FREE DISCUSSION

音声認識におけるrobustnessの新paradigmをさぐる

提案： 河合 剛(東京大学 工学部、goh@kawai.com)
パネリスト： 小林哲則(早稲田大学 理工学部、koba@tk.elec.waseda.ac.jp)
武田一哉(名古屋大学 工学部、takeda@nuee.nagoya-u.ac.jp)
西 宏之(NTT HI研、nishi@nttspch.hil.ntt.co.jp)

1. 討論の経緯と形態

本テーマは今回の研究会のパネルディスカッションのトピックの一案として提案したものである。これは私の直接の専門分野でないが、以前から興味を持っている問題である。提案した時点ではしかるべき専門家にパネルディスカッションを主導していただくのを期待していたのだが、研究会幹事から討論をまとめるように依頼されたので、今回の討論を自分が日ごろ抱いている疑問を言語化するチャンスにしようと思う。

今回はパネリストの方々も問題を提起なさるけれども、むしろ質問者集団として機能して、会場フロアの皆様から発言していただくという形態を検討しているので、研究会に出席される方々全員の積極的かつ気軽な意見交換をお願いしたい。

また、パネリスト以外の、今回の研究会に出席できない方々、出席が未定の方々、あるいは聴衆として参加なさりたい方々など、いろいろな方々からもメールにて事前にご意見をいただき、議論の視点が多角化するように努めたい。あいにく私の準備不足ゆえにこの原稿を書いている時点ではご意見がまだ集約できておらず、このためにお名前やご意見を予稿集に掲載できず、私は著しく恐懼している。ご意見は研究会会場にて資料として配付する。どうかご寛恕をたまわりたい。

討論の主題は次節に述べる通りである。まだまだ荒削りであるし、在来のアプローチと異なる工学的手法に私の主張をどうやったら結びつけられるのかといった批判もあろう。要点をお汲み取りいただき、robustnessを専門とされる方々におかれては、自由討論を通じてrobustnessの動向をとらえなおす機会として今回の討論に価値を見出していただければ光栄である。また、取り組んでおられるプロジェクトにrobustnessが直接関係ない方々も、同じく非専門家である私の素朴な疑問を契機に、一般教養を得る場としてご利用いただければ幸甚である。

2. 問題の背景

1997年12月号の「音響学会誌」の小特集(とくに「編集後記」)にもあったように、robustな音声認識の要望が高い。いままでのrobustnessへの戦略は、騒音や目的外話者の重畳などの雑音、マイクロホンや通信回線などの音声伝送経路、語彙やperplexityなどの言語情報などを勘案して進められてきた。それぞれの要素技術を別個に改良して、システム全体の性能を高めてrobustnessを追求してきた。要素技術の分類の仕方から分かるように、これは音声認識装置の開発・処理プロセスのモジュール分類をなぞったもので、各要素技術の総和がシステム全体の性能を定めるといういわば「分割統治的」アプローチである。

しかしrobustでない状況(つまりfragileな状況)は、必ずしも上記の分類の明瞭な組み合わせといえない場合が多い。たとえば、システムの開発者が緊張すると、さっきまで動いていたデモがなぜこわれるのか。音声対話セッション冒頭の音声は対話言語モデルの制約が厳しいために必ず認識できるはずなのに、naive userに使わせるとなぜ認識されないのか。これらの問題は音声言語処理の要素技術のどれかひとつにわりふっても解決できないように思われる。

fragilityを複数のmoduleの相互交渉や調整によって解決するのはad hocなmodelingとimplementationを行なうだけでは難しそうである。そこで、もしもmoduleごとにわけたシステムの開発手順を現実的条件として受け入れるならば、システムのcomponent technologyの間を何らかの方法で取り持つ枠組、すなわちthroughputとしてrobustnessをもたらすようなmeta componentが必要になる。このようなmeta componentの問題を、applicationに依存したuser interfaceやimplementationの個別的な課題とかたづけるのは間違いだと思う。meta componentの性質の決定はたしかにapplicationの要求仕様によって決定される問題だが、システム全体のrobustnessを論ずる場合はこのようなmeta componentの良否が論ぜられてもいいはずだ。いままでfine-tuningやknow-howとみなされてきた知見を、もうそろそろformalizeできる段階に来ているのではないか?いままでに培ったknow-howをもしもまとめあげられるならば、robustnessの追求の方法論のひとつになりはしないか?

また、システムの開発手順を無視して、fragilityを引き起こす原因のみに注目すれば、fragilityの類型を考えることを通じて、fragilityの原因とその思い切った扱い方を見出せるかもしれない。たとえば、ユーザの行動形態を「緊張・不慣れ」「怒り・不満・いらだち」「迷い・優柔不断」などと分類して、それぞれの行動様式が持つ性質を音声言語システムの観点から列挙するといった方法である。

「このようなアプローチはhuman factorsの研究であってrobustnessの研究ではない」と判断するのは容易だけれども、今日の技術はnoise robustnessなどを扱うだけではなくてuser robustnessをも対象としうる水準に達しつつあるのではなかろうか。少なくとも、ユーザの行動形態のtypologyを通じて、robustnessが要求される要素技術の一部を発掘できるだろう。いままでは、たとえば白色雑音への対応のように問題点をひとつの要素技術に限定してrobustnessを求めてきたし、こうするのは研究の橋頭堡を築くうえで最善策でもあった。これからは現実的な条件によりあてはまる研究開発が望まれるはずで、そのためにも現実的に役立つ観点からfragilityの原因を系統的に求めるのがいいかもしれない。

具体例をあげて説明すると、たとえば「優柔不断」なユーザがいて、そのために(1)ユーザがなかなかしゃべらない、(2)しゃべってもfalse startが多かったり、意見・希望が発話途中に変わる、(3)システムから情報をどのように取り出すかを決められず、受身に待っている、(4)システムからの提案を採用するか却下するかをすぐさま判断できない、などの行動をしめすとしよう。システム開発者にとってみれば思い通りにしゃべってくれない困ったユーザであるが、現実のお客様とはこうしたものである。そこで「優柔不断」「あきっぽい」「わがまま」などの切口から非協力的なユーザのもつ行動形態を分類し、fragilityの要因を探りなおし、現況の音声対話システムの改良を考えてはどうだろう。たとえば、ユーザがしゃべらないと音声区間の検出がむつかしくなったり、雑音を音声だと勘違いしたりする。となると音声区間の検出そのものが正しいやり方かどうかを問いなおす必要が出てくる。音声区間を検出して音声をまるごと認識すること自体の可否を問うべきだとの結論に至れば、次時代の音声認識技術はwhole-utterance recognitionではなくverbal intent detectionが適しているかもしれない。このあたりにrobustnessへの糸口がひそんでいるように私は感じている。

以上、日ごろ感じている疑問をなるべく具体的に述べた。皆様の忌憚のないご意見をお願いする。