音声言語情報処理 34－21
(2000. 12. 22)

〔サーベイ〕 話し言葉音声認識の概観

河原 達也

京都大学 情報学研究科
〒606-8501 京都市左京区吉田本町

あらまし 読上げ音声の認識は数万語彙でもかなりの認識精度を達成しているのに対して、自然な話し言葉の音声については、タスクドメインを限定した場合でないと十分な性能が得られていないのが現状である。本稿では、話し言葉音声認識の困難さについて分析を行った上で、音響モデル・発音モデル・言語モデルなどのアプローチについて概観する。

キーワード 音声認識, 話し言葉, 音響モデル, 発音モデル, 言語モデル

Toward Spontaneous and Conversational Speech Recognition

Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

Abstract While large vocabulary continuous speech recognition systems achieve high accuracy in read speech as in dictation systems, recognition performance on the spontaneous and conversational speech is still poor unless the task domain is limited. In this review, the problems and approaches in acoustic, pronunciation and language models are addressed.

key words speech recognition, spontaneous speech, conversational speech, acoustic model, language model

1 はじめに

ディクテーションシステムに代表されるように、人間が機械に "聞かせてあげる" ように発声した音声の認識は、かなりの精度でできるようになった。数万語彙の認識においても、実時間で約90%の認識精度が得られるようになっている [1]。

その一方で、人間どうしがふだん話しているような音声に対しては、大きく性能が低下するのが現状である。ロボットや擬人化エージェントのように、人間に近いインタフェースにおいてはより自然な発声を扱える必要があるし、講演の書き起こし [2] や会議の議事録作成 [3] のためには、そのような音声を認識する必要がある。

これまでに音声対話システムや音声翻訳システムを指向して、対話音声の認識の研究は数多く行われている。しかしながらそれらは、きわめて限定されたタスクドメインを想定しており、読上げではないものの、機械に話すことを前提としている。

本稿では、ドメインを限定しない大語彙で、かつ人間相手に話されるような真の「話し言葉」音声を対象として、その分類を行い、認識のためのアプローチを概観する。

2 「話し言葉」の分類

まず、音声認識の対象としての「話し言葉」について分類を行ってみたい。

Spontaneous Speech

話し言葉に明確に対応する英語表現は定かでないが、[1] "spontaneous speech" という言葉はよく用いられる。これは、もっぱら "read speech" への対比で、発声内容やテキストをあらかじめ用意しないで、自発的に発声された音声のことである。

ただし、"spontaneous speech recognition" とうたった論文の大半は、音声対話システムか音声翻訳システムを想定しているものである。これらは、自発的な発声であるが、音声認識装置を意識しながらのものであり、間投語や言い淀みなどはあるが、基本的には人間どうしとの会話に比べれば丁寧である。

またこのようなアプリケーションにおいては、タスクやドメインを限定するのが一般的である。すなわち、フライトや列車の情報検索であるとか、旅行とかスケジュール管理といった類いである。このようにドメインを限定することで語彙サイズはおおむね数千以下におさまり、また当該の対話のデータベースを収集すればパープレキシティが数十の統計的言語モデルが構築できる。この結果、テストセットに対してはおおむね80～90%程度の認識率を得ている。ただし、システムの想定外の発話に対してはほとんど認識できないので、本質的に解決されているとはいえない。

本稿では、そのようなタスクやドメインが限定された "spontaneous speech" は対象としない。対象とするのは、ドメイン非限定の大語彙連続音声である。

[1] "colloquial" という単語は音声認識の論文では使われていない。

Conversational Speech

また話し言葉音声として、"conversational speech" という用語もしばしば用いられる。これは、会話調・口語調の音声であり、必ずしも電話音声である必要はないが、事実上 Switchboard[4] と Call Home の両コーパスを対象としたものがほとんどすべてである。このような会話音声は、当然 "spontaneous" であると考えられるが、その最大の特徴は、話者が認識装置を意識していないことである。また、これらには情報検索といった明確なタスクはなく、会話自体が目的である。個々の会話には話題が与えられるが、話される内容は多岐にわたり、個々の話題に特化したモデルは構成できない。

上記コーパスを対象とした認識は、DARPA の Hub-5 で行われてきたが、おおむね30～40%の誤り率で推移してきた。今年最高で20%を下回る結果が得られたが、これは主にテストセットを (Switchboard-2 から Switchboard-1 に) 変更したためであり、実質的には過去数年間若干の向上しか得られていない [5]。

講演音声

これに対して我が国では、昨年度より始まった開放的融合研究「話し言葉工学」(代表：古井貞熙教授) のプロジェクト [6] において、独話による講演を主な対象として、700万形態素を目標に大規模なコーパスの構築を行っている。講演音声は自動書き起こしという明確なマーケットが存在し、技術的には国会や裁判所等の記録にも応用できると考えられる。講演は一応フォーマルな場でのスピーチであり、"conversational" ではないし、多くの人はリハーサルをするので、真に "spontaneous" かどうかも疑問である。ただし、原稿を読上げる人は10%未満であり、すべてをそらんじて話す人は少ない。メモ程度も準備しない人が30%強おり、こうした熟練した人はかなり自然な口調で話すことが多い。また、原稿などを用意した場合としない場合とを比較すると、フィラーや言い

誤りの出現には大差なく、発話速度にのみ有意な差がある (用意しない方が速い) ことが報告されている [7]。これらの理由から、講演音声は十分に話し言葉に近いといえる。特定の話題に沿って話す点で Switchboard-1 に類似しているかもしれないが、話題や専門用語の数は講演の方がはるかに多い。また、同プロジェクトでは原則としてヘッドセットマイクを用いて広帯域で収録しており、音響的な雑音や歪みの問題はあまりないので、純粋に話し言葉の問題を扱うには適している。

比較的フォーマルな講演以外に、大学での講義 [8] や会議 [9] などの音声も考えられる。これらの方が原稿を用意したりリハーサルを行ったりしない点で、より話し言葉らしいといえる。これらは講演のすぐ次のターゲットと考えられる。ただし、講義や会議の性格上、模擬でないデータを公開を前提に収集するのは非常に困難である。

3 話し言葉音声認識へのアプローチ

次に、話し言葉音声の困難さを分析した上で、話し言葉を指向した音声認識のアプローチについて概観する。

3.1 セットアップ

まず見落とされがちな大きな問題として、従来の書き言葉の音声認識ではあまり問題とならないような下準備の困難さが挙げられる。

コーパスの作成はどのような場合でも容易ではないが、話し言葉の場合は書き起こし作業が困難を極める。読上げ音声の場合は正しく発声されたかのチェックだけであり、対話音声の場合も個々の発声はおおむね短く明瞭であるので、それほど困難ではなかった。しかし、講演のような音声においては、不明瞭な区間が多数あり、何度聞いてもわからない箇所がある。[2]

次に何とか書き起こせたとして、形態素解析の問題がある。「おっきい」「白っぽい」といった口語的な言い方に対応する必要がある。特に、「〜じゃない」「見とく」「こりゃ」といった複数の形態素にまたがる変形の扱いは注意を要する。形態素解析を拡張するか、元の形を併記するといった解決が必要であるが、書き起こしの段階から一貫性を保持するには、基準の整備と熟練を要する。

このような理由から、融合研究プロジェクトでもコーパスの整備に多大な作業を要している。

上記とは別に、音声認識実験を行う際に音声データのセグメンテーションの問題が生じる。Viterbi アルゴリズムは入力の終端を検出しないと認識結果が確定しないし、マルチパス探索の場合は必ずどこかで入力を中断する必要があるが、講演音声などでは文を単位として発声が行われておらず、文の途中でも長いポーズが入ったり、逆にポーズがほとんどなく連続的に文が発声される場合がある。認識結果を逐次的に確定させたり [10]、言語モデル計算のための単語履歴を引継ぐなどの処理が必要になる。特に、講演録作成支援などのアプリケーションを考えると、後処理の修正において複数候補が必要となるので、効率的に N-best 候補を求めることも要求される。また、文の境界を検出するために、ポーズだけでなく F0 などの情報を利用することも考えられる [8]。

以下に述べるように、音響的・言語的な曖昧性が大きいために、ディクテーションや放送ニュースの書き起こしでは実時間の数倍で動作するシステムでも、話し言葉の認識には実時間の数十倍を要する。Switchboard タスクでは数百倍が一般的である。このように評価実験が効率的に行えないことも困難な要因の一つに数えられよう。

3.2 音響モデル

話し言葉のように、変形が大きく、ときに不明瞭な音声をどのようにモデル化するかは大きな問題である。

セグメントモデル [12] は、HMM のように状態とその分布のみではなく、もっと音響的特徴量の動的な変化の自由度を大きくするもので、発声の変形をモデル化するには自然なアプローチと考えられる。ただし、多大な計算量と音素などへのセグメンテーションを要するため、主にリスコアリングにしか用いられず、その効果が十分に示されているかは不明である。

また、quinphone や機能語モデルのように、より長いコンテキストや単位でモデル化することも変形を吸収するアプローチとして自然な考え方であるが、コンテクストの組合せが膨大になるため、学習量が問題になる。

最近のいくつかの研究では、発話速度 (とその変動) が最も認識精度に関係するのではという考えに基づいて、話速別にモデルを用意したり [13]、決定木による状態クラスタリングの際に発話速度も考慮する [14] などの方法が提案されている。

Switchboard タスクにおいては、ケプストラムの声道長正規化 (VTLN: Vocal Tract Length Normalization)[15][16] と、音響モデルの話者適応型学習 (SAT: Speaker Adaptive Training) [17][18] がほぼすべてのシステムで採用されており、ベースラインモデルからの最大の改善はこれらによるものと推測される。VTLN や MLLR は発話毎に第一パスの認識結果を用いて実行されており、処理時

[2] 著者も「専門家が聞けばまだわかるのでは」と日本音響学会の講演のチェックを頼まれたが、不明瞭なところはやはりわからなかった。

間があまり重要でないオフラインでマルチパスの認識が可能なアプリケーションの効用ともいえる。

3.3 発音モデル

話し言葉においては、正しい発音 (baseform) 通りに発声されるとは限らないので、音響モデルだけでなく、サブワード単位の設定や発音辞書の記述が重要である [19]。

単純に可能な変形パターンを記述していくと、Switchboard のようなタスクでは、ある単語に数十種類の発音エントリが登録されることになるし、例えば、'and' に [ae n] というエントリを追加すると 'an' と混同しやすくなるなどの副作用が生じる。

そのため、データから音素モデル等を用いて発音辞書を自動的に学習・構成する研究が多数行われている。また、発音辞書と音響モデルを統合的に最適化したり [20]、機能語モデルと音素モデルを併用する方式 [21] の研究も行われている。

しかし一般的には、単純に発音エントリを増加させると曖昧性が増大するので、発音変形の出現をモデル化する必要が生じる。単純に発音の生起確率を求めることもできるが [22][23]、これは unigram モデルなどに相当し、若干の効果しか期待できない。したがって、その他のコンテクストを利用したモデル化 (dynamic pronunciation model) が必要である。発話スタイルや発声速度 [13] の他に、音韻論的な素性や語彙的強調の有無などの言語的な特徴が関係すると考えられ [24]、数々の試みがなされているが、どのように有効な特徴を効率よくモデル化するかは今後の (現在の) 課題である。

3.4 言語モデル

話し言葉は必ずしも文法的でないので、新聞記事などの文に比べておおむねパープレキシティが大きい。また、講演音声などを対象にすると数万語彙は必要である。

しかし話し言葉のコーパスは、前述のように構築が大変であるので、新聞記事などの書き言葉のテキストデータに比較して、はるかに少量である。現在構築が進められている融合研究コーパスに期待される。学習データを補完するために、書き言葉の言語モデルと混合したり、新聞記事などから話し言葉に近いテキストを選択・抽出するなどの処理も考えられる。

話し言葉に固有の言い淀みなどのモデル化も問題である。フィラー (間投詞) は忠実な書き起こしテキストがあれば一応のモデル化はできるが、その前後の単語の連鎖が中断されるので、予測はするが履歴には用いないといった透過単語扱いにするなどの処理 [11] が必要である。なおディクテーションにおいては、ポーズは句読点と対応させることもできたが、話し言葉ではポーズが文やフレーズの切れ目とは必ずしも関係なく出現するので、ポーズもフィラーと同様に透過単語にするなどの扱いが必要である。

また講演や講義、会議においては、特定の分野や話題が存在し、専門性も高いことから、専門用語を追加したり、分野や話題に適応するといった処理も考えられる [2]。

4 話し言葉処理の応用

講演や会議などの音声の自動書き起こしを実用面から考えると、人手による修正・編集作業が不可欠であるので、それを支援できることが望ましい。認識結果として N-best 候補を求めるだけでなく、信頼度を付与しておけば、誤り訂正の負荷が軽減されるであろう。また、フィラーや言い淀みなどにマーカを付けたりするのも役に立つかもしれない。さらに、編集や要約の (半) 自動化の研究も行われている。

5 「話し言葉」の「音声理解」

このように講演録の作成や要約までを考えると、話し言葉の一字一句を残さず書き取ることの意味、すなわち単語認識率で評価することの妥当性に疑問が生じる。特に認識が難しい曖昧な発声区間は、(人間どうしの意志伝達における情報理論の観点からも) あまり重要でない可能性が高い。音声対話システムなどでは、タスクを遂行するのに必要なキーワード集合というのが比較的明確であるので、意味理解率という尺度を定義することも可能であるが、講演音声などに対する理解率を定義し、そのような評価基準に基づいてシステムを設計・構築することはできるだろうか。しかし、そこまでできて初めて真の「音声理解」といえると思われる。

参考文献

[1] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価. 情報処理学会研究報告, SLP-31-2, NL-137-7, 2000.

[2] 加藤一臣, 李晃伸, 河原達也. 講演ディクテーションのための話題独立言語モデルと話題適応. 情報処理学会研究報告, SLP-26-2, 1999.

[3] 秋田祐哉, 河原達也. 会議音声の自動アーカイブ化システム. 情報処理学会研究報告, SLP-34, 2000.

[4] J.J.Godfrey, E.C.Holliman, and J.McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. IEEE-ICASSP*, Vol. 1, pp. 517–520, 1992.

[5] J.Fiscus, W.M.Fisher, A.Martin, M.Przybocki, and D.S.Pallett. 2000 NIST evaluation of conversational speech recognition over the telephone. In *DARPA Speech Transcription Workshop*, 2000.

[6] S.Furui, K.Maekawa, and H.Isahara. Toward the realization of spontaneous speech recognition – introduction of a Japanese priority program and preliminary results –. In *Proc. ICSLP*, Vol. 3, pp. 518–521, 2000.

[7] 籠宮隆之, 菊地英明, 小磯花絵, 前川喜久雄. 大規模話し言葉コーパスにおける発話スタイルの諸相. 音講論, 2-Q-9, 秋季 2000.

[8] 野村和弘, 河原達也, 堂下修司. F0 パターンに基づく講義音声の文単位へのセグメンテーション. 電子情報通信学会技術研究報告, SP99-13, 1999.

[9] H.Yu, T.Tomokiyo, Z.R.Wang, and A.Waibel. New developments in automatic meeting transcription. In *Proc. ICSLP*, Vol. 4, pp. 310–313, 2000.

[10] 今井亨, 小林彰夫, 佐藤庄衛, 安藤彰男. 逐次 2 パスデコーダを用いたニュース音声認識システム. 情報処理学会研究報告, 99-SLP-29-37, 1999.

[11] 西村雅史. 日本語ディクテーションシステムの現状と今後の課題. 情報処理学会研究報告, 99-SLP-29-2, 1999.

[12] M.Ostendorf, V.V.Digalakis, and O.A.Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech & Audio Process.*, Vol. 4, No. 5, pp. 360–378, 1996.

[13] J.Zheng, H.Franco, and F.Weng. Word-level rate of speech modeling using rate-specific phones and pronunciations. In *Proc. IEEE-ICASSP*, pp. 1775–1778, 2000.

[14] C.Fugen and I.Rogina. Integrating dynamic speech modalities into context decision trees. In *Proc. IEEE-ICASSP*, pp. 1277–1280, 2000.

[15] L.Lee and R.C.Rose. Speaker normalization using efficient frequency warping procedures. In *Proc. IEEE-ICASSP*, pp. 353–356, 1996.

[16] S.Wegmann, D.McAllaster, J.Orloff, and B.Peskin. Speaker normalization on conversational telephone speech. In *Proc. IEEE-ICASSP*, pp. 339–342, 1996.

[17] J.W.McDonough, T.Anastasakos, G.Zavaliagkos, and H.Gish. Speaker-adapted training on the switchboard corpus. In *Proc. IEEE-ICASSP*, pp. 1059–1062, 1997.

[18] D.Pye and P.C.Woodland. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In *Proc. IEEE-ICASSP*, pp. 1047–1050, 1997.

[19] M.Ostendorf. Moving beyond the beads-on-a-string model of speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1999.

[20] T.Holter and T.Svendsen. Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 199–206, 1997.

[21] M.Bacchiani and M.Ostendorf. Using automatically-derived acoustic sub-word units in large vocabulary speech recognition. In *Proc. ICSLP*, pp. 1843–1846, 1998.

[22] B.Peskin, M.Newman, D.McAllaster, V.Nagesha, H.B.Richards, S.Wegmann, M.Hunt, and L.Gillick. Improvements in recognition of conversational telephone speech. In *Proc. IEEE-ICASSP*, pp. 53–56, 1999.

[23] H.Schramm and X.Aubert. Efficient integration of multiple pronunciations in a large vocabulary decoder. In *Proc. IEEE-ICASSP*, pp. 1659–1662, 2000.

[24] E.Fosler et al. Automatic learning of word pronunciation from data. In *Proc. ICSLP*, 1996.