一般口演
# 一般口演2
# 病院情報システム２（ハードウェア，インターフェース）
2018年11月23日(金) 10:15 〜 11:45  F会場 (5F 502+503)

## [2-F-1-6] 病院情報システムに関する問い合わせ対応効率化のための自動音声応答システムの構築

○南部 恵理子[1], 野崎 一徳[1], 玉川 裕夫[1], 林 美加子[2] （1.大阪大学歯学部附属病院医療情報室, 2.大阪大学大学院歯学研究科口腔分子感染制御学講座）

1.背景　病院情報システムや病院情報端末に関する院内各診療室からの問い合わせ対応を実施するコールセンター的業務は、煩雑且つ冗長な部分が多い。問い合わせ内容は、カルテの入力方法からシステムの不具合、病院情報端末の故障・不具合に関するものまで様々である。2017年度の病院情報システム関連の Q&A件数は255件であり、そのうち4割は病院情報システム運用マニュアルに記載のある、システム操作方法に関する質問であった。これら問い合わせへの応答業務により、深刻で調査が必要な問題に対する解決が遅れ、その結果、外来患者の会計待ち時間にまで影響が及んでいるのが現状であり、業務効率化と患者の会計待ち時間の短縮を検討する必要がある。2.目的　問い合わせ自動音声応答システムを構築し、システム導入前後での担当者による応答業務件数の比較と、システムによる回答の正確度を検証する。3.方法　病院情報システム運用マニュアルの記述と、過去の Q&A集をデータベース化する。質問者から医療情報担当者へ電話した際、自動応答システムが質問内容を聞き出す。自然言語処理によって質問者の音声を認識し、質問をテキスト化、意味解析を行う。そこから抽出したワードを元に、マニュアル DBと Q&A集 DBを検索し、該当したデータから適切な回答のテキストを構築し、音声化して質問者に伝達する。回答内容に対して、質問者から正確度を判定してもらい，学習することにより回答の正確さを向上させる。データベース上に該当するデータが存在しなかった場合のみ、担当者へ電話を繋ぎ、質問者と担当者間の対話内容を学習用 DBに蓄積することで、回答可能な事例を増やす。4.結果　病院情報システムのイントラネットワーク上にてデータベース構築と、自動応答システムの構築を行う。

# 病院情報システムに関する問い合わせ対応効率化のための自動音声応答システムの構築

## -第一報　自動音声認識システムの開発-

南部 恵理子[*1]、岡 真太郎[*2]、野崎 一徳[*1]、玉川 裕夫[*1]、林 美加子[*2]
*1 大阪大学歯学部附属病院医療情報室
*2 大阪大学大学院歯学研究科口腔分子感染制御学講座（歯科保存学教室）

# Construction of an automatic voice response system for inquiries on hospital information systems to improve efficiency

## - First Report：Development of automatic speech recognition system -

Eriko Nambu [*1], Shintaro Oka [*2], Kazunori Nozaki [*1], Hiroo Tamagawa[*1], Mikako Hayashi[*2]
*1 Division for Medical Information, Osaka University Dental Hospital,
*2 Department of Restorative Dentistry and Endodontology, Osaka University Graduate School of Dentistry

The call center service, which carries out correspondence inquiries from the clinics in hospitals related to hospital information systems and hospital information terminals, there are many complicated and redundant parts. The contents of the inquiry are system defects due to charting patient information, it also varies to failure and malfunction of hospital information terminals. The number of Q & A related to the hospital information system in 2017 is 255, of which 40% are written in the hospital information system operation manual, it was a question about system operation method. Responding to these inquiries led to a delay in solving the serious problem that needs to be investigated, as a result, it is currently the case that outpatient's waiting time for medical fee calculation has been affected, so it is necessary to consider improving work efficiency and shortening patient's accounting waiting time. In constructing an automatic response system that improves this, we developed a speech recognition method that can process at high speed with high accuracy, real time, and compared and verified the accuracy of speech recognition in our new method and the existing method. Based on those results, we examine the application of the developed method to automatic voice response system.

Keywords: speech recognition, information system, object detection

## 1. 背景

病院情報システムや病院情報端末に関する院内各診療室からの問い合わせ対応を実施するコールセンター的業務は、煩雑且つ冗長な部分が多い。問い合わせ内容は、カルテの入力方法からシステムの不具合、病院情報端末の故障・不具合に関するものまで様々である。2017 年度の病院情報システム関連の Q&A 件数は 255 件であり、そのうち 4 割は病院情報システム運用マニュアルに記載のある、システム操作方法に関する質問であった。これら問い合わせへの応答業務により、深刻で調査が必要な問題に対する解決が遅れ、その結果、外来患者の会計待ち時間にまで影響が及んでいるのが現状である。

図 1 に示すような構造の自動音声システムを構築するためには、音声認識の精度を向上し、音声認識誤りに対応する必要がある [1]。そこで、本研究では自動音声応答システムのうち、最初のステップである図 1[2]の①に示した音声認識の部分に焦点を当てた。既存の音声認識の手法と比較して、より高精度かつリアルタイムで高速処理が可能な音声認識システムの開発を行い、既存の音声認識手法と比較し、正解率を求めて認識精度を検証する。その結果を元に、本研究で開発したシステムが、将来的に自動音声応答システムへの応用が可能であるかを評価する。

## 2. 方法

汎用大語彙連続音声認識エンジン Julius[3]は、音声データを周波数領域の情報に変換した後に、人間の聴覚の周波数に対する敏感度を考慮に入れた尺度（メル尺度）を元に逆変換を行い、バンドパスフィルタを利用することで 20 程度の係数ベクトルとして処理し、そのベクトルを混合ガウス分布の期待値として正規確率を算出する。さらに隠れマルコフモデルを用いて音素の遷移確率をモデル化することにより音声を音素のリストとして識別する。本研究では、音声データのスペクトログラム画像に対して任意にラベル付けをし、畳み込みニューラルネットワークを利用した深層学習ベースの物体検出手法である YOLO v3（You Only Look Once）[4]を用いて、スペクトログラム画像と音素との関係を学習させ、確率分布や状態遷移確率に寄らない手法の確率を目指す。YOLOv3 を用いた自動音声認識システムの概要を図 2 に示す。YOLOv3 は、GPGPU を利用することによって高速・高精度で計算実行出来、それによりリアルタイムでの物体検出が可能である。

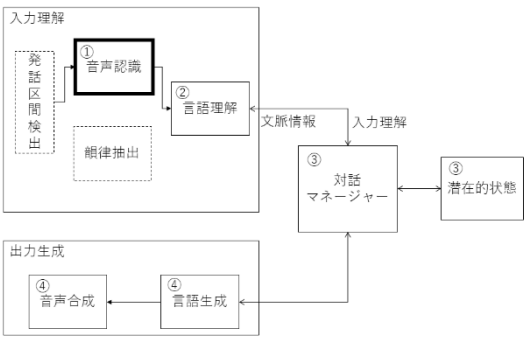


図 1.自動音声応答システムの概要図

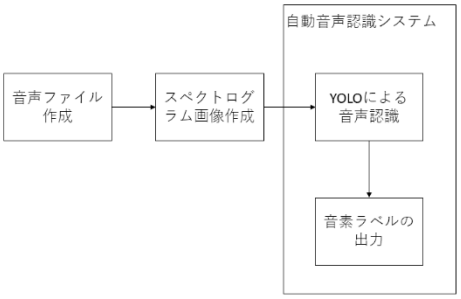


図 2.自動音声認識システムの概要図

## 2.1 音声データの作成

典型的な院内各診療室から 1 問い合わせ内容に含まれる、一文を全部で 5 種類選択し、各々音声を 30 回、同一話者により読み上げ、IC レコーダー（ICD-SX1000、SONY、東京）で録音した。音声の録音はサンプリング周波数 44.1 kHz、ベクトル量子化 16 bit で行った。尚、1 件あたりの発音時間は前後の無音部分を含め 5 秒以内とした。

## 2.2 音声ファイルのラベル付け

ラベル付けには、音声データを分析するためのフリーソフトウェア Wavesurfer（version 1.8.8）[5]を用いた。Wavesurfer で、2.1 で作成した WAV ファイルを開き、音声を聴きながら一音素ずつラベル付けを行う。この時、波形とスペクトログラムを表示することで、ラベル付けの参考とした。ラベル付けの単位は「あいうえお」の発話の通りとし、表記は一般的なローマ字のヘボン式を採用した。この時、発話前後の無音とノイズは含めず、発話中の無音やノイズは前後の音に含めてラベル付けをした。

Wavesurfer でラベル付けしたファイルをテキストデータとして保存する。実際にラベルを付与したデータファイルは図 3 のフォーマットとなっており、一行ごとに各ラベルと、ラベルが付与されている音声の開始時間と終了時間が書き込まれている。

本研究において使用した各質問文と、質問文の音声ファイルに付与したラベルは表1の通りである。

```
0.0907080 0.1681416 a
0.1681416 0.3163717 shi
0.3163717 0.4469027 ta
0.4469027 0.6725664 wa
0.6725664 0.7433628 ha
0.7433628 0.8606195 re
0.8606195 1.0221239 ma
1.0221239 1.1946903 su
1.1946903 1.4048673 ka
```

図 3. Wavesurfer でラベル付けしたテキストファイル記述例

表1．質問文とラベル

| No | 質問文 | ラベル |
|---|---|---|
| 1 | これはなんですか？ | ko/re/wa/na/n/de/su/ka |
| 2 | 明日は晴れますか？ | a/shi/ta/wa/ha/re/ma/su/ka |
| 3 | 電話をかけてもいいですか？ | den/wa/wo/ka/ke/te/mo/i/i/de/su/ka |
| 4 | あれはあなたの車ですか？ | a/re/wa/a/na/ta/no/ku/ru/ma/de/su/ka |
| 5 | このあと予定はありますか？ | ko/no/a/to/yo/te/i/wa/a/ri/ma/su/ka |

## 2.3 音声のスペクトログラム化

Python のグラフ描写ライブラリである matplotlib を使用して、音声ファイルの音声をフーリエ変換し、スペクトログラムの画像として出力するプログラムを作成した。画像サイズを統一するため、録音開始から 3.0 秒間の長さに編集した音声ファイル（以下、音声ファイル）を使用し、スペクトログラム画像の出力を行った（図 4）。画像に表示する時間軸の目盛りは 0.010 秒単位とした。

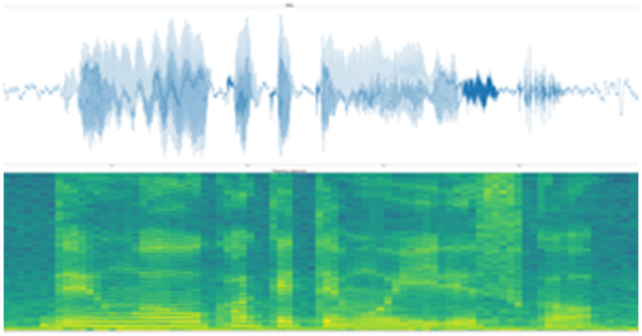

図 4. 音声ファイルを元にスペクトログラム画像を出力

## 2.4 スペクトログラム画像のアノテーション

YOLOv3 の学習用として、画像に音素をタグ付けするため、アノテーションツールである labelImg を用いた。labelImg では、スペクトログラムの画像に対してラベルの領域を矩形ボックスで指定でき、YOLOv3 の学習データとして使用可能なデータをテキストファイルとして生成することができる。アノテーションにはクラスファイルを必要とするため、Wavesurfer で付与したラベルを元に、ローマ字ヘボン式表記で 86 クラス記述したクラスファイルを作成した（図 5）。

```
a i u e o ka ki ku ke ko sa shi su se so
ta chi tsu te to na ni nu ne no ha hi fu he ho
ma mi mu me mo ya yu yo
ra ri ru re ro wa wo n
ga gi gu ge go za ji zu ze zo
da di du de do ba bi bu be bo bya byu byo
pa pi pu pe po pya pyu pyo
sha shu sho ja ju jo cha chu cho
```

図 5. labelImg で使用したクラス

YOLOv3 の学習データセット作成 Wavesurfer で作成したテキストデータの開始時間および終了時間を、スペクトログラムの画像に 0.010 秒単位でアノテーションする。矩形で指定

する際、スペクトログラムの縦軸である周波数の範囲は制限せず、最小値から最大値の幅とした。

## 2.5 YOLOv3 での学習とテスト

YOLOv3 の学習には、スペクトログラムの画像と、アノテーションした YOLOv3 学習用のテキストファイルを使用する。画像とデータセットのファイルは 1:1 で、両ファイル名は拡張子の前までは一致させる。準備した音声ファイルとデータセット 150 組のうち、本研究では 1 問につき 9 割にあたる 27 件を学習用データ、残り 1 割の 3 件をテスト用データとして振り分けた。学習結果の重みファイルが作成された後、テスト用データとして振り分けた画像を用いて精度を確認した。

実行環境として、OS に Ubuntu 18.04.01 LTS、GPGPU に NVIDIA TITAN V を使用した。

## 2.6 既存の手法との比較

本研究の手法である YOLOv3 を用いた解析手法の精度を、既存の手法と比較する。比較対象には、汎用大語彙連続音声認識エンジンの Julius を用いた。Julius で扱う音声ファイルは、モノラル音声かつサンプリングレートが 16 kHz の音声ファイルに限定されているため、iTunes を用いてステレオ音声を Julius で扱える形式に変換した。その後、音声標準パッケージとして提供しているディクテーションキットを使用して、音声ファイルを読み込み、音声認識を行った。これによって求められた結果を元に、開発した自動音声認識システムと既存の音声認識手法との精度比較を行った。

# 3. 結果

## 3.1 テスト用データを用いた YOLOv3 でのテスト結果

YOLOv3 での学習結果の重みを利用して、テスト用データの認識を行った。実行結果は図 6 の通りである。実際に検出された結果は、画像ファイルとして出力される。出力された画像には、検出結果を枠で囲われ、対応したラベルが表示される。元の質問文と画像中に示されたラベルを比較したところ、テスト用データの正解率は 100.0 %であった。

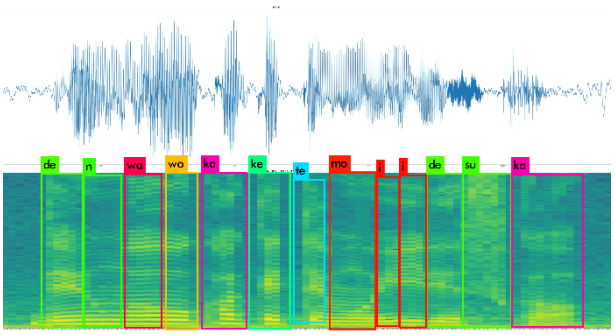


図 6. YOLOv3 のテスト実行後に作成された画像の出力例

## 3.2 YOLOv3 で学習済みの質問の認識

テスト用データでの精度が確認できたため、学習に用いられていないデータとして、新たに表 1 の Q1～Q5 の質問文の音声を各々20 回録音し、音声ファイルとスペクトログラム画像を作成し、YOLOv3 での音声認識を行った。出力された画像に表示されたラベルを一音素ずつ確認し、最初から最後までのラベル全てがアノテーションした通り正しく出力されていれば、認識された質問文は正解とし、一部でも誤りがあれば不正解と判定して、質問文全体の正解率を求めた(表 2)。その結果、Q1～Q5 の質問文ごとに差は見られたものの、平均 80.0 %の正解率となった。

なお、認識処理に要した時間は、1 件あたり平均 0.04 秒であった。

表 2.文章としての正解・不正解の判定

| 正解の質問文 | 出力結果 | 判定 |
|---|---|---|
| ko re wa na n de su ka | ko re wa na n de su ka | 正解 |
| ko re wa na n de su ka | a re wan a n de su ka | 不正解 |

YOLOv3 と Julius での認識精度比較のため、YOLOv3 での認識に用いた音声ファイルを使い、Julius を実行した。Julius で出力された音素並びから判定した文章としての正解率は、平均 38.0 %となった。YOLOv3 と Julius との正解率の比較は表 3 の通りである。

表 3. YOLOv3 と Julius の正解率比較(文章)

| No. | 質問文 | YOLO | Julius |
|---|---|---|---|
| Q1 | これはなんですか？ | 95.0 % | 65.0 % |
| Q2 | 明日は晴れますか？ | 95.0 % | 5.0 % |
| Q3 | 電話をかけてもいいですか？ | 65.0 % | 30.0 % |
| Q4 | あれはあなたの車ですか？ | 65.0 % | 40.0 % |
| Q5 | このあと予定はありますか？ | 80.0 % | 50.0 % |
| 平均 | | 80.0 % | 38.0 % |

誤答となった結果を見ると、YOLOv3 の場合、「ko re wa wa na n de su ka」のように、同じ音を連続で重複して出力されていたのに対し、Julius の場合は単語単位での誤り、助詞の誤り、声の抑揚および大きさや、発話前後のノイズの影響による認識誤りが多く見られた。そこで、より詳細に認識精度を確認するため、ラベルおよび音素単位での評価を行った。

表 4.に、ラベルおよび音素の判定方法を示した。表中の①～④の判定基準については、正解の音素と突き合わせた結果により次の通りとした。TP(True Positive):認識されるべき音が全て正しく認識、FN(False Negative):音が一部誤って認識、FP(False Positive):認識されるはずではない音を認識、TN(True Negative):認識されるべき音素が認識されず欠落

表 4.ラベルおよび音素の判定

| | こ | れ | は | な | ん | で | す | か | | 判定 |
|---|---|---|---|---|---|---|---|---|---|---|
| 正 | ko | re | wa | na | n | de | su | ka | | - |
| A | ko | re | wa | na | n | de | su | ka | | TP |
| B | ka | re | wa | na | n | de | su | ka | | FN |
| C | ko | re | wa | wa | na | n | de | su | ka | FP |
| D | ko | re | | na | n | de | su | ka | | TN |

各ラベルおよび音素単位で質問文 Q1～Q5 の YOLO と Julius で認識した結果を判定し、集計した。

表 4 の判定基準からクロス表(表 5)を作成し、判定結果を算出した。YOLO と Julius での算出結果は各々表 6、表 7 の

通りとなった。

表 5.クロス表

| | | 音声認識結果 | |
|---|---|---|---|
| | | 認識 | 認識不可 |
| 予測結果 | 正解 | TP | FN |
| | 不正解 | FP | TN |

表 6. Q1～Q5 評価結果（YOLO）

| | 敏感度 | 特異度 | 有効度 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|---|---|
| Q1 | 100 | 100 | 10 | 1 | 1 | 1 |
| Q2 | 100 | 0 | 0.99 | 1 | 0.99 | 1 |
| Q3 | 100 | 0 | 0.97 | 1 | 0.97 | 0.98 |
| Q4 | 99.62 | 0 | 0.97 | 1 | 0.97 | 0.98 |
| Q5 | 100 | 0 | 0.98 | 1 | 0.98 | 0.99 |
| 平均 | 99.92 | 20.00 | 0.98 | 84.16 | 54.10 | 0.83 |

表 7.Q1～Q5 評価結果（Julius）

| | 敏感度 | 特異度 | 有効度 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|---|---|
| Q1 | 89.21 | 16.67 | 0.86 | 0.89 | 0.96 | 0.93 |
| Q2 | 63.69 | 0 | 0.62 | 0.64 | 0.95 | 0.76 |
| Q3 | 84.16 | 53.85 | 0.82 | 0.84 | 0.97 | 0.90 |
| Q4 | 90.70 | 100 | 0.91 | 0.91 | 1 | 0.95 |
| Q5 | 93.04 | 100 | 0.93 | 0.93 | 1 | 0.96 |
| 平均 | 84.16 | 54.10 | 0.83 | 0.84 | 0.98 | 0.90 |

## 3.3 YOLOv3 で学習済みの音素から作成した質問の認識

質問文 Q1～Q5 を用いて、YOLOv3 で学習済みの音素のみを使用し、新たに質問文 Q6 を作成した（表 8）。

表 8.学習済み音素から作成した質問文

| No. | 質問文 | ラベル |
|---|---|---|
| Q6 | あの薬はありますか？ | a/no/ku/su/ri/wa/a/ri/ma/su/ka |

Q6 の質問文を 20 回、同一話者により読み上げ、音声ファイルとスペクトログラム画像を作成し、YOLOv3 と Julius で音声認識し、音素単位での評価を行った。結果は各々表 9、表 10 の通りとなった。

表 9.Q6 評価結果（YOLOv3）

| | 敏感度 | 特異度 | 有効度 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|---|---|
| Q6 | 73.61 | 100 | 0.76 | 0.74 | 1 | 0.85 |

表 10.Q6 評価結果（Julius）

| | 敏感度 | 特異度 | 有効度 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|---|---|
| Q6 | 94.09 | 0 | 0.94 | 0.94 | 1 | 0.97 |

YOLOv3 で学習済みのラベルのみ使用したが、YOLOv3 に比べ、Julius の結果の方が正解率は高くなった。

## 3.4 YOLOv3 で未学習の音素から作成した質問の認識

YOLOv3 で学習していない音素を含んだ質問文 Q7 および Q8 を新たに作成した（表 11）。ここでは、病院情報システムでの質問を想定し、医療情報に関係する質問文を作成した。

表 11.未学習の音素から作成した質問文

| No. | 質問文 | ラベル |
|---|---|---|
| Q7 | 病名を慢性歯周炎に変更できますか？ | byo/u/me/i/wo/ma/n/se/i/shi/syu/e/n/ni /he/n/ko/u/de/ki/ma/su/ka |
| Q8 | 患者情報を修正したいのですが | ka/n/jya/jyo/u/ho/u/wo /syu/u/se/i/si/ta/i/no/de/su/ga |

Q1~Q6 と同様に、Q7 および Q8 の質問文とも各々20 回、同一話者により読み上げ、音声ファイルとスペクトログラム画像を作成した。Q7 は他の質問文と異なり、音素数が多く、発音時間が長い文章だが、音声ファイルの編集をせずにそのまま音声認識に用いた。

YOLOv3 と Julius での音声認識結果は、各々表 12、表 13 の通りとなった。

表 12.Q7～8 評価結果（YOLO）

| | 敏感度 | 特異度 | 有効度 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|---|---|
| Q7 | 29.51 | 100 | 0.72 | 0.30 | 1. | 0.46 |
| Q8 | 23.85 | 100 | 0.78 | 95.56 | 1 | 0.96 |

表 13.Q7～8 評価結果（Julius）

| | 敏感度 | 特異度 | 有効度 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|---|---|
| Q7 | 87.17 | 65.22 | 0.86 | 0.87 | 0.97 | 0.92 |
| Q8 | 95.56 | 100 | 0.96 | 0.96 | 1 | 0.98 |

YOLOv3 と Julius ともに、他の質問に比べて正解率は下がったが、YOLOv3 では著しく正解率が低下した。

## 4. 考察

3.2 の結果から、YOLOv3 で学習済みの質問文については、従来の波形データを用いた音声認識に比べ、YOLOv3 を用いてスペクトログラムの画像から音声認識する手法の方が、正解率が高いという評価ができた。Julius では同じ質問文を同一話者が発話した音声ファイルであっても、声の抑揚や大きさ、ノイズに影響を受けやすく、誤認識もしくは音素が欠落することが多かったのに対し、YOLOv3 ではそれらの影響に左右されず認識することができたと考えられる。ところが、3.3 の結果に見られるように、既に学習済みの音素であっても正解率が低下したことから、前後の音素との関係が認識精度に影響する。さらに、3.4 で確認した未学習の音素が含まれた音声については、YOLOv3 での正解率が Julius を大きく下回ったことから、YOLOv3 の精度向上のためには、全音素と前後の音素関係を網羅した発話パターンの学習が必要であると思われる。

ただし、YOLOv3 については、学習済みの質問については 1 件あたりの認識速度が短時間にも関わらず、平均 80 %の確率で正解を認識できたことから、スペクトログラム画像から音声認識を行うことは、質問応答システムの利用に有効であると考えられる。さらに、本研究で、YOLOv3 の学習用データとして準備した音声ファイルが 1 問あたり 30 件と少ないデータ

数で高精度の結果が得られたことから、同一話者でより多くの発話パターンを学習させる際、準備する音声ファイルは少なくて済む。複数の話者の発話を学習させる場合であっても、従来の物体検出手法に比べ、採取するデータ数ははるかに少なくて済むため、質問応答システムの構築が容易である。

## 5. 結論

本研究の結果、YOLOv3 を用いた音声認識手法が、音素を学習済みであるという条件下では有用であることが明らかとなった。今後、この手法で検出した音声ラベルをテキスト化し、図 1 に示したシステム全体の構築を行うことによって、質問文に対応した答えを返すシステムの構築を目指す。

## 6. 謝辞

本研究の一部は日本電気株式会社共同研究費「スマートデンタルホスピタルに関する研究」の助成による。

## 参考文献

1) 駒谷和範, 河原達也. 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理. 情報処理学会論文誌. 2002 ; 43 : 3078-3086.
2) 奥村学, 中野幹夫, 駒谷和範, 船越孝太郎, 中野有紀子. 対話システム(自然言語処理シリーズ). コロナ社. 2015.
3) 大語彙連続音声認識エンジン Julius http://julius.osdn.jp/
4) Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi,“You only look once: Unified, real-time object detection”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp.779-788.
5) Wavesurfer http://www.speech.kth.se/wavesurfer/