

=.53zw plus 3pt minus 3pt =.53zw plus 3pt minus 3pt

# 目次

第 1 章 序論	3
1.1 研究の背景	3
1.2 研究の目的	4
1.3 論文の構成	5
第 2 章 飲食店での接客について	6
2.1 声とは	8
2.2 各社音声認識用ソフトの調査	9
2.3 音声認識とは	10
2.4 各社音声合成用ソフトの調査	11
2.5 音声合成とは	12
第 3 章 検討用のシステムについて	14
第 4 章 音声認識実験	16
4.1 実験目的	16
第 5 章 結言	22
5.1 本研究のまとめ	22
5.2 今後の課題	22
参考文献	23
謝辞	24

# 第1章

## 序論

### 1.1 研究の背景

近年ソフトバンク社の孫社長が Pepper を開発した。また、長崎県に存在するテーマパークで宿泊する際に利用すると考えられる「ハウステンボス」の受付には女性タイプと恐竜タイプの接客ロボットや話しかけることで照明のオンオフを行ってくれるロボットなどが配備されている。本研究は飲食店にて接客業務での人件費削減を行うため、ロボットを制作してほしいという要望があった。そこで音声でのやりとりをするためのシステムを究明する必要があると考え、本研究を開始した。飲食店で注文を取るロボットに使用した音声を認識させるシステムについて今から述べていく。図 1.1 に pepper、図 1.2 にホテルの接客ロボットの外観を示す。

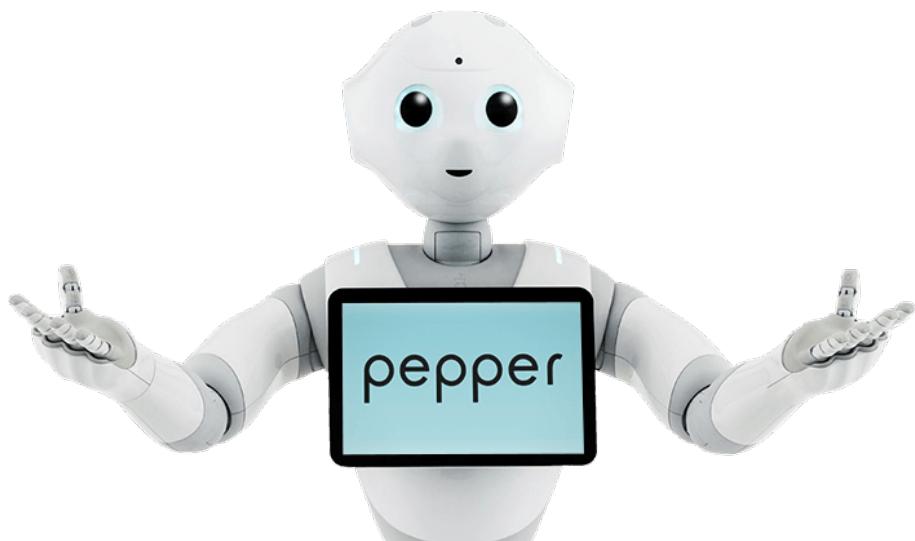


図 1.1 様子



図 1.2 様子

## 1.2 研究の目的

飲食店にて PC を実装したロボットに音声認識をするため Julius, Open-JTalk, 認識すべき単語, マイクを準備し飲食店で禁煙か喫煙かの判断や「あちらの席へどうぞ」「こちらの席へどうぞ」のようなレベルの案内をすることを目的とする。図 1.3 に音声認識をしようとしたときのイメージ図を示す。

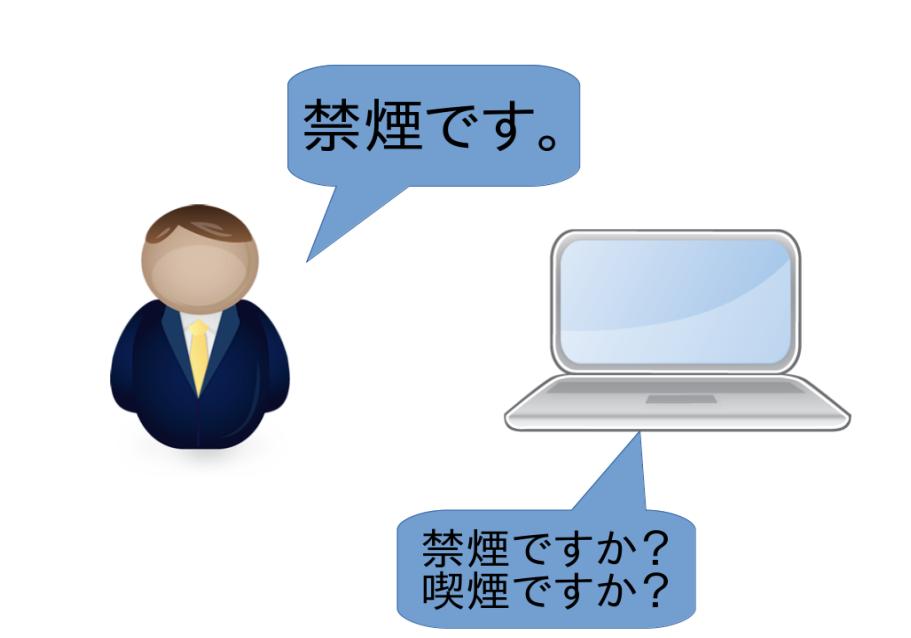


図 1.3 様子

### 1.3 論文の構成

この論文では以下のように構成されている。

1章では、音や音声について、研究背景について考察したものである。

2章では、本研究に使用する、音声を認識するための道具及び発生する音声について研究したものである。

3章では、本研究のために構築した検討用のシステムについて研究したものである。

4章では、前章で述べた検討用のシステムを使用して行った実験について研究したものである。

5章では、本研究のまとめや今後の課題について研究したものである。

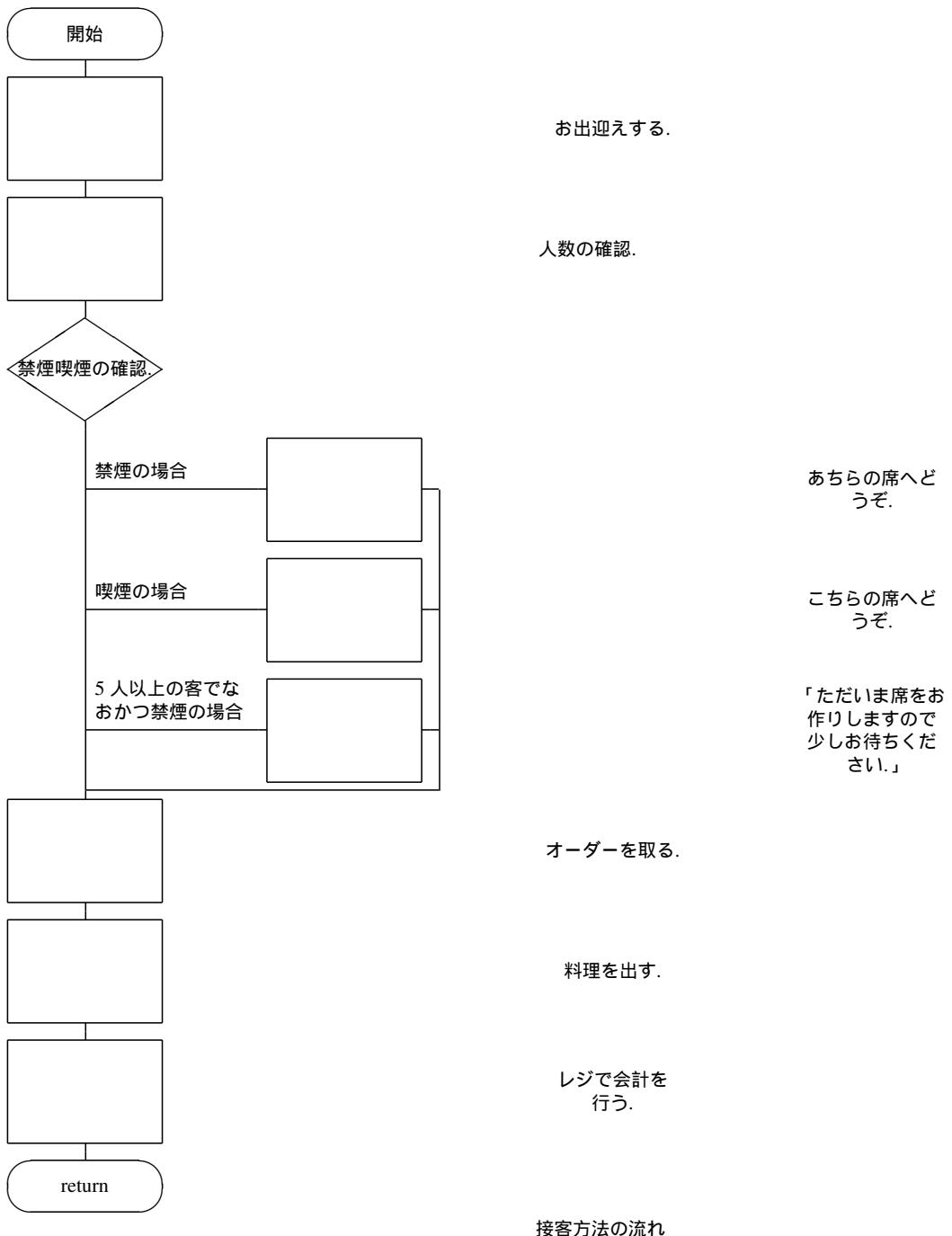
6章では、本研究において大きな力となった関係者各員に謝辞を記す。

## 第2章

# 飲食店での接客について

### 2.0.1 接客の業務内容

お客さまから注文を取り、出来上がった料理や飲み物を運び、空いた皿の片付け、レジ業務を行うのが主な仕事である。その接客の業務内容の流れをフローチャートを図 2.0.1 に示す。



## 2.0.2 研究として検討する接客の業務内容

前述では接客の仕事内容をフローチャートとして挙げたが今回、接客をするロボットの開発を行ったにあたり、店の入口付近で客がタバコを吸うかどうかの判断や「あちらの席へどうぞ」「こちらの席へどうぞ」の案内をすることが研究として検討する接客の業務内容となる。

## 2.1 声とは

声とは空気の振動であり波である。つまり  $\sin$  カーブで表すことができ、人間が声として発する「あ」という音一つから「卒業研究」のような単語、「私の職業は学生であるため勉強をすることが仕事である。」と言った文章を声で発している限り  $\sin$  カーブで表すことができる。人間の声をマイクに入れてそのデータをグラフのように近いグラフを書くためには特別な手段を用いる必要がある。それは周波数成分をチェックすることであり、音声にフーリエ変換を掛けることである。

### 2.1.1 高速フーリエ変換とは

前述の通り声は空気の振動のため実際に視覚的に見ようとした場合、特別な手段を用いなければならない。ここではマイクを使い視覚的に表示することで音声の波形を見る能够と考え、実際に audacity という音声の編集ソフトを使用して音の波形を視覚的に確認することに成功した。ここで、音の波形を高速フーリエ変換にかける。高速フーリエ変換は英語で Fast Fourier Transform と呼ばれていることから頭文字を取って FFT とも呼ばれている（以下 FFT とする）。フーリエ変換に各音専用の係数をかけることで FFT を行うことができる。図に「あ」の音声波形、図に「い」の音声波形、図に「う」の音声波形、図に「え」の音声波形、図に「お」の音声波形、図に「あ」に FFT をかけた波形、図に「い」に FFT をかけた波形、図に「う」に FFT をかけた波形、図に「え」に FFT をかけた波形、図に「お」に FFT をかけた波形を示す。

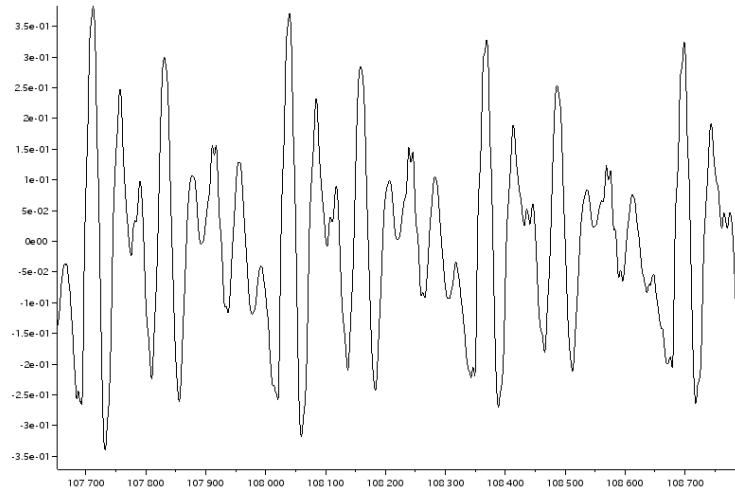


図 2.1 あの音声波形

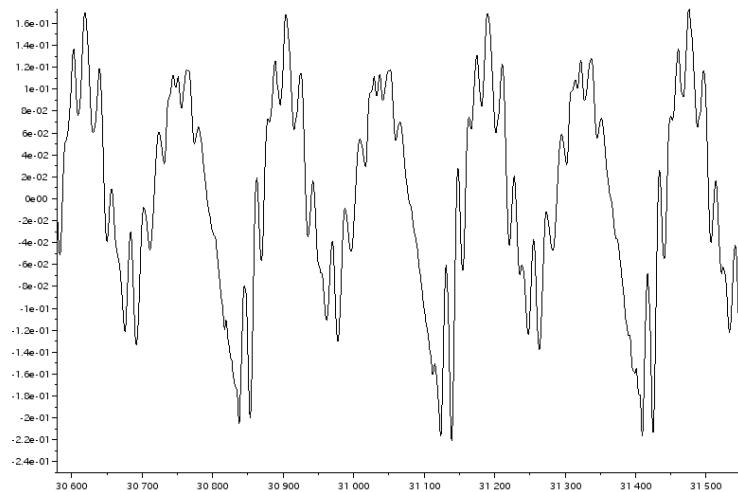


図 2.2 いの音声波形

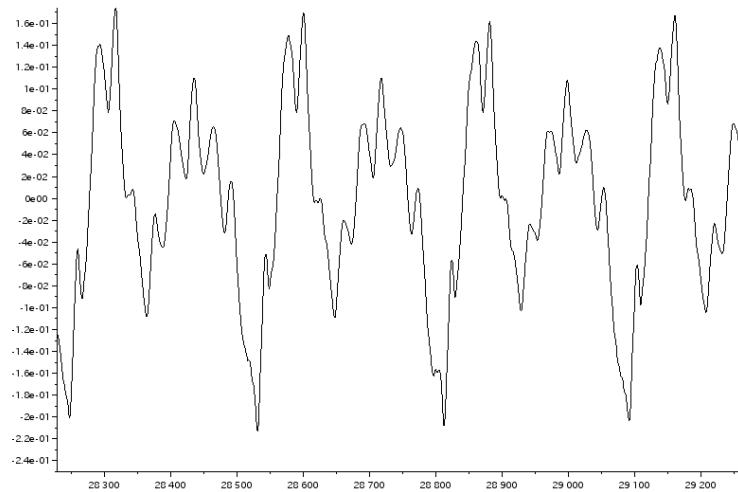


図 2.3 うの音声波形

## 2.2 各社音声認識用ソフトの調査

音声によるロボットを制作するにあたり、まずは音声を認識することが必要条件である。そのため、音声を認識するプログラムを作成し使用することが考えられるのだが時間的コストがかかることが予想された。そのため音声を認識するためのソフトを調査した。調査した結果,Julius, ドラゴンスピーチ,docomo 音声認識 API,Ami Voice,Google Speech API の 5 つが候補として挙がった。本研究では

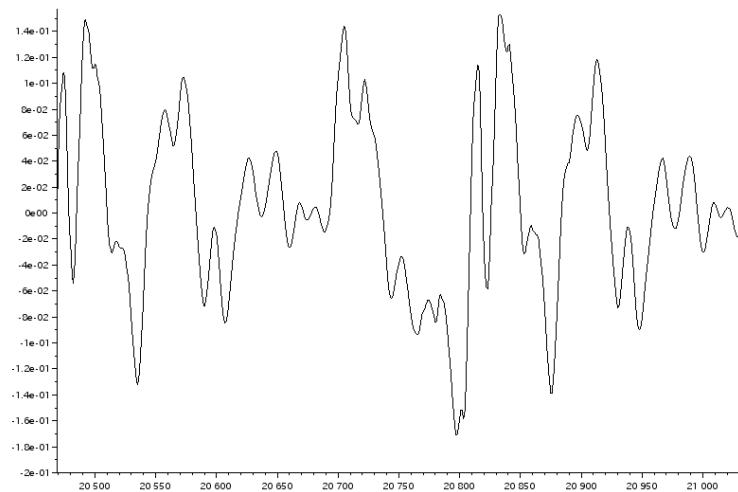


図 2.4 えの音声波形

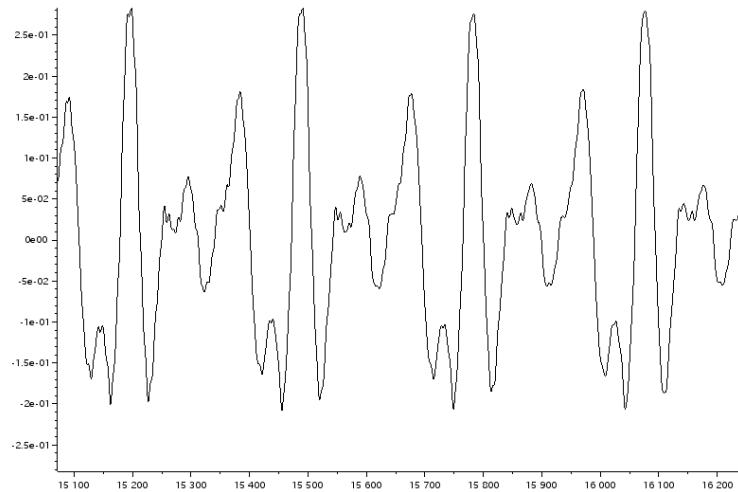


図 2.5 おの音声波形

linux でも使用でき無料かつ使用期間の制限がない Julius を選択した. 表 2.1 に音声認識一覧を示す.

## 2.3 音声認識とは

音声認識とは音声をマイクより入力し, 入力された音声を文字として表すことを指す.

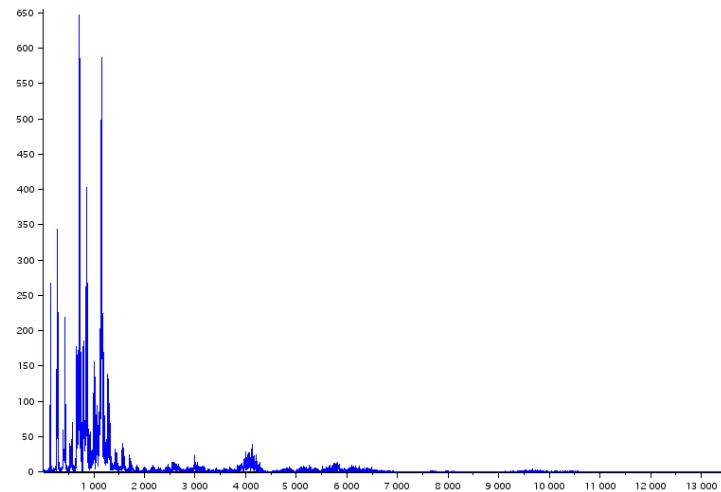


図 2.6 あの fft 後の波形

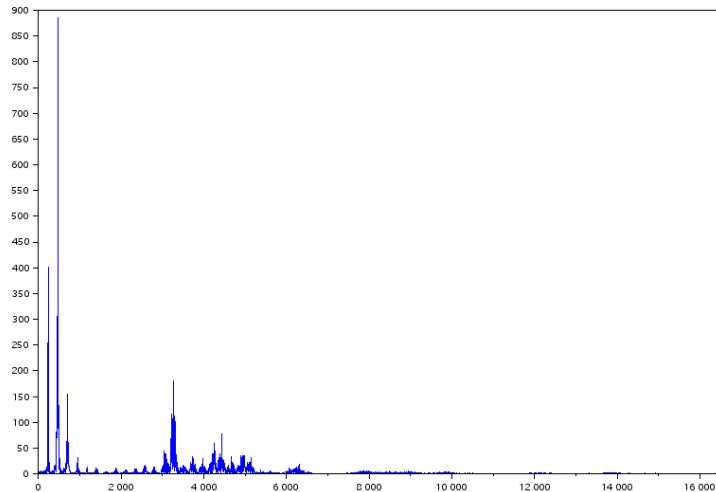


図 2.7 いの fft 後の波形

## 2.4 各社音声合成用ソフトの調査

接客ロボットは音声を認識した後に音声を発して店内での案内も行わなければならない、よって音声合成ソフトを使用することで音声を作成し c 言語を用いて発声させることを考えた。そこで音声を作成するためのソフトの調査を行った結果,Open-JTalk,Vocaloid,AITalk,VoiseText, テキストオーディオ,Clipboard Translator, 棒読みちゃん, 歌声合成ツール UTAU の 8 つのソフトが候補に挙がった。本研究ではまず音

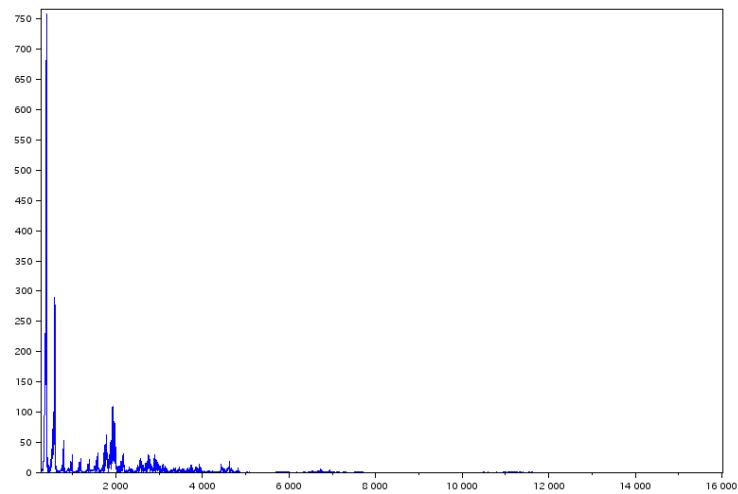


図 2.8 うの fft 後の波形

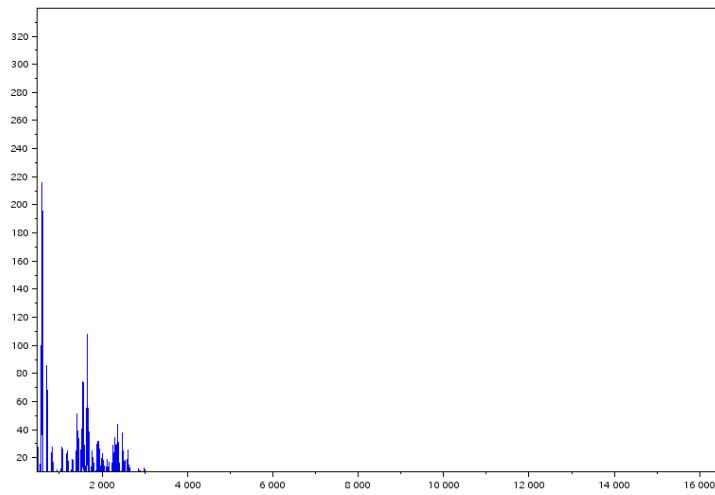


図 2.9 えの fft 後の波形

声を発声させることができ、なおかつ音声の速さが変更できる Open-JTalk を選択した。表 2.2 に音声合成一覧を示す。

## 2.5 音声合成とは

音声合成とは人間の音声を人工的に作り上げることを指す。

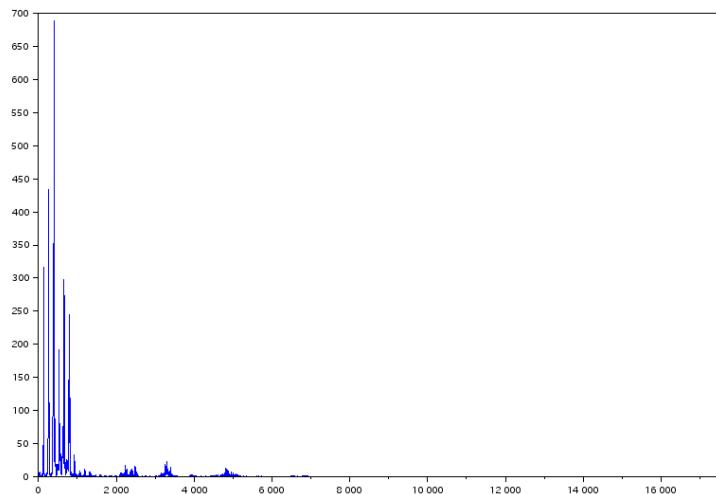


図 2.10 おの fft 後の波形

表 2.1 音声認識一覧

ソフト名	無料である	使用期限がない	認識率
× 不明 heightAmi Voice	heightdocomo 音声認識 API × ×	90 % height ドラゴンスピーチ × ×	不明 heightG
		不明	

表 2.2 音声合成一覧

ソフト名	音の早さが変更出来る
heightOpen-JTalk なし heightVocaloid 歌声合成ソフト heightAITalk 海外の言語は 36 種類 heightVoiseTextv なし height テキストトーク Open-JTalk と提携した場合にのみ女性の声が使用可能 heightClipBoard Translator Windows のみ使用可能 height 棒読みちゃん なし height 歌声合成ツール UTAU 歌声合成ソフト	不明 不明 不明 不明 不明

## 第3章

# 検討用のシステムについて

今年度のシステムは図 3.1 に示すシステムを使用して音声認識を行う。



図 3.1 使用したシステム

### 3.0.1 実験で使用するマイク

本研究では buffalo 製のマイク (BSHSM04BK) を使用した。その図を図 3.2 に示す。

### 3.0.2 Julius

Julius は、音声認識システムの開発・研究のためのオープンソースの高性能な汎用大語彙連続音声認識エンジンである。数万語彙の連続音声認識を一般の PC 上でほぼ実時間で実行できる。また、高い汎用性を持ち、発音辞書や言語モデル・音響モデルなどの音声認識の各モジュールを組み替えることで、様々な幅広い用途に応用できる。<sup>[1]</sup>Julius はコマンド 1 つで起動させることができ、オプションでサー



図 3.2 使用したマイク

バモードで使用できるため、今回はサーバモードで起動し、クライアント側で認識した単語に対応した音声を発生させることを目指す。また本研究で使用した Julius は最新版としてバージョン 4.4.2 が平成 28 年 9 月 12 日に Github にて更新されている。しかし、研究を始めた時点でバージョン 4.3.1 が最新版だったためこれを継続して使用することとした。

#### Julius の仕組み

Julius の音声認識アルゴリズムは、ツリートレリス探索方式を基礎とするアルゴリズムである。全体は 2 パス構成となっており、2 段階に分けて認識処理を行う。まず、第 1 パスでは入力全体に対して荒い認識処理を行い、有望な候補の集合をある程度絞り込む。このとき、簡便なモデルや近似計算を用いることで高速に処理を行う。第 2 パスでは詳細な認識処理を行うが、その際に第 1 パスの結果を参照しながら探索を行うことで、必要な部分にだけ精密な再計算を行って、最終的な最尤解を求める。[2]

#### 3.0.3 Open-JTalk

Open-JTalk は、日本語テキストを音声に変換するシステムである。ここでは案内のため使用する音声の内容を WAV ファイルとして保存するために使用する。今回はあちらの席へどうぞ、こちらの席へどうぞの 2 つを WAV ファイルとして保存した。

## 第 4 章

# 音声認識実験

### 4.1 実験目的

今回 Julius を使用して音声を認識させ, 100 回分の認識結果を集計し平均を取ることを目的とする.

#### 4.1.1 実験方法

マイクを実装したパソコンを使用し, Julius を端末より起動させる. Julius を使用し, 禁煙です, 吸わないです, 吸いません, 喫煙です, 吸いますの認識を行う. このとき 100 回分の認識したときと認識しなかったときの結果を集計する. 100 回分の認識結果の平均を式 4.1.1 より算出する.

$$x = a/100 \quad (4.1)$$

#### 4.1.2 実験結果

#### 4.1.3 Julius を用いて行った実験結果

今回 Julius を用いて実験を行い, 前節に示した式より算出した認識率は禁煙が 99 %, 喫煙が 100 %, 吸いますが 100 %, 吸いませんが 30 %, 吸わないですが 100 % という結果が得られた. 表 4.1 に禁煙, 表 4.1.3 に喫煙, 表 0 に吸いますが, 表 0 に吸いません, 表 0 に吸わないですがの認識結果, に認識結果一覧を示す. また, そのときのグラフを図 4.1 に示す.

表 4.1 禁煙の認識結果

回数	禁煙です	回数	禁煙です	回数	禁煙です	回数	禁煙です
1	1	26	1	51	1	76	1
2	1	27	1	52	1	77	1
3	1	28	1	53	1	78	1
4	1	29	1	54	1	79	1
5	1	30	1	55	1	80	1
6	1	31	1	56	1	81	1
7	1	32	1	57	1	82	1
8	1	33	1	58	1	83	1
9	1	34	1	59	1	84	1
10	1	35	1	60	1	85	1
11	1	36	1	61	1	86	1
12	1	37	1	62	1	87	1
13	1	38	1	63	1	88	1
14	1	39	1	64	1	89	1
15	1	40	1	65	1	90	1
16	1	41	1	66	1	91	1
17	1	42	1	67	1	92	1
18	1	43	1	68	1	93	1
19	1	44	1	69	1	94	1
20	1	45	1	70	1	95	1
21	1	46	1	71	1	96	1
22	1	47	1	72	1	97	1
23	1	48	1	73	1	98	1
24	1	49	1	74	1	99	1
25	0	50	1	75	1	100	1

回数	喫煙です	回数	喫煙です	回数	喫煙です	回数	喫煙です
1	1	26	1	51	1	76	1
2	1	27	1	52	1	77	1
3	1	28	1	53	1	78	1
4	1	29	1	54	1	79	1
5	1	30	1	55	1	80	1
6	1	31	1	56	1	81	1
7	1	32	1	57	1	82	1
8	1	33	1	58	1	83	1
9	1	34	1	59	1	84	1
10	1	35	1	60	1	85	1
11	1	36	1	61	1	86	1

吸いますの認識結果

回数	吸います	回数	吸います	回数	吸います	回数	吸います
1	1	26	1	51	1	76	1
2	1	27	1	52	1	77	1
3	1	28	1	53	1	78	1
4	1	29	1	54	1	79	1
5	1	30	1	55	1	80	1
6	1	31	1	56	1	81	1
7	1	32	1	57	1	82	1
8	1	33	1	58	1	83	1
9	1	34	1	59	1	84	1
10	1	35	1	60	1	85	1
11	1	36	1	61	1	86	1
12	1	37	1	62	1	87	1
13	1	38	1	63	1	88	1
14	1	39	1	64	1	89	1
15	1	40	1	65	1	90	1
16	1	41	1	66	1	91	1
17	1	42	1	67	1	92	1
18	1	43	1	68	1	93	1
19	1	44	1	69	1	94	1
20	1	45	1	70	1	95	1
21	1	46	1	71	1	96	1
22	1	47	1	72	1	97	1
23	1	48	1	73	1	98	1
24	1	49	1	74	1	99	1
25	1	50	1	75	1	100	1

吸いませんの認識結果

回数	吸いません	回数	吸いません	回数	吸いません	回数	吸いません
1	1	26	1	51	1	76	1
2	1	27	1	52	1	77	1
3	1	28	1	53	1	78	1
4	1	29	1	54	1	79	1
5	1	30	1	55	1	80	1
6	1	31	1	56	1	81	1
7	1	32	1	57	1	82	1
8	1	33	1	58	1	83	1
9	1	34	1	59	1	84	1
10	1	35	1	60	1	85	1
11	1	36	1	61	1	86	1
12	1	37	1	62	1	87	1
13	1	38	1	63	1	88	1
14	1	39	1	64	1	89	1
15	1	40	1	65	1	90	1
16	1	41	1	66	1	91	1
17	1	42	1	67	1	92	1
18	1	43	1	68	1	93	1
19	1	44	1	69	1	94	1
20	1	45	1	70	1	95	1
21	1	46	1	71	1	96	1
22	1	47	1	72	1	97	1
23	1	48	1	73	1	98	1
24	1	49	1	74	1	99	1
25	1	50	1	75	1	100	1

#### 吸わないですの認識結果

回数	吸わないです	回数	吸かないです	回数	吸わないです	回数	吸わないです
1	1	26	1	51	1	76	1
2	1	27	1	52	1	77	1
3	1	28	1	53	1	78	1
4	1	29	1	54	1	79	1
5	1	30	1	55	1	80	1
6	1	31	1	56	1	81	1
7	1	32	1	57	1	82	1
8	1	33	1	58	1	83	1
9	1	34	1	59	1	84	1
10	1	35	1	60	1	85	1
11	1	36	1	61	1	86	1
12	1	37	1	62	1	87	1
13	1	38	1	63	1	88	1
14	1	39	1	64	1	89	1
15	1	40	1	65	1	90	1
16	1	41	1	66	1	91	1
17	1	42	1	67	1	92	1
18	1	43	1	68	1	93	1
19	1	44	1	69	1	94	1
20	1	45	1	70	1	95	1
21	1	46	1	71	1	96	1
22	1	47	1	72	1	97	1
23	1	48	1	73	1	98	1
24	1	49	1	74	1	99	1
25	1	50	1	75	1	100	1

認識結果一覧

単語	禁煙です	喫煙です	吸います	吸いません	吸わないです
計	99	100	100	30	100
平均	0.99	1	1	0.3	1

#### 4.1.4 考察

実験結果より認識率を算出し認識率は禁煙が 99 %, 喫煙が 100 %, 吸いますが 100 %, 吸いませんが 30 %, 吸わないですが 100 % という結果が得られたが認識率が 99 %, 30 % の単語が存在するという点から雑音が混入した結果ではないかと考える。また 100 人のうち, 70 人の方が誤認識をする可能性も考

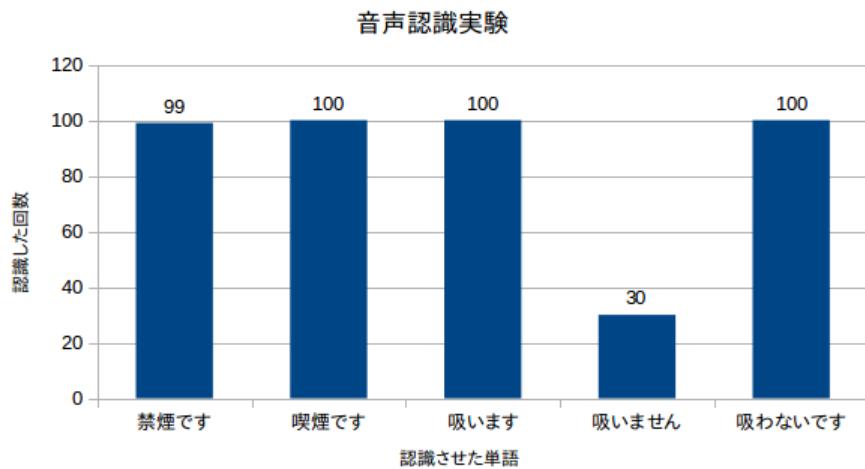


図 4.1 認識結果グラフ

えられるため、ノイズカットも必要になると考えられる。

# 第5章

## 結言

### 5.1 本研究のまとめ

音声を認識する接客ロボットのシステムを構築するにあたり Julius を用いることで音声認識を行うことが可能となった。また、音声の認識結果より平均を算出することで Julius の有用性を確かめることができた。

### 5.2 今後の課題

今後は研究者だけでなく研究者以外の人間を実験対象とし、認識率を算出する。また認識率を算出後にノイズカットを行い認識率を向上させ、音声を認識したあとに認識した単語に対応した音声を発生させるプログラムを TCP/IP を介して組み合わせることにより、音声を認識し、認識した結果より対応した音声を発声させるためのシステムとして扱いやすいシステムの確立ができるのではないかと考える。また、このシステムを用いることで音声認識を用いた接客ロボットの開発を進めることができるのではないかと考える。

# 参考文献

- [1] 第7章 言語モデル・[https://julius.osdn.jp/juliusbook/Ja/desc\\_lm.html](https://julius.osdn.jp/juliusbook/Ja/desc_lm.html)・2016/09/02 閲覧
- [2] 第8章 認識アルゴリズムとパラメータ [https://julius.osdn.jp/juliusbook/ja/desc\\_search.html?id=2538692](https://julius.osdn.jp/juliusbook/ja/desc_search.html?id=2538692)・2017/01  
閲覧

# 謝辞

本論文作成にあたりテーマの決定, 研究の考え方, 方法のまとめ方など全てにおいて長期にわたって厳しくも熱意のあるご指導, ご鞭撻していただいた, 伊藤恒平教授に厚く御礼申し上げます.

特に分析においても論文の書き方においても論文を何度も読んでいただき, 指導していただいた伊藤恒平教授に大変ご苦労をかけてしまいましたことにも心よりお詫び申し上げたいです.

同級生のメンバーには論文の作成, 修正にご協力いただき心より感謝しております. その他, 助けていただいた多くの皆様に心から感謝しております. ありがとうございました.