A STUDY OF MULTI-SPEAKER DIALOGUE SYSTEM FOR MOBILE INFORMATION RETRIEVAL

Hsien-Chang Wang, Chieh-Yi Huang, Chung-Hsien Yang and Jhing-Fa Wang

Department of Computer Science and Information Engineering
Department of Electrical Engineering
National Cheng-Kung University, Tainan, Taiwan, R.O.C.
{Wangsj,wangjf}@csie.ncku.edu.tw

ABSTRACT

The advances of speech processing technologies make it possible to build spoken dialogue systems (SDS) which may provide people with useful information. However, most current SDS can only deal with one speaker at a time, improvements should be made before the SDS can be applied to domains where multiple speakers interact with the system. This paper discusses research issues for developing a multi-speaker dialogue system (MSDS) which is able to retrieve various mobile information in the car environment.

The differences between traditional (single speaker) and multi-speaker dialogue system are first addressed. Then, two research topics are studied. 1) Speech source identification, which determines the active speaker. 2) Multi-speaker dialogue management, which interpreter the intention and maintain the dialogue history of the speakers to keep the interaction smooth.

Many testers in a car environment attended the experiment for active speaker detection and multi-speaker dialogue system. The experiments showed an encouraging result that the proposed approach of MSDS did work properly.

1. INTRODUCTION

It has been several decades since the first development and release of spoken dialog system (SDS). Currently, most SDSs only deal with the interaction between the system and one speaker. In some situations, interaction may occur between several speakers and the system. Speech signal processing and the dialogue management should be reconsidered to deal with multiple-speaker interaction. This motivates us to study the development of multi-speaker dialogue system (MSDS).

In a human-to-human multi-speaker dialogue, speakers may cooperate to accomplish a common goal, or negotiate to solve conflict opinions to achieve a same goal. The same cases may occur and should be considered in a MSDS. We defined two types of goal in a MSDS, i.e., the individual and global goal. The individual goal is that a speaker wants to achieve. Since individual goals may conflict with each other, system should maintain a global goal to integrate the individual goals. The following examples demonstrate different cases where individual goals do and do not conflict with each other. These examples of multi-speaker dialogue took place in the car environment as shown below:

(Cooperative, speakers have the same goal)
User1: Find a place for me to eat something.
User2: I want to have hamburger and coke.
System: There is a MacDonald 300 meters ahead.

In the first example, the individual goal of User1 is to find a restaurant; and User2 is to eat hamburger. The dialogue manager should detect and integrate these individual goals to form the global goal, that is, a place where hamburger is supplied. Another example, in which the individual goals are conflicted, is given below:

(Conflict, speakers have conflict goals)
User1: Is there any Chinese-Oil gas station?
User2: I think the Taiwan-Plastic gas station is better.
System: please confirm which gas station do you want, the Chinese-Oil or Taiwan-Plastic?

In this example, the global goal should be adjusted when speaker User2 has a different individual goal from that of User1.

Previous study [1] of MSDS could deal with one goal only; speakers were not allowed to make cross-domain queries. To deal with multi-speaker dialogue interaction, techniques such as multiple microphone recording, source identification, noise canceling, multi-speaker voice activity detection and separation, and multi-speaker dialogue management should be studied first. Figure 1 is the flowchart of a multi-speaker dialogue system (four speakers in this figure). Due to the capacity limit of this publication, we focus on the major component of the MSDS, i.e., speaker source identification and multi-speaker dialogue management.

Figure 1. The flowchart of a multi-speaker dialogue system.

This paper is organized as follows. Section 2 describes how to identify the active speaker, which includes the issues about choice and placement of multiple microphones in the car environment. Section 3 illustrates the algorithm of multi-speaker dialog manager, together with several examples. Section 4 is the experimental results. The concluding remarks are given in Section 5.
2. SPEECH SOURCE IDENTIFICATION

2.1. Multiple microphone placement

In a noisy environment, multi-microphone is usually used to enhance the input signal. We had studied many kinds of multi-microphone placement to determine the best one for multi-speaker purpose. Table 1 is the brief comparison of various kinds of microphone placement.

![Table 1. Comparison of various types of microphone placement.](image)

Concluded from many tests and surveys, we suggest that using multiple unidirectional microphones is the most suitable placement for MSDS in car environment currently. We place a unidirectional microphone in front of each speaker. Totally four microphones are used in our study.

2.2. Speech source identification

The block diagram of the multi-microphone processing is shown in Figure 2. We use a matched filter to identify the speech source (i.e., to identify the active speaker) in a multi-microphone placement; and an adaptive filter to estimate the enhanced (noise-free) signal.

![Figure 2. Block diagram of microphone processing](image)

The matched filter is to detect speech in noisy environment, while a user speaks to a microphone [2]. The output of matched filter from each microphone is compared with a threshold to decide from which microphone the speech comes, i.e., the primary channel. The impulse response $h_i(t)$ of the matched filter in the $i$-th microphone for a signal $y_i(t)$ is given by

$$h_i(t) = y_i(N-1-t) \quad 0 \leq t \leq N-1$$  \hspace{1cm} (1)

The decision of speech detection is made as

$$D_i(t) = \begin{cases} 1 & \text{if } z_i(t) \geq \text{Threshold} \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (3)

We mark a microphone as a primary channel if its $D_i(t)$ is equal to one, and the others are then classified as secondary channels. These secondary channels are used to estimate the noise $\hat{n}(t)$. The enhanced signal is obtained by subtracting $\hat{n}(t)$ from the primary channel $y_p(t)$. We make a common assumption that the signal and noise are uncorrelated. For the secondary channels $y_s(t)$, the minimum mean square error (MMSE) filter $h(n)$ is used to estimate the noise $\hat{n}(t)$.

Applying the orthogonality principle to secondary channels [3], we have the following equation:

$$E\left\{ y_p(t) - \sum_{j=0}^{N-1} h(j)y_s(t-j) y_s(t-k) \right\} = 0$$  \hspace{1cm} (4)

Let $r_{ys}$ and $r_{yp}$ denote the auto-correlation of $y_s$ and $r_{yp}$ denote the cross-correlation between $y_s$ and $y_p$. We can rewrite Eq. (4) into the following form

$$r_{yp} = \sum_{j=0}^{N-1} h(j)r_{ys}(k-j) = 0$$  \hspace{1cm} (5)

where $k=0, ..., N-1$. For simplicity, the matrix form of Eq. (5) is given by

$$r_{yp} y_s^T = R_{yp} h = 0$$  \hspace{1cm} (6)

The gradient method is applied to solve Eq. (6).

$$\epsilon(h) = E\left\{ y_p(t) - \sum_{j=0}^{N-1} h(j)y_s(t-j) \right\}^2$$  \hspace{1cm} (7)

$$\frac{\partial \epsilon(h)}{\partial h} = 2(r_{yp} y_s^T - R_{yp} h)$$  \hspace{1cm} (8)

The Eq.(8) can be used to derive the coefficients of the filter $h(n)$, and thus we have the enhanced speech signal for the active speaker.
Algorithm 1: Management of multi-speaker dialogue

Input: word graph of speech recognition for each speaker, \( W_{G1}, W_{G2}, \ldots, W_{Gn} \), where \( n \) is the number of total speakers.

Output: response to speakers.

Step 1: Initialization
Initialize the semantic vectors \( SV_i \) to be NULL.
\[ SV_i = (d_{i1}, d_{i2}, d_{i3}, \ldots) \]
each element \( d_{ij} \) is an integer. For speaker \( i \),
\( d_{i1} \) represents the domain that speaker mentioned;
\( d_{i2} \) is the primary attribute for this domain; and
\( d_{i3} \) is the secondary attributes, where \( j \) varies with domain.
Initialize the dialogue history lists, \( H_s \), for each speaker and \( H_t \) for system to be NULL.

Step 2: Determine semantics vector of each speaker
Apply NLP techniques to \( W_{Gi} \) to determine the corresponding semantic vector \( SV_i \).

Step 3: Determine accomplished goal(s)
Semantic vector \( SV_i \) for this turn is copied to the history \( Hi \). \( SV_i \) and \( Hi \) are integrated to form the \( Hs \).

Step 4: Decision.
If any goal is completed, go to step 5; else go to step 6.
Note that, under what condition a goal was complete is definable for each domain.

Step 5: Response.
Perform database query and generate response to the user according to the goal(s) found in step 3.
Go to step 6.

Step 6: Iteration
Accept next input \( (WG_i) \) and go to step 2.

3. MULTI-SPEAKER DIALOGUE MANAGEMENT

Once the system is able to determine the active speaker, the next important task is to maintain and keep smooth the interaction between system and multiple speakers. In a MSDS, each speaker may have his own “goal” for information retrieval, which is defined as individual goal in this paper. Contrasted to the individual goal, the global goal is the integration of each individual goal. The management of multi-speaker dialogue is to interpret intentions and semantics of individual speaker, to detect if there was conflict between speakers, to integrate individual goals into global goals, to determine whether a specific goal is completed, and to generate the response. In the next sections, we illustrate how the management of MSDS works by giving an algorithm and some examples.

3.1. Algorithm of multi-speaker dialogue management

The algorithm of multi-speaker dialogue management is shown in Algorithm 1. Each time the system receive the input

<table>
<thead>
<tr>
<th>Time index</th>
<th>Action</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>User1 input</td>
<td>“I want to go to the city hall”</td>
</tr>
<tr>
<td>2</td>
<td>User2 input</td>
<td>“Tell me the weather of Taipei”</td>
</tr>
<tr>
<td>3</td>
<td>Generate SVi</td>
<td>( \text{SV}_1=\text{“GUIDE”, “DESTINATION”, “city hall”, Null…} ) \text{SV}_2=\text{“WEATHER”, “LOCATION”, “Taipei”, Null…} )</td>
</tr>
<tr>
<td>4</td>
<td>Check goal completeness</td>
<td>User1=TRUE User2=TRUE</td>
</tr>
<tr>
<td>5</td>
<td>Check if conflict</td>
<td>NO</td>
</tr>
<tr>
<td>6</td>
<td>Generate response</td>
<td>“The city hall is about 450 meters away, please follow my instruction” “The weather in Taipei is rainy.”</td>
</tr>
</tbody>
</table>

Example 1. Speakers have different individual goals.

Example 2. Speakers have conflict individual goals.

<table>
<thead>
<tr>
<th>Time index</th>
<th>Action</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>User1 input</td>
<td>“Find me a Chinese food restaurant”</td>
</tr>
<tr>
<td>2</td>
<td>User2 input</td>
<td>“No, I want Italian specialties”</td>
</tr>
<tr>
<td>3</td>
<td>Generate SVi</td>
<td>( \text{SV}_1=\text{“GUIDE”, “RESTAURANT”, “Chinese food”, Null…} ) \text{SV}_2=\text{“GUIDE”, “RESTAURANT”, “Italian specialties”, Null…} )</td>
</tr>
<tr>
<td>4</td>
<td>Check goal completeness</td>
<td>User1=TRUE User2=TRUE</td>
</tr>
<tr>
<td>5</td>
<td>Check if conflict</td>
<td>YES</td>
</tr>
<tr>
<td>6</td>
<td>Generate response</td>
<td>“Please specify again, do you want Chinese food or Italian specialties”</td>
</tr>
</tbody>
</table>
Example 3. Speakers have common goal.

<table>
<thead>
<tr>
<th>Time index</th>
<th>Action</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>User1 input</td>
<td>“I want to know the route to …”.</td>
</tr>
<tr>
<td>2</td>
<td>Generate SVi</td>
<td>SV_i=<code>(&quot;GUIDE&quot;, &quot;DESTINATION&quot;, Null…)</code></td>
</tr>
<tr>
<td>3</td>
<td>Check goal completeness</td>
<td>User1=FALSE, (no DESTINATION)</td>
</tr>
<tr>
<td>4</td>
<td>Generate response</td>
<td>“Please specify the destination.”</td>
</tr>
<tr>
<td>5</td>
<td>User1 input</td>
<td>“To the nearest gas station”</td>
</tr>
<tr>
<td></td>
<td>User2 input</td>
<td>“And, how far is it”</td>
</tr>
<tr>
<td>6</td>
<td>Combine new SVi with old ones</td>
<td>SV_i=<code>(&quot;GUIDE&quot;, &quot;DESTINATION&quot;, &quot;gas station&quot;, &quot;nearest&quot; Null…)</code></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SV_i=<code>(&quot;DISTANCE&quot;, &quot;DESTINATION&quot;, &quot;gas station&quot;, &quot;nearest&quot;, Null…)</code></td>
</tr>
<tr>
<td>7</td>
<td>Check goal completeness</td>
<td>User1=YES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User2=NO</td>
</tr>
<tr>
<td>8</td>
<td>Check if conflict</td>
<td>NO</td>
</tr>
<tr>
<td>9</td>
<td>Generate response</td>
<td>“A gas station is 1.5 kilometers ahead, please go straight.”</td>
</tr>
</tbody>
</table>

4. EXPERIMENTS

We set up an experimental automobile which contains 4 unidirectional microphones in front of each seat for the active speaker identification. The recording device is a notebook computer together with a PCMCIA multi-channel recording card. Totally 24 testers, grouped into eight groups, attended our experiment. Testers were informed with the system capability briefly. The domains, i.e., route guide, weather forecast and stock prices etc., which the system may provide information for, were also informed to the testers. Two types of experiments were carried out in our study, i.e., the speech source identification and the dialogue completeness rates.

4.1. Speech source identification

Several stable speeds of the vehicles are carefully maintained to control the decibel of the noise. Attendants of this experiments talked to the system, the recorded data are used for the speech source identification. The experimental result is shown in Table 2, which showed that the adopted approach achieves good performance for active speaker detection.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Correct rate</th>
<th>Speed (km/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Silent</td>
<td>93 %</td>
<td>91 %</td>
</tr>
<tr>
<td>Radio on</td>
<td>91 %</td>
<td>90 %</td>
</tr>
<tr>
<td>Windows opened</td>
<td>91 %</td>
<td>89 %</td>
</tr>
</tbody>
</table>

Table 2. Experimental result of active speaker identification.

4.2. Dialogue completeness success rate

The evaluation of MSDS is done by checking the dialogue completeness rate for different situations. Experiments include single speaker, cooperative multi-speaker, and conflict multi-speaker. The results show only the part that speech recognition output the correct result in top 3 candidates. Table 3 shows the experimental results. Some notations used are described below.

- \( S_c \): number of correct turns for single speaker.
- \( G_c \): correct turns for cooperative speakers.
- \( C_c \): correct turns for conflict speakers.
- \( S_a \): total number of turns for single speaker dialogues.
- \( G_a \): total number of turns for cooperative dialogues.
- \( C_a \): total number of turns for conflict dialogues.

<table>
<thead>
<tr>
<th>Type of multi-speaker dialogue</th>
<th>Correct rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single speaker (Sc/Sa)</td>
<td>91.2 %</td>
</tr>
<tr>
<td>Cooperative speakers (Gc/Ga)</td>
<td>89.5 %</td>
</tr>
<tr>
<td>Conflict speakers (Cc/Ca)</td>
<td>86.4 %</td>
</tr>
</tbody>
</table>

Table 3. Evaluation of the MSDS.

5. DISCUSSIONS AND CONCLUSIONS

In this paper, we have addressed important issues for the development of a multi-speaker dialog system, especially in the car environment where every passenger may want to interact with the system. Deciding microphone placement, locating speaker position, and managing multi-speaker dialogue context are essential topics and should be further studied.

In fact, two kinds of interaction may occur in the car environment, the interaction between speaker and system \((\text{interaction})\), and between speaker and speaker \((\text{intra-action})\). This paper discussed the former only. How to model both the interaction and intra-action in a MSDS is a more difficult task and requires more studies. Research of multi-speaker spoken dialog system (MSDS) is in the initial stage, we hope this paper can give rise to the research in the techniques of MSDS.

6. REFERENCES