

コミュニケーションロボットのための 遠隔音声認識システムの構築と実証実験

Establishment and Experimental Demonstration of Distant Speech Recognition System
for Communication Robot

山 本 俊 一*
Shunichi YAMAMOTO

住 田 直 亮**
Naoaki SUMIDA

中 臺 一 博**
Kazuhiro NAKADAI

要 旨

人と自然な距離でコミュニケーションするためのロボット用遠隔音声認識技術を実現した。課題は二つあり、一つ目は周囲の残響や雑音の影響によって認識性能が低下することであり、これまでに開発した技術に加えて、残響抑圧、音源同定技術を導入して解決を試みた。もう一つは、人によって音声の周波数特性や言葉の言い回しが異なることであり、音響モデルに Deep Neural Network を、言語モデルに音声認識タスクの性質を考慮したクラス N-gram を用いることで認識性能向上を試みた。また、これらの技術をロボットシステムとして統合し、受付取次を題材とした実証実験を行った。結果として、受付取次達成率 98.3% を達成するとともに、音声誤認識の原因の分析により、性能向上のためのさらなる課題を明らかにした。

ABSTRACT

Distant speech recognition technology for a robot was implemented to communicate with humans at a natural distance. There were two issues to address. First, surrounding reverberation and noise act to reduce recognition performance. In an attempt to resolve this, dereverberation and sound source identification technology were introduced in addition to the existing developed technology. Second, speech frequency response characteristics and wording differ from person to person. To enhance recognition performance, a deep neural network was used in the acoustic model, in addition to class-based N-grams accounting for the properties of speech recognition tasks in the language model. These technologies were then integrated into a robot system to perform a themed experimental demonstration taking reception calls.

The demonstration achieved a reception success rate of 98.3% and, through analysis of the causes of erroneous speech recognition, clarified further issues to address for improving performance.

1. まえがき

人と共存し、人にサービスを提供するロボットに必要な能力の一つはコミュニケーション能力であり、日常最も使用され自然で直感的なコミュニケーション手段の一つはやはり音声でのやりとりである^{(1),(2)}。本稿では、音声認識を用いたバーバルなコミュニケーションに焦点をあてる。音声認識に関しては、昨今、スマートフォンによる情報検索など、接話マイクロホンの利用を前提とした音声入力インタフェースが実用レベルで提供されており、社会に受け入れられつつある⁽³⁾。人とコミュニケーションすることが求められるロボットにとっても、音声認識は、様々な応用

が期待できる重要な機能である⁽¹⁾。しかし、人とロボットが自然な距離で、音声でやりとりするためには、ロボットに搭載されたマイクから離れた距離で発話された音声を認識(遠隔音声認識)できる必要がある。遠隔音声認識では接話マイクロホンを使用した場合と比較して、雑音と残響により S/N 比が低下して音声認識の性能を著しく低下させるという課題がある^{(3),(4)}。本文で現れる雑音は、対象話者の音声以外のすべての音のことを表すものとする。また、日常の環境で音声認識を行う場合には、雑音など対象話者の音声以外の音が入力されるので、正しく応答するために対象話者の音声とそれ以外の音を聞き分けるという課題がある⁽⁵⁾。特に遠隔音声認識では、雑音など対象話者の

* R&D センターX

** 株式会社ホンダ・リサーチ・インスティテュート・ジャパン

音声以外の音が入力される頻度が高くなるため重要な課題となる。その他、音声認識での一般的な課題である、声質や言い回しといった個人差にも対応する必要がある⁽³⁾。

遠隔音声認識の個々の要素技術は数多く研究されており、CHiME CHALLENGE という共通のベンチマークを使用した遠隔音声認識のコンテストも行われている⁽⁶⁾。このコンテストでは、遠隔音声認識で課題となる雑音や残響への対応に焦点を当てている。一方、我々は、ヒューマノイドロボット ASIMO⁽⁷⁾を用いて、人と音声でやりとりを行うアプリケーションを題材として、遠隔音声認識技術を実用レベルで実現することを目指している。ロボットでの遠隔音声認識を実現するためには、遠隔音声認識の個々の課題を解決するだけでなく、それらの要素技術をロボット上のシステムとして統合する必要がある。また、そのシステムを長期間運用し、そのデータを解析することで、そのアプリケーションでの性能を評価するとともに、実際のお客様に使ってもらった際に起こる現場ならではの課題を明らかにすることが重要となる。このような研究例は他ではほとんど見受けられず、実証実験で取得したデータ及び結果は今後の研究開発に大きく寄与すると考える。本稿では、実現した遠隔音声認識の技術及び効果確認のために行った実証実験結果について報告する。

2. 研究のねらい

本研究のねらいは、実用レベルの遠隔音声認識の性能を向上させることと、実際に使用してトータルなシステムの中で遠隔音声認識が使用されるときに課題を明らかにすることである。遠隔音声認識を実現するには、前節で述べた通り様々な課題がある。その中で、本稿では以下の四つの課題を解決することで性能向上を試みた。

- (1) 残響による音声認識性能低下
- (2) 非目的音の音声誤認識
- (3) 声質の個人差による音声認識性能低下
- (4) 利用者の自由な言い回しによる音声認識性能低下

(1)については、残響時間が音声認識処理単位(本システムでは 25 ms)を超えると認識性能劣化が顕著になるといわれている⁽⁸⁾。我々がターゲットの一つとしている公共施設では、これ以上の残響時間が想定される。さらに、遠隔マイクは接話マイクと比較して、観測される直接音のパワーは小さくなるが、残響成分のパワーはほとんど変わらないので残響の影響が大きくなる。したがって、残響成分を除去するために Gomez らが提案した残響抑圧手法⁽⁸⁾を導入した。

(2)については、雑音を間違えて音声認識処理をすると性能が低下するため、音声認識の前処理として、目的音声か非目的音かを識別する混合正規分布 (Gaussian Mixture Model: GMM) による音源同定を導入した⁽⁵⁾。

(3)については、音声の各音素をモデル化(以下、音響モデル)する場合に、なるべく多くの人の特徴を反映する必要がある。このモデル化には従来、混合正規分布による隠れマルコフモデル (Gaussian Mixture Model based Hidden Markov Model: GMM-HMM)⁽⁹⁾が用いられてきたが、より緻密なモデル化が可能な深層学習を用いた研究⁽¹⁰⁾が近年盛んである。本研究では、Deep Neural Network による隠れマルコフモデル (Deep Neural Network based Hidden Markov Model: DNN-HMM)⁽¹⁰⁾を導入した。

(4)については、例えば、企業受付で担当者呼び出す場面を考えてみる。受付の人に「中臺様お願い致します」という人もいれば、「山本と申しますが、住田さん呼んでもらえますか」という人もいるだろう。このように、利用者が様々な言い回しで発話しても認識できる必要があり、音声認識が利用されるアプリケーションの制約もある。これらを考慮してクラス N-gram 言語モデルを導入した⁽¹¹⁾。

我々は、本当に使える技術にするためには、上記技術の個別評価だけではなく、ロボットシステムとして統合し、実際にお客様に使っていただいて効果確認を行うことが大事であると考えた。それを行うことで、本質的な課題の把握とさらなる技術開発によって性能向上がはかられていくのである。

遠隔音声認識を用いたロボットシステムの実証実験例としては、株式会社日立製作所の EMIEW3 を用いた空港案内⁽¹²⁾や Haneda Robotics Lab での案内ロボット⁽¹³⁾などがあるが、いずれも 1 週間程度の短期間であり、分析を含む結果は公表されていない(2017 年 4 月現在)。我々は、約 2 ヶ月間に渡り受付取次を題材として実証実験を行い、そのデータを分析し、課題を明らかにした。

3. 達成手段

本章では、課題を解決するために用いた個々の技術について説明する。

3.1. 残響抑圧

3.1.1. スペクトルサブトラクションに基づく残響抑圧

本節では、スペクトルサブトラクションに基づく残響抑圧手法⁽⁸⁾をオンライン処理可能に変更したアルゴリズムについて述べる。本手法では入力音声のうち残響の成分を雑音とみなし、その雑音をスペクトルサブトラクションで抑圧することで残響抑圧処理を実現している。残響成分は一般的にその性質の違いから初期反射と後期残響に分けて扱われる。本手法では、初期反射は抑圧せず、後期残響の成分を推定してスペクトルサブトラクションにより抑圧する。後期残響のみを抑圧する理由は、初期反射は直接音が到来してからそのあと比較的早い時間に到来する反射音であり、初期反射の大部分は後段の音声認識の処理単位であ

る窓長に収まるので、特徴量正規化手法であるケプストラム平均正規化やスペクトル平均正規化で抑制できるからである。

後期残響は、入力音声の短時間フーリエ変換 (short-term Fourier transform: STFT) $X(f, t)$ と後期残響の伝達関数 $H_L(f)$ から推定する。 $X(f, t)$ の時間窓に重なる過去 D フレームの STFT も考慮して、周波数ビン f , 時刻 t の後期残響 $\hat{X}_L(f, t)$ の振幅スペクトル推定値を以下の式で近似した。

$$|\hat{X}_L(f, t)| = \left\{ \frac{1}{D+1} \sum_{d=0}^D |X(f, t-d) H_L(f)|^\gamma \right\}^{\frac{1}{\gamma}} \quad (1)$$

ここで、 γ は後期残響の推定値を大きくしすぎないためのパラメータで 2 以下の値に設定し、本研究では $\gamma = 1.5$ に設定した。このように STFT でフレームに分解して過去フレームを利用することにより遅延の小さいオンライン処理を実現できている。

直接音と初期反射を含むパワースペクトル $|X_E(f, t)|^2$ を、後期残響の推定値を用いたスペクトルサブトラクションにより以下のように計算する。

$$|\hat{X}_E(f, t)|^2 = \begin{cases} |X(f, t)|^2 - \delta(f) |\hat{X}_L(f, t)|^2 & (0 \text{ より大きい}) \\ \beta |X(f, t)|^2 & (\text{それ以外}) \end{cases} \quad (2)$$

ここで、 β はパワースペクトルを負の数にしないためのパラメータである。 $\delta(f)$ は事前に測定したインパルス応答を用いて、後期残響成分の推定値 $\delta(f) \hat{X}_L(f, t)$ と実際値 $X_L(f, t)$ から最小平均二乗誤差基準最適化により計算する。

出力のスペクトル $\hat{X}_E(f, t)$ は、入力 $X(f, t)$ の位相を用いて以下のように計算する。

$$\hat{X}_E(f, t) = \sqrt{|\hat{X}_E(f, t)|^2} e^{j\phi_X(f, t)} \quad (3)$$

ここで、 $\phi_X(f, t)$ は $X(f, t)$ の位相を表す。

3.1.2. 評価実験

残響抑圧の効果を確認するために、残響抑圧を適用した場合と適用しない場合の両方で DNN-HMM の音響モデルを学習させてベンチマークに対する単語正解精度の比較を行った。DNN-HMM 音響モデルの学習と学習データの詳細については 3.3. 節を参照されたい。評価用ベンチマークデータとして、Reverberation Time 20 (RT20) 法による残響時間が 1.02 秒の受付でロボットを用いて収録した受付取次の発話、例えば、「中臺商事の山本と申しますが、住田さん呼んでもらえますか」という発話を 26 人、受付取次 258 回、総単語数 5629 語を収録した。評価は単語正解精度 (Word accuracy) で行い、すべての単語の平均と各発話に

出現する各クラス (訪問者名 - 担当者名, 訪問者 - 組織名, 担当者 - 所属名, 以下, 取次必要語集) ごとの結果を Fig. 1 に示す。残響抑圧を適用した場合としない場合では、平均単語正解精度は 2 ポイント向上しており、2 項分布の差の検定⁽¹⁴⁾により危険率 5% で有意な差があった。各クラスの単語正解精度を見ると訪問者名 - 担当者名, 訪問者 - 組織で特に効果が高かったことがわかり、平均単語正解精度と同様に有意な差があった。訪問者名 - 担当者名, 訪問者 - 組織はクラス内単語数が多く音声認識が容易ではない部分なので、残響抑圧の適用による認識率向上の効果が高かったと考えられる。

3.2. 音源同定

3.2.1. 音源同定モデル

音源同定では、入力音声区間ごとに目的音声か非目的音 (雑音, 目的音声以外の音声) かを判別し、高い再現率で目的音声を判別し、その次に雑音を棄却できることを目的とする。特徴量として入力音声から抽出した Mel-Scale Log Spectrum (MSLS) 特徴量 27 次元⁽¹⁵⁾ を利用し、目的音声と雑音を GMM でモデル化して判別に使用した。入力音声区間は音声を取りこぼさないために区間の先頭と後尾にマージンを取っており、先頭と後尾は音源同定にとって不必要である。したがって、対象としたい部分と比較してパワーが小さいことを考慮して、入力音声区間の平均パワーより大きい時刻の特徴量のみ音源同定に使用した (Fig. 2)。

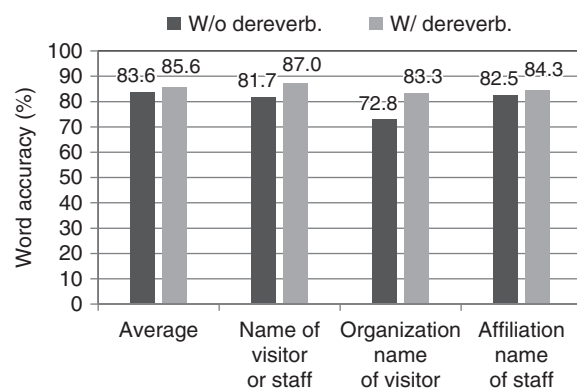


Fig. 1 Advance of word accuracy by using dereverberation

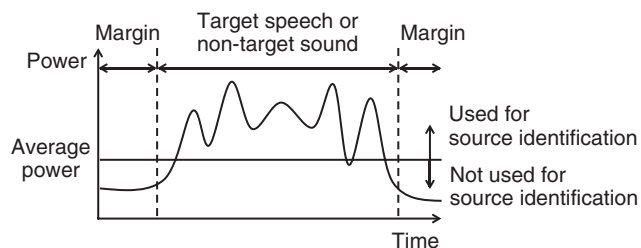


Fig. 2 Frames used for source identification

目的音声 (target) と非目的音 (non-target) の GMM のモデル化について説明する. 時刻 t の入力特徴量を x_t とすると, クラス m (target または non-target) の確率密度関数 $p_m(x_t)$ は以下のように表せる.

$$p_m(x_t) = \sum_{k=1}^K \lambda_{mk} N(x_t; \mu_{mk}, \sigma_{mk}^2) \quad (4)$$

ここで, K は混合数, λ_{mk} は混合重み, $N(x; \mu, \sigma^2)$ は平均 μ , 分散 σ^2 の正規分布の確率密度関数を表す. パラメータ λ_{mk} , μ_{mk} , σ_{mk}^2 は, 学習データより EM アルゴリズムにより推定した.

音源同定する際は, 入力音声の平均パワー P_{avg} より大きい時刻の尤度 $p_m(x_t)$ を使用して, 時刻ごとに最も尤度が大きいクラス m に投票する. 全体の合計票数で正規化したものをスコアとし, スコアが最も高いクラス m を, その発話区間の音源同定結果とする. スコア S_m は以下のように計算した.

$$S_m = \frac{1}{|T_P|} \sum_{t \in T_P} \text{Presence}_m(x_t) \quad (5)$$

$$\text{Presence}_m(x_t) = \begin{cases} 1, & m = \underset{i}{\operatorname{argmax}} p_i(x_t) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

ここで, $P(t)$ は時刻 t の入力音声のパワーを表したとき $T_P = \{t | P(t) > P_{avg}\}$ と定義し, 入力音声の平均パワーより大きい時刻の集合を表す.

3.2.2. 学習データ

目的音声と非目的音の GMM の学習データは, 受付取次の使用環境でロボットに被験者が話しかけた音声と周囲の雑音を, 音源定位, 発話区間検出, 音源分離して使用した. 各クラスの学習データ数は, 目的音声の話者 104 人の 26,164 発話, 非目的音が 1,577 個である. 正解ラベルは音源分離した音ごとに目的音声と非目的音のラベル付けを行った.

3.2.3. 評価実験

本システムの音源同定手法の評価実験では, 受付取次の使用環境でロボットに話者が話しかけた音声を収録した 2 種類の評価セットを使用した. 一つ目は前述の話者 26 人による受付取次の評価用ベンチマークデータ, 二つ目は実際にお客様に利用していただいた受付取次実証実験 (詳細は後述) のデータである. それぞれの評価セットの目的音, 非目的音の内訳を Table 1 に示す.

音源同定結果の性能を発話単位の再現率と適合率で評価した. 再現率は, 実際に目的音声 (非目的音) である発話のうち, 目的音声 (非目的音) であると予測された発話の割合を表す. 適合率は, 目的音声 (非目的音) と予測した発話のうち, 実際に目的音声 (非目的音) である発話の割合を表す. 各評価セットでの音源同定結果の再現率と適合率を

Table 1 Evaluation set for source identification

Evaluation set	Class	Number of sounds
Benchmark	Target speech	1,006
	Non-target sound	239
Reception experiment	Target speech	829
	Non-target sound	7,978

Table 2 Evaluation results of source identification

Evaluation set	Class	Recall (%)	Precision (%)
Benchmark	Target speech	96.5	91.1
	Non-target sound	94.5	80.5
Reception experiment	Target speech	87.1	34.7
	Non-target sound	95.7	98.4

Table 2 に示す. 受付取次では, 目的音声に対して高い再現率を重要視しており, 各評価セットで高い再現率を実現できた. しかし, ベンチマークと比較して受付取次運用では目的音声の再現率が低くなっている. これは, 受付取次実証実験評価セットの音の総数に対して非目的音の割合が高いことが原因であり, 非目的音の総数が多い理由は, 長時間実験を行っていたため話者がいないときに周囲の雑音を多数検出したからである. このような多数を占める雑音に対しては, 95.7% という高い再現率で非目的音と判別し棄却することができた.

3.3. 音声認識—音響モデル

音声認識では認識率が飛躍的に向上することが報告されている DNN に基づく手法のうち DNN-HMM をロボットの遠隔音声認識に適用した. ロボットの遠隔音声認識への DNN-HMM の適用に際しては, 以下の課題がある.

- (1) 大量の学習データを必要とすること
- (2) 入力音声と音響モデルのアンマッチ

一般的に, 深層学習では大量の学習データを必要とするため, ロボットの遠隔音声認識への DNN-HMM の適用においても大量の学習データを必要とすると考えられる. 従来, 我々は 82 時間の読み上げ音声コーパスを使用して GMM-HMM による音響モデルの学習を行っていたが, DNN-HMM の適用にあたっては 904 時間の音声コーパスを収集し学習に利用した.

また, ロボットに搭載されたマイクアレーを用いる遠隔音声認識では, 周囲の雑音, ロボットの自己雑音, 残響により入力音声の S/N 比が低下する. 雑音除去や残響抑圧などの前処理を行ったとしても入力音声から雑音や残響を完全に除去することはできず, 入力音声と音響モデルとのアンマッチにより音声認識率が低下する. 音声認識利用時の環境でロボットのマイクアレーを使用して音声を収録すること (以降, 実機録音と呼ぶ) によって, 入力音声と音響

Table 3 Training data of our acoustic model

Speech corpus	Style	Speech	Hours
Commercially available speech corpus	Reading	Synthesized	329
Reception speech corpus	Reading and spontaneous speech	Synthesized	453
Reception speech corpus recorded with ASIMO	Reading	Recorded with ASIMO	122
Total			904

モデルをマッチさせることが望ましいが、そのような方法で大量の学習データを準備することは容易ではない。したがって、我々は実機録音音声だけでなく、事前に測定したインパルス応答とロボット自己雑音を用いて音声認識利用時の入力音声を模擬した音声をクリーン音声から合成して音源分離と雑音抑圧の処理をした音声（以降、合成音声と呼ぶ）を学習データとした。

3.3.1. Deep Neural Network による音響モデル

DNN-HMM 音響モデルの学習とデコードには、オープンソースソフトウェアの音声認識ツールキットである Kaldi⁽¹⁶⁾ を利用した。入力特徴量は MSLS 特徴量 27 次元とその前後 5 フレームを結合した合計 297 次元とし、出力をトライフォン HMM の各状態とする、全結合型の隠れ層 6 層、2,048 ユニットの DNN を利用した。DNN の学習基準には state-level Minimum Bayes Risk を利用した。

3.3.2. 学習データ量による性能検証

本研究で作成した音響モデルの性能を確認するために、学習データ量を変えたときの単語正解精度の比較を行った。学習データには Table 3 に示した合計 904 時間を利用した。受付音声コーパスと ASIMO 受付音声コーパスについては、本受付取次タスクを想定して収集した音声コーパスである。受付音声コーパスは、受付取次タスクを想定したシナリオの読み上げ音声と Wizard of Oz 法を用いて収録したタスク発話で構成される。タスク発話の収録では、話者に対してタスクは与えられるが、どのように発話するかは話者に委ねられており、話者がシステムと対話している状況を再現しているのでシステム利用時に近い音声を収録することができる。ASIMO 受付音声コーパスは、受付取次タスクを想定したシナリオの読み上げ音声の実機録音である。評価には前述の評価用ベンチマークデータを使用し、合計 5,629 単語の単語正解精度で評価を行った。学習データ量に対する単語正解精度の関係を Fig. 3 に示す。

学習データ 904 時間に対して、88.9% の単語正解精度を達成した。比較のため、評価用ベンチマークデータを従来の GMM-HMM 音響モデルで評価すると単語正解精度は 29.2% であり、大幅に性能が向上していることがわかる。

3.4. 音声認識一言語モデル

受付取次のための言語モデルで満たすべき要件を以下のように設定した。

- (1) 話者の自由な言い回しを許容すること
- (2) 担当者名、担当者所属名を漏れなくカバーすること
- (3) 訪問者名、訪問者組織名については実証実験中に収集したデータから随時追加登録できること

これらの要件を満たすために、クラス N-gram 言語モデルを使用した。話者の自由な言い回しを許容するために、多様な受付取次の言い回しを収集した言語コーパスからクラス N-gram 言語モデルを学習させた。一方、担当者名や訪問者名などをクラスにすることで、漏れなくカバーすることができ、また新しい訪問者名などの追加が容易になるようにした。クラスは、担当者名、担当者所属の会社名、担当者所属の事業所名、担当者所属の部署名、担当者所属のブロック名、訪問者名、訪問者組織名の合計 7 種類である。

作成した受付取次用言語モデルの単語数は、実証実験開始当初で 1,586 語、新しい訪問者名などを追加して実証実験終了時には 1,681 語となった。

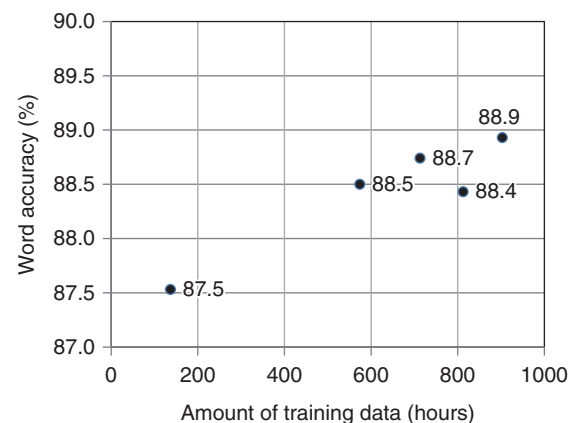


Fig. 3 Advance of word accuracy with increased training data

4. 実験

前章の技術をロボット上で統合し、一般のお客様に利用していただき、性能評価を行った。

4.1. ターゲット設定

実際に性能を評価するにあたり、タスクを明確に定め、それを一般のお客様に利用していただき評価を行った。これは、実際に一般の方に使ってもらうことで、実用に向けた現状性能を把握でき、分析を通してさらなる課題も明確になると考えたからである。ここでは、企業受付での取次タスクをターゲットとした。受付に来たお客様に担当従業員 (Person in charge: PIC) を呼び出すようにロボットへ

話しかけてもらい、実際に従業員を呼び出してもらった。音声による呼び出しの流れを Fig. 4 に示す。なお、この実証実験は弊社生命倫理委員会の承認を受けたのちに行っている。

4.2. 実験環境

実験は R&D センター X (旧基礎技術研究センター) 内の受付にロボットを設置して行った。実際に来所したお客様に音声で話しかけていただき、従業員を呼び出していただいた。

4.3. システム構成

実験で使用したシステムの構成を Fig. 5 に示す。点線枠で示した部分が本研究で取り扱った技術トピックである。実証実験では、残響抑圧、音源同定、音響モデル、言語モデルに関して、既述のアルゴリズム及びデータを適用した。

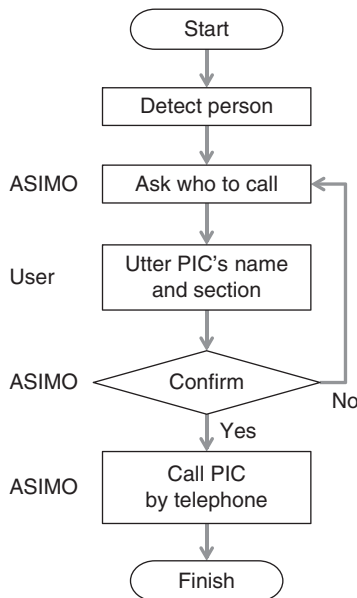


Fig. 4 Reception task flow

後段の言語理解部に関しては、ルールベースで取次必要語彙を抽出した。

ロボット対話に使用する遠隔音声認識システムは応答が速く同時に複数入力される音声进行处理できる必要がある。早い応答を実現するために、マイクアレーへ音声入力してから音声認識で処理するまでオンライン処理を実現した。同時に複数入力される音声进行处理できるように、複数の音源の方向を追跡し音源ごとに一意の ID (音源 ID) を振って管理し、音声認識システムも同時に複数の音声を認識処理できるような構成にした。

音声入力から特徴量抽出までの処理はロボット聴覚オープンソースソフトウェア HARK⁽¹⁷⁾ を用いてオンライン処理を実現した。特に、残響抑圧手法は発話単位で音声入力しなければならないアルゴリズムであったので、3.1.1. 節で述べたとおりオンライン処理可能なアルゴリズムに改良して実装した。また、音声認識システムの DNN 音響モデルを用いた音響尤度計算と重み付き有限状態トランスデューサデコーダーを用いた単語列候補探索の実装については前述の Kaldi を利用し、本研究では入力特徴量の逐次処理とスレッドによる複数音源同時処理を新たに実装した。10 ms 単位で入力された特徴量を入力された分だけ逐次的に処理でき、入力された音源 ID と特徴量列のペアごとにスレッドに分けて同時に処理できる実装にすることで、速い応答と複数音源の同時処理を実現した。

4.4. 実験シナリオ

実験シナリオに関しては、Fig. 4 の流れに基づいてシナリオを構成した。実証実験ではシステムからの聞き直しを 3 回まで許容し、正しい担当者に取次げたら取次成功とし、それ以上音声誤認識等でシナリオが進行しない場合は取次失敗とした。

4.5. 結果

2016 年 1 月から 2 月にかけて計 37 日間、受付で実験

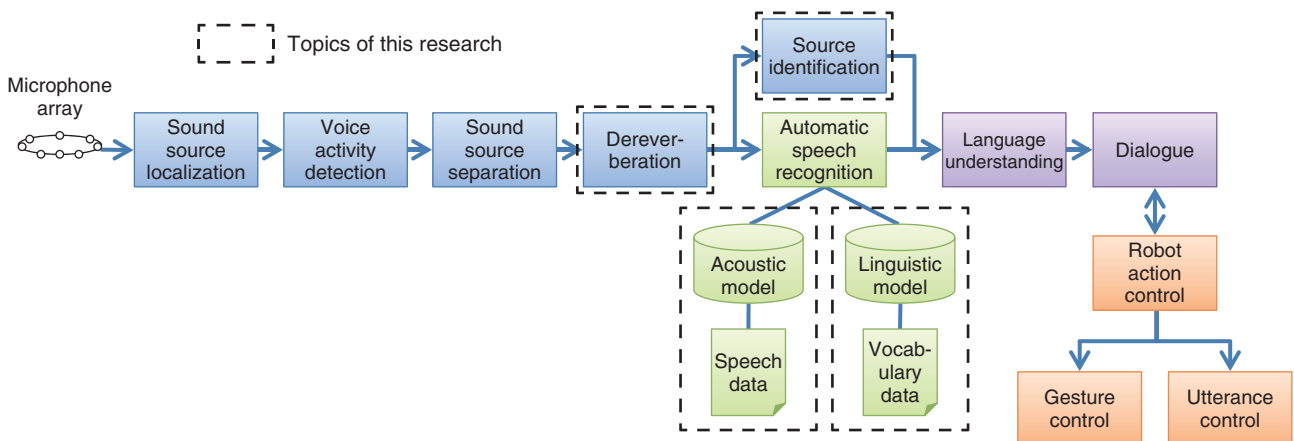


Fig. 5 Receptionist system

Table 4 Cause of error in reception experiments

Cause of error	Word error rate (%)
Unknown words	10.9
Spontaneous speech	8.8
Non-target sound and VAD	5.7
Lack of training data	4.5
Overlapping speech	3.3
Total	33.2

を行った。実際に音声で受付をした回数は延べ 303 回であった。実際にお客様が発話し、音声認識の評価に使った単語数は 4,881 語である。評価は単語正解精度と正しい担当者を実際に呼び出せた率(取次達成率)で評価した。全体の単語正解精度は 66.8% であった。その結果、取次機会 303 回中、取次に成功したのは 298 回で取次達成率 98.3% であった。

4.6. 考察

受付取次が失敗した 5 回の原因を分析してみると、置換誤りによる誤認識が 2 回、音源同定による誤同定(目的音声を非目的音として棄却)が 3 回であった。置換誤りに関しては、実際の担当者名に類似した姓名に誤認識してしまった。音源誤同定に関しては、モデルが GMM で十分記述できていない可能性があり、別のモデル化手法が必要と思われる。

音声認識に関しては、ベンチマークでは単語正解精度が 88.9% だったのに対して、実証実験では 66.8% まで低下した。実証実験での誤認識の原因を分類すると Table 4 の通りであった。これらの誤認識原因のうち以下の三つは、実用での課題を表している。一つ目は、受付取次特有であるが初めて来るお客様の名前や会社名が未知語になることによる誤認識の増加である。二つ目は、一般のお客様ではフィラーや言い誤りなど自発発話特有の現象が増加したことによる誤認識の増加である。三つ目は、ロボットが発話するタイミングがユーザにとってわからない等の原因によりロボットの発話と重なってしまうことによる誤認識の増加である。ロボットは例えお互いの発話が重なっても認識できるように対応してあるが、発話が重なってしまうと人は発話を中断してしまう場合が多く、誤認識の原因となってしまう。これは、ロボットが発話を始めるタイミングがわかりづらいということが原因であろう。但し、これらの誤認識の大部分は、取次タスクの遂行にとって致命的な誤認識ではなく、最も重要な担当者名と担当者所属名を高い精度で認識できたことにより、高い受付取次達成率が達成できたといえる。今後、受付取次より複雑なタスクを音声対話で行うためには、未知語の自動追加、自然発話特有の発話現象への対応、ユーザにとってわかりやすいターンテイキングの実現も必要になる。

5. まとめ

人と自然な距離でコミュニケーションするための遠隔音声認識技術を構築した。

本研究で提案した残響抑圧、音源同定、音響モデルを評価用ベンチマークデータで評価し、導入した技術が性能向上に効果があることを確認した。特に DNN を用いた音響モデルでは、904 時間のデータを学習させることで性能が 88.9% まで向上することが確認できた。提案した手法をロボット上で統合し、実際の受付でお客様に使っていただき、単語正解精度全体 66.8%、取次達成率 98.3% を達成した。

実証実験のデータを分析することで、さらなる性能向上のために対策が必要な課題を明らかにした。

- (1) 未知語への対応
- (2) 自発発話特有の発話現象への対応
- (3) ターンテイキング等のロボット表現

実際に毎日継続的にお客様に使ってもらうことで、データを集め、分析及び検証した本研究の実証実験は、今後、実証実験を行う上での一つのテストケースになると考えている。

今回は受付という比較的静かな場所を題材としたが、今後は、より多くの人が集まる公共施設などで雑踏音や館内放送等、高雑音への対応も重要な課題となるであろう。

参考文献

- (1) 経済産業省: ロボット新戦略ービジョン・戦略・アクションプランー, ロボット革命実現会議, (2015)
- (2) 奥乃博: ロボット聴覚の現状と展望, 日本ロボット学会誌, Vol. 28, No. 1, p. 2-5, (2010)
- (3) 河原達也: 音声認識技術の現状と将来展望, 電気学会誌, Vol. 133, No. 6, p. 364-367, (2013)
- (4) Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Hori, T., Nakatani, T., Nakamura, A.: Linear Prediction-based Dereverberation with Advanced Speech Enhancement and Recognition Technologies for the Reverb Challenge, REVERB Workshop, (2014)
- (5) Lee, A., Nakamura, K., Nishimura, R., Saruwatari, H., Shikano, K.: Noise Robust Real World Spoken Dialogue System Using GMM Based Rejection of Unintended Inputs, Proceedings of International Conference on Spoken Language Processing, Vol. 1, p. 173-176, (2004)
- (6) Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Yu, M. F. C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S., Nakatani, T.: The NTT CHiME-3 system: Advances in speech enhancement

- and recognition for mobile multi-microphone devices, IEEE ASRU, (2015)
- (7) 重見聡史, 川邊浩司, 中村孝広: 新型 ASIMO の開発 – 自律機械の実現 –, Honda R&D Technical Review, Vol. 24, No. 1, p. 34-41
- (8) Gomez, R., Kawahara, T.: Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 7, p. 1708-1716, (2010)
- (9) 鹿野清弘, 伊藤克巨, 河原達也, 武田一哉, 山本幹雄: 音声認識システム, オーム社, 198p., (2001)
- (10) Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Processing Magazine, Vol. 29, No. 6, p. 82-97, (2012)
- (11) Brown, P. F., Pietra, V. J. D., de Souza, P. V., Lai, J. C., Mercer, R. L.: Class-Based n-gram Models of Natural Language, Computational Linguistics, Vol. 18, No. 4, p. 467-479, (1992)
- (12) ヒューマノイドロボット EMIEW3 を活用した実証実験, 日立評論, Vol. 98, No. 12, p. 742-743, (2016)
- (13) <https://www.tokyo-airport-bldg.co.jp/hanedaroboticslab/>
- (14) 中川聖一, 高木英行: パターン認識における有意差検定と音声認識システムの評価法, 日本音響学会誌, 50 巻, 10 号, p. 849-854, (1994)
- (15) 西村義隆, 篠崎隆宏, 岩野公司, 古井貞熙: 周波数帯域ごとの重みつき尤度を用いた音声認識の検討, 日本音響学会 2004 年春季研究発表会講演論文集, p. 117-118, (2004)
- (16) Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J.: The Kaldi Speech Recognition Toolkit, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, p. 1-4, (2011)
- (17) 奥乃博, 中臺一博: ロボット聴覚オープンソースソフトウェア HARK, 日本ロボット学会誌, Vol. 28, No. 1, p. 6-9, (2010)

■ 著 者 ■



山本 俊一



住田 直亮



中臺 一博