Voice Based Crowdsourcing using Google Home Smart Speaker

Leonardus Elbert Putra/755379

Demo Video

https://youtu.be/uzNC-wa7Cvg

Aim

To create a voice based crowdsourcing system on google home smart speaker.

Abstract

With the rise of crowdsourcing as a data gathering method around the world, many people are looking for more efficient ways to do it in order to get data even faster and cheaper. The result of this is the rise of many online crowdsourcing platform in recent years. In this study, we want to go even further and try to implement a voice based crowdsourcing system using the latest popular smart device google home. Since smart speaker is interacted with a voice command which is a more natural and ubiquitous way of interaction, it has the potential to make crowdsourcing to become less burdening leading to a more efficient and effective method of gathering data. The resulting application is proven to be able to handle simple crowdsourcing tasks that has multiple predefined answer choices but a bit struggling in handling tasks which answer has no restriction for example language translation. However, the application is successful in lessening the burden of crowdsourcing within our participants as some of them managed to do other activities while doing the tasks.

Introduction

Online crowdsourcing has proven to be an efficient and cheap method of collecting data from user input that is crucial as a base for many professional practices on various industries. It can reach a lot of people faster and cheaper than hiring a professional and certainly more effective than recruiting a limited number of participants because it capitalises on everyday people spare time (Howe, 2006). As a result, professionals such as businesses and academics rely on this method for example through online questionnaires or surveys. However, the success rate of this is usually affected by the motivation of the participants when doing the task (Hossain, 2012). This problem leads to the creation of online paid crowdsourcing platform such as Amazon Mechanical turk which reward participants with small amount of money for each task that they did. This is a brilliant idea, but a recent research shows that most participants are only able to gather 2 USD/hour on average, despite spending a large amount of time and focus in doing the tasks (Ross et al., 2009). Therefore, there is a need to improve the efficiency of this platform. To address this problem, we recommend the use of smart speaker as a crowdsourcing platform. We hope this can reduce the amount of focus needed as people can do the tasks simultaneously with other activity. Thus, we aim to create a voice based crowdsourcing platform on google home and evaluate its performance in gathering user input through demonstrating several suited crowdsourcing task with the application.

Related Study

Currently, there is no crowdsourcing platform yet that uses smart speaker as its main input device. However there are a few studies that has applied voice crowdsourcing, which are; Respeak (Ashista, Sethi & Anderson 2017) and Bespeak (Ashista, Sethi & Anderson 2018). These studies uses a smartphone instead as its media of a smart speaker. Both studies are done by the same authors with a similar purpose, which is to improve crowdsourcing accessibility to people with lower typing skills or disabilities. Respeak and Bespeak are used for transcribing an audio file using voice input by making participants repeat the words that they heard from the audio files that is played by the device. Then, using speech recognition, the spoken words are turned into text containing the script for each corresponding audio. Both studies receives good feedback from its participants due to its simpler UI design and ease of use compared to the Amazon Mturk. This provides a good background and great expectations for this research, since it will use a smart speaker which provides a more natural voice interaction experience compared to a smartphone.

Study Design Consideration

Originally, the project specification intended one of the tasks to be a collaboration task, but after careful consideration, we decided to focus more on the regular task instead due to several reasons. According to Ikeda et al. (2016), collaboration in crowdsourcing can be differentiated into 3 categories which are sequential that involves several workers improving each other result at a different time, simultaneous that involves several worker working together simultaneously, and hybrid that incorporates the 2 previous category together. Firstly, smart speaker can only receives one input at a time making it hard to do simultaneous tasks together. This eliminates the simultaneous and hybrid options that the collaboration has. Next, sequential task is essentially the same as adding another task on top of the previous task which is the same as multiple regular tasks. Hence, this experiment will instead focus more on 5 regular tasks.

Study Design

In order to evaluate this system for crowdsourcing, there is a need to select a few tasks that are suitable to be completed via voice input. This is because the majority of tasks that are currently available in most crowdsourcing platform are screen based. Some examples of this are interpreting information from a picture, transcribing a video/audio, and accessing contents which are some of the common task that can be seen on online crowdsourcing platforms such as Amazon Mechanical Turk (Amazon Mturk). According to Gadiraju, Kawase, & Dietze (2014), there are currently 6 top “goal oriented” classifications for crowdsourcing tasks which are:

- **Information Finding (IF):** tasks that require workers to find information about specific things e.g. company information, cheapest price

- **Verification and Validation (VV):** tasks that require workers to check the validity of certain information e.g. checking if a particular website is reliable, validify a social account

- **Interpretation and Analysis (IA):** tasks that require its workers to use their own knowledge in order to interpret useful information e.g. sentiment analysis

- **Content Creation (CC):** tasks that require workers to generate a new content based on the instruction e.g. translating a language, transcribing a piece of information

- **Surveys (SU):** tasks that gather user opinion and experience in regards to something e.g. user evaluation

- **Content Access (CA):** tasks that require workers to access a particular online content e.g. clicking a link to an online content given.

From these categories, CA is not really suitable for the smart speaker due to its necessity of accessing an online content which is a feature that the smart speaker currently don't have. Furthermore, tasks in the category of IF and VV can theoretically be solved by audio only, but it often also require access to online content. In the case of IF, access to online content is used for gathering the information needed for the task while for VV it is used to verify the information that is provided by the task. Both of these can certainly be solved via voice but it needed the use of other devices for that online access which is counter productive with the aim of making crowdsourcing be simpler to do. Therefore, this experiment select tasks from the 3 category that is left which are IA, CC , and SU. From the 5 tasks that are chosen to be used in this study, 3 tasks are from IA, 1 from CC, and 1 from SU. These tasks are:

- **Class IA/Sentiment Analysis** : This task requires the participant to determine whether a sentence intention is positive or negative toward a specific subject. In this task,movie reviews will be read to participants where they must choose whether its sentiment is positive or negative. According to Natsukawa & Yi (2003), the ability to analyse positive and negative opinions can provide many benefits for various application. It is usually used as part of marketing analysis or managing risks for companies. A lot of money is usually spent on this research in order for company to know the state of its market and customers satisfaction rate of their product and services. This task also require high intelligence and deep understanding of text that usually require human intelligence to perform. However, since the data is often too many and doing it manually will require a lot of investment, therefore a lot of researchers have started to study ways of doing it automatically (Kouloumpis, Wilson, & Moore, 2011). Machine learning algorithm is usually the answer for this kind of problem but a study by Medhat, Hassan & Korashy (2014) recently stated that one of the difficulties of doing this is the rarity of reliable datasets that can be used to train the machine. Therefore, crowdsourcing is usually one of the methods that is used to gather this datasets because it is considered to be more efficient and cost effective. This task is perfect to test the voice crowdsourcing application since it can be done via voice only and the answer is restricted to only 2 choices which are positive or negative. Furthermore, varying the question difficulty based on indirectness/sarcasm and sentence length can test the capability of the voice crowdsourcing application in tasks that have long questions and require much focus since the participants need to analyse every word in order to know the sentiment and not be misled by the sarcasm. The dataset for this set is taken from a movie reviews dataset by Maas et al., (2011).

- **Class IA/Emotion Analysis** : This task requires the participant to determine what is the emotion that is depicted by an audio provided. Participants will hear an audio recording of a person speaking and choose whether the speaker was expressing happiness, anger, disgust, sadness, or fear. According to Busso et al (2004), by incorporating emotions, human and machine interaction could potentially be more natural. They also stated that tone of speech is one of the important signals that humans use to detect and express emotion. Even a subtle change in voice due to expressing different emotions can be picked up by human sense. Juslin & Scherer (2008) mentioned that, one of the aspects of emotion analysis is acoustic analysis where emotion is detected based on its sound and tone of speech only. This task have 5 predetermined answer choice and only need a human intuition in order to answer. Therefore, this task is perfect for testing the voice crowdsourcing system if it can handle tasks that have multiple choice questions. This task will have 10 questions where 5 contain an audio spoken by a young person (26 year old) and 5 by an old person (64 years old). All audio will say the same word but with different tones of speech portraying different emotions. This task will try to replicate a study done by Dupuis & Pichora-Fuller (2011) and see if it can produce similar results. The dataset that is used also belongs to the same author.

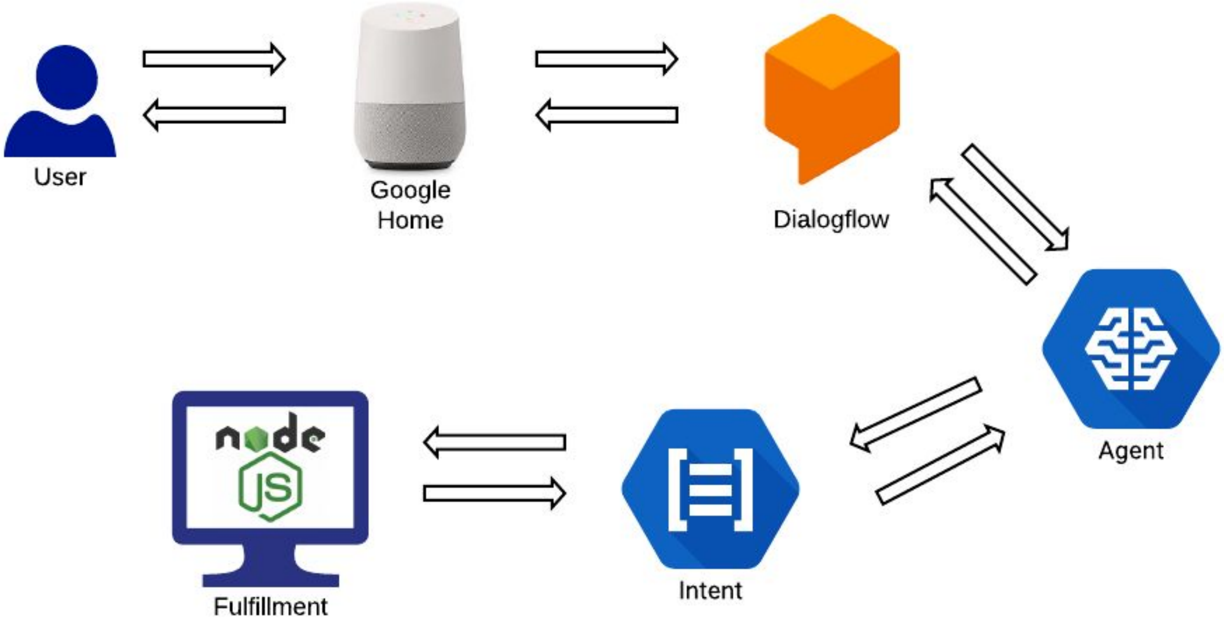
- **Class IA/Gender Recognition** : This task requires the participant to detect whether the speaker in an audio recording is male or female. In this task participants will hear a voice recording and they will determine whether the speaker is a male or female. According to Childers & Wu (1991), humans can easily detect many kinds of information through speech only but the same is not true for man-made machine. It requires a lot of data to create machines that are able to do this automatically. However gathering the data manually will take a lot of time and cost. Therefore, a simple task such as this is usually crowdsourced to many people since it is also simple enough and very rarely produce invalid data. This task will test the voice crowdsourcing application capability of handling simple crowdsourcing task. The task will be composed of 5 questions where it will increase in difficulty in terms of the voice tendency to be male or female.The dataset for this task is taken from the website VoxForge.org.

- **Class CC/Language Translation** : This task requires the participants to translate sentences from one language to another. In this task, participants will hear an Indonesian sentences being read by the smart speaker and reply its translation in English. Language translation has been a popular example of crowdsourcing application. This is because a lot of data is required to build a machine translation system and gaining it through linguists or language experts are often less variable, expensive and takes a lot of time (Kunchukuttan et al., 2014). Furthermore, this task carry significant difficulty due to it being open-ended in nature where participants can input any translation that they want without restrictions. In addition to that, this task can potentially be done via voice only which will be a great test for the voice crowdsourcing platform to see if it can handle tasks with long and unexpected answer. The questions in this task will increase in difficulty from task 1 to 5 which is based on sentence length and harder vocabulary. The answers will be rated by a Indonesian and English bilingual speaker on a scale from 1 to 3 where 1 is for wrong translation, 2 is for close enough but some words do not fit the sentence, and 3 for a usable translation. Dataset from this task is taken from an Indonesian grammar book by Dyenar (2003).

- **Class SU/Survey** : This task requires participants to answer several questions based on their experience in using this device. According to Difallah et al., (2015) survey is one of the recently most popular tasks to do in online crowdsourcing platforms such as Amazon Mechanical Turk. Therefore, in this task we want to see if this task is possible to be done in voice and at the same time gather data to further evaluate the system. In this task, there are 5 questions that are going to be asked which include whether the participant has done crowdsourcing task before, rating the use difficulty of the app, whether the participant choose computer or smart speaker to do the same tasks, will they do it if paid the same with an average earning on amazon MTurk, and lastly what is the most difficult task to do. Unlike other tasks which order to do not matter, this task will be instructed to be done last to the participants.

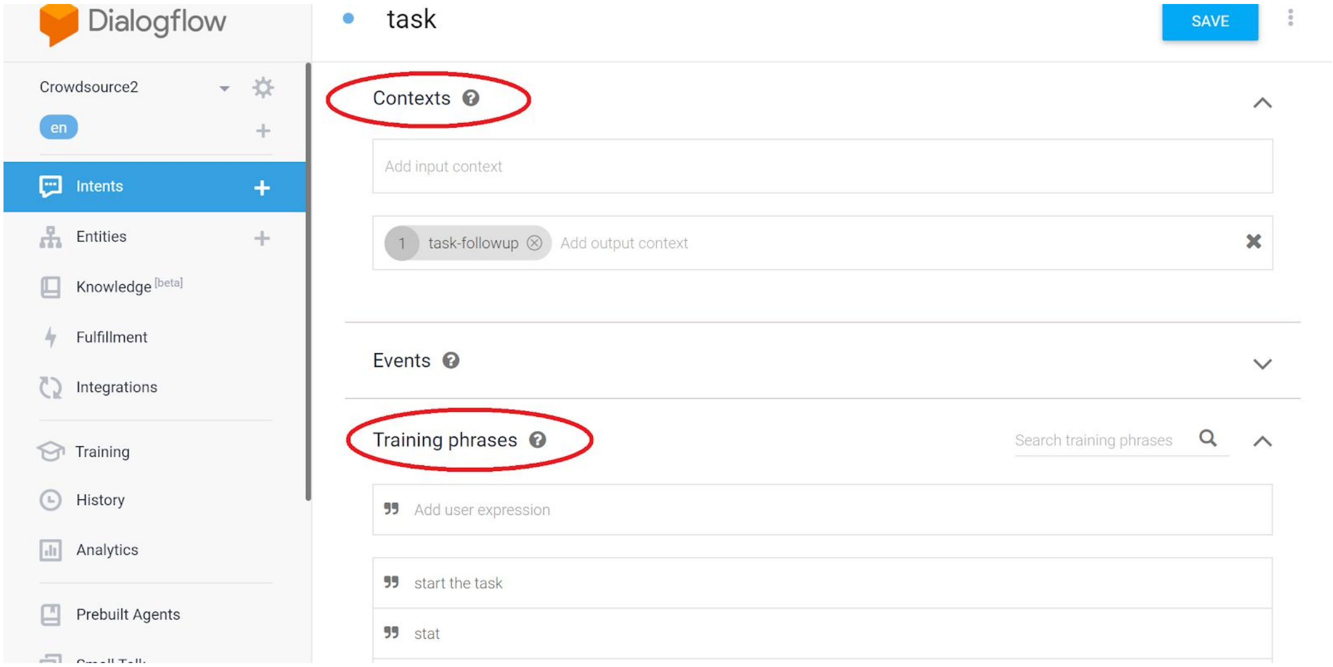
Application Architecture

The structure of the application:



To begin with, the structure of the voice application follow the previous diagram. User will talk to the smart speaker google home where their speech will be processed by the Dialogflow Platform. Dialogflow has an AI agent which can map user’s speech input to a matching intent. In this platform, intent represent an event that will be triggered if user speech’s input match or closely resemble the training phrase that has already been prepared when creating the intent. Then, the intent will call its associated function through an external webhook server which is programmed in node JS. This call can be made by switching on the enable webhook button on the bottom part of the intent menu on

Dialogflow. To understand better, several basics of intent will be explained in more detail. The following image represents an example of an intent in Dialogflow.



task

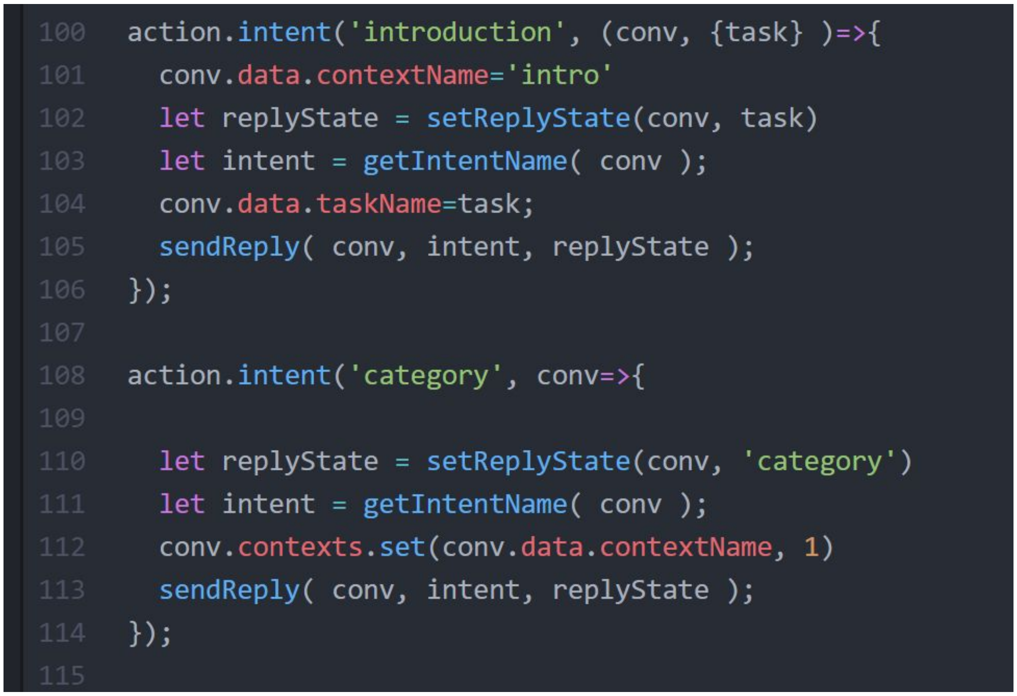
1 task-followup

start the task

stat

In here, there are several important concepts that need to be understood first which are contexts and training phrases. Firstly, contexts are used to represent the current state of conversation and also used to carry on information from 1 intent to the others. It is usually used to control the conversation These contexts also has specified active time that can be modified to make it lasts more than a turn of conversation. In this example, it is set at 1 which can be seen by the number 1 beside the “task-followup” context. Furthermore, these contexts are divided into 2 categories which are input and output contexts. Input contexts are used to make sure that certain intents can only be triggered when that context is available while output context is used to set up a new context at the end of the intent. Using both of these, the programmer can manipulate the conversation with users to some extent. Next, training phrases are predictions of what the user might say to trigger than intent. In this part, several predictions can be made and the agent will be trained based on those predictions so it will become smarter the more training phrases it has.

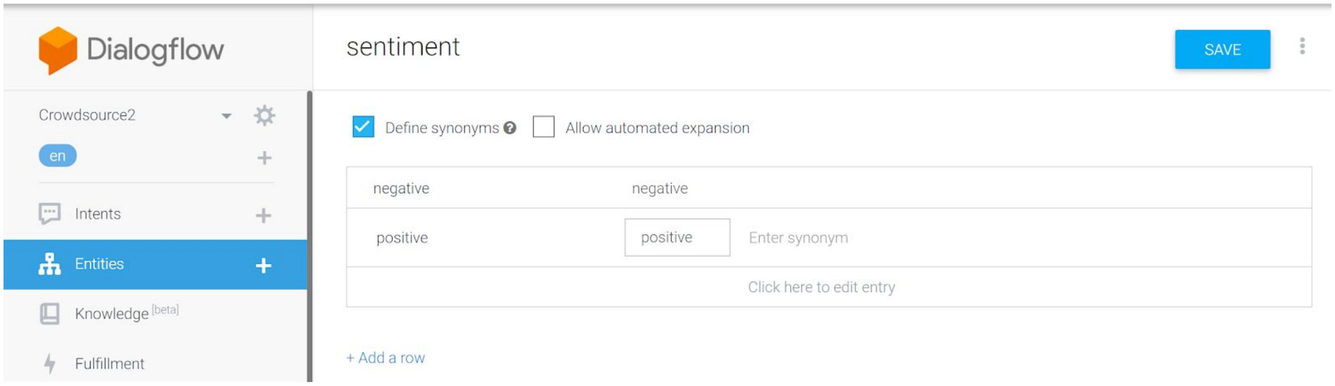
Next is the fulfillment, which are the functions that are mapped to each intent call. All of these functions are specific and can only be triggered if its corresponding intent is triggered as well in Dialogflow. Below is an example of 2 fulfillment functions that is mapped to intent ‘introduction’ and intent ‘category’.



```javascript
action.intent('introduction', (conv, {task} )=>{
  conv.data.contextName='intro'
  let replyState = setReplyState(conv, task)
  let intent = getIntentName( conv );
  conv.data.taskName=task;
  sendReply( conv, intent, replyState );
});

action.intent('category', conv=>{

  let replyState = setReplyState(conv, 'category')
  let intent = getIntentName( conv );
  conv.contexts.set(conv.data.contextName, 1)
  sendReply( conv, intent, replyState );
});

```

Since these functions are specific and can only be invoked according to what intent the user's speech is matched against, there must be some special considerations when designing the software of this application.

Finally, there is also a feature to define and include parameters to better predict what the user going to say in the training phrase. To do this, an entity containing variation of the parameters must be made first in the entity tab on Dialogflow console. The following diagram will show the entity "sentiment" for sentiment analysis task.



sentiment

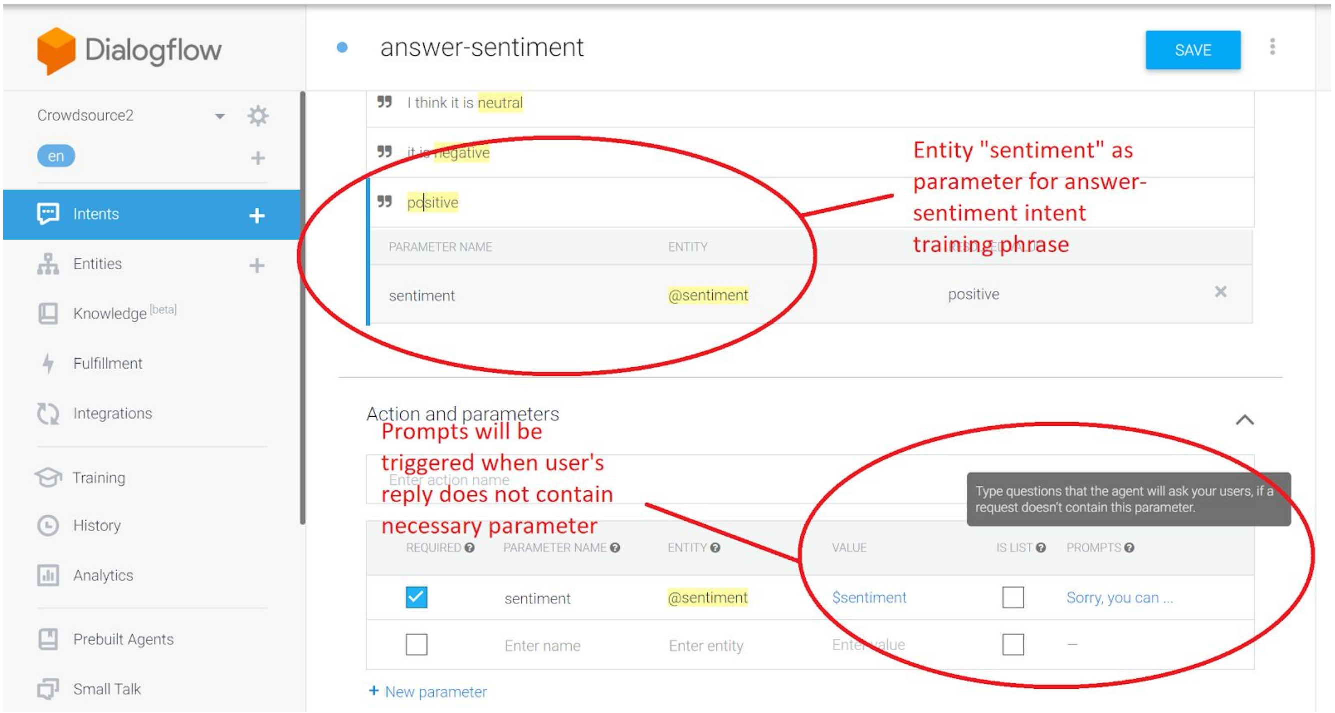
☑ Define synonyms □ Allow automated expansion

negative	negative
positive	positive Enter synonym

Click here to edit entry

+ Add a row

After creating this, it can be included in the answer-sentiment intent which are used to answer sentiment analysis task.



answer-sentiment

- I think it is neutral
- it is negative
- positive

PARAMETER NAME	ENTITY	
sentiment	@sentiment	positive

Entity "sentiment" as parameter for answer-sentiment intent training phrase

Prompts will be triggered when user's reply does not contain necessary parameter

Type questions that the agent will ask your users, if a request doesn't contain this parameter.

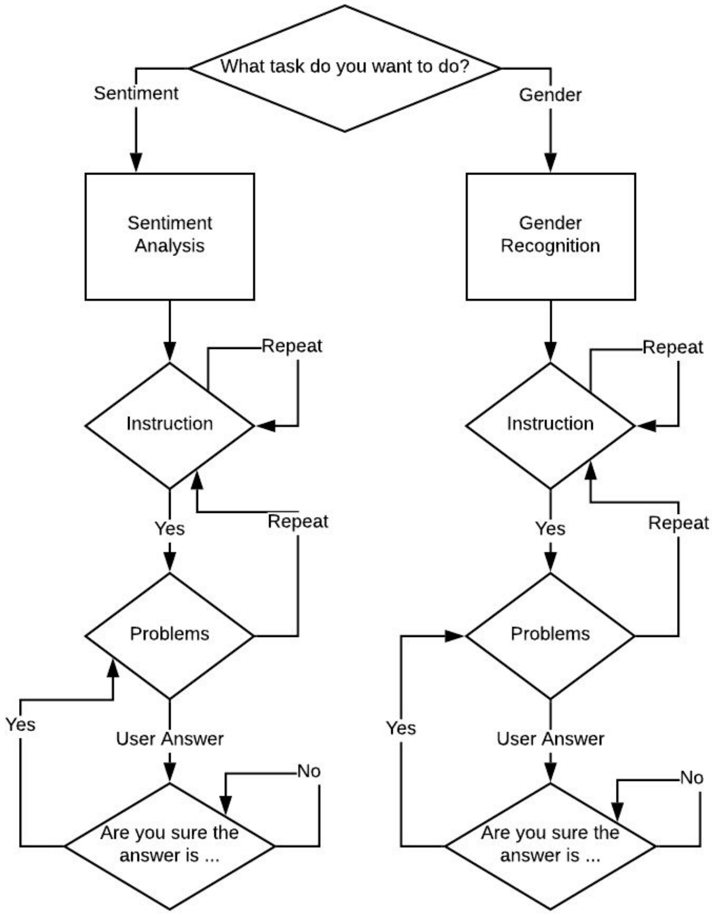
REQUIRED	PARAMETER NAME	ENTITY	VALUE	IS LIST	PROMPTS
☑	sentiment	@sentiment	$sentiment	□	Sorry, you can ...
□	Enter name	Enter entity	Enter value	□	–

By doing this, it made sure that participants' input will contain the necessary parameter therefore preventing invalid input and error. In this case, participants can only answer if their input contains 'positive' or 'negative' statement.

All of these concepts are then used to design the smart speaker software which will be shown in the next section.

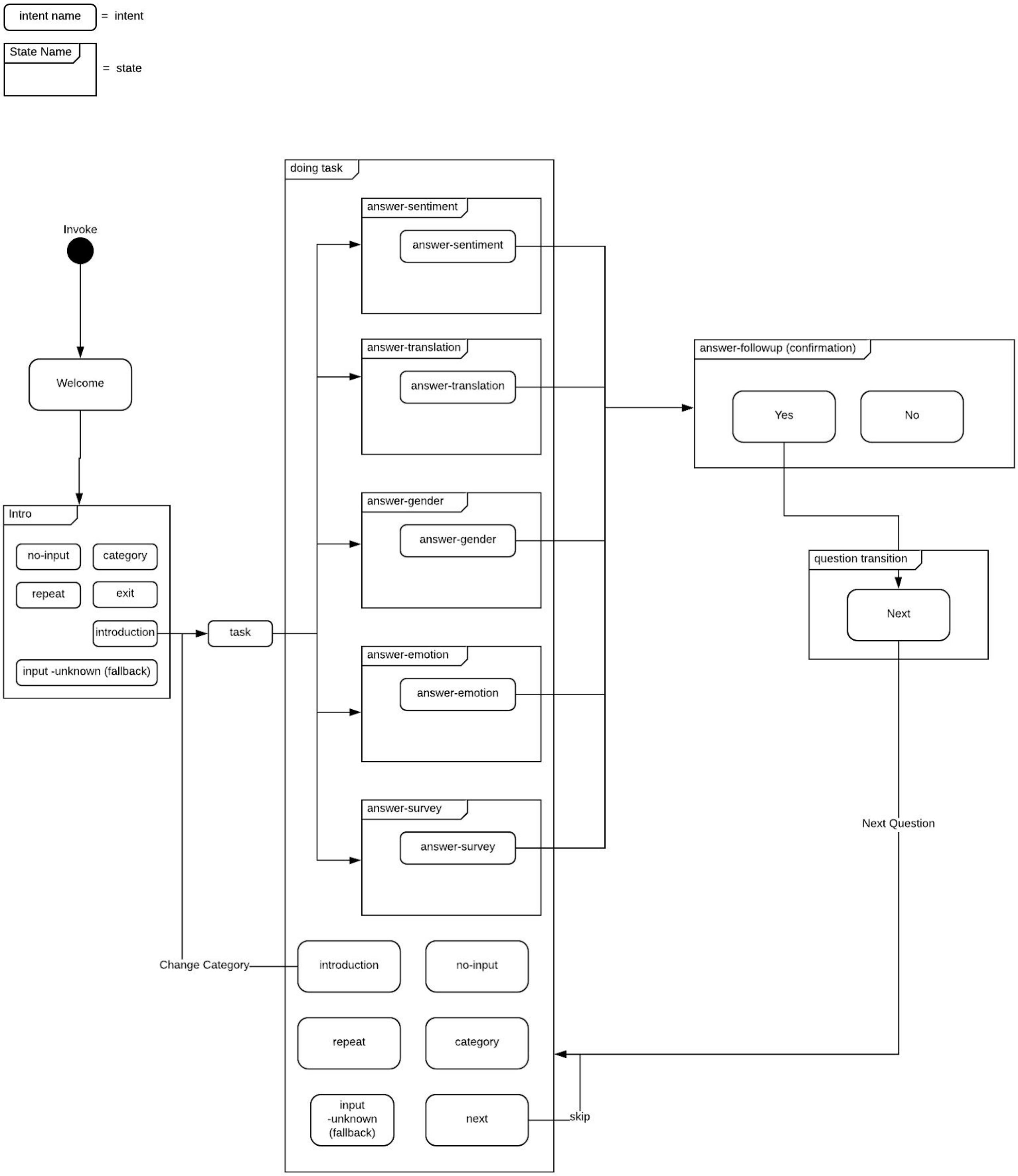
Design Phase

The design for the software is a bit different than that of usual software programming as it requires a different way of thinking compared to normal coding. According to a Google Developer Expert, Firstenberg (2018), coding a google home application is more similar to designing a conversational exchange rather than making a usual software. Designing a regular software usually follows similar patterns such as making a flowchart in order to gather similar parts to make functions that can be used multiple times in the programs. He mentioned that this way of thinking is not suited for building voice application as it mimics human conversation that needs to handle varied responses and does not necessarily follow certain logic flows. To visualize this, I present the first flowchart attempt at creating this application below.



This figure follows the basic patterns of coding a general application. However when designing a voice application, several problems can be seen with this diagram. Firstly, from the figure above, it can be seen that the logic flow between the two task categories, sentiment analysis and gender recognition looks identical and might be able to be put on the same functions. However, this will create problems as the input training phrases that we expect are different between the two tasks. Sentiment analysis expect a positive or negative answer while the gender recognition task expect a male or female answer from the users. Even though technically it can be put together by force using conditionals, the function will grow exponentially complicated if we add more task categories. In addition to that, user might also do unexpected things such as changing task category in the middle of another category, giving an unpredictable answer to a task question, etc. These will be hard to handle if we put it all into the same functions.

To solve this, Firstenberg (2018) advises to focus more on what the user is saying and use the webhook fulfillment to control the conversation by changing the state of the conversation using contexts. With the use of contexts, the conversation can be controlled to some extent by making some intents available only during certain conversation state which is differentiated by what contexts are currently active during that period. This leads to the prototype design that is reflected in the following diagram.



This diagram represents the logic sequence of the voice crowdsourcing software. To begin with, after invoking the application, the user will be greeted with the welcome and short introduction of the system in the ‘welcome’ intent. Then the state of the application will change to ‘intro’ where participants can start replying to the speaker Note that in this state there is no active context yet so all the intent here are available during all state of the application. In this state, users can call the desired task name for example ‘sentiment analysis’ and will be taken to the task introduction in the ‘introduction’ intent. In this intent they will be briefed with the instruction of what to do in the current task and finally prompted to say ‘start’ in order to start doing the task. When users say ‘start’ in this state, the ‘task intent will be triggered and will select the first task of the current task category and read it to the user. From here on, the state is changed to ‘doing task’. In this state, there are 5 sub-state corresponding to each task category: ‘answer-sentiment’ for sentiment analysis, ‘answer-emotion’ for emotion analysis, ‘answer-language’ for language translation, ‘answer-gender’ for gender recognition, and ‘answer-survey’ for survey. In addition to that, users can also change to another task category or skip the current question with the ‘next’ intent. In each of these sub-states, user input need to contain answer parameters that corresponds to the current task. In this case, since the task is ‘sentiment analysis, the available sub-state is answer-sentiment where users’ answer must contain the word ‘positive’ or ‘negative’. If the input does not contain either of those words, the speaker will prompt the user until they give a valid input. Next, after the user successfully answer the question, the speaker will repeat their answer in the ‘confirmation’ state and prompt them to confirm they answer with a yes or no in order to avoid accidental input by user. After confirming their answer, the speaker will prompt the user that their answer has already been submitted and they can say ‘next’ in order to move on to the next question through the ‘next’ intent or change to another category by saying the category name if the questions have run out. If this happens, the state will go back to the ‘intro’ state again repeating the same sequence for the next task category.

Experiments

There are 2 parts of the experiment that is done for this application to test its performance as a crowdsourcing platform. The first part of the experiment is an evaluation study that is used to test the flow of the application and add some improvements before the real test. The second experiment is the real test where all participants must answer all questions and their responses will be recorded for the data analysis.

A total of 18 participants are recruited for this study. The participants are all undergraduate Indonesian students that never used any smart speaker before. All of them have English as their second language but are quite proficient in them. For the first evaluation experiment, 3 participants will try 1 of the task categories in order to evaluate the flow of the program since the flow for all tasks are similar. Next, the second experiment will involve 15 participants to answer all tasks with any kind of order except survey task which is scheduled to be done last.

The recruitment process uses the snowball sampling method to make sure that the participants are reliable and provide valuable input for the study. This is because in this study there is no incentive for the participants yet so in order to make sure that the input that we get is valid and usable, there is a need to select participants that is willing to help with the study with no reward. Furthermore, one of the required criteria to this study is participants that understand and proficient in both English and Indonesian language as one of the tasks that is required to be done is translating Indonesian sentences to English. This method is perfect for this case as the participants will be recruited by its peers with similar criteria effectively reducing cost and time in gathering participants (Sadler et al, 2010).

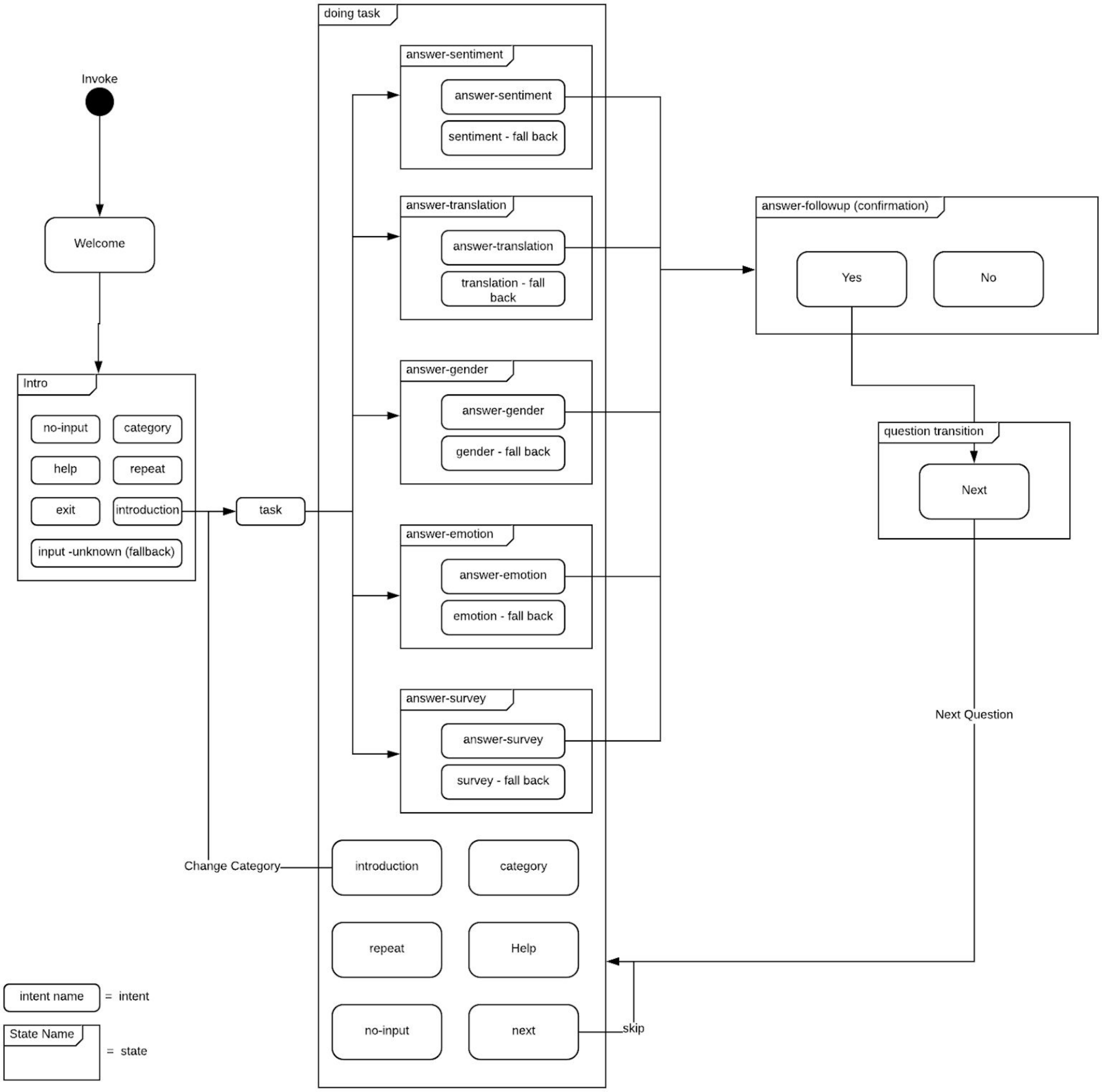
FInally, for the experiments itself. The first experiment will use 2 methods which are user observation and interview which are good to engage and encourage participants to give more details input (Lazar, Feng, & Hochheiser 2017). This method is chosen because participants input is necessary in order to fix problems and improve the flow of the system early in the design before moving on to the phase 2 of the experiment. The second experiment uses mainly user observation to gather data while participants are interacting with the device. For the result, a combination between quantitative and qualitative data from the observation and participants' task answer will be analysed to evaluate the smart speaker performance as a crowdsourcing platform.

Results and Discussion

Experiment 1

In experiment 1, most participants are overwhelmed with the amount of information spoken by the speaker especially during instructions at the start of every task. 2 out of 3 test participants felt that due to a lot of information that is given in each turn of the conversation, they do not know when they are supposed to reply to the speaker. Furthermore, all participants stated that the smart speaker talk too fast so they are having a hard time digesting what is said. The third problem is about the fallbacks. All participants felt annoyed that they return to the start of the conversation when they answer one of the questions wrong.

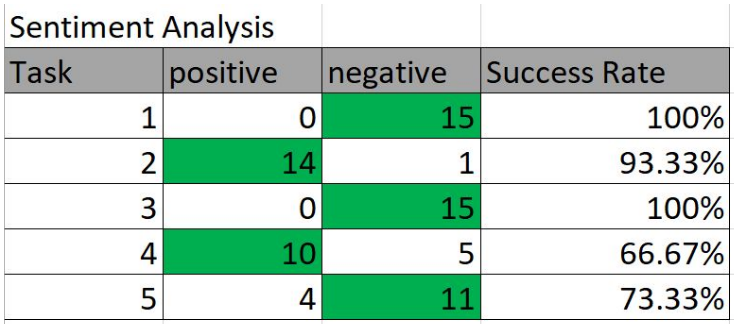
These problems lead to several improvements to the flow of the program. Firstly, instead of putting all the instruction in every start of the task, a help intent is added so participants can optionally call for it when they want to know what commands are possible to call. This will also help more experienced users in the future that has already used the device before, to skip some of the instructions as an accelerator increasing its flexibility and efficiency of use (Nielsen, 1994). Next, for the fast audio problem, a feature called Speech Synthesis Markup Language (SSML) was introduced. Using this feature allow for slight modification to the speech of the speaker. It made it possible to add natural pauses at the end of each sentence so participants can understand the smart speaker speech better. In this case, a 300ms pause between each sentence were added for every speaker response. Finally, the third problem which is related to handling participants wrong response is quite tricky to solve. This is because of how the smart speaker system is built. Initially, this problem is expected to be solved by the entity parameters mentioned in the system architecture section. However, during the evaluation testing this failed because the intent itself is not triggered since the user's input does not contain any word that is similar with the training phrase. Because of this, the parameter prompt which is inside the answer intent cannot be called by the speaker. This event results in the speaker moving to the 'input-unknown fallback' intent which is the default fallback intent when the speaker does not find any matching intent for the user's input. To solve this issue, custom fallback intent for every answer sub-states are created. It is done so that instead of falling to the default fallback intent when giving an invalid response, the speaker will call the custom fallback intent that act as the prompt substitution for each answer sub-states. After fixing all the issues, the design structure is updated in the following diagram.



Here it can be seen that there are new intents that are available in each state. The first intent that is added is the 'help' intent which is available on every state of the application and its content is updated accordingly to help user during anytime. Then, for every answer sub-states in the 'doing task' state, custom fallbacks are created for the purpose of handling invalid user input while answering the task questions. This updated design is the one that will be used in the next phase of the experiment.

Experiment 2

- **Sentiment Analysis**

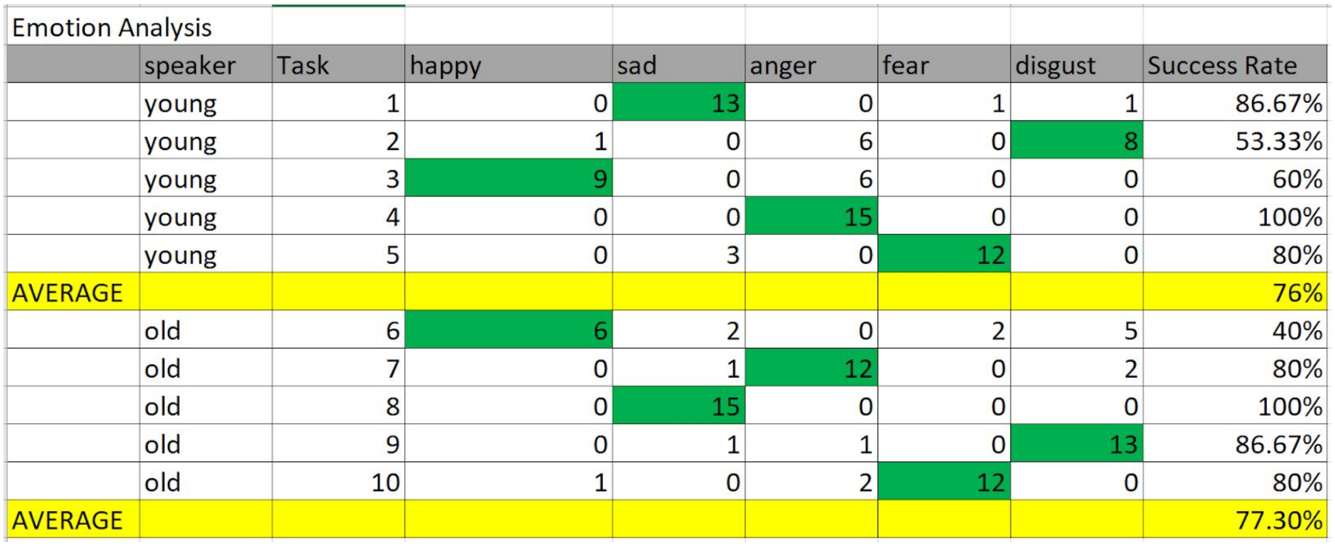


Sentiment Analysis

Task	positive	negative	Success Rate
1	0	15	100%
2	14	1	93.33%
3	0	15	100%
4	10	5	66.67%
5	4	11	73.33%

In this task, Most participants were able to pick the right choice. However the performance seems to be getting lower on harder difficulty as seen in task 4 and 5. On task 1 to 3, the success rate is over 90% while tasks 4 and 5 is 66.67% and 73% respectively. This is understandable since the tasks are growing in difficulty in terms of its complexity and sentence length. From task 1 to task 3, the difficulty varies in terms of indirectness of the sentence while task 4 and 5 focus more on the length differences. This shows that in this experiment, sentence length affects more of the participants performance rather than indirect meaning or sarcasm. This might be related to participants attention span while doing the task. When doing the test, most of the participants lost focus on task 4 and task 5 due to the length of the movie reviews. When asked, most of them relied on finding the first sentence that signifies any sentiment and answer based on that due to losing interest because of the length of the reviews. This can be seen by the difference in the success rate of task 4 and 5. In task 5, the last sentence of the reviews reveal its sentiment therefore gaining higher success rate compared to task 4 because participants can spot it easier. The result of this experiment shows that voice crowdsourcing may not be very effective for tasks that require participants to listen to a long narration and a lot of attention as the participants might lose focus easily.

- **Emotion Analysis**

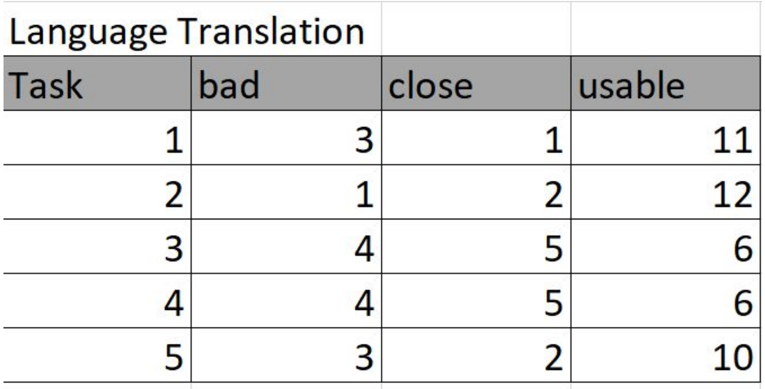


Emotion Analysis

	speaker	Task	happy	sad	anger	fear	disgust	Success Rate
	young	1	0	13	0	1	1	86.67%
	young	2	1	0	6	0	8	53.33%
	young	3	9	0	6	0	0	60%
	young	4	0	0	15	0	0	100%
	young	5	0	3	0	12	0	80%
AVERAGE								76%
	old	6	6	2	0	2	5	40%
	old	7	0	1	12	0	2	80%
	old	8	0	15	0	0	0	100%
	old	9	0	1	1	0	13	86.67%
	old	10	1	0	2	12	0	80%
AVERAGE								77.30%

In this task, we compare the audio emotional expression between young people (26 years old) and old people (64 years old). The result shows that there is not much different on the overall performance between the two age groups with young people showing 76% success rate and old people showing 77.3% success rate. However, there is a difference between the success rate of each emotion analysis task between the old and young speaker. Participants seem to be able to predict sad and anger best in the young people while find disgust emotion to be the hardest to analyse. This is reflected by the success rate comparison of the three were sad and anger score 100% and 86.67% respectively while disgust only score 53.33%. This result is similar to a study done by Dupuis & Pichora-Fuller (2011) where they found that sad and anger emotion that is expressed through speech is easier to be recognised than other type of emotions. They also found that disgust is one of the hardest emotions to analyse. The emotion disgust is mainly mistaken to be anger by the participants in the case of young speaker. This similarity may be due to the speaker expressing “cold anger” rather than “hot anger” (Hammerschmidt & Jürgens, 2007). However, in the case of old person as speaker, only happiness has the lowest success rate compared to the other emotions. This is interesting as most participants confused it with disgust which are a completely opposite emotion type. Happiness is considered a positive emotion and disgust is considered a negative emotion. In this case there is no research that correlates to this finding and therefore this may be caused by an error during the sampling. The result of this experiment shows that voice crowdsourcing can be used for tasks that have multiple choice questions with several predetermined answer.

- **Language Translation**

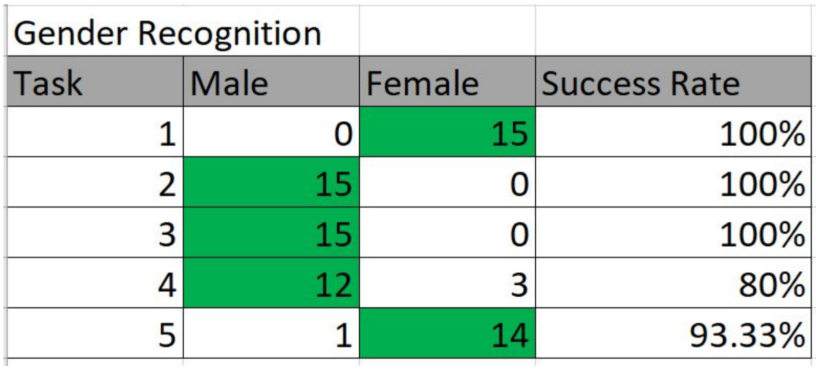


Language Translation

Task	bad	close	usable
1	3	1	11
2	1	2	12
3	4	5	6
4	4	5	6
5	3	2	10

In this task, participants seems to struggle giving the right response to the speaker, especially on the harder difficulty oh higher task numbers. This can be seen by the number of usable translation that keeps on decreasing on higher task numbers, especially on task 3 and 4 where only 6 translations are usable. However, task 5 which are supposed to be more complicated than the others, receives 10 usable translation which are interesting. This may be caused by the difference in translation vocabulary that is used in each question. Often, the lower quality translation is caused by some words getting misheard by the speaker into different similar sounding words due to different types of pronunciation. The problem is usually solved by confirming the input first with the participants before putting the data into the database which has already been done in this study. However, similar sounding words may slip through this error prevention method. This is the case for task 3 and 4 where the translation use many words that are prone to be misheard. Some examples are “birth” = “bird”, “they’re” = “their” = “deer”, “large” = “lodge”. Furthermore, depending on the accent it might get worse. Some participants try to get around this by translating the sentence differently but it usually resulted in lower quality translation which is not good enough to be usable. These findings show that voice crowdsourcing is not suited for tasks which answer has no restriction since it is difficult to check whether the input is exactly what the user wants.

- **Gender Recognition**



Gender Recognition

Task	Male	Female	Success Rate
1	0	15	100%
2	15	0	100%
3	15	0	100%
4	12	3	80%
5	1	14	93.33%

In this task, overall success rate of gender recognition by voice is high. Most participants are able to guess the right gender just by listening to a short audio even with increasing difficulty especially in task

4 and 5 where the speaker voice. The lowest success rate for this task category is on task 4 which is 80% where only 3 out of 15 participants failed to guess the gender of the speaker in the audio. This is expected as human listeners are generally capable of analysing various information including gender from only an acoustic signals (Childers & Wu, 1991). This result shows that voice crowdsourcing is perfectly capable of being used in simple task such as this which only have 2 predetermined answer and does not require much focus.

- **Survey**

Firstly, only 2 people out of 15 agreed to have experience doing intentional crowdsourcing task like this. In terms of difficulty of use, the voice crowdsourcing platform on average is rated 3.6 out of 5 by all participants which are still considered difficult but understandable since this is the first prototype. Next, the task that are considered the most difficult one by the participants are Language Translation followed by Emotion Analysis. 9 out of 15 people picked Language Translation due to the difficulty in getting their response heard by the speaker during the task. The speaker often misheard some of the words causing participants to be annoyed since they need to repeat their answer again until the speaker gets it right. In addition to that, when participants are repeating their answer slowly in hope that the speaker will better understand them, the microphone closes in the middle of the sentence resulting in an unfinished sentence when the speaker is confirming the input to the user. All of this caused some participants to be frustrated during the task resulting in 4 out of 15 people confirming their answer because it is 'close enough' even though some words are still wrong. For the emotion analysis task, 6 people chose this as the most difficult one due to the amount of questions it had which are a lot compared to other tasks. Next, when asked whether to choose computer or a smart speaker to do these tasks, only 2 people choose smart speaker because they realized that they can do other things while doing these tasks. Finally, following the research of Ross et al., (2009) which stated that an average Mturk workers only gain on average $2/hour, all participants declined when offered the same pay showing that the average crowdsourcing payment are too small compared to the effort of the participants.

- **Overall Usage Observation Analysis**

Several interesting things can be noticed when the participants are interacting with the smart speaker. Firstly, regarding the 'help' intent that are added to minimize the amount of information spoken by the speaker. Only 1 participant invoke this intent during their interaction while the other just use the speaker with trial and error based on their own intuition. This raises a question in whether the help intent is really necessary for this kind of application because of its rare use. Next, 4 participants are taking notes to keep up their hearing especially on task such as emotion analysis which have many answer options. This shows that there must be special consideration when doing tasks that require long instruction or have multiple answer options such as multiple choice survey. Furthermore, 11 people accidentally exited the application in the middle of the test due to the smart speaker mistakenly heard the 'start' command because it is similar to the word 'stop' which is the default exit command. This should be considered in future application to pay attention to similar sounding words in order to reduce the possibility of miscommunication between the user and the system such as this. Finally, several participants are also observed doing other things such as playing a phone game or online chatting when interacting with the speaker. Although not all participants were seen doing this, it can infer that one of the aims to create a voice crowdsourcing application that can be done simultaneously with other activities is a success.

Limitations

There are several limitations for the first prototype of this voice crowdsourcing device.

- **Microphone Quality**: Although the smart speaker microphone is considered better compared to other devices for capturing user audio input, it still have many difficulties capturing input from users with diverse accent and pronunciation, especially those whose first language is not English as seen in this study. An example of this is shown during the task language translation where participants need to repeat several times before getting their answer heard right by the speaker.
- **Speech Composition**: Since this study is more focused to test whether the smart speaker is capable of doing data crowdsourcing, there is not enough research on making a good speech composition to be used in the application. Therefore, only natural pauses at the end of each sentence are added to the smart speaker speech. This may be led to some increase to the difficulty in using the speaker as the speeches are still sound unnatural. Future studies may focus more on this part to create a guide for future application.
- **Time and Resource**: Since this is a short study, there is not enough time and equipment to recruit many participants in order to gain more data for testing. Future studies should have more participants or more equipment in order to gain more varying data.

Conclusion

In the end, we concur that it is possible to do a voice based crowdsourcing using a smart speaker. From all of the tasks and observations during the use of the device, we can conclude several things about the use of voice crowdsourcing via smart speaker. Through the tasks, we found out that voice crowdsourcing is perfectly suited for doing short and simple tasks that have answer restriction such as gender recognition and emotion analysis. Tasks that have long question or need a lot of focus such as sentiment analysis may poses some problems since most participants tend to lose focus if the questions are too long. Finally, tasks that have no restriction on the answer such as language translation is not really suited for this application yet as the smart speaker still have a lot of trouble to get the right input of the user. Future studies may compare the performance of the same tasks on the computer to see if the performance is comparable.

Bibliography

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U. & Narayanan, S. (2004, October). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205-211). ACM.

Childers, D. G., & Wu, K. (1991). Gender recognition from speech. Part II: Fine analysis. *The Journal of the Acoustical society of America*, *90*(4), 1841-1856.

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., & Cudré-Mauroux, P. (2015, May). The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th international conference on world wide web* (pp. 238-247). International World Wide Web Conferences Steering Committee.

Djenar, D. (2003). Students Guide to Indonesian Grammar.

Dupuis, K., & Pichora-Fuller, M. K. (2011). Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto Emotional Speech Set. *Canadian Acoustics*, *39*(3), 182-183.

Firstenberg, A. (2018). Thinking for Voice: Design conversations, not logic. Retrieved from https://medium.com/google-developer-experts/thinking-for-voice-design-conversations-not-logic-7c60869451f1

Gadiraju, U., Kawase, R., & Dietze, S. (2014, September). A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media* (pp. 218-223). ACM.

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice*, *21*(5), 531-540.

Hossain, M. (2012, May). Users' motivation to participate in online crowdsourcing platforms. In *2012 International Conference on Innovation Management and Technology Research* (pp. 310-315). IEEE.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1-4.

Ikeda, K., Morishima, A., Rahman, H., Roy, S. B., Thirumuruganathan, S., Amer-Yahia, S., & Das, G. (2016). Collaborative crowdsourcing with crowd4U. *Proceedings of the VLDB Endowment*, *9*(13), 1497-1500.

Juslin, P. N., & Scherer, K. R. (2008). Speech emotion analysis. *Scholarpedia*, *3*(10), 4240.

Kunchukuttan, A., Roy, S., Khapra, M., Cancedda, N., & Bhattacharyya, P. (2014). *U.S. Patent Application No. 13/592,736*.

Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In *Fifth International AAAI conference on weblogs and social media*.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann.

Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., & Potts, C. (2011). *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*(pp. 142 - 150). Portland, Oregon, USA: Association for Computational Linguistics.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

Nielsen, J. (1994). 10 Heuristics for User Interface Design: Article by Jakob Nielsen. Retrieved from https://www.nngroup.com/articles/ten-usability-heuristics/

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2009). Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep*.

Sadler, G. R., Lee, H. C., Lim, R. S. H., & Fullerton, J. (2010). Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. *Nursing & health sciences*, *12*(3), 369-374.

Gender Recognition Datasets Retrieved from http://www.voxforge.org/