learning rates. After convergence, we found that the F-measure of our model optimized by Adam is 0.924, which is slightly better than that of SGD (i.e., 0.911). We thereby chosen Adam as our optimizer.

In addition, we studied the quantitative impact of the number of positive samples for pattern factorization, by keeping the other parameter settings of our model fixed. Figure 6(b) shows the F-measure of our model with different number of positive samples for pattern factorization under different fusion schemes. We can see that our model with different fusion schemes consistently reaches

(a) Word cloud over all terms. (b) Word cloud over nouns.

Figure 7: Word cloud generated from positive transcripts.

the best performance, when all the positive samples (1,429 in total) were incorporated. This suggests that the more positive examples were used, the more accurate performance our model will reach.

Besides, we explored the effect of the number of hidden topic patterns by fixing other parts of our model. The experimental results are shown in Figure 6(c). It can be seen that with the increasing number of hidden topics from 10 to 90, the performance of two fusion strategies first increases fast and then reaches the peak value at $K = 30$, followed by continuous decrease. This implies that the number of hidden topic patterns does affect our performance, and it is not the more the better. One possible explanation is that the number of common topic patterns interacted between human and dialog robots are around 30.

Intuitively, to verify the existence of the talking pattern, we visualize the content of our positive transcripts with talking intention (i.e., those in $\mathcal{D}_{\{c+,t+\}} \bigcup \mathcal{D}_{\{c-,t+\}}$) by word cloud. We first translated the Chinese transcripts with Google Translate API[6]. Figure 7 shows the word cloud over all terms and that with only nouns. Based on the word clouds, we noticed that the positive transcripts do contain the common human-robot interaction patterns, such as ask the robot to turn on/off devices, turn up/down volume, open apps, check weather and news, and look for hotels and restaurants. Interestingly, we found that people interacted with the dialog robots with more casual words like "want", rather than the formal ones like "what" and "where".

4.4 Overall Performance Comparison (RQ2)

To justify the effectiveness of our model, we compared it with following advanced methods we designed.

- **MSA_TFN**: This method is originally devised for multimodal sentiment analysis, and applicable in our context. In particular, we used tensor-fusion network [42], where a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions across three modalities is employed. Openface2.0, librosa, and the pre-trained Chinese BERT base word embedding were used to encode the visual, audio, and textual modalities, respectively.
- **MSA_LMF**: We replaced the tensor-fusion network in MSA_TFN with a lowrank multimodal fusion method [21], which learns both modality-specific and cross-modal interactions with modality-specific low-rank factors. The modality encoding is the same with MSA_TFN.
- **MSA_MFN**: We derived MSA_MFN from MSA_TFN by replacing the fusion method with memory fusion network [43] that explicitly accounts for both interactions in a neural architecture and continuously models them through time.
- **MSA_EF-LSTM**: Derived from MSA_TFN, this method [40] first concatenates initial inputs of three modalities and then uses LSTM to capture their dependencies in the sequence.
- **MSA_LF-DNN**: In contrast with MSA_EF-LSTM, we adopted the later fusion DNN to learn unimodal features first and then concatenated these features before classification.
- **MAS_Frame**: In order to check the effectiveness of our fine-grained landmark embedding, we replaced it with the frame-level visual representation, which is a 1,024-D visual features extracted from the sequence of visual frames by the pre-trained I3D model [3].
- **MAS_Face**: In this baseline, the granularity of visual representation is between the frame-Level and our landmark-level. We first performed face detection via the dlib face recognition tool and then embedded the detected face to a 1,024-D visual feature by the pre-trained I3D model [3].

[6]https://translate.google.com/.

Table 2: Performance comparison among different models. The symbol * means statistically significant improvement over the strongest baseline with p < 0.05.

Method	Fusion	Precision	Recall	F-measure
MSA_TFN	-	0.900	0.903	0.901
MSA_LMF	-	0.890	0.940	0.914
MSA_MFN	-	0.749	0.865	0.803
MSA_EF-LSTM	-	0.893	0.901	0.887
MSA_LF-DNN	-	0.883	0.927	0.904
MAS_Frame	Late	0.851	0.897	0.874
MAS_Face	Late	0.863	0.899	0.880
MAS	Late	0.884	0.905	0.894
MAS_Frame	Early	0.899	0.932	0.915
MAS_Face	Early	0.901	0.936	0.918
MAS	**Early**	**0.903***	**0.947***	**0.924***

The comparison results are summarized in Table 2. From this table, we obtained the following observations: 1) Overall, our method performs better with the early fusion scheme as compared to the late one. One possible explanation is that the objective function of the early fusion scheme is devised to directly address the task of activation. By comparison, the late fusion is introduced to solve the sub-problems, whereby the overall task is solved as a by-product. 2) MAS_Early outperforms all the baselines, which demonstrates the effectiveness of our multimodal activation scheme. In particular, the reason why our model surpasses all derivatives of MSA, i.e., MSA_TFN, MSA_LMF, MSA_MFN, MSA_EF-LSTM, and MSA_LF-DNN, may be due to the fact that these methods overlook both the audio-visual consistency and talking intention inference in the multi-modal activation context. And 3) MAS performs better than both MAS_Face and MAS_Frame, and MAS_Face is slightly better than that of MAS_Frame. This verifies the necessity of the fine-grained landmark feature extraction in the audio-visual consistency detection. Meanwhile, it reveals that the finer the visual cues are characterized, the better performance the model will achieve.

4.5 Component-wise Evaluation (RQ3)

4.5.1 Ablation Study. In this part, we derived two derivatives from our MAS: MAS_w/o_C and MAS_w/o_T, where we removed

Table 3: Experimental results on ablation study.

Method	Precision	Recall	F-measure
MAS_w/o_C	0.831	0.898	0.863
MAS_w/o_T	0.883	0.877	0.880
MAS_Early	**0.903**	**0.947**	**0.924**

the audio-visual consistency detection and the semantic talking intention inference component from our MAS_Early, respectively. Table 3 summarizes the experiment results of ablation study. As can be seen, MAS_Early outperforms both MAS_w/o_C and MAS_w/o_T, which indicates the importance of each component in the context of multimodal activation. To intuitively show the impact of both components, we further illustrated the comparison among MAS_w/o_C, MAS_w/o_T, and MAS_Early with several case studies in Figure 8. 1) In the first case, the speech heard by the dialog robot does deliver the talking intention that the user may want to ask the dialog robot for recommending some nice restaurant to enjoy crayfish. The heard speech is, however, not consistent with the captured visual signal, i.e., the speech does not belong to the detected user. In such a case, equipped with the audio-visual consistency detection module, MAS_Early can accurately learn that there is no need to activate the robot, while MAS_w/o_C cannot. Similar observation can be found in the second case. And 2) as to the third and fourth cases, although these two samples meet the audio-visual consistency, their transcripts indicate that the users have no talking intention with the dialog robots. Our MAS_Early correctly classified these samples as negative ones, while MAS_w/o_T failed. This reflects the importance of the semantic talking intention inference.

4.5.2 Component-wise Evaluation. We also justified the effectiveness of our two key components in their corresponding tasks, where they were separately cast to the binary classification problems regularized by the cross entropy losses. Based on our dataset, there are 2,577 positive and 10,164 negative samples for audio-visual consistency detection task, while 3,574 positive and 9,167 negative samples for semantic talking intention detection. In total, we had 12,741 video samples for each component evaluation, and according to the 80%/10%/10% split manner, the number of training, validation and testing samples for these two tasks are respectively 10,192, 1,274, and 1,275.

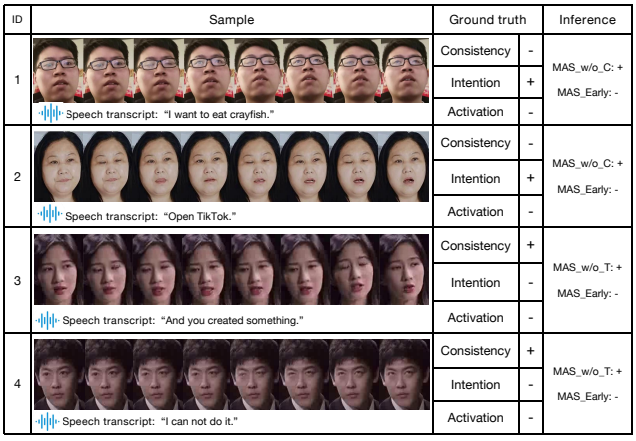


Figure 8: Case study of each component.

Table 4: Performance comparison on audio-visual consistency detection. The symbol * means statistically significant improvement over the strongest baseline with p < 0.05.

Method	Precision	Recall	F-measure
AVE-NET	0.672	0.684	0.678
MMS	0.683	0.772	0.725
MAS-C	**0.831***	**0.898***	**0.863***

Table 5: Performance comparison on semantic talking intention inference. The symbol * means statistically significant improvement over the strongest baseline with p < 0.05.

Method	Precision	Recall	F-measure
Bert	0.770	0.755	0.761
XLNet	0.830	0.846	0.837
MAS-T	**0.883***	**0.877***	**0.880***

As for comparison, we adopted baselines AVE-NET [2] and Multi Matching Syncnet (MMS) [7] for our audio-visual consistency component, referred as MAS-C. Meanwhile, we selected the Bert [9] and XLNet [41] baselines to compare with our talking intention inference component, dubbed as MAS-T. For fair comparison, we adopted the pre-trained Bert and XLNet, while only fine-tuned the last fully-connected layer for classification. Tables 4 and 5 show the performance comparison of two components of our scheme, respectively. As can be seen, each component surpasses the baselines on the corresponding task, which suggests the effectiveness of our two components. In addition, our MAS-C outperforms both AVE-NET and MMS that use the frame-level features, reconfirming the importance of the fine-grained landmark feature in multi-modal activation. Besides, our MAS-T surpasses Bert and XLNet, demonstrating the advantages of the latent human-robot talking pattern modeling.

5 CONCLUSION AND FUTURE WORK

In this work, we define a new research task of multimodal activation without wake words. To accomplish this task, we present a Multimodal Activation Scheme named MAS, comprising of audio-visual consistency detection and semantic talking intention inference. The former is devised to figure out whether the heard speech comes from the detected user in front of the dialog robot, where the fine-grained landmark features are incorporated. The latter is introduced to identify whether the speech delivers the user's talking intention, where the matrix factorization is adopted to uncover the latent human-robot talking patterns. Extensive experiments over our newly created dataset demonstrate the effectiveness of our scheme and the importance of each component. In addition, we observe that the fine-grained landmark-level features obviously promote the model performance as compared with the frame-level and face-level features. Meanwhile, the two key components of our MSA also achieve the superior performance on the tasks of audio-visual consistency detection and talking intention inference, respectively. We plan to extend our model to handle more sophisticated scenarios, like more users appear in the dialog robot's filed of view.

6 ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China, No.:U1936203.

REFERENCES

[1] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison. 2020. Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Device Ecosystems. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. ACM, 1121–1131.

[2] Relja Arandjelovic and Andrew Zisserman. 2018. Objects that Sound. In *Europeon Conference on Computer Vision*. Springer, 451–466.

[3] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6299–6308.

[4] Mengli Cheng, Chengyu Wang, Xu Hu, Jun Huang, and Xiaobo Wang. 2020. Weakly Supervised Construction of ASR Systems with Massive Video Data. *arXiv preprint arXiv:2008.01300*.

[5] Zhiyong Cheng, Xuanchong Li, Jialie Shen, and Alexander G Hauptmann. 2016. Which Information Sources are More Effective and Reliable in Video Search. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 1069–1072.

[6] Joon Son Chung and Andrew Zisserman. 2016. Out of Time: Automated Lip Sync In The Wild. In *Asian Conference on Computer Vision*. Springer, 251–263.

[7] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. 2019. Perfect Match: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3965–3969.

[8] Thierry Desot, Francois Portet, and Michel Vacher. 2019. Towards End-to-End Spoken Intent Recognition in Smart Home. In *International Conference on Speech Technology and Human-Computer Dialogue*. IEEE, 1–8.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 4171–4186.

[10] Y. Fan, J. W. Kang, L. T. Li, K. C. Li, H. L. Chen, S. T. Cheng, P. Y. Zhang, Z. Y. Zhou, Y. Q. Cai, and D. Wang. 2020. CN-Celeb: A Challenging Chinese Speaker Recognition Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7604–7608.

[11] Asif A. Ghazanfar and Daniel Y. Takahashi. 2014. Facial Expressions and the Evolution of the Speech Rhythm. *Journal of Cognitive Neuroscience* 26, 6, 1196–1207.

[12] Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. Speech Intention Classification with Multimodal Deep Learning. In *Canadian Conference on Artificial Intelligence*. Springer, 260–271.

[13] Xianjing Han, Xuemeng Song, Jianhua Yin, Yinglong Wang, and Liqiang Nie. 2019. Prototype-Guided Attribute-Wise Interpretable Scheme for Clothing Matching. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 785–794.

[14] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1, 221–231.

[15] Jintao Jiang, Abeer Alwan, Lynne E Bernstein, Patricia Keating, and Ed Auer. 2000. On the Correlation between Facial Movements, Tongue Movements and Speech Acoustics. In *International Conference on Spoken Language Processing*. ISCA, 42–45.

[16] Veton Kepuska and Gamal Bohouta. 2018. Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *IEEE Annual Computing and Communication Workshop and Conference*. IEEE, 99–103.

[17] VZ Këpuska and TB Klein. 2009. A Novel Wake-Up-Word Speech Recognition System, Wake-Up-Word Recognition Task, Technology and Evaluation. *Nonlinear Analysis: Theory, Methods & Applications* 71, 12, e2772–e2789.

[18] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1746–1751.

[19] Barrett E Koster, Robert D Rodman, and Donald Bitzer. 1994. Automated Lip-Sync: Direct Translation of Speech-Sound to Mouth-Shape. In *Proceedings of the Asilomar Conference on Signals, Systems and Computers*. IEEE, 583–584.

[20] John Lewis. 1991. Automated Lip-Sync: Background and Techniques. *The Journal of Visualization and Computer Animation* 2, 4, 118–122.

[21] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2247–2256.

[22] Takashi Maekaku, Yusuke Kida, and Akihiko Sugiyama. 2019. Simultaneous Detection and Localization of a Wake-Up Word Using Multi-Task Learning of the Duration and Endpoint. In *Annual Conference of the International Speech Communication Association*. ISCA, 4240–4244.

[23] Etienne Marcheret, Gerasimos Potamianos, Josef Vopicka, and Vaibhava Goel. 2015. Detecting Audio-Visual Synchrony Using Deep Neural Networks. In *the Annual Conference of the International Speech Communication Association*. ISCA, 548–552.

[24] David F McAllister, Robert D Rodman, Donald L Bitzer, and Andrew S Freeman. 1997. Lip Synchronization of Speech. In *Workshop on Audio-Visual Speech Processing*. ISCA, 133–136.

[25] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama–a Gaze Activated Smart-Speaker. In *Proceedings of the ACM on Human-Computer Interaction*. ACM, 1–26.

[26] Liliane Momeni, Triantafyllos Afouras, Themos Stafylakis, Samuel Albanie, and Andrew Zisserman. 2020. Seeing Wake Words: Audio-visual Keyword Spotting. In *British Machine Vision Conference*. BMVA, 1–13.

[27] Shigeo Morishima, Shin Ogata, Kazumasa Murai, and Satoshi Nakamura. 2002. Audio-Visual Speech Translation With Automatic Lip Syncqronization and Face Tracking Based on 3-D Head Model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2117–2120.

[28] Arpan Mukherjee, Shubhi Tiwari, Tanya Chowdhury, and Tanmoy Chakraborty. 2019. Automatic Curation of Content Tables for Educational Videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1329–1332.

[29] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

[30] Yishuang Ning, Jia Jia, Zhiyong Wu, Runnan Li, Yongsheng An, Yanfeng Wang, and Helen Meng. 2017. Multi-Task Deep Learning for User Intention Understanding in Speech Interaction Systems. In *AAAI Conference on Artificial Intelligence*. AAAI, 161–167.

[31] Patryk Pomykalski, Mikołaj P Woźniak, Paweł W Woźniak, Krzysztof Grudzień, Shengdong Zhao, and Andrzej Romanowski. 2020. Considering Wake Gestures for Smart Assistant Use. In *the CHI Conference on Human Factors in Computing Systems*. ACM, 1–8.

[32] J Robin Rohlicek, William Russell, Salim Roukos, and Herbert Gish. 1989. Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 627–630.

[33] Enrique Argones Rúa, Hervé Bredin, Carmen García Mateo, Gérard Chollet, and Daniel González Jiménez. 2009. Audio-Visual Speech Asynchrony Detection Using Co-Inertia Analysis and Coupled Hidden Markov Models. *Pattern Analysis and Applications* 12, 3, 271–284.

[34] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp. 2007. Audiovisual Synchronization and Fusion Using Canonical Correlation analysis. *IEEE Transactions on Multimedia* 9, 7, 1396–1403.

[35] Paraic Sheridan, Martin Wechsler, and Peter Schäuble. 1997. Cross-Language Speech Retrieval: Establishing a Baseline Performance. In *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 99–108.

[36] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync From Audio. *ACM Trans. Graph.* 36, 4, 95:1–95:13.

[37] Yusheng Tian and Philip John Gorinski. 2020. Improving End-to-End Speech-to-Intent Classification with Reptile. In *the Annual Conference of the International Speech Communication Association*. ISCA, 891–895.

[38] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur. 2020. Wake Word Detection with Alignment-Free Lattice-Free MMI. In *the Annual Conference of the International Speech Communication Association, Virtual Event*. ISCA, 4258–4262.

[39] Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken Language Understanding. *IEEE Signal Processing Magazine* 22, 5, 16–31.

[40] Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language*. ACL, 11–19.

[41] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*. MIT, 5754–5764.

[42] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1103–1114.

[43] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 5634–5641.

[44] Chenbin Zhang, Congjian Luo, Junyu Lu, Ao Liu, Bing Bai, Kun Bai, and Zenglin Xu. 2020. Read, Attend, and Exclude: Multi-Choice Reading Comprehension by Mimicking Human Reasoning Process. In *Proceedings of the International ACM SIGIR conference on research and development in Information Retrieval*. ACM, 1945–1948.