

ロボットのための音声認識

河原達也
(京都大学)

1. ロボットと音声認識

人間のようなロボットを実現する上で、音声で会話する能力は必要不可欠なもの1つである。音声対話[1]を実現する上で、(1)音声認識、(2)音声合成、(3)認識してから返答・合成するまでの処理、の3つの要素が必要になる。(3)については究極の人工知能の研究ともいえるが、実際にはロボットにどのような機能(タスク)を持たせるかに依存する。(2)の音声合成についてもアンドロイドロボットのように人間と同等レベルが要求される場合もあるが、見た目がメカニカルなロボットだと、あまり人間らしい声はかえってミスマッチなので、多少自然性が悪くても現状の音声合成システムで十分とも考えられる。したがって、(1)の音声認識が最大の問題になっている。ペット型ロボットであれば、3回に1回程度正しく認識できれば許容されるかもしれないが、介護や案内などを想定した人間型ロボットでは当然人間に近い認識能力が期待される。

音声認識技術はこの10年余の間で飛躍的に進歩し、自動電話応答システム、カーナビ・携帯電話などの機器、講演や会議の書き起こしなど様々な応用への実用化がなされたが、ロボットへの搭載に関してはデモンストレーションの域を超えるものはほとんどなく、実現されたとは言いがたい。ロボットの音声認識は、音声認識の研究者にとっても非常に挑戦的なテーマとなっている。これは、人間が音声を認識する過程がいかに高度であるかを示唆するものであるが、一方で、我々が外国に行くと街角で様々な人に話しかけられてもうまく聞き取れない状況にも通じる。

2. 音声認識の原理と課題

2.1 音声認識の原理[2]

図1に音声認識の基本的な原理を示す。入力された音声は、信号処理によりMFCC(Mel-Frequency Cepstrum Coefficient; メル周波数ケプストラム係数)などのスペクトル包絡を表す特徴ベクトルに変換される。この信号処理において、雑音抑圧や話者・環境正規化なども行われる。抽出された特徴ベクトルは、HMM(Hidden Markov Model: 隠れマルコフモデル)に代表される統計的な音響モデルと照合される。この音響モデルは、音素単位で用意されることが一般的であるが、認識対象語彙とその発音を規定した単語辞書も同時に照合され、単語辞書に現れる音素系列のみがマッチングの対象となる(すなわち、単語辞書に登録されていない未知語は認識できない)。文発声などの連続音声認識の場合は、単語の接続を規定する文法・言語モデルも照合され、音響モデルの尤度と総合的に評価された結果最尤となる仮説が認識結果として出力される。

このように音声認識は、音響レベルの処理と言語レベルの知識及びアプリケーションの制約を統合的に利用して行われる。これは認識の最適化・高精度化の点で優れているが、どこか1つでも問題があると性能が大きく低下することとなり、しかもその問題を特定しにくくしている。

ここで、信号処理と音響モデルは主に入力環境に依存し、電話音声やカーナビ、ロボットなど利用される音響条件に合わせて構成し、話者が限定されていれば話者への適応も行う。一方、単語辞書と言語モデルはアプリケーション・タスクに応じて構成する。例えば、天気を案内するシステムであれば、それに関連した単語や言い回しを用意する。タスクが単純な場合は単語辞書や文法を手で記述できるが、そうでない場合は想定される発話文を多数収集して単語N-gramといった統計モデルを学習する。

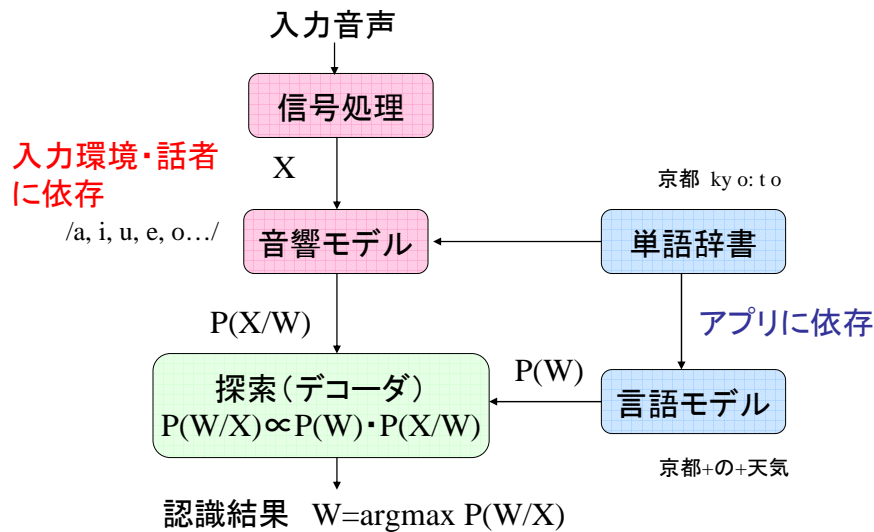


図1 音声認識の原理

2. 2 音声認識ソフトウェア Julius[3][4]

著者らは、約 10 年前からフリーの連続音声認識エンジン Julius の開発と公開を行っている (Web サイト <http://julius.sourceforge.jp/>)。これは、オープンソースでライセンス上の制約も緩いことから、多くのロボットにも搭載されている (と伺っている)。Julius の最大の特徴は、音響モデル・単語辞書・言語モデルを明確なインターフェースで分離しており、それらを置き換えたり、更新したりすることが容易になっていることである。そのため、国内外の多くの研究機関で基盤ソフトウェアとして用いられている。

ただし、Julius とパッケージ形式で公開している音響モデルは、無音室で接話マイクを通して発話されたデータ (“clean” と表記) で学習されたものであり、汎用性は高いものの、後述するように mismatches があると大きく性能が低下する。そのような場合は、音響モデルを環境に応じて構成し直す必要がある。また、同様に添付されている統計的言語モデルは、新聞記事や Web 上のテキストから学習されているので汎用性はあるが、基本的にはアプリケーションに応じて構成した方がよいだろう。文法の記述や言語モデルの学習は、音声認識の知識がなくても比較的容易である。音響モデルや言語モデルの構成法については、文献[2]の付録 CD-ROM にも演習があるが、著者らが毎年夏に講習会を行っているので、興味がある方は参加頂きたい。

2. 3 音声認識の技術的課題

最初に述べたように音声認識技術は飛躍的に進歩したが、本質的に解決していない課題として、話し言葉への対応と実環境への頑健性が挙げられる。話し言葉は発音の変形や言い回しの多様性が大きく、一般的なモデル化が困難であり、講演や会議などの個別の状況ごとにコーパスを収集して解決を図っている。家の中や車内・屋外などの実環境における雑音等への対応は、古くから多くの研究が行われているが、依然として音声認識を利用する上で大きな障壁になっている。

ロボットの音声認識においても、これら 2 つはいずれも重要な問題と思われる。しかし、実際に人間がロボットと会話する際に、人間どうしのような自然な話し言葉にならないのも事実であり、実環境への対応の方が大きな問題となっている。多くのロボットは動き回

るので、自分自身が非定常な雑音源となり、また周囲の音響特性も変化しうるという点で、非常に困難な音響条件を呈している。特に、ロボットに装着されたマイクと話者との距離が1m以上離れると、残響の影響も大きくなり、たとえ静かな条件でも認識は容易でない。

3. 実環境ロボットにおける音声認識技術

ロボットを指向した音声認識の実環境への対応について述べる[5]。ここでは、研究開発されている技術的な方法だけでなく、実際的な対策についても言及する。

3. 1 音声入力

ロボットのどこにいくつマイクを装着させるかは、悩ましい問題である。

1~3m程度の遠隔発話を音声認識するアプローチとしてマイクロフォンアレイがある。多数(4~8個程度)のマイクロフォンを一行に配置して、音が到来する時間差を揃える(Delay-and-Sum)ことで強調したり、逆に雑音源に対して死角を形成したりするものである。SN比を10dB程度改善する効果が知られているが、多チャンネルADなどの装置を含めて大規模・高価になるのが難点である。そのため2チャンネル以下で対応する研究開発が多い。

マイクロフォンは話す人間に対してできるだけ近く、指向性があることが望ましいので、「ジャーナリストロボット」[6]のようにロボットがマイクを向けるというのが(必ずしも一般的ではないが)理想的である。

マイクロフォンさらにデジタル化に至るまでの配線は、ロボット自身のモータなどの騒音源からできるだけ隔離されていることが望ましく、またロボット自体が音を反射しにくい材質であることが望ましい。実際の状況で収録された音声を聞いてみると、我々が想像するよりかなり音質が悪いことがしばしばある。

また、入力のゲイン制御が適切でなかったり、デバイスドライバやOSでブーストされている場合があるので、これらにも注意が必要である。

3. 2 雑音・残響抑圧

信号処理レベルにおける雑音抑圧の古典的な方法として、スペクトル減算法(SS: Spectral Subtraction)やウィーナフィルタ(WF: Wiener Filtering)がある[7]。これらは、雑音のみの区間から雑音のスペクトルの推定を行いながら、入力から差し引くものである。これらは単純で、特段の事前学習もなく適用できるので、広く用いられている。ただし、効果が限定的であり、また定常的でない雑音には対応しにくいという問題がある。

同様に、ケプストラム領域で長時間平均ケプストラムを差し引くケプストラム平均正規化(CMN: Cepstrum Mean Normalization)も一般的に用いられている手法である。これは、話者・チャンネル正規化とともに、一定範囲の残響抑圧や雑音抑圧の効果も有する。ただし、ロボットのようなリアルタイムの音声認識では、CMNは直前の数発話を用いて行われるのが一般的であるので、話者が交代したり、音響環境が変化した際には逆効果となる場合がある。話者交代時には、まず「こんにちは」などの認識しなくてもよい無難な発話をしてもらうのが望ましい。

3. 3 音響モデル適応

一般に最も有効と考えられる実環境への対応法は、実際の利用状況で音声を収録して音響モデルを再構成することである。少量の音声で適応できれば簡易であるが、モデル適応は話者に対しても同時に行われるので、不特定話者を対応としたシステムでは難しい。したがって、雑音や残響特性を収集して、元の学習データベースに重畳することになるが、これはかなりの手間・コストと専門的知識を要する。しかも、ロボットのように使用する部屋や音環境が事前に特定できない場合には、事実上不可能である。それでも、ある程度想定されるいくつかの典型的な音環境のデータを収集して、音響モデルを構成しておくこと(マルチコンディション学習)は意味がある。

3. 4 音声区間検出 (VAD: Voice Activity Detection)

通常、音声認識は、認識したい意味のある発話区間に対して行われるが、様々な騒音がある実環境では、このような音声区間を正しく切り出すこと自体が容易でない。音声認識システムの認識率などの評価は、切り出された音声に対して行われていることが多いので、実環境に適用して著しく動作しない場合は、(音声入力系の問題か) 音声区間検出に問題があることが多い。ユーザにスイッチを使用してもらうのも不自然なので、ロボットが発話し終わってから音声入力を受け付けるという制御を行うのが一般的であるが、ロボットの発話が完全に終わる前から話し始める人も多い。この発話区間検出の問題も、切り出された音声をファイルに保存して聞いてみれば確認できる。

音声区間検出を前処理と捉えずに、雑音抑圧や音響モデルと組み合わせて解決を図る研究が進められている[8]。また、騒音や笑い声などの認識対象としない入力も棄却する必要があるので、同様の枠組みで扱える。

4. テストケース

著者らが最近行った実験データを図2に示す[9]。2万語彙の新聞記事文(JNAS)のディクテーションのタスクにおける評価である。語彙サイズは大きいですが、言語モデルの効果も大きいので、実効的には数十～百語程度の単語認識と同程度と考えてよい。

“clean”の条件では90%を超える認識率であるが、騒音のない部屋で1~2m離れた入力条件(“遠隔”)で約70%まで低下している。これに加えて、計算機のファンなどの騒音があまり気にならないレベル(+SNR25dB)、少しうるさいレベル(+SNR15dB)の条件では、順にさらに大きく低下することがわかる。これに対して、上記のSSなどの雑音・残響抑圧手法や音響モデルの更新を行うことで、かなり改善されている。それでも、遠隔発話+SNR15dBでは、60%程度にしか到達していない。このデータはあくまで一例であるが、参考になれば幸いである。

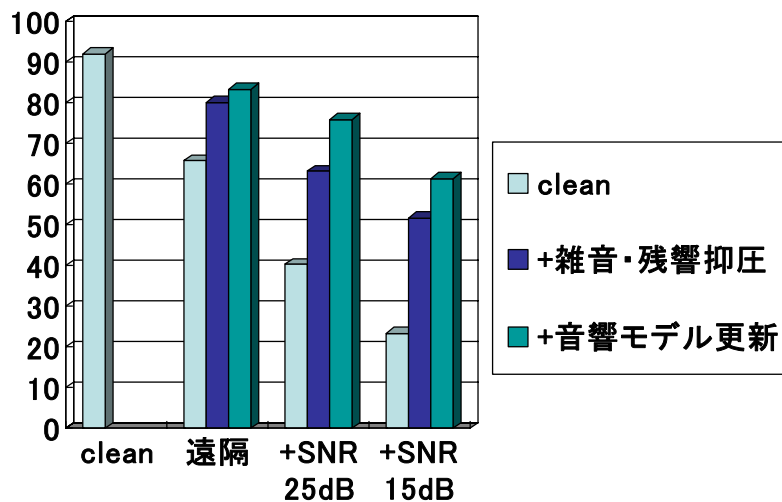


図2 音声認識精度
(2万語彙ディクテーション)

我々が外国語でもそれなりにやりとりできているように、文脈や状況などの様々な情報を総合的に活用しながら、対話できるようにするための研究も重要であろう。ロボットには様々なセンサが搭載されているので、このような観点の研究には適している。

参考文献

- [1] 河原達也, 荒木雅弘. [音声対話システム](#). オーム社, 2006.
- [2] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. [音声認識システム](#). オーム社, 2001.
- [3] 李晃伸, 河原達也. Julius を用いた音声認識インタフェースの作成. ヒューマンインタフェース学会誌, Vol.11, No.1, pp.31--38, 2009.
- [4] 河原達也, 李晃伸. 連続音声認識ソフトウェア Julius. 人工知能学会誌, Vol.20, No.1, pp.41--49, 2005.
- [5] 奥乃博. ロボット聴覚の現状と課題 --特集「ロボット聴覚」, 日本ロボット学会誌, Vol.28, No.1, 2010.
- [6] R.Matsumoto, H.Nakayama, T.Harada and Y.Kuniyoshi. Journalist Robot System: Robot System Making News Articles from Real World, Proc. IROS, pp.1234--1241, 2007.
- [7] 北岡教英. 音声認識におけるロバストネス. 日本音響学会誌, Vol.66, No.1, pp.23-28, 2010.
- [8] 石塚健太郎, 藤本雅清, 中谷智広. 音声区間検出技術の最近の研究動向. 日本音響学会誌, Vol.65, No.10, pp.537-543, 2009.
- [9] R.Gomez and T.Kawahara. Speech enhancement optimization based on acoustic model likelihood for noisy and reverberant environment. 人工知能学会研究会資料, Challenge-A902-9, 2009.

著者紹介： 河原達也 (Tatsuya Kawahara)

1989年京都大学大学院工学研究科修士課程修了。京都大学工学部助手，同助教授などを経て，2003年より同大学学術情報メディアセンター教授。音声言語処理，特に音声認識及び対話システムに関する研究に従事。京大博士（工学）。情報処理学会 SIG-SLP 主査。日本音響学会，情報処理学会 各代議員。IEEE Senior Member。