

音声認識が人間に近づく・人間を超える

ディープラーニングと信号処理の進化がもたらす音声認識のひろがり

Avenues toward super-human speech recognizers

Advances in deep learning and signal processing that are making speech recognition leap forward



メディア情報研究部
吉岡 拓也
Takuya Yoshioka

プロフィール

NTT コミュニケーション科学基礎研究所 メディア情報研究部 研究主任。京都大学博士(情報学)。IEEE ASRU2015 Best Paper Award Honorable Mention、日本音響学会独創研究奨励賞 板倉記念等受賞。音声認識、マイクロホンアレイの研究に従事。

音声認識のカンブリア紀

私たちが話した声をコンピュータが自動的に認識する—それを可能にする音声認識技術の実用化が今、急速にすすんでいます。AppleのSiriやGoogleのGoogle Now、NTTドコモのしゃべってコンシェルに代表される音声アシスタント、各種検索エンジンに搭載されている音声検索、カーナビやゲーム端末の音声操作、官公庁における会議録作成支援。音声認識技術は今や多様な場面で利用されており、まさにカンブリア紀を迎えています。その背景には、近年の音声認識精度の劇的な改善があります。本講演では、この革新を生み出したディープラーニングと呼ばれる技術と、音声認識の応用領域をさらに拡大する上で重要となる信号処理技術について我々の取り組みを交えて紹介し、音声認識の未来を展望します。

ディープラーニングの台頭

ディープラーニングは階層構造をもつニューラルネットワークを用いた機械学習技術です。機械学習では、大量のデータ(学習データと呼ばれます)をコンピュータに与えて、そのなかから有用な規則を自動的に抽出させます。音声認識ならマイクロホンから得られる人の声の波形から文字列に変換する規則を自動的に学習します。こうして得られた規則を用いることで、音声を文字列に自動的に変換できます。

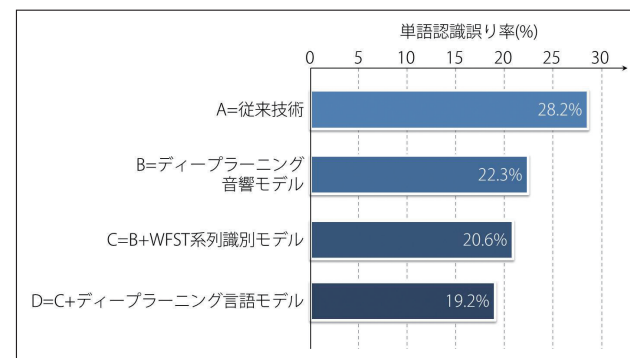


図1:ディープラーニングの発展がもたらした音声認識精度の改善(NTTでの実験結果)

ディープラーニングの音声認識への導入は2009年頃にはじまりました。2011年にはSwitchboardと呼ばれる本格的な音声認識タスクで、従来技術と比べて音声認識誤りを約3割削減できたという報告がなされました[1]。この大成功をきっかけに、多くの研究機関がディープラーニングを用いた音声認識技術の研究にいつせいに取り組み、系列識別学習、話者適応化、畳み込みニューラルネットワーク、長短期記憶ネットワーク、コネクションスト時系列分類、大規模並列学習など、重要技術が次々と発表されました。その結果、前述のSwitchboardタスクでは、かつては考えられなかった認識誤り率8%が報告され[2]、人間の認識誤り率といわれる4%に急速に近づいています。データによっては人間並みの認識精度も報告されています。

「どこでも使える音声認識」を実現する信号処理技術

音声認識の前処理である信号処理技術もまた、音声認識の世界に劇的な変化をもたらしています。現在の音声認識技術は、声の波形から文字列への変換規則を学習データから抽出するため、学習データの性質が実際の利用シーンと異なる場合はどうしても認識精度が劣化してしまいます。このことは、マイクロホンで録音された音声に不要な雑音や残響が含まれている場合に特に問題になります。信号処理技術は、あらかじめ不要な音を取り除くことで、こうした性能劣化を防ぎます。

特に、複数のマイクロホンを用いて雑音を除去するマイクロホンアレイ信号処理技術が実際に使われるようになってきています。これによって、例えばMicrosoftのKinectやAmazonのEchoでは、手元にマイクロホンがなくても声を使った操作が可能になっています。さらに、我々の研究グループが昨年参加してベストスコアを達成したCHiME3と呼ばれる音声認識評価国際イベントでは、タブレットに取り付けられた複数のマイクを組み合わせることで騒がしい環境でも音声認識精度を劇的に向上させることが示されています[3]。このような技術が進展すれば、音声認識がどこでも使えるようになり、応用範囲は飛躍的に広がるでしょう。

人間に近づく・人間を超える音声認識

音声認識技術は今後ますます発展し、近い将来に多くの

タスクで人間並み・もしくは人間以上に正しく認識できるようになるでしょう。インターネットでの買い物、天気の確認、ビデオの視聴、ロボット掃除機の動作。こういったことが室内に端末を一台置いておくだけでできるようになり、私たちの日常生活は一層便利になると期待されます。街に出れば、ヒューマノイドが案内を務めたり、音声応答システムが券売窓口やドライブスルーで動作したりしているでしょう。我々は、そのような未来を目指してこれまで以上に研究をすすめると同時に、社内外の機関と連携して研究成果を世の中にタイムリーに提供できるよう取り組みをすすめていきます。



図2: 社会に溶け込む音声認識技術

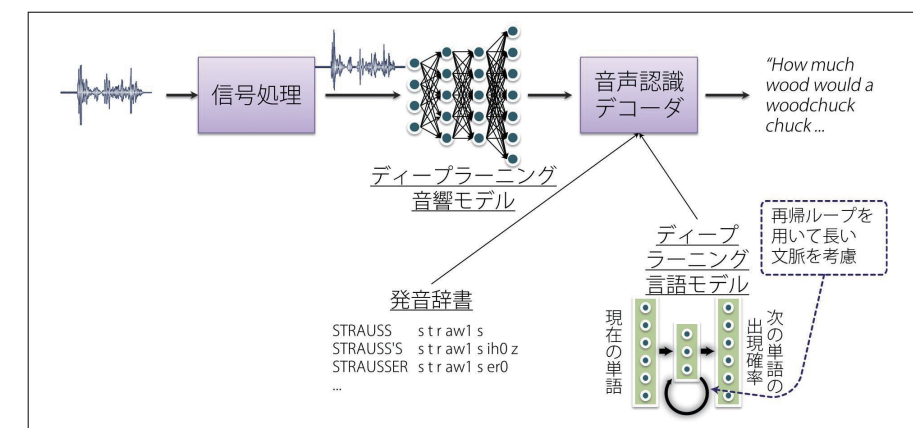


図3: 信号処理とディープラーニングを統合した音声認識システム

【関連文献】

- [1] F. Seide, G. Li, D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. 12th Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 437-440, 2012.
- [2] G. Saon, H.-K. J. Kuo, S. Rennie, M. Picheny, "The IBM 2015 English conversational telephone speech recognition system," arXiv:1505.05899, 2015.
- [3] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, T. Nakatani, "The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 436-443, 2015.