Vol.2009-SLP-78 No.11
2009/10/23

実世界に存在する音声・音響を対象とした認識技術

大　淵　康　成[†1]

UI としての音声認識技術は，かなり高い精度を実現してはいるものの，ユーザーの要求水準も高く，必ずしも広く普及しているとは言えない．一方，実世界には既に多くの音が存在しており，これらのデータを用いた分類・検索技術が実現すれば，大きなメリットが得られる．本稿では，音声と非音声の判別，ログデータからの音声検索，音響イベントのセンシングといった応用の観点から，音声・音響処理の新しい方向性について述べる．

Recognizing Commonly Existing Speech and Sound

Yasunari Obuchi[†1]

Speech recognition as a human-machine interface has already achieved reasonable accuracy. However, it is yet to be used widely due to the users' high-level requirements. Meanwhile, the world is filled with a lot of sounds. Classification and search techniques for these sound data would provide various benefits. In this paper, a new paradigm of speech and sound processing from the viewpoint of speech/non-speech discrimination, spoken term detection, acoustic event sensing, etc.

1. はじめに

音声認識技術は,「人とコンピュータが話をする」というメタファーで語られてきた．SF 小説のように流暢に話す装置を期待したユーザーは，試しに使ってみると，その性能が完璧なものではないことに気づき，多くの場合その後はゆっくりと丁寧に話すようになる.「ボタンを押して，ピーッという音がしたら話しはじめて下さい」というようなガイダンスがあれば，それに従ってくれる．このような協力的なユーザーを念頭に置いての研究開発の結果，最新の音声認識システムは，90%を超えるような高い認識率を示すようになった．カーナビの音声入力や電子カルテ作成など，いくつかの分野で音声認識は一定の存在感を示している．

しかし，このように協力的な態度を強いられるユーザーにとって，音声認識の協力なライバルとなるのが，キーボードやタッチパッドである．携帯電話のテンキーでの文章入力は大変で，音声認識への期待が高いと言われて久しいが，音声入力が主流となるには至っていない．正しく押しさえすれば認識率 100%であるキーボードとの比較においては，音声認識に求められる要求性能も極めて高くなるわけである．

一方，音声認識装置の存在などとは関係なく，現実世界は様々な声や音に満ち満ちている．人と人との対面での対話音声や，電話などを通じての通話音声，動物や風，川などの自然の音，音楽や電子音など，それぞれ人間にとって有用な情報を内包している．これらの音を認識したり分類したりすることは決して簡単ではないが，もともと機械で処理することを前提としたものではないため，ユーザーの視点で見ると,「駄目で元々」と考えることもできる．たとえば，コールセンターに蓄積される大量の通話音声データの中から特定の製品名の発話頻度を分析することができれば，その値の信頼度が多少低いものであったとしても，マーケティングの重要なツールとすることが可能であろう．あるいは，窓の割れる音や人間の悲鳴などを検知することができれば，安全安心な生活を送るための大きな助けとなる．

もちろんユーザーの協力的な使用が期待できない以上，認識システムの側にとって難しい課題も多い．遠隔マイクロフォンが前提となるため，聞きたい信号は常に雑音と共に入力されると考えなければならない．人の声においては，発声の明瞭性も確保されないし，発話内容も文法的に統制されていないケースがほとんどである．しかし，耐雑音音声認識や話し言葉認識などの研究成果の積み上げにより，このような困難な条件下でも，有用な情報を取り出すことが可能になりつつある．

最後に，こうしたアプローチの研究は,「人とコンピュータの対話」という旧来のスキームにも役立つことがある．例えば,「コンピュータに向かって話した声」と「人間同士で話している声」を分別することができれば，後者による妨害の影響を減らすことができ，結果として前者の認識によるインターフェースの効率を向上させることができる．以下では，様々な種類の声や音の認識技術を通じて，音声・音響処理の新しい方向性を紹介したい．

†1 日立製作所中央研究所
Central Research Lab., Hitachi Ltd.

Vol.2009-SLP-78 No.11
2009/10/23

2. 様々な音声・音響認識技術

2.1 講演や会議音声の認識

連続文章を対象とした音声認識は，いわゆるディクテーションの自動化として研究が始められた．この場合，話者は機械による認識を前提としており，協調的な入力スタイルが期待できる．また，当初のディクテーションソフトでは，話者適応のためのエンロールメント操作を行うことが一般的であった．その後，認識を前提としない発話の認識へと興味が移り，例えばテレビのニュース番組の字幕作成自動化などが試みられた．しかし，ニュース番組のような比較的好条件の音声を対象とした場合でも，認識を前提としないことでの認識率の低下は顕著であり，それを補うためのリスピーク方式が用いられることもあった[1)]．

認識を前提としない発話の中では，ニュース番組の他に，会議や講演の音声の認識などが挙げられる．こういった分野では，話し言葉認識のための音声コーパスの整備などにより認識率が向上し，70%から 90%程度の単語認識率が得られるようになってきている[2)]．今後は，同じようなドメインでの認識率を上げる研究と並行して，同程度の認識率を維持したまま，より自然なスタイルの発話へと応用を広げていくことが期待されている．

2.2 会話ログデータからの音声検索[3)]

数千〜数万時間の音声データがタグ付けされずに蓄積されているような状況では，そこから何らかの情報を抽出することのメリットは大きく，ある程度の誤りであれば許容されうる．具体的な用途としては，コールセンターのログからトラブルを未然に防いだり，顧客の潜在要求を抽出したりといったことが挙げられる．また，コンシューマー用途では，録り貯めたテレビ番組やビデオ映像，ネット上の動画コンテンツの検索といった応用も考えられる．

音声検索の実現形態としては，対象データをすべてディクテーションにより文字化しておき，その結果に対してテキスト検索をかけるというのが最も一般的である．しかし，固有名詞や新語など，認識用辞書や言語モデルの整備が困難な言葉の検索は，ディクテーションをベースとした手法では難しい．そこで近年，音素などのサブワードをベースとした音声検索が注目を集めている[4)]．特に最近ではストレージやネット空間の大規模化により膨大なデータを対象とした検索が求められており，大量データから瞬時にキーワードを見つける技術の重要性が増している．

2.3 音響イベントの認識・分類

人間の声に限らず，様々な音響イベントの自動分類により，危険や故障の発見などに役立てることができる．特に，近年発展が著しいマイクロフォンアレイによる音源方向推定の技術と組み合わせることにより，詳細な音源マップの作成が可能となり，応用の広まりが期待されている[5)]．コーパスに基づく学習を基本とする場合，種別既知の音源の識別は比較的容易であるが，種別未知の雑音が混入する可能性がある場合でも頑健なアルゴリズムの開発が，今後の重要な課題であろう．

2.4 認識を意図した発話の検知

実世界に存在する音声・音響の認識技術の研究からは，従来型の音声インターフェースにフィードバックされるものも多い．音声インターフェースの使用を意図しない音を排除するための技術が発展すれば，余分なノイズを排除することが可能になり，結果として音声インターフェースの誤作動を減らすことができる．たとえば，音声取り込みが常時オンになっている家電向け音声インターフェースを考えた場合，ユーザーの日常会話に反応しないということは極めて重要である．音声/非音声判別・感情認識・音声認識信頼度尺度などの基準を統合することにより，このような意図的な発話の検知が可能となることが報告されている[6)]．

3. おわりに

音声認識の研究は，人と機械が対話するためのインターフェースの開発として進められてきたが，そこでの成果を活用して，実世界にもともと存在している声や音の認識という新しい応用が生まれている．人間社会から声や音というものが無くなることは考えられない以上，そこには常に新しい技術が入り込む余地がある．既に実用が見えつつあるいくつかのテーマを足がかりとして，今後も様々な応用を広げていくことが期待されている．

参考文献

1) Imai, T., et al.: Speech Recognition with a Re-speak Method for Subtitling Live Broadcasts, *Proc. ICSLP 2002*, Denver, CO, USA.
2) 秋田祐哉他: 会議録作成支援のための国会審議の音声認識システム, 情報処理学会音声言語情報処理研究会, SLP-74-21.
3) 秋葉友良他: SLP 音声ドキュメント処理ワーキンググループ活動報告, 情報処理学会音声言語情報処理研究会, SLP-74-20
4) Kanda, N., et al.: Open-Vocabulary Keyword Detection from Super-Large Scale Speech Database, *Proc. IEEE MMSP 2008*, Cairns, Australia.
5) 西浦敬信他: マイクロホンアレーを用いた HMM に基づく音源識別の評価, 電子情報通信学会技術報告, SP2000-80.
6) Obuchi, Y., et al.: Intentional Voice Command Detection for Completely Hands-Free Speech Interface, *Proc. INTERSPEECH 2009*, Brisbane, Australia.