音声言語情報処理 32－6
（2000．7．14）

音声認識技術の今後の10年について –予測調査–

河原 達也　(京都大学)

あらまし

音声認識技術の今後の10年について予測調査を行う。特にここ数年は、応用面への期待が大きくなっている半面、今後の長期的な研究目標が見えにくい面もある。そこで本稿では、(1) 応用面での展開、(2) 技術面での展開、(3) 社会の環境の点から質問項目を設定し、可能性を考察した。これを基に今後の研究指針について考えたい。

Perspective of Speech Recognition Technologies in the Cominig Decade

Tatsuya Kawahara　(Kyoto University)

Abstract

Perspective of speech recognition technologies in the coming decade is discussed. While a tremendous number of their applications are being expected, the next long-term research goal is not clearly agreed. The author has set up a questionnaire with respect to (1) application issues, (2) technical issues, and (3) social environments. This will hopefully be helpful for considering future direction of research.

はじめに

1990年代は音声認識技術が最も進展し、実用化への展開が進んだのではないかと思われる [1]。これには、統計的なモデルの高精度化とその学習を可能にした大規模なデータベースの蓄積によるところが大きい [2]。

10年前に本研究会の前身ともいえる第二種研究会で、「音声認識研究 –これからの10年」というパネル討論が行われたが、“HMMの枠組みでデータを増やしていけばかなりのところまでは行く”という予測が何名かによってなされており、その通りになったといえる。しかしこれは逆にいうと、本質的なアプローチやアルゴリズムの点で10年間変わっていないともいえる。

また6年前に本研究会が発足した際に、嵯峨山先生を中心に、「なぜ音声認識は使われないか」[3] と議論されたのが一転して、現在はビジネスへの展開が米国を筆頭に著しい。本当に使われているかは別として、音声認識を用いたアプリケーションが身の回りに存在するようになった。少なくとも“ようやく使える”というムードが漂っているように思える。しかし、こうした応用面への期待が大きくなっている半面、今後の長期的な研究目標が見えにくい面もある。

時代の変化は年々激しくなり、数年後のことさえ予測するのが困難であるが、今後の研究開発の指針とする上で、音声認識技術の今後の10年について、(1) 応用面での展開、(2) 技術面での展開、(3) 社会の環境の点から予測を行いたい。なお応用面では、語学教育支援 (CALL) とかゲーム・オモチャといったEdu-tainmentの分野があるが、こういったいわば副産物的なものは除外している。

この予稿の段階においては、著者が興味をもち設定した質問項目について、種々の可能性を鑑みて選択肢を用意した。研究会当日までに電子メール等で調査を行い、集計した上で、コメントや議論を持ちたいと考えている。[1]

予測調査項目

(1) 応用面での展開

5年後あるいは10年後において、音声認識を用いた以下のアプリケーションができていると思いますか。なお、「できている」というのは、実用化され、一般の人に浸透している(それなりに購入・利用される)レベルに到達していることを意味します。例えば、現在のディクテーションソフトや電話機の名前ダイヤル機能などが挙げられます。

選択肢の後に、「当然!」という以外の理由(論点)を挙げています。

Q1 講演音声の自動書き起こしが実用化されている
(a)5年後　(b)10年後　(c)10年後もされていない

- 70〜80%程度の認識率でも、最初から人手で書き起こすより効率がよい
- 話し言葉の認識は難しく、それほど進まない
- 個々の話題や話者への対応/適応が困難である
- 技術的にできても、ビジネスとして成立しない

Q2 裁判所や国会の記録に、速記の代わりに音声認識が導入されている
(a)5年後　(b)10年後　(c)10年後もされていない

- 速記官のなり手不足という要因からも望まれる
- 話し言葉の認識は難しく、それほど進まない
- 個々の話題や話者への対応/適応が困難である
- 技術的にできても、法的・制度的な整備が遅れる

Q3 駅の券売機やコンビニの端末での音声入力が一般的になっている
(a)5年後　(b)10年後　(c)10年後もなっていない

- 騒音などへの対応ができない(発話の検出を含む)
- 語彙外発話への対応ができない
- 人々は公衆の面前では音声入力を使いたがらない
- 新たにコストが生じるので導入されない

Q4 家電製品(照明やエアコン)の音声操作が標準的なオプションになっている
(a)5年後　(b)10年後　(c)10年後もなっていない

- 遠隔マイクでの認識ができない
- 語彙外発話への対応ができない
- 最初は面白がられても、コスト面を含めて結局は使われない

Q5 電話による情報案内は音声認識が主流になっている
(a)5年後　(b)10年後　(c)10年後もなっていない

[1] 調査票のオンライン版と回答の結果は、http://winnie.kuis.kyoto-u.ac.jp/~kawahara/asr-forecast/

- 語彙外・文法外発話が多すぎて、対応できない
- 日本のサービスでは不完全な技術は好まれない
- WWW や携帯端末が普及すれば電話音声サービスは使われない

Q6 電話による商品の注文やバンキングも音声認識が主流になっている
(a)5 年後　(b)10 年後　(c)10 年後もなっていない

- 語彙外・文法外発話が多すぎて、対応できない
- お金の決済を伴う場合は音声認識は使用されない
- WWW や携帯端末が普及すれば電話音声サービスは使われない

Q7 旅行の相談や法律の相談が音声対話システムでできるようになっている
(a)5 年後　(b)10 年後　(c)10 年後もできていない

- わかるものに答えられるだけでも、有用である
- 内容が多岐にわたりすぎて、対応できない
- ビジネスとして成立しない

Q8 携帯電話／端末において音声入力による E-mail 作成が標準装備されている
(a)5 年後　(b)10 年後　(c)10 年後もされていない

- 騒音や入力系に対応できない
- 計算パワー (CPU・メモリ) が足りない
- 人々は公衆の面前で音声入力を使いたがらない

Q9 日常会話の音声翻訳システムができている
(a)5 年後　(b)10 年後　(c)10 年後もできていない

- これくらいには、できていて欲しい?
- そんなに早くできると困る?

Q10 特定のタスクでは、機械と話しているのがわからない程度の音声対話システムができている
(a)5 年後　(b)10 年後　(c)10 年後もできていない

- 今でも存在している!? → そのシステム名は?
- 認識はできても合成が難しい??

(2) 技術面での展開

10 年後において、音声認識の基本的な技術はどのようになっていると思いますか。

Q11 依然として、MFCC もしくは LPC のパラメータが主流である。そうでない場合、他の候補は ?

(0) はい
(1) RASTA, PLP
(2) 聴覚モデル
(7) その他 (　　)
(8) 何かはわからない／言えないが、別のものになっているだろう
(9) わからない

音響特徴量については様々な研究が行われているものの、世の中の認識システムで採用されているものは10年間変わっていません。むしろ、以前より画一化されてきた様相です。フロントエンドをよくしなければ、という意見もありますが。

Q12 音響モデルは、依然として現在のようなHMMが主流である。そうでない場合、他の候補は？

(0) はい
(1) セグメントモデル、トラジェクトリモデル
(2) ニューラルネットのようなモデル
(7) その他 (　　)
(8) 何かはわからない／言えないが、別のものになっているだろう
(9) わからない

10年前には現在のHMMの枠組み・アルゴリズムはできていました。ただし、10年前は離散HMMが主流だったような気がしますし、ニューラルネットも全盛でした。なおニューラルネットのようなモデルとは、サポートベクトルマシンなどの識別型モデルを包含します。

Q13 言語モデルは、依然として現在のようなN-gramが主流である。そうでない場合、他の候補は？

(0) はい
(1) 文法モデルと統合されたモデル
(2) 意味や文脈を反映したモデル
(7) その他 (　　)
(8) 何かはわからない／言えないが、別のものになっているだろう
(9) わからない

10年前は、日本語はN-gramでモデル化できないという意見が多数でした。言語学 (文法) 的な知識が見直される時は来るのでしょうか。また、3-gramモデルを優位に向上させる意味や文脈のモデル (LSAやキャッシュモデル?) は実現されるのでしょうか。これもデータ量の問題でしょうか。

Q14 タスクやドメイン毎に言語モデルを用意する必要はなくなっている。
(0) いいえ　　(1) はい　　(9) わからない

音響モデルの性能が向上し、汎用的な言語モデルが構築できれば、こうなるかもしれません。なおメモリの制約は考えないことにします。話題を制御する場合も動的であれば、汎用的としましょう。人間の言語認知はどうしているのでしょうか。

Q15 話者適応はもはや不要 (手間の割に効果がない) となっている。
(0) いいえ　　(1) はい　　(9) わからない

非母国語話者や極端な方言はとりあえず除きます。また、モデルの選択は適応には含みません。音響モデルがあらゆる話者層をカバーしてしまえば、こうなるかもしれません。これも人間の認識はどうしているのでしょうか。

Q16 人間と同程度に、マイクとの距離は問題ないレベルになっている。
(0) いいえ　　(1) はい　　(9) わからない

マイクロフォンアレイあるいはスペースダイバシティ型の音声認識で実現されるのでしょうか。

(3) 社会の環境

10年後において、音声認識を取り巻く社会環境はどのようになっていると思いますか。

Q17 「音声認識はできた」というのが世の中の認識になっている。
(a) はい　(b) いいえ　(c) わからない

Q18 音声認識の研究を行っている大学などの研究室は現状より増えている。
(a) はい　(b) いいえ　(c) わからない

Q19 音声認識技術をビジネスにしている企業は現状より増えている。
(a) はい　(b) いいえ　(c) わからない

Q20 携帯端末 (i-modeのような) が広く普及した後も、電話音声による自動応答サービスは必要とされていると思いますか。
(a) はい　(b) いいえ　(c) 端末での単語認識ですむようになっている

Q21 人間以外のもの (エージェント含む) に対して、人格を見いだして自然に話しかけることが社会で一般的に受容されているでしょうか。
(a) はい　(b) いいえ

Q22 音声言語情報処理研究会は、この名称で存続していると思いますか。
(a) はい　(b) いいえ　(c) わからない
(b) の場合、その名称は＿＿＿＿＿＿＿＿

(4) 最後に、研究目標について

自由回答でお願いします。今後10年の話と限りません。

Q23 流暢な第二言語 (外国語) 話者と同程度の音声認識はいつ頃実現できると思いますか。

[コメント] TOEICでは800〜900点以上でしょうか。英語のディクテーションや放送音声認識のシステムをみると、我々を上回っているような気もします。ただしここでは、ドメイン非限定で、環境へのロバストネスも含みます。母国語話者なみというと気が遠くなるほど先ですし、工学的にはこのあたりが目標と思われます。生きているうちに実現したいものです。

Q24 話し言葉 (Switchboard DBなど) に対する認識率は依然として低いですが、今後データを増やしていけば、現在のHMMと単語N-gramの枠組みで解決されると思いますか。あるいは、どのあたりに抜本的な改良が必要と思いますか。

[コメント] 話し言葉は読上げ音声に比べてバリエーションが大きすぎて、単にデータが足りないのか、モデルの自由度が足りないのかが明らかではありません。前者であれば回答はYesです。私にはそもそも、 $p(W|X) = p(W)p(X|W) = \prod p(w_k|w_1..w_{k-1})p(x|w_k)$ でデコードするという前提 (W:単語列,X:音声) がおかしいような気がします。

Q25 話者層や入力環境へのロバストネスが大きな課題となっていますが、今後データを増やしていけば、現在考えられているような適応の枠組みとあわせて解決されると思いますか。あるいは、どのあたりに抜本的な改良が必要と思いますか。

[コメント] 10年前のパネルでも、6年前の討論でも、「ロバストネスが問題である」と言われました。パターン認識の本質ともいえますが、音声認識の使用される状況に限れば、ほぼ対応される日が来るのでしょうか。

予稿作成段階で寄せられたコメント

この調査項目の妥当性の検証を兼ねて、数名の方に事前に回答をお願いした。寄せられた回答からポイントと思われる箇所(主にQ24-25の回答)を抜粋した。なお各先生方には、研究会当日に来られれば、コメンテータをつとめて頂く予定である。(以下敬称略)

中村 哲 (ATR)

音響モデルのミスマッチも大きいが、非常に自由度の大きい対話の認識にはかなりかかると思う。基本的には、言語モデルをかなり変えないといけないと思う。やはり、文法や意味、文脈などの高次の情報を使う方向に進む。また、柔軟かつ統計的な対話のモデルが出現するだろう。話者について、年齢や静的な差は解決される。方言などの問題は残る。入力環境への頑健性は、音声以外の情報の利用、環境のデータベースの整備で研究が進むと思う。このあたり、すこし人間の情報処理を学んだ方が良いかもしれない。

西村 雅史 (日本IBM)

Q6の質問に関してですが、"電話の音声認識によるサービス"って5年後や10年後にもまだ使ってますかね？ みんなi-modeでやってるんじゃないかって気がずっとしているのですが。私としては、"5年後・10年後、i-modeのような機器が広く普及した後も電話音声による自動応答サービスは必要とされていると思いますか?" というような質問を社会の環境あたりの項目に加えていただけたらと思います。
[著者注] この指摘に基づいて質問項目を当初より追加しました。

山下 洋一 (立命館大)

音響情報、言語情報、韻律情報 (+マルチメディアシステムであれば画像情報) など、個別の情報の取捨選択 (どこには何が有効か) および統合がやはり今後の課題の一つになりそうです。きれいな読み上げ音声の認識では、音響情報と言語情報を確率の枠組で統合し、成功を収めたと言えますが、会議の音声や対話音声などのように多様性が広がると、母集団の特徴をカバーできるだけのデータを収集することが難しいのではないかという気がします。そうなると、別の情報統合の枠組が必要ということになりますが、果たしてどこまで「データ収集＋確率モデル」で挑むのでしょうか？

武田 一哉 (名古屋大)

いかにHMMやN-gramといえども、モデルの大規模化には限界がある。(どれだけデータがあっても、電話とマイクで同じモデルというわけにはいかないでしょう)。様々なモデルを適宜切り替えながら利用する情報統合の原理が必要ではないか。

小林 哲則 (早稲田大)

(1) かなりのことは10年後に解決している。(2) そのとき、基本的には今ある技術 (HMM+Ngram) がベースになっている。(3) ただし、データだけの問題ではなく、それぞれの確率モデルも進歩を遂げている。(4) 適応というよりは、データの正規化の技術が進む。(5) ディスタントマイクの問題など、ニーズのある技術は数年で急速に発展する。

参考文献

[1] 河原達也. ここまできた音声認識技術. 情報処理, Vol. 41, No. 4, pp. 436-439, 2000.

[2] 伊藤克亘, 河原達也, 武田一哉. どうすればデータ共有を成功させることができるか–音声認識分野での事例–. 情報処理, Vol. 41, No. 7, 2000.

[3] 嵯峨山茂樹. なぜ音声認識は使われないか・どうすれば使われるか？ 情報処理学会研究報告, 94-SLP-1-4, 1994.