特別寄稿

すべての動画に字幕を

京都大学　情報学研究科　教授　河原（かわはら）　達也（たつや）さん

NHKの連続テレビ小説「ひよっこ」の最初の方で，「奥茨城」にいる主人公の母親が東京に出稼ぎに出ている父親の安否が気になって，隣家で電話を借りる場面があったが，茨城から東京に電話するのに申し込んでから30分くらい待たなくてはならなかったということに大変驚いた．ドラマのこのシーンは，明治でも昭和初期でもなく，先の東京五輪があった1964年を描いており，日本の高度成長の時代でも長距離電話をするのが大変だったのである．

時代は経過し，次の東京五輪を迎えようとしている今ではほとんど電話回線がIP化されて，固定電話でも携帯電話でも全国一律料金が一般的になりつつある．そもそも冒頭で紹介したような状況でも，ほとんどの方は携帯電話ですぐに連絡が取れる．

この50年間で回線容量が何万倍になったのか何億倍になったのか見当もつかないが，この膨大となった回線容量が何に使われているかというと，大半が動画配信・視聴のようである．電車やバスの中で，スマートフォンやタブレットで動画を気軽に見ることができるのは，かつて自宅のパソコンで動画を見るのが大変だったことを思うと隔世の感がある．誰でも動画を配信できるようになり，その数はすごい勢いで増えているので，視聴が増えるのも必然である．私が所属している京都大学でもOCW（OpenCourseWare）やMOOC（Massive Open Online Courses）などで多数の講演や教育コンテンツを配信しているが，同様のサービスは他でも多数提供されている．収録された動画だけでなく，学会などの講演のライブ配信（生放送）も行われている．

ところで，このようなネット上で配信されている動画に字幕がどの程度付与されているかというと，ほとんど皆無といってよい．テレビ番組では国の行政指針もあり，生放送も含めてかなりの割合で字幕が付与されるようになった．一方，ネット上のコンテンツに関しては，米国ではオバマ前大統領による「21世紀の通信と映像アクセシビリティ法（CVAA：The 21st Century Communications and Video Accessibility Act）」の制定があったが，我が国ではそういう機運もない．字幕は言うまでもなく，聴覚に障害のある方にとって不可欠なものであるが，そうでなくても電車やバスなどで視聴する際にあると大変便利である．にもかかわらず字幕が付与されないのは，ひとえにコスト・手間の問題である．

この字幕付与を効率的に行うために，我々は自動音声認識に関する研究開発を行っている［1］．音声認識は，今ではスマートフォン上の検索・翻訳・対話アプリなどで身近になってきたが，講演などの長い話し言葉を高い精度で書き起こすのは容易でない．これは外国語の旅行会話とテレビ番組の聞取りを対比すると分かりやすい．そこで音声認識システムは，大規模な講演のデータベースを用いて深層学習により構成し，さらに個別の話者への適応を行う．

前述の京都大学OCWの講演や放送大学で配信されている「オンライン授業」に上記音声認識システムを適用し，認識精度や字幕としての可用性の評価を行っている［2］［3］．京都大学OCWの講演は会場も話題もさまざまで，山中伸弥先生のiPS細胞に関する講演などは発声も収録も明瞭であるため，単語辞書をカスタマイズすることで実現できたが，普通の教室でビデオカメラ付属の遠隔マイクで収録された品質だと厳しい．一方，放送大学の講義はスタジオ収録なので，音響条件は良く，おおむね90％の認識率が得られた．ただし，このようにあらかじめ収録して配信する動画の字幕には高い正確性が求められ，人手による修正を想定しても，かなり高い音声認識精度でないと実用的でない．これまで音声認

Profile

1987年京都大学工学部情報工学科卒業．1989年同大学院修士課程修了．同大学助手・助教授を経て2003年より現職．音声情報処理，特に音声認識および対話システムに関する研究に従事．2012年度　科学技術分野の文部科学大臣表彰，ドコモ・モバイル・サイエンス賞，前島密賞などを受賞．IEEE Fellow．APSIPA理事．情報処理学会，日本音響学会，電子情報通信学会，人工知能学会，言語処理学会各会員．

識を試行した事例は多数あるが，結局人手で一からタイプ入力するのと変わらないという結果が多い．そこで今回，放送大学の多数の講義で評価したところ，音声認識精度と修正・編集に要する時間の間には明確な相関，すなわち認識精度が高いほど編集時間は短くなる傾向が確認された．また，人手で一からタイプ入力する場合と比較すると，認識率約87%以上で優位性が見られ，93%になると1/3以上の時間短縮効果が確認できた．これは逆にいうと，87%以下の精度では音声認識の優位性はない（使いモノにならない）ということである．ちなみに，我々が以前に研究開発した国会の会議録作成システムでも同様の傾向を確認している．ただし，87%の認識率の実現は容易でなく，高いレベルの技術に加えて，講師の話し方，収録環境，単語辞書の準備などすべての条件がそろってはじめて実現できるのである．

同様に，情報処理学会の研究会などで，動画配信を含めてライブで字幕を付与する実証実験も行っている．アクセシビリティ研究会（SIG-AAC：Special Interest Group Assistive & Accessible Computing）では，聴覚障害の方も多く聴講に来られるので，人手による修正・編集を行い，できるだけ品質の高い字幕を出すようにしている．そのためには前述の通り多くの条件が必要であり，どのような講演でもできるわけではない．事前に予稿を提供してもらうことが望ましいが，さらに話し方に気を使ってくれる講演者もいる．これは一般の方も聞きやすくなるので良いことである．一方，音声言語情報処理研究会（SIG-SLP：Special Interest Group Spoken Language Processing）では，聴覚障害者の方の聴講がほとんど無いので，音声認識結果をそのまま提示している．誤りは散見されるが，だいたいは分かるレベルである．

このような字幕が付与されたコンテンツを見ると，字幕がないものに比べて明らかに理解が深まる感じがする．これは，聴覚（音声）と視覚（文字）の相乗効果に加えて，日本語では表意文字の漢字を用いている効果もある．例えば，「カンサイボウ」と聞くだけより，「幹細胞」と見る方が概念をイメージしやすい．一方で，我々は毎年「聴覚障害者のための字幕付与技術シンポジウム」を開催して，聴覚障害者の方にも意見を伺うのだが，「表示のタイミングが早くて情報量も多くて良い」という意見もあれば，「五月雨式に文字がたくさん表示されても理解がついていかない」という意見も多い．字幕は単に発話を文字にすれば良いというものでなく，分かりやすいようにレイアウトする必要がある．しかしこれは，（聴覚障害者の方には申し訳ないが）その場で冗長さを含めてやりとりする音声言語とそもそも遠隔に伝えるための文字言語の本質的な違いに起因する問題である．このような点も考慮しながら，字幕がもっと普及することを願い，研究開発を行っている次第である．

長距離電話が高価だった頃は簡潔に話しただろうし，葉書や手紙では推敲して文章をやりとりしていたが，現代のようにSNSでも動画でも五月雨式に何でも配信するようになるとかえって効率が悪いのではないのかと思うのは古い世代だろうか．

文献

[1] 河原 達也：“ICT・音声認識の活用による講演・講義の字幕付与，”情報処理，Vol.56，No.6，pp.543-546，Jun. 2015.

[2] 河原 達也，秋田 祐哉，広瀬 洋子：“自動音声認識を用いた放送大学のオンライン授業に対する字幕付与，”情報処理学会研究報告，SIG-AAC-2-5，Dec. 2016.

[3] 京都大学　情報学研究科　河原研究室：“音声認識技術を用いた字幕付与支援．”
http://sap.ist.i.kyoto-u.ac.jp/jimaku/