

これは一九六二年と書いておりますので、五十五年ぐらい前ですけれども、今写っている恩師の堂下先生が、音声タイプライターを世界で初めてつくられました。その当時は、今から五十年前です。その時から、パソコンも当然ありませんし、コンピュータ自体が日本に数台しかなかった時代ですけれども、そのときから京都大学ではこういう音声認識の研究を世界に先駆けてやっていたという歴史がございます。

〔写真提示〕

体どのくらいのものができているのかを御紹介させていただきます。これは京都大学で開発しているものですけれども、多分市販のものよりもまさるとも劣らないものであると思いますので、それを御紹介させていただきます。最後に会議録作成に対しまして、どういう方法で貢献できるか、どういう方法があるかについて、ざっとこういう流れでお話いたします。

第58回全国議事記録議事運営事務研修会

音声認識技術の現状と 会議録作成への適用可能性(1)

京都大学教授 学術情報メディアセンター 河原 達也氏

- * 昨年10月18～20日、毎年恒例の日本速記協会主催の研修会が開
- * 催されました。講演、分科会討議、情報交換会など多彩なプログ
- * ラムにより議事記録や議事運営に対する理解が深まったものと思
- * われます。今回は、参加者アンケートでも関心の高かった京都大
- * 学教授・河原氏の講演を御紹介します。

ただいま紹介にあずかりました河原と申します。本日はどうもお招きいただきましてありがとうございます。

今、御紹介いただきましたように、私は音声認識の研究に約二十年携わっておりますが、この音声認識の研究というのも、随分歴史があるのです。

きょう使いますスライドの方は、お手元の資料に大体ございますが、一部差しかえもございますので、その辺はご容赦いただけたらと思います。

簡単に自己紹介をさせていただきますと、私は京都大学の学術情報メディアセンターというところにおります。大学院では情報学研究科というところになりまして、大学も昔に比べて随分組織がややこしくなっております、このようなどころになります。

研究のテーマは、音声認識。音声対話システムというのは、機械とお話ができるコンピュータ、あるいはロボットみたいなものをつくっております。

真空管かトランジスタか、パツと見てわかりませんが、何せ先生の身長からすると、今の大型コンピュータぐらいの装備で音声認識装置をつくっておられた。お金もすごくかかっていたと思います。

その当時から、こういうものができると、速記の方にかわるのではないかと、冗談とも本気ともつかないことがマスコミをにぎわしたと伺っております。

もちろん、当時はNHKのアナウンサーのような方が「わ・た・し・は」と発声すると、「わ・た・し・は」と出てくるようなイメージであります。

・用語の解説

技術的な内容ですので、基本的な用語だけ若干の説明をさせていただきます。片仮名語が多くて恐縮ですが、何「タスク」、「ドメイン」という用語が何回か出てきます。

それから、大学におるものですから、どうしても語学が非常に重要でして、語学学習というと、昔はLL教室というのがありましたけれども、最近ではIT、マルチメディアの流れで全部コンピュータ化されまして、昔のテープでやっていたLLは、今は全部CD・ROMになっております。コンピュータで音声認識を使って、学習者の発音を自動的に判定したり指導します。なかなかそこまではいきませんが、そういうことを研究しているということでもあります。

きょうのお話は四部構成になっておりまして、最初に音声認識技術の現状、概要をお話しします。それから、ちよつとだけテクニカルな原理の話もいたしまして、議会向けのこと。今の音声認識技術は、何でも使える万能なものではありませんので、どうしても議会向けにいろいろつくらないといけないところがございます。その辺の課題でありますとか現状をお話しします。途中、今、大

用語の解説

- ・タスク： 音声認識システムの目的
 - － 会議を書き起こす、カーナビを操作する
- ・ドメイン： 音声認識の対象分野
 - － 議会での話題、カーナビの機能(地名、施設・店)
- ・コーパス： 音声言語データベース
 - － 音声を集め、書き起こしたもの
- ・認識率： 単語単位で計数

(正解)	京都	の	天気	は	晴れ
(認識)	今日	殿	天気		晴れ
	S	S	C	D	C →

 2/5=40%

ここでいう「タスク」は、直訳すると「仕事」ですけれども、音声認識システムの「目的」になります。例えば会議を書き起こす、カーナビを操作するといったことを「タスク」と呼びます。

「ドメイン」といいますのは、音声認

識の対象分野でございまして、例えば議会でしたら、話題であるとか、財務委員会でありましたか財務委員会、法律の話産業の話といったことを指します。カーナビでしたら、機能。地図を案内するとかレストランに案内することを「ドメイン」といいます。話題とか分野ということだと思っています。

「コーパス」という言葉もよく出てきます。これは音声言語のデータをたくさん集めたものでありまして、具体的には音声を集める。議会の場合ですと、実際に議会で話されている音声を収集し、それを忠実に書き起こしたものです。これが音声認識システムを構築する上で重要な位置づけとなっております。

それから、「認識率何%」という言い方もよくします。こういうお話をすると、認識率はどう計算するんですかと聞かれる方もおられるのですけれども、単語単位で計算しております。日本語の場合は、単語はそれほど自明ではございませ

んけれども、大体国語辞典に載っている単語を想定していただければと思います。

この例でいきますと、「京都の天気は晴れ」が正解だとすると、下の方も、読み方によっては、「きょう都の」も音韻的には合っているのですけれども、会議録を作成する上では当然これではだめですから、エラーとカウントしています。「S」はサブステイテューションの略で、置換誤り、「D」はデリーション、削除、脱落誤り。正解の単語五つのうち三つ誤りがあるので、五分の二で四〇％という計算になります。

・語彙サイズと発声スタイルによるアプリケーションの分類

音声認識は、時々マスコミに登場しますが、どのくらいまでできているのかと素朴に思われる方も多いと思います。実際にマスコミの方も、ちよつと失礼な言い方ですが、そんなにわかって全部フ

このグラフは、この横軸が扱える単語の数、語彙サイズ、ボキャブラリーの数です。二十、二百、二千、二万とプロットしております。当然、単語の数が少ない場合は簡単になります。我々が日常的に知っている単語は大体数万と言われておりますし、ちよつと大きい辞典でも数万から十万ですので、その辺が上限と考えられています。

縦軸は、発声の仕方でありまして、例えば「音声コマンド」は、「開けゴマ」と言ったら、とびらが開くというものがあります。あるいは、「何とかさん」と発声すると、その人に電話がかかるといったものもそうです。このようなものは非常に簡単で、八〇年代、銀行のシステムで使われたのが最初の事例でした。

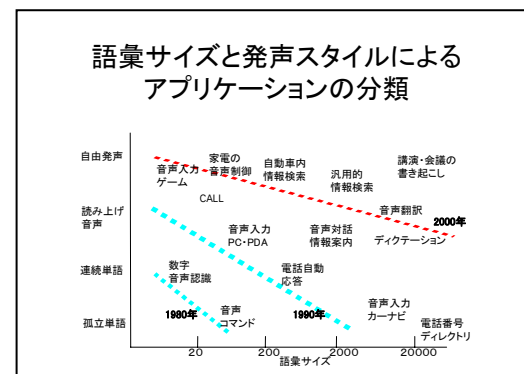
八〇年代の途中からもう少し自然な文章が言えるようになってきます。世の中に出てきましたのが九〇年代からでありまして、身近な例ですと、カーナビがあります。行き先を、「国会議事堂前」

と例えば、国会議事堂前にナビしてくれるものでありますとか、電話の応答システム。日本の場合にはそれほどありませんけれども、アメリカでは、大抵の場合、企業のカスタマーセンターとか予約センターに電話をしますと機械に応答されます。昔は、プレス1、プレス2と押していたのですが、今は大抵音声認識が入っています、「予約」の場合は「リザーベーション」と言うと、予約の窓口につながるということです。

この辺は結構世の中に入ってきておりまして、もう少し話し言葉に近いのが最近の研究開発の中心になっております。具体的には翻訳システムでありますとか、きょうお話しする議会や会議の書き起こしです。

ディクテーションというのは、皆さんにとつて身近なものだと思いますけれども、これが九〇年代の後半にIBMのVia Voiceであるとか「ドラゴンスピーチ」のシステムが商用化されまして、数

オローして報道しているというよりは、メーカーが発表するたびに、おもしろければ報道するという感じですので、本当に技術のトレンドがどこまでいつているかは、一般の方には必ずしも伝わっていないところもございします。



万の語彙で読み上げの発声を扱っております。

このグラフを見ていただいて興味深いのは、八〇年代、九〇年代、二〇〇〇年代。これは私の主観で大ざっぱに引いておるのですけれども、三つの直線が平行ではないことがおわかりいただけるかと思ひます。これはどういうことかといひますと、ボキャブラリー、語彙サイズは既に十万でもオーケーという段階にきているのですけれども、発声の方が「自由発声」と書いておりますような自然的な話し言葉に持っていくところが、技術的にはまだ難しい。

唯一、一般消費者向けに出ていると思われるのが、ビデオゲーム、テレビゲームです。最近幾つかのメーカーで音声認識が入っております。これなどは語彙は少ないですが、ゲームの場合は、間違えても許されるというものです。ちゃんと発声するところにゲーム性が出てくるという側面もありまして、本当の意味での

アプリケーションといえないかもしれないかもしれません。

先ほど申し上げましたように、会議の書き起こしは、かなり難しいところに位置づけられております。

・音声認識の話し言葉への展開

世の中の現状、一般的に音声認識のアプリケーションの紹介をざっといたしますと、先ほど申しましたディクテーションソフトが数万語彙であっても、使ったことのある方もおられると思います。が、基本的にはかなりきつちりしやべらないと、認識ができません。ちよつとサボっていいますと、とたんに認識率が悪くなります。それから、当然静かなところで話すことが前提になっております。これは技術的にはほとんど完成しておりまして、商用的にも出ているのですけれども、実際のところそんなに普及しているとはいえません。もちろん、視覚障害の方には非常に重宝していただい

ておりますけれども、一般の方がそれほど使っているということではありませんで、この商品自体も、いつときはパソコンショップに並んでおりましたけれども、今は見かけるのも難しくなっております。

その理由といたしましては、仕事に使うには疲れるとか、日本のオフィスの場合は大体相部屋になっておりますから、周囲に迷惑になるといった理由があるのではないかと思います。

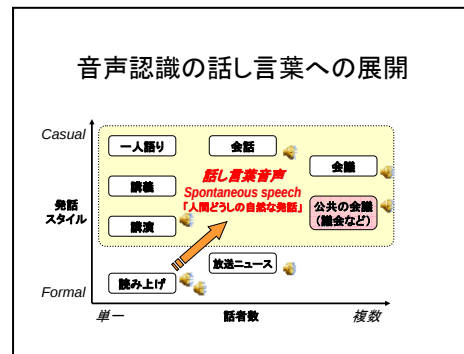
・書き起こし技術の現状

ディクテーションといえますのは、本的には言いたい文章を機械に話しかけることを前提にしています。これに対して、きょうのお話の会議とかの書き起こしでは、話す人、つまり議会で発言される方は、機械に認識されるということは直接は意識されていないと思います。その辺で難しさがありまして、ディクテーションソフトのように、しゃべるそば

から認識結果が出てきて、ちよつとサボってしゃべると調子が悪くなるなどというのがわかると、気合いを入れて話すというフィードバックがかかりますけれども、議会でありまして裁判所、あるいは今私が話しているような講演、本当の意味での話し言葉では、そういうことが難しい。この場合は、現状、万能なのは技術的には難しいということで、先ほどちよつと解説しましたけれども、タスクドメインごとにチューニング開発が必要になってきます。言うなれば、前者は既製品が可能で、後者は、現状は特注するしかないと言えらると思います。

ほかのアプリとしましては、先ほどちよつと出ましたけれども、カーナビ。音声によるコマンド入力としてカーナビでありますとか携帯端末、携帯電話にも一部入っております。こちらは発声する内容が基本的には「国会議事堂前」といった単語であります。発声する内容は簡単でも、環境が車の中でありまして、か

屋外でありますとか、かなりうるさいところで話されるという問題が大きいということ。一応商品は出ておりますけれども、まだ技術的にも完成しているとは言えないところもあります。それから、先ほどちよつとお話ししましたが、アメリカで一番音声認識技術が成功しているといわれているのは電話の応答サービスでありまして、かなり導入が進んでおります。日本では、残念ながらそれほど普及していません。日米でなぜこれだけ違いがあるのかは、私が推測するには、日本の方がサービスに対する要求が高い、あるいは先ほど申しましたけれども、オフィスが相部屋である。また、日本人は電車です。また、アメリカ人は大体車で通います。車の中では端末は扱えませんが、音声電話でしたらオーケーです。日本のように電車や公共交通機関では、話すのがなかなか難しい状況もあるのかもしれない。



今までが音声認識の世間一般の概要ですけれども、きょうの話題であります話し言葉の書き起こしに関して申し上げますと、幾つかのジャンルに分類できるかと思えます。横軸は、話者が一人か、二人か、たくさんかという分類です。縦軸は話し方のスタイルで、フォーマルかカジュアルか。きちんと話すか、気軽な雰

囲気になるかということです。まず、話し言葉の書き起こしに該当するのが講演や講義。今、私が話しているのは講演に相当しますが、かなり緊張して話している状況です。逆に、大学の講義になりますと、学生が相手ということもありますし、録音されていないということもありますので、話し方がもつと気軽になります。

それから、話者がたくさん存在する状況が会議とかミーティングでありまして、いわゆるビジネスの会議やミーティングもあります。議会の会議となると、発話スタイルはフォーマルだろうと考えられます。当然フォーマルの方がやりやすくなります。この観点からいいますと、議会の方がかなりやりやすいと言えます。

どういう研究開発が行われているかという紹介をいたしますと、まず放送です。プロのアナウンサーが原稿のある状況でしゃべっているということで、字幕

を作成するアプリがあります。この研究が九〇年代の半ばから結構行われているし、身近なところではNHKの七時のニュースなど一日三回ぐらいメインのニュースがありますけれども、それに対して音声認識技術を使って字幕がつけられております。認識率は、NHKの場合はアナウンサーの方が随分訓練されておりますので、九五%と言われています。もちろん、残りの五%は人手で瞬時に直してオンエアされています。

次が、私が今しゃべっているような講演です。これは、アナウンサーほどではありませんけれども、ある程度話した人が丁寧に発声するという状況で、講演録をつくる。きょうは人手でつくっていただいていますけれども、これを機械化することが目標になります。

こちらに関しても、二〇〇〇年代の前半に、我が国で日本語話言葉コーパスプロジェクトが行われてまして、こういう講演のデータベースを二千五百件ぐら

い集めたものがございます。こちらの方は、接話型マイクという話し手に近いマイクが前提です。接話マイクは大体二十センチ以内と考えていただけるといいのですけれども、そういう前提で八〇%ぐらいの認識率になっております。

その次に我々が取り組んでいますのが、大学の講義でありまして、特に最近ユニバーサル、機会均等ということで、聴覚障害の方に対してもきちんとノートマイクを提供しようということがどの大学でも言われております。今はもちろん人間のボランティアの方がやっておられますけれども、これは随分大変でありまして、これもできれば機械化したい。全部機械でできるわけではありませんけれども、会議録作成と同じように半自動にできないかということで取り組んでおります。

ただ、大学の講義の場合は、実は接話マイクはそれほど使いません。と申しますのは、京都大学の場合は九十分講義で

すが、三十分ぐらいすると、半分ぐらいの学生が寝てしまいますので、どうしても教員は動き回ったり、途中で学生に当たります。何かしないと、学生の集中力を維持するのがなかなか難しい。教員の集中力の問題もあるのですが、そうになると、お願いできるのはせいぜいワイヤレスのピンマイクということになります。ピンマイクとは、タイプイン型のマイクで、口とマイクの距離が三十センチぐらいになりますし、ピンマイクはスーツやシャツにこすれたりして雑音が入ったり、音をちゃんと拾えないことがあります。あとワイヤレスで飛ばすというところで難しいし、話し方も講演に比べれば随分くだけてきますので、認識率も下がる。また途上ではありますけれども、六割程度ということで、実用レベルではありません。

それから、先ほど一人語りというのがありましたけれども、アメリカではMALACHというプロジェクトがありま

す。これは第二次大戦のホロコーストの生存者のインタビューを集めたアーカイブを書き起こす。日本でいうと戦争とか原爆の語り部に相当すると思いますけれども、そういった貴重なアーカイブをデジタル化しようという話です。

次に電話の会話を書き起こす。

アメリカでは、特に9・11以降は、ナショナル・セキュリティ(安全保障)の関係から、これに関して一番お金が出ております。といっても、怪しい人の電話を全部盗聴するわけにはいきませんので、自動でスポットすることを目指す。それがまだ動いているわけではありませんけれども、そういうのを目指してやっています。電話では、普通の人が家族とか友達と話すことが多いですから、認識もかなり難しくなります。

最後が、会議、ミーティングです。

ここで言うミーティングは、一般の人がビジネスでやるようなところですけれども、議会と一番違うのは、接話マイ

クの問題です。ビジネスの会議では、八人ぐらいいて、八人ともにマイクをセッティングするという状況はなかなか難しいということ、卓上に一本マイクを置くことが多いのですけれども、非常に雑音が多くなったり、人間が聞いても聞き取りにくい場合もあって、音声認識も難しくなる。これも、接話マイクを使えば七割ぐらい行くのですが、卓上マイクでは六割でもいい方という感じになってまいります。

最後は、きょうのトピックになります。公共の議会。これは政治家の方が接話マイクを使って話している状況です。話し方は丁寧に、音響的にも一応雑音は少ないという前提であります。

これはヨーロッパで、TC・STARプロジェクトが数年前から始まっています。ヨーロッパはいっぱい国がありまして、言葉も英語、フランス語、ドイツ語といろいろありますし、EU議会は、それらの代表が全部集まってくる

で、通訳を入れなければならないということで、自動通訳、自動翻訳を目指して研究しています。

今やっているのは本会議だけでして、EUの本会議というのは、私もビデオを見ただけですけれども、ほとんど原稿をその場で読み上げているだけです。かなりやりやすいということで、九〇%ぐらいの認識率になっています。

あとでお話ししますけれども、我々の京都大学では数年前から衆議院の会議に取り組んでおりまして、現状、これは予算委員会のデータですけれども、八〇%ぐらいの認識率になっています。

それから、皆さんの方がお詳しいかと思いますが、我が国の地方議会の一部、北海道議会などで幾つか導入がなされておりまして、これは北海道議会の山崎さんの資料を写しただけですけれども、本会議で八〇%から九〇%、委員会では〇%程度と伺っています。

・ステノマスク+ボイスライティング

これは余談ですが、アメリカでは、これもきょうお見えではないですけども、兼子さんから聞いた話で、ステノマスクという装置を口につけて話すことが普及しているらしいです。ステノマスクというのは、話している原告とか被告の声をそのまま認識するのは難しいので、裁判所の速記の方が、音声認識を使って復唱します。

(cf.) ステノマスク+ボイスライティング

- ・口に覆い(stenomask)をした上で、復唱入力(リスピーク)
- NHKのスポーツ中継と同じ
- ・市販のディクテーションソフトを利用
- ・米国の裁判所で一部の速記者が利用
- ・技術的には簡便な解決
- ・一定速度で正確に復唱するために、かなりの訓練が必要



© Talk Inc.

データをたくさん集めてきて、それに合った状況にチューンして、特注でシステムをつくるのが重要になります。

これはデータ・コーパスをたくさん集める必要が生じます。さっきいろいろ紹介しましたが、議会システムでありますとか電話用システム、インタビューシステムなど、それぞれのタスク、ターゲットごとに何十時間、何百時間というデータを集めて書き起こしてから、ようやく先ほどの認識率ができています。ですから、一つの音声認識システムがあつて、全部あいうえうにできるわけではない。それらに関して一年から二年の研究投資、開発をして、ようやくそこまでできているという現状であります。先ほどの認識率は結構いいと感じられた方もいらっしゃるかもしれませんが、かなり専門の方が相当頑張つてそこまでできているという状況です。

ここであつと息抜きというかCMみたいなものですが、どんな感じ

これはNHKでも、相撲やオリンピックなど一部のスポーツ中継でなされているらしいです。相撲は、アナウンサーの方もそんなに単調にはしゃべりませんので、そのままでは音声認識は難しいので、実は別のアナウンサーが状況をより冷静に要約して発声し、それを字幕にするそうです。これも全部やればいいのですけれども、なかなか大変ですので、要望が多い番組だけやっているということ。これは、市販のVia Voiceやドラゴンのシステムでも動きます。人間の側がディクテーションソフトに認識できるように丁寧に、外国人に話しかける感じでゆつくりしゃべる。でも、裁判所とか議会でしゃべっていると周りにも迷惑ですので、マスクをする。これは技術的には話し言葉とか雑音という問題を解決しています。

ただ、これはこの方に負担を全部押しつけているわけでして、一定速度で正確に復唱するためには、ある程度要約もし

かを皆さんに体感してもらうために、例をお聞かせします。

これを日本語でやると、皆さんは日本人です。何でも聞き取れますので、英語でやります。英語の苦手な方もおられるかも知れませんが、ちよつと我慢してください。

〔サンプル1 (読み上げ音声) 視聴〕
これは英語のセンター試験や、TOEICとかTOEFLなどの検定試験に出てくる例に近いものです。これで九〇%から、いいところで九五%といわれています。

〔サンプル2 (ニュース) 視聴〕
次は、ニュースです。最初にBGMが入っています。これは、ちよつと英語ができる方や、ふだんCNNやBBC、BSとかでやっている英語のニュースを聞いて練習していると、聞き取れるようになります。機械でも、アナウンサーの方と最後の一般の方とちよつと違いますが、八五%ぐらいは認識でき

ないといけませんので、かなりの訓練が必要になります。日本でこれが入っているというのは、私は聞いたことがございません。

・音声認識技術の直面している課題

以上、いろいろ申しましたけれども、まとめますと、今、音声認識技術が直面している技術的な課題と申しますのは、大きく分けて二つあります。

一つは発声スタイルの問題。話し言葉でありますとか感情的な音声、早口、誤り、ささやき声などは難しい。もう一つは、実環境。使用環境、うるさいところは、雑音でありますとか、マイクが遠い状況になってくると難しい。こういう状況に対しては、現状では場当たり的、もう少し正確に言うと、タスクごとに解決するしかないということです。

・実用的な認識率確保の鍵

今の裏返しで、タスクドメインごとに

る。NHKのアナウンサーぐらい鍛錬された方ですと九五%という水準です。

〔サンプル3 (会議) 視聴〕

次が、普通の会議とか講演です。これはかなり難しく感じると思いますが、外国へ行つて商談するとか、外人と仕事するととなると、このぐらいの英語能力が必要になってくるといわれています。TOEICで言うと七百三十点とか八百点という水準ですが、機械でもこれで七〇%ぐらい。全部聞き取れるわけではないですが、大体何を話しているかがわかるという水準です。

〔サンプル4 (電話での会話) 視聴〕

最後が、電話の会話です。これはほとんど聞き取れないかもしれませんが、僕らも、アメリカに行つて、向こうの人間同士が、ちやくちや話しているのは、なかなか聞き取れません。それと同じようなものです。ただ、話している内容はすぐたわいもないことでありまして、我々も日常、家ではそんなに

難しいことは話しません。ボキャブラリーで言うと百語ぐらいでほとんど済んでしまうのですけれども、それでもくだけたといいますか、口語的な話し言葉は難しいということを実感していただけるかと思います。

これは機械でやっても難しくくて、六〇％。最近になってもまだ難しいというのが現状であります。

今のを見ていただくとわかりますように、この英語の認識システムは、平均的な日本人よりはるかに能力が高いし、外国語話者としても、かなりレベルが高いですけれども、まだネイティブの母国語話者のレベルまでは、到底行っていません。

それから、何回も繰り返しますけれども、万能ではないということ。つまり、それぞれのアプリケーション、タスクドメインごとにオーダーメイドしないといけないというのが、今の音声認識技術の現状ではないかと言えます。

以上が第一部です。

（以下次号）