Over the AI ―― AIの向こう側に（14）：

開き直る人工知能 ～「完璧さ」を捨てた故に進歩した稀有な技術

http://eetimes.jp/ee/articles/1708/29/news024.html

音声認識技術に対して、長らく憎悪にも近い感情を抱いていた筆者ですが、最近の音声認識技術の進歩には目を見張るものがあります。当初は、とても使いものにはならなかったこの技術は、なぜそこまでの発展を遂げられたのか――。そこには、「音声なんぞ完璧に聞き取れるわけない！」という、ある種の“開き直り”があったのではないでしょうか。

2017年08月29日 11時30分 更新

[江端智一, EE Times Japan]

今、ちまたをにぎわせているAI（人工知能）。しかしAIは、特に新しい話題ではなく、何十年も前から隆盛と衰退を繰り返してきたテーマなのです。にもかかわらず、その実態は曖昧なまま……。本連載では、AIの栄枯盛衰を見てきた著者が、AIについてたっぷりと検証していきます。果たして”AIの彼方（かなた）”には、中堅主任研究員が夢見るような”知能”があるのでしょうか――。⇒連載バックナンバー

音声認識技術 vs. 江端（音声案内編）

　私は、基本的に言語以外のコミュニケーション（身振り、手振り）で、多くの国で何とかしてきました（関連記事：「TOEICを斬る（後編） ～“TOPIC”のススメ～」）。

　もちろん、こんなやり方はビジネスでは通用しませんが、旅行程度であれば、世界中どこでもなんとかしてみせる ―― という、漠然とした自信が、私にはあります。ですので、特に英語によるコミュニケーションについては、私は「語彙数」を重視して、「発音」など本当にどうでもよいと思っていました。

　しかし、最近、私は、英語のコミュニケーションの「相手」を見誤っていたのではないか、と思っています。

それは、

―― 英語のコミュニケーションの相手は、「人間」ではなく「機械」である

ということです。

米国赴任中、電話で飛行機の予約をしようとしたとき、英語をしゃべる自動電話応答システムが、私に次々と質問をしてきました。

『ご希望のボタンを、次の1〜5の中から選んでください』と、まあ、この程度はよいのですが、『どの空港から、どの空港へ、何日の何時の飛行機に乗りたいのかをおっしゃってください』に対して、私は、手書きのメモに目を落としつつ、できるかぎりゆっくり、かつ、できうる限り正確な発音を心掛けて、自動電話応答システムに語りかけました。

ところが、その自動音声システムは、いつも私にこう言ってきました。

『申し訳ありません。聞き取れませんでした。もう一度おっしゃってください（I am sorry, but I cannot understand what you said now. Would you please try it again?）』

しかも、同じ質問を3回繰り返された揚げ句、"Sorry, See you again." と言われて、一方的に電話を切られたのです。そして、そこには、電話の受話器を握り締めながら、ぼうぜんと立ちすくむ私がいました。

私は、以前、コラム「エンジニアが英語を放棄できない「重大で深刻な事情」」で、「日本の製造メーカーですら、日本語のマニュアルを作らなくなった」というお話をしましたが、このようなトレンドが、自動電話応答システムにまで拡張しないと、誰が断定できるでしょうか。

コールセンターの無人化（自動電話応答システムの導入）は、経費削減の観点から、今後も進み続けます。これは、システムのエンジニアとして断言できます。そして、その自動電話応答システムの「多言語対応」は、絶対に期待できません。多言語対応は膨大なコストが掛かるからです。どんな会社だって、対応言語を「1つ」にしたいと思うのは自然なことです。

今後、生命保険、医療、その他の分野のサービスが日本への流入してくることは、時代の流れです。TPP（環太平洋パートナーシップ協定）の主要な加盟国であった某大国の大統領は「永久に離脱する」と明記した大統領令に署名しましたが、彼の就任期間が終わった翌日からTPP、あるいはTPPと実質的に同じ内容の協定交渉が再開するだろうと、私は踏んでいます（参考記事（外部媒体に移行します））。

私たちが今考えなければならない問題は、今後、日本で展開されていく海外のサービス形態と、その海外のサービスに、既存の国内のサービスは勝てるか否か、ということです。

そして、もし海外のサービスが勝者となる場合には、（1）膨大な経費をかけても「日本語」に対応する企業が勝つか、（2）「日本語」には対応しないが安価なサービスを提供できる企業が勝つかですが、現時点では、私には判断できません。

音声認識技術 vs. 江端（スマホのナビ編）

私には、『スマホ（スマートフォン）なんぞは、しょせん、携帯するパソコンにすぎない』という思い込みがあり、徹底してガラケー（携帯電話機）にこだわってきましたが、ある日のこと、とある理由（著者のブログ）から、月額500円のデータ通信専用スマホと、音声通信用のガラケーの

併用を始めました。

　そして、今年（2017年）のゴールデンウイーク、親戚の結婚式に出席するため、スマホの「Googleマップ」のカーナビ機能を使って結婚式場に向かったとき、本気の本気でビックリしました。


画像はイメージです

　超低速、月額500円のデータ通信専用スマホで、どうして、これだけ自然なリルートやバードビュー表示、自動地図拡大や縮小ができるのか ―― 今でも魔法を見ていたかのような気がします。

　ですが、それ以上に私が驚愕（きょうがく）したのは、Googleの音声認識機能でした。後述しますが、私は1999年に、ある事件に遭遇して以来、「音声認識」という技術を、全く信じなくなっていたからです。

　しかし、今回スマホに、その「結婚式場」の名称は当然として、母の介護老人保健施設の名称「老健 スギタ（ロウケン スギタ）」と、スマホに語りかけたとき、端末には「介護老人保健施設 スギタ」の名称が表示され、スマホの地図上にその場所がドンピシャで示されていました。

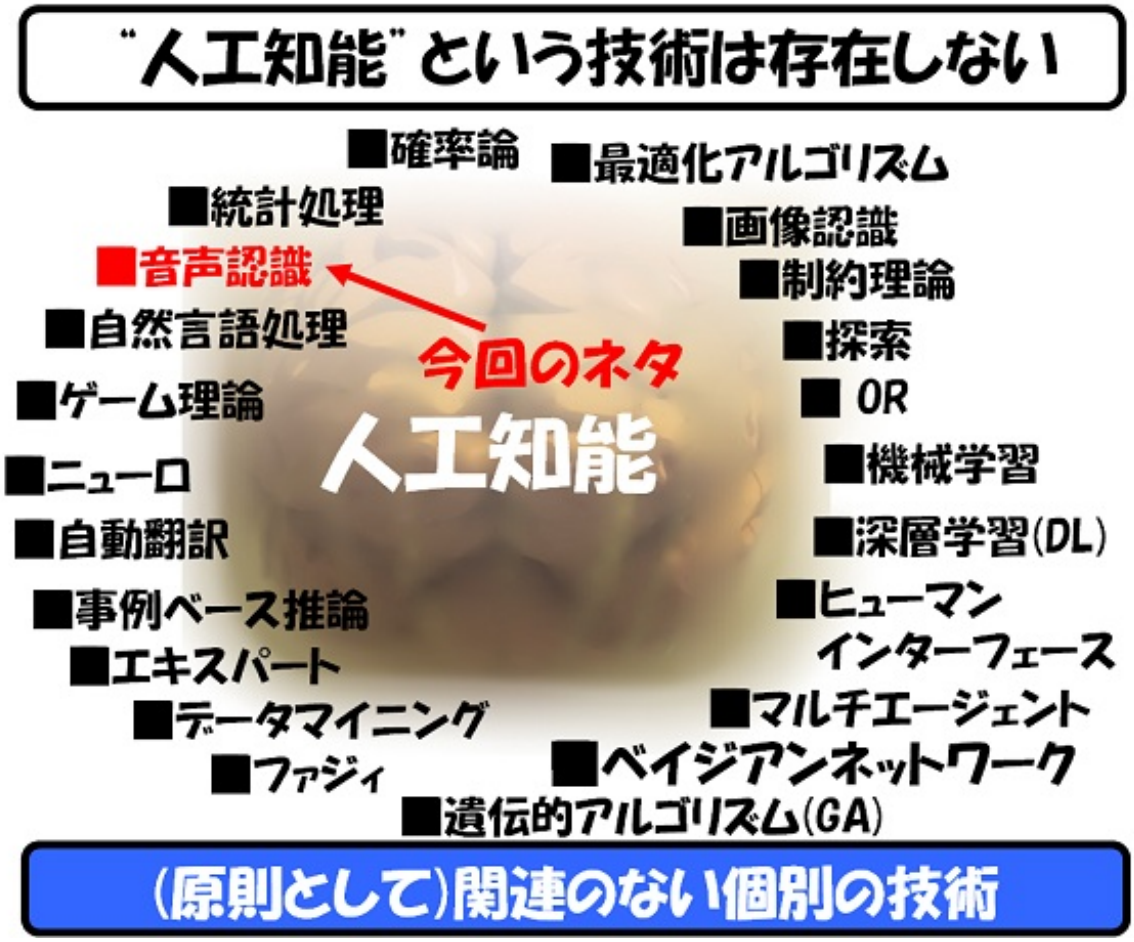
　その時の私の反応は、「びっくりした」などというレベルではなく、「腰が抜けた」といっても良いものでした。1999年の段階であれば、ロウケン スギタ」は「老犬過ぎた」と表示されていたはずです。

　―― 私が、「音声認識技術は使いものにならない」と決めつけた後の世界に、一体何があったのだ？

　では、今回の「Over the AI ―― AIの向こう側に」、始めていきたいと思います。

何かが「変」な、音声認識技術のトレンド

　こんにちは、江端智一です。今回は、「音声認識技術」についてお話したいと思います。



定番のフレーズなので、もう飽き飽きしている人も多いと思いますが、まあ、これも私なりの「けじめ」として今回も申し上げます。

「音声認識技術が“人工知能”なのかどうか」の判定については、『江端AIドクトリン』に基づき、「江端が『AI』と思ったものであれば、誰がどう反論しようが、それはAIである」を押し通します*)。

*) 正直なところ、"音声認識技術"を、"AI技術"と言い張ることには、さすがの私も抵抗があるのですが、"AI技術"に必須の"インタフェース技術"ということで押し通します。

さて、冒頭で紹介したように、私が「腰が抜ける」ほど驚いた「音声認識技術」のパラダイムシフトが、いつごろ発生したのかを調べるために、今回も定番の"Googleトレンド"を使って、そのトレンドを分析しました。

でも、何か変なんです。

「音声認識」を見出しとする記事は、どんどん減っています。グラフに表示されている期間中、音声認識に対する大きな技術革新あった事実は読み取れません（2012年に、ちょっとしたトリガが観測できますが）。

次に、「音声認識」に関する歴史をご覧いただきたいと思います。

SFの世界においては、対話できるコンピュータとして2001年宇宙の旅の「HAL9000」、対話できる自動車「ナイトライダー」などがありました。しかし、その当時の音声認識技術は、まだ実用化には程遠く、研究室での研究レベルにとどまっていました。

実用化が始まったのは1990年あたりで、音声認識が製品として適用された最初のモノは、カーナビゲーションシステム(カーナビ)だったと思います。

もちろん、カーナビの音声による操作は、安全運転に貢献するという面もあったでしょうが、音声認識技術をハードウェアに実装して、研究開発費を回収できるような製品は、カーナビくらいしかなかったと思います(例:自宅のドアの鍵を「開けドア」のようにすることもできたかもしれませんが、鍵を使った方が安全な上に安価です。パソコンのパスワード入力についても同様のこ

とが言えます）。

　要するに、音声認識技術を社会インフラとして使うニーズ、あるいは、音声認識技術の研究開発コストを回収できる対象製品がなかったのです。

“福音”のような技術だったはずなのに……

　音声認識は、その後、ソフトの世界に移行し、エンタメ系（ゲーム）で適用されたこともありましたが、これも、次のバージョンに引き継がれることなく消滅していきました。何度言い直しても、ゲームキャラクターが、言葉を全然理解してくれないゲームなんか、楽しいわけがありません。

　そして、音声認識技術に対する不信感を決定づけた製品が、1999年に発売された、パソコン用音声認識ソフトウェア「ViaVoice」でした。

　これは日本国内でも、大々的にテレビCMが放映されました。男性アイドルグループのメンバーの1人が、音声だけで電子メールの文面を作成し、音声で、そのままメール送信をするというCM*）でしたが、"Windows95"のブーム後、パソコンのキーボード入力が不得意な人にとっては、「ViaVoice」は福音のようなソフトウェアとして映ったものでした。

*）元SMAPの香取慎吾さんが、ヘッドセットで『先日いただいたべったら漬け、あなたの温もりがこもっていた、あなたの愛がこもっていた、まる』と言うだけで、メールの文章が作成される、という内容のCMでした。

　しかし、その希望は、絶望という最悪の形で終息します。「ViaVoice」は、とても実用レベルで使えるようなものではなかったのです。私も試したのですが、認識精度は極めて悪く、私の記憶している限り、文章の体を成すフレーズは、ただの1回も作ることができませんでした。

　おまけに、音声認識の結果を、ディスプレイで確認して、キーボードで修正しなければならない、と ―― まさに、本末転倒を絵に書いたような製品でした。

　ここに、

―― 音声認識は使えない

という認識が世界中で共有され、「音声認識技術、冬の時代」が始まるのです。

　そして、その後、冒頭の、米国での不愉快な自動電話応答システムの体験（2000～2002年）を経て、私の「音声認識技術」に対する不信感は、憎悪にまで発展します。

　私は、この「音声認識技術、冬の時代」を終わらせた（のか？）Appleの「Siri」については全く知りませんでした。"Siri"のつづりを"Silly（浅はかな）"と思い込んでいたくらいです*）。

*）憎悪もここまで徹底すれば、ある種の美学だと思います。

「音声認識技術」と「江端」、歴史的和解の瞬間

私が「変だな」と感じるようになったのは、2年くらい前です。次女が、スマホに変な呪文『"ハイ! シリー"』と唱えていた時です ―― が、私は、あまり気にしていませんでした。ティーンエージャーの奇行をいちいち気にしていては、保護者はやっていけません。

また、嫁さんが、スマホのGoogleの検索エンジンに対して、検索キーワードをしゃべり始めた時には、「まあ、音声のいくつかでも拾えれば、全部入力するよりはラクだろうしね」くらいにしか考えていませんでした。

しかしながら ―― 皆さん、ついに、今年(2017年)のゴールデンウイークに「その時」がやってまいります(NHK歴史情報番組「その時歴史が動いた」の松平定知アナウンサー風に)。

史実(江端のブログ)には、江端が自ら発したフレーズ「ロウケン スギタ」が、「老犬過ぎた」ではなく、「介護老人保健施設スギタ」と表示されたのを見て、愕然(がくぜん)としている様子が、明確に記載されております。

実に20年近い月日を経て、「音声認識技術」と「江端」は、歴史的和解に至ったのです。

「音声」とは、最低最悪の情報伝達信号である

ここからは後半になります。この連載の後半は、「私の身の回りの出来事」を使った、「数式ゼロ」のAI解説になります。

本日は、前半に引き続き「音声認識技術」についての概要をお話したいと思います。

まず、音声認識技術は、音声合成技術とは違います。

音声合成技術とは、「声を作る」技術です。具体的には、自分の作った楽曲を、パソコンに歌わせたり、原稿を読み上げさせたりするものです。ボーカロイド「初音ミク」(参考記事(外部媒体に移行します))や、ボイスロイド「結月ゆかり」(参考記事(外部媒体に移行します))などが該当します。

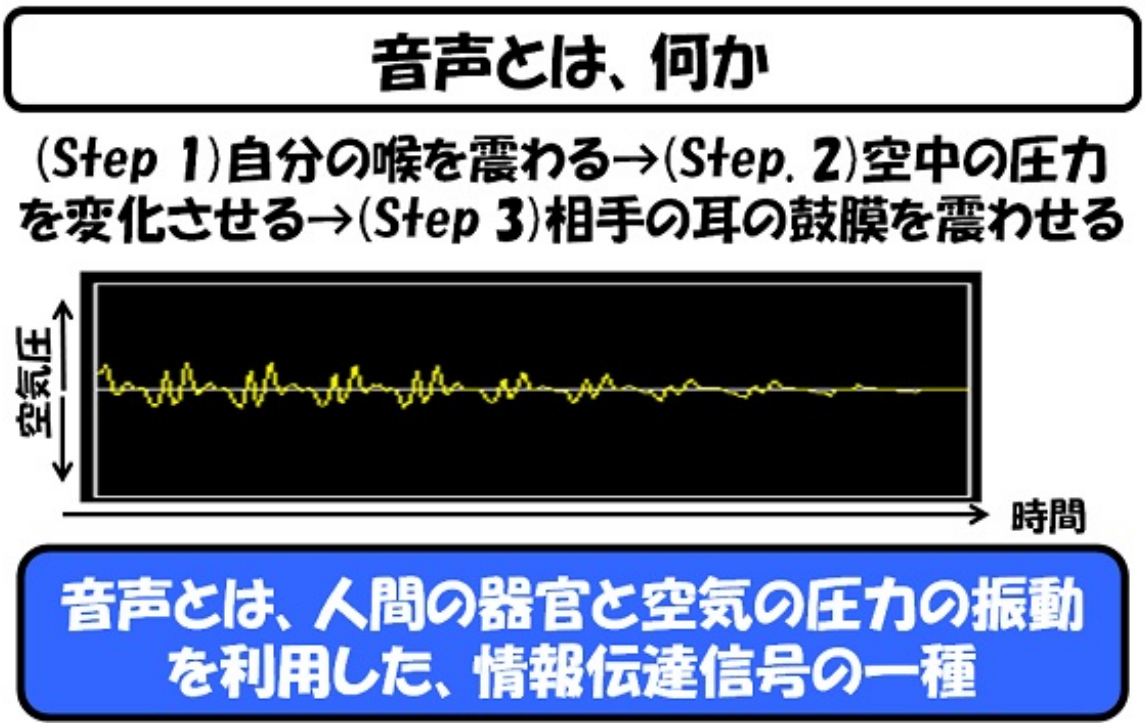
音声認識技術は、その逆です。

私たち人間が喋っている音声を、文字に変換する技術です。これは、「声を作る」よりもはるかに難しいです。しかも、特定の人物の音声だけでなく、不特定多数の人物に対応しなければならないので、そのハードルはさらに高くなります(まあ、1950年から研究され続けて、70年の時を経て、ようやく最近、検索エンジンなどで使えそうなものになった、という点から見ても、その難しさは明らかです)。

まず、音声認識技術の前に、「音声」について簡単に説明します。

「音声」とは、情報量が少なく、エラー率は高く、おまけに効率の悪い、最低最悪の情報伝達

信号です。



さらに、情報の送受信システムとしても、最悪と断言できます。

信号送信装置である「のど」は、呼吸器官と連動させながら、のど（声帯）を振動させなければなりません。その振動回数は、1秒間に数百回程度にもなり、その振動をミリ秒単位で変化させ続ける必要があるのです。

信号伝達媒体である「空気」は、情報の伝達方向をコントロールすることができません（セキュリティ性ゼロ）、そして、その伝達距離はせいぜい数メートル程度、20mを超えるのは難しいでしょう（例：2人の人間が、25mプールの両端に立って会話することを考えれば、明らかです）。

信号受信装置である「鼓膜」は、空気の微小の圧力変化を感知しなければならないので、受信エラーが頻繁に発生します。また、「鼓膜」には、チューナー（選択受信機能）が内蔵されていないので、特定の話者を選択して受信することはできません。どんな音声であれ受信してしまうので、情報処理のやりにくいこと、この上もありません。

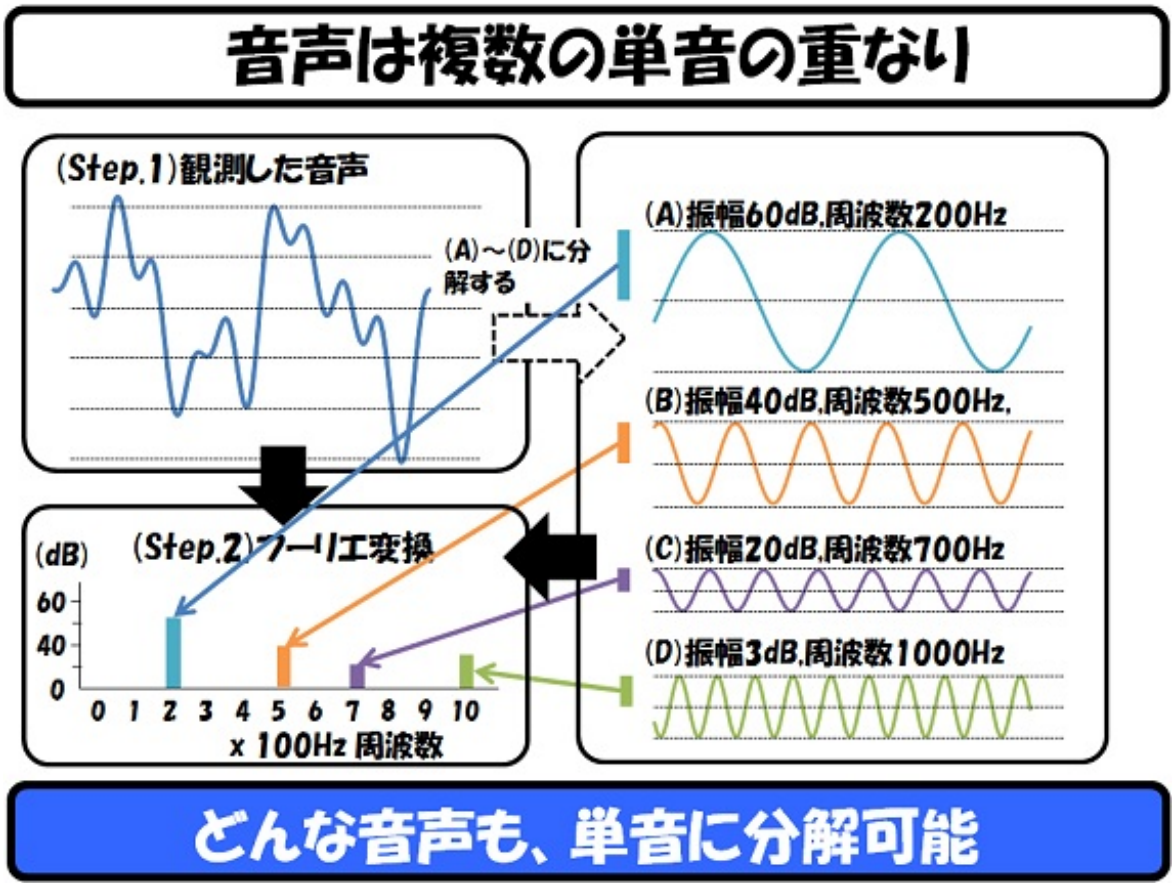
さらに、1時間の深夜ラジオを、1秒に圧縮して伝達し、後で、脳の中で解凍、再生するなどの技も使えません ―― はっきりいって不便です。

これだけでも、「音声」が、取り扱いが難しい情報伝達信号であることが分かります。

それでも、音声から話者を特定できるのは、人間の音声には、それぞれバラバラの特徴があるからです。これを「声紋」といいます（潜水艦のモーター音から、潜水艦を特定する場合には「音紋」と言われますが、基本的に同じモノです）。

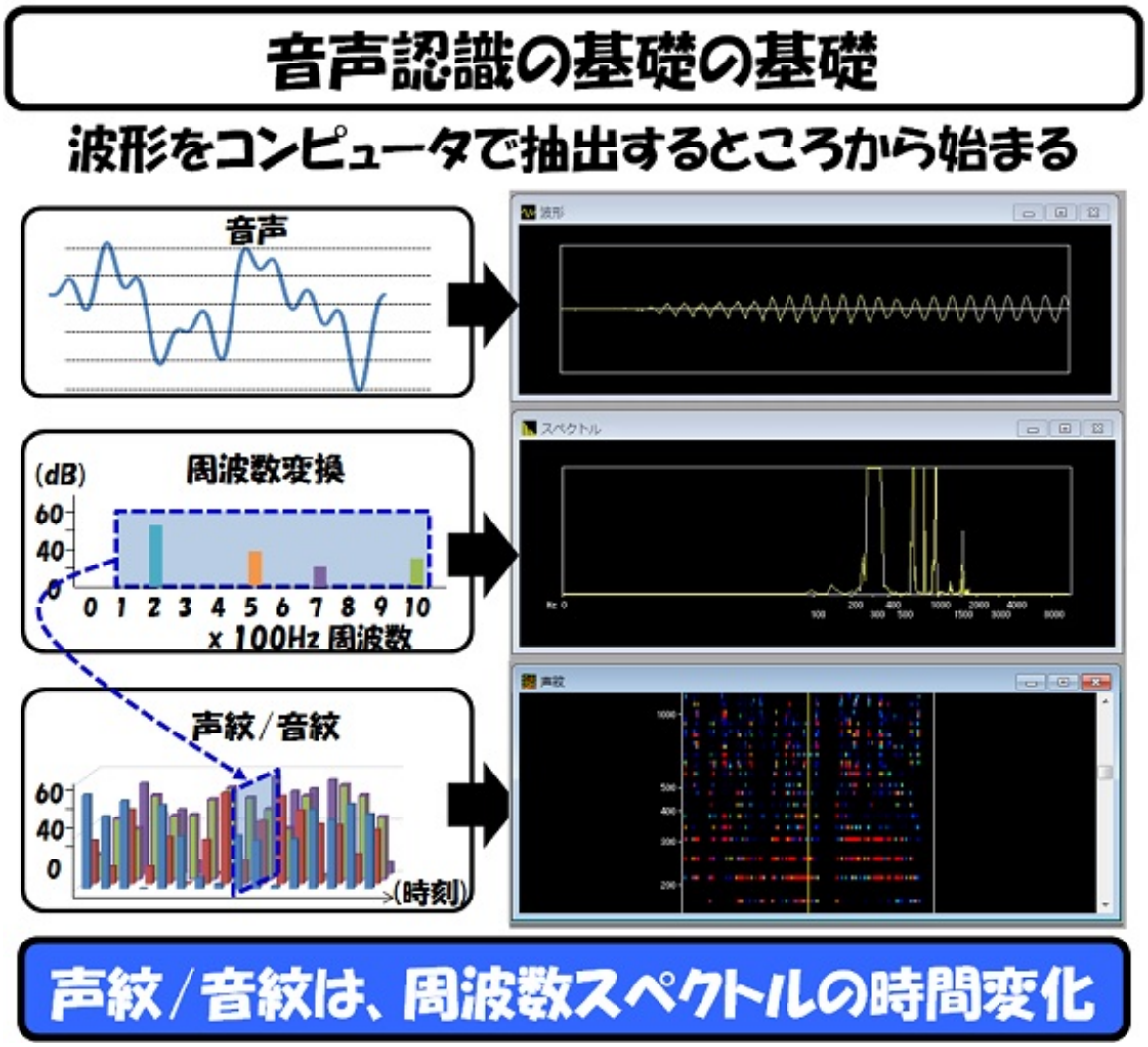
音声というのは、基本的に複数の基本周波数の音（音声スペクトル）の集合でできています*）。

*）正確には「どんな音声であれ、基本周波数の音に分解することができる」が正しいです。



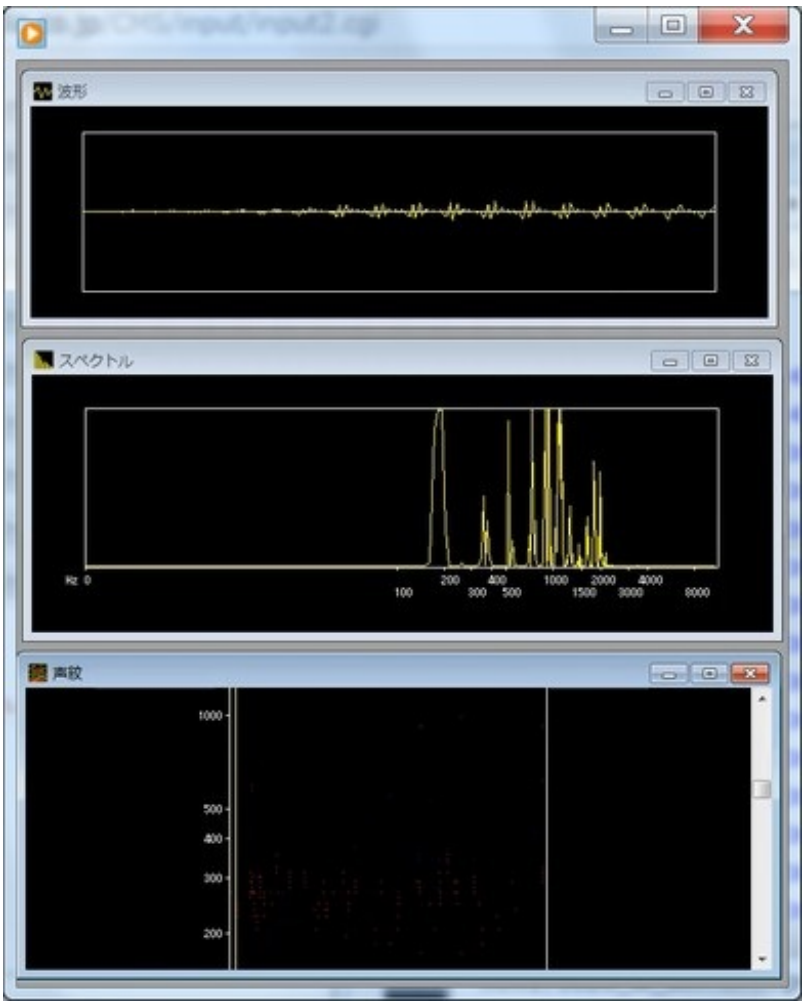
同じ曲をピアノで演奏してもらって演奏者を特定することと、同じ歌を歌ってもらって歌手を特定することの、どちらが簡単かは言うまでもありません。私たちは生まれながらにして、“世界でたった1つの異なる楽器”を持っているからです。

下図は、音声を声紋画像に変換したり、スペクトルを表示したりできるフリーソフトウェア「声紋」を使って、ボイスロイド「結月ゆかり」で作った声を、解析している様子を示しています。



音節単位（例：「エ」とか「バ」とか「タ」）の音声波形（上図）が、周波数スペクトルに変換され（中図）、さらにそれを時間軸方向に並べてられたもの（下図）があります。この一番下の図を「声紋（または音紋）」と言います。

実際に、このソフトウェアが、音声から「声紋」を作成している様子を動画でご覧いただきます。



「声紋」を作成している様子（クリックすると音声が流れます）

　このように、音声というのは、（1）周波数単位に分解できる音の波（音波）の集合体であって（2）人間ごとに、その音声の性質は全く異なるものになる（別々の楽器から出てくる音のようなもの）、ということを覚えておいてください。

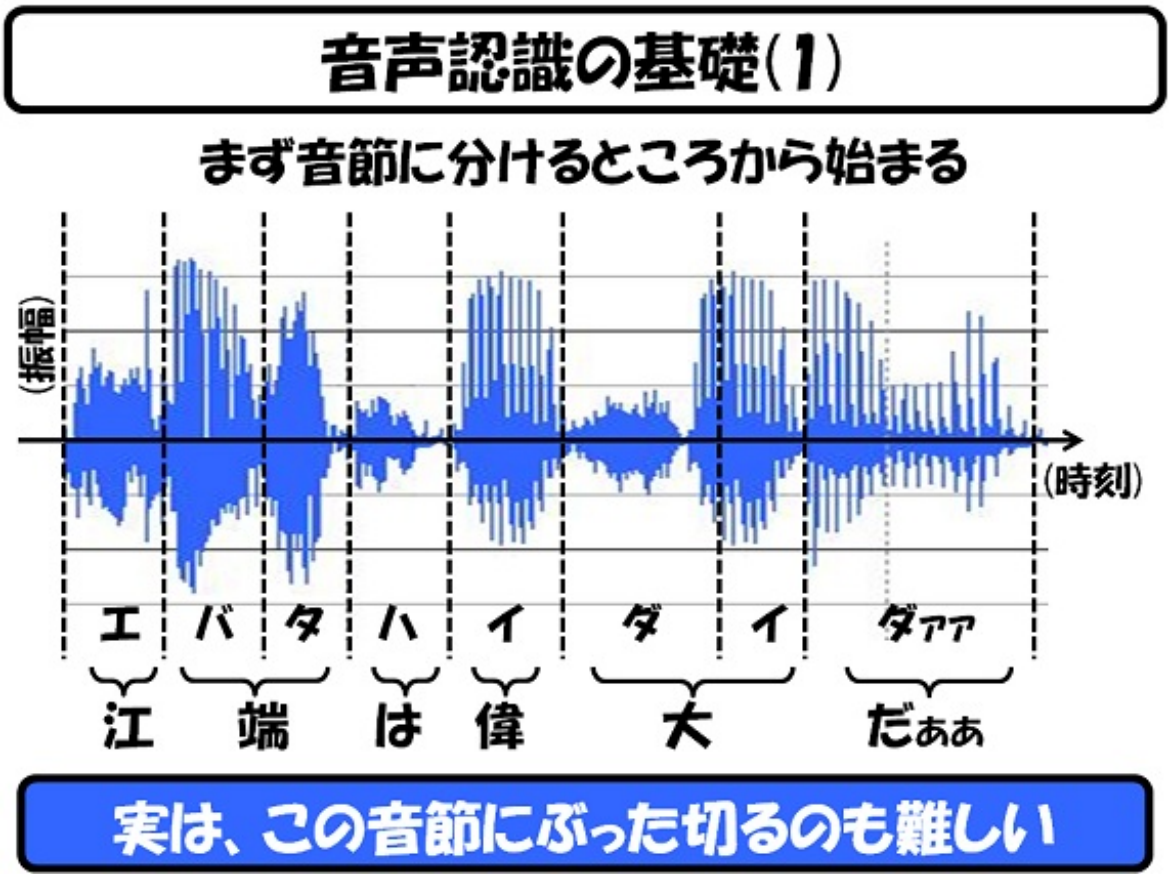
　そして「音声認識技術」とは、いわば、別々の楽器から出てくるどんな音であっても、そこから、文字に変換しなければならないという ―― 人間にとっては朝飯前*）のことであっても ―― コンピュータにとっては、とてつもなく難しい技術なのです。

*）英語などの外国語は聞き取れないから、私には、朝飯前でもないですが。

音声認識技術の基礎

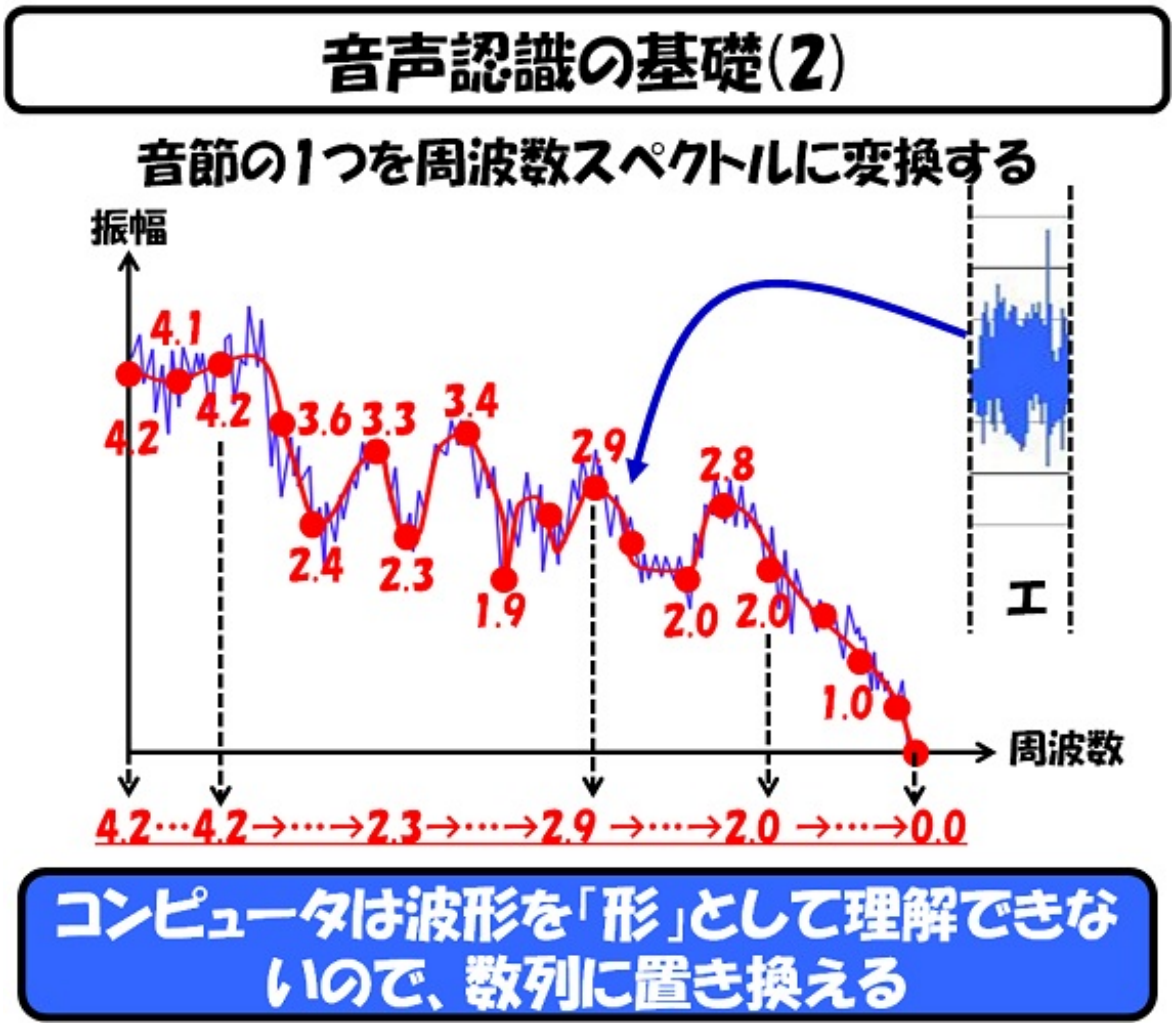
　さて、ここからは、音声認識技術の1つとして、メル周波数ケプストラム係数（MFCC）を利用した方式を、恐しく簡単に説明します（なお、MFCCの名前は忘れていただいて結構です）。

　最初に、どんな会話であろうとも、まずは「音節」としてぶった切らなければ、話は始まりません。どんな言語で会話しようとも、結局のところは音節の集合体だからです。



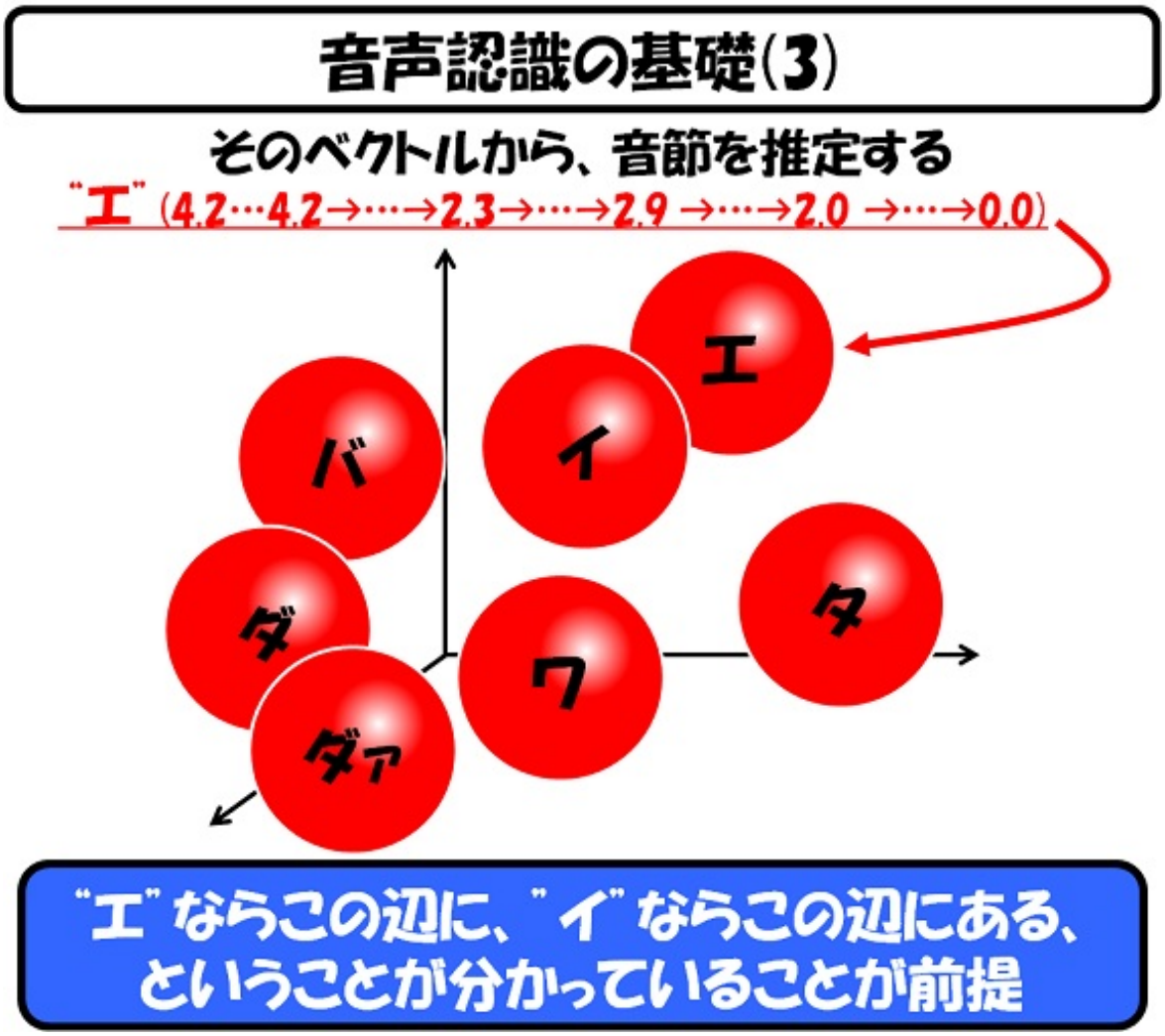
ところが、この音節にぶった切るのが、結構難しいのです。上記の「ダァァ」が「ダァァァァァァァ」と長くなった場合に、コンピュータは、どこで音節が切れるのかを判断することができません。この問題に対処する技術の1つにHMM（Hidden Markov Model（隠れマルコフモデル））というものがあるのですが、今回は割愛します（面倒だから）。

次に、この1つの音節を取り出して、この音節の周波数スペクトルを読み取ります。



先ほど申し上げた通り、音声は単音の集合体ですが、単音の周波数は連続値（×離散値）をとなるので、結局のところ、こんな形になってしまいます。そして、コンピュータは波形を、波形のまま理解することができないので、適当な周波数の間隔の値を拾って覚えておきます。

上記の例で、12個の値を拾う場合には、12次元のベクトル値として記憶しておきます。



さて、この12次元のベクトルを、12次元上の過去のデータと照合します。上記は3次元のイメージで記載されていますが、実際は12次元あるものとして考えてください。

そのベクトルが"エ"の球の中に含まれているなら、「多分、その音節は"エ"に違いない」という風に、一つ一つ音節を決定していきます。

当然の事ながら、このような球を作るためには、膨大な"エ"の発音を、多くの人に発音してもらって、データベース化しなければ、何も始めることができません。「音声認識技術」では、このようなサンプルを集める地道で苦痛な作業を避けて通れないのです。

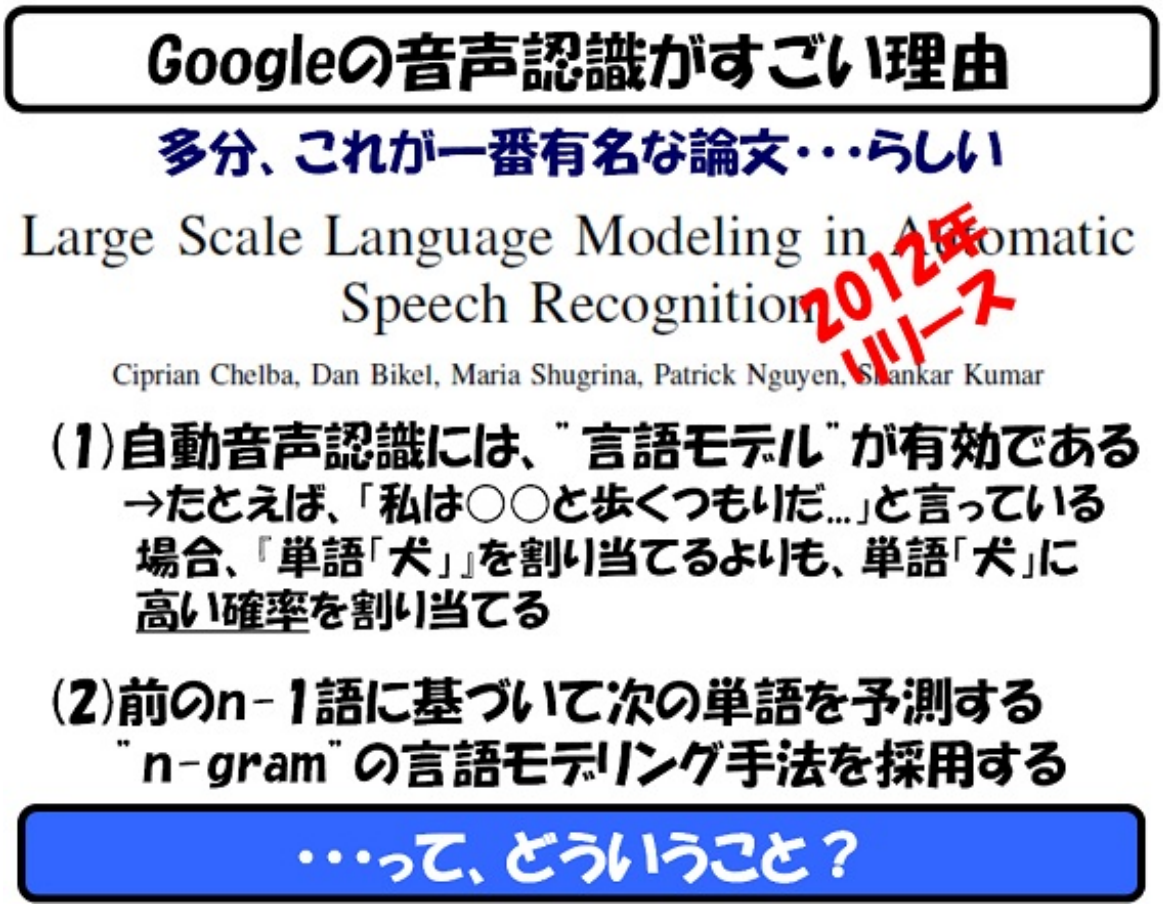
しかも、その"エ"が、その球の中に必ず入るという保証もありません。そのような場合、この方法による音節の抽出は、100%失敗します。

このように、「音声認識技術」は、恐しく面倒で地道な努力を続ける割に、その努力に比例してその認識率が向上する訳でもないという、研究者やエンジニアにとっては、苦痛に満ちた研究開発が必要となっていたのです。

Googleが音声認識技術を開発できた理由

では、一体どうやって、Googleは、私に「腰が抜けた」と言わしめる程の、音声認識技術を開発できたのか？

今回、いろいろ調べてみたのですが、どうも、これについて丁寧な解説をしている日本語の文章を見つけられませんでした。仕方がないので、今回ばかりは江端方式（関連記事：論文や特許明細書の英語は“読まない”で“推測する”）を使って、Googleの出している論文を一通りざーっと読んでみました。

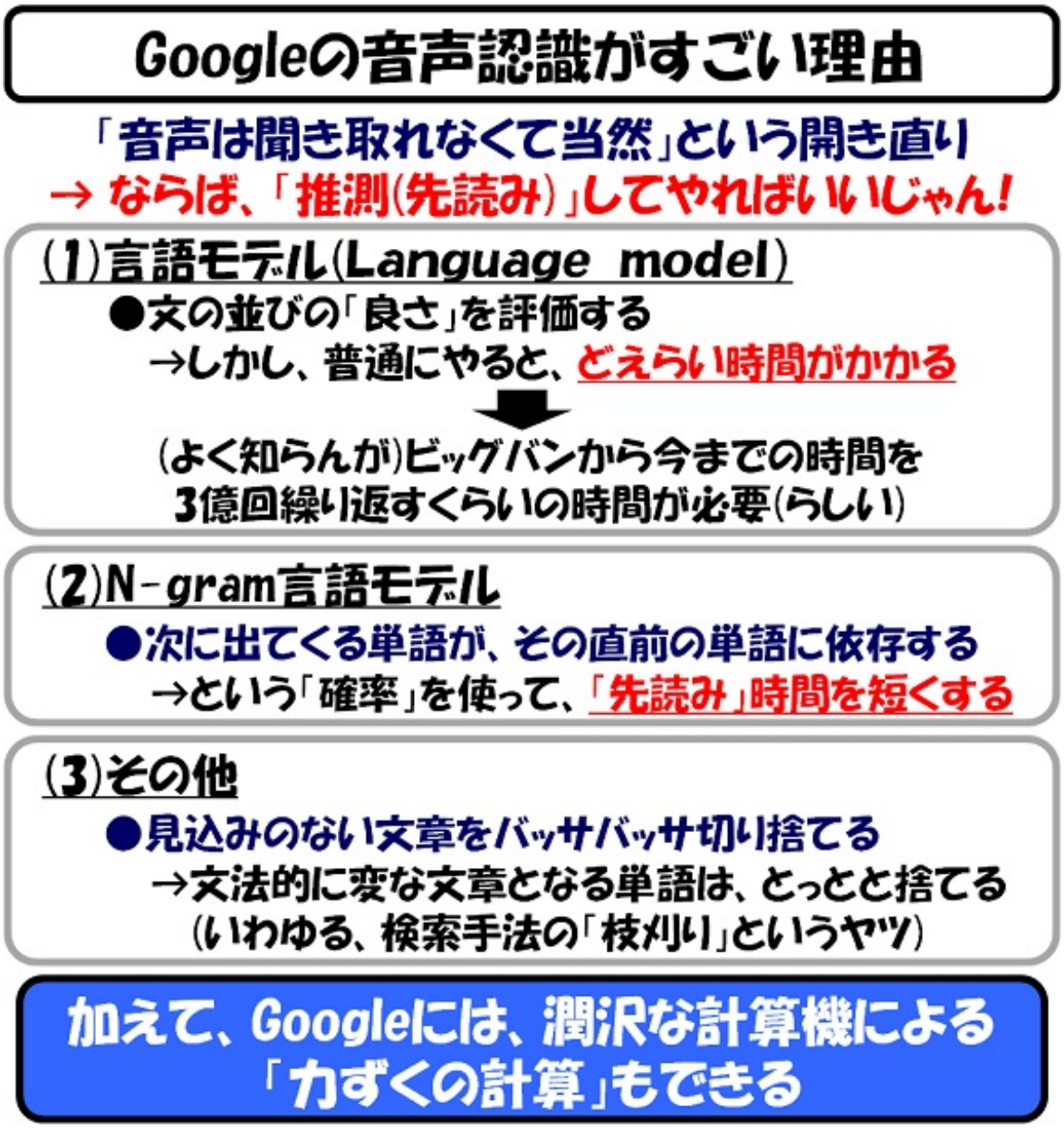


この論文の発表が、冒頭で紹介した、Googleトレンドで2012年にトリガとして観測された原因の可能性があります（ウラは取れていません）。

まあ、それはさておき、はっきり言って論文の内容が「分からん」。

エンジニアとしては失当の発言かもしれないのですが、そもそも、何で、音声認識に「予測」なんぞが登場してくるのか、さっぱり理解できなかったのです。

で、私なりに、この論文の根拠となっていることは、こういうことなんじゃないかなーと、考えを纏めてみたのが、以下の図です。



私なりに読み解いた結果、Googleは、

―― 音声なんぞ、そもそも聞き取れないものだろう?

という、ものすごい開き直りをしている、と私は理解しました。

そのような「聞き取れない音声」を推測するためには、「言語モデル(Language Model)という「文の並び」のテンプレートを用意しておき、そこに当てはまるかどうかを評価してやればいい、という考え方があります。

しかし、「文の並び」のテンプレートを当てはめて、真面目にその良さを評価した場合、その評価に必要な時間は、宇宙の年齢を3億倍する時間よりもかかる、という文献も見つけました。

そこで、登場してきたのが、N-gramモデルです。N-gramモデルは、次の「音節」が聞こえてくる「前」に、過去の1つ以上の音節を使って、次に登場する可能性の高い音節の幾つかを取り出

して「待ち構えていればいい」という、従来の音声認識技術では使えなかった、逆転の発想を行います。

さらに、見込みのない文章を作ってしまう音節なら、早々に切り捨てる（文法的に、ハチャメチャなフレーズになるなら、切り捨てるとか）などして、「待ち構えている」数をさらに減らします。

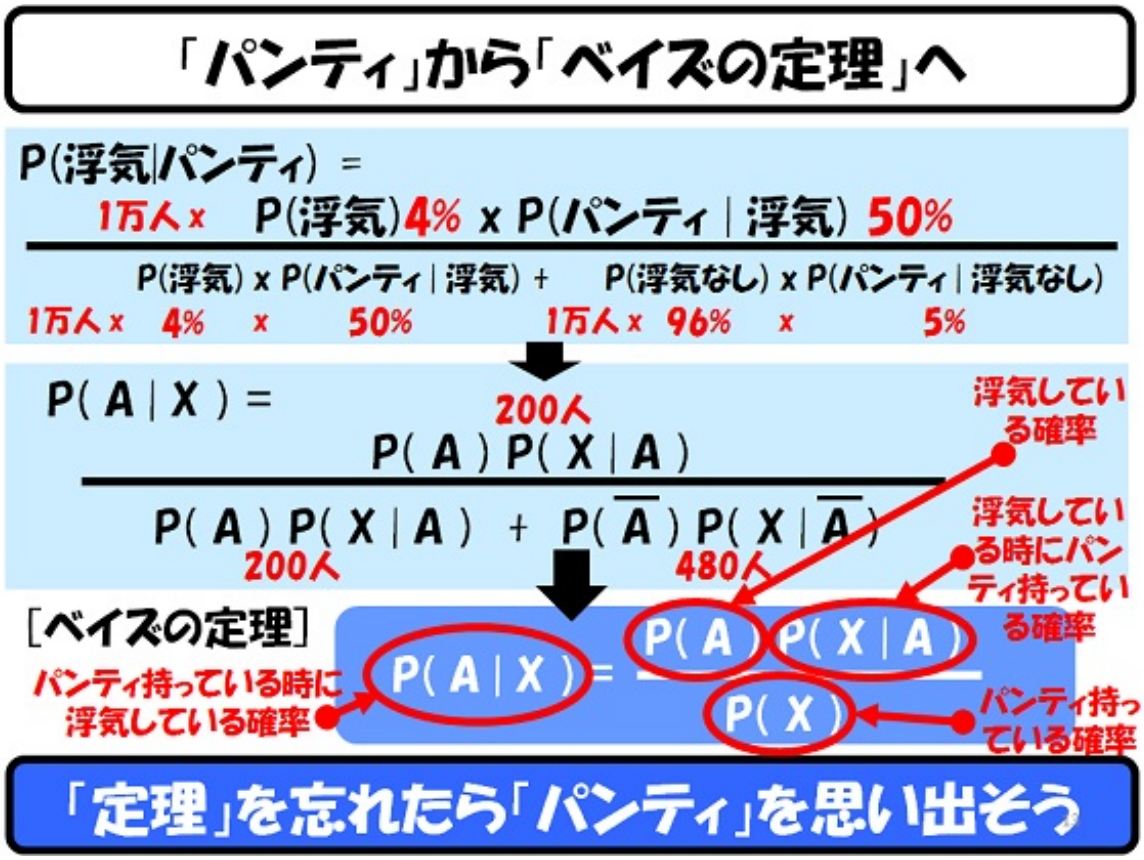
なぜこの技術が従来使えなかったかというと、この方法を使うためには、世界規模の膨大な（そんじょそこらの「膨大」とはケタの違う「膨大」な）音声データとそれに対応するテキストデータの両方を同時に収集し、そんな膨大な情報を、「本当に」で処理できる計算リソースとエンジニアの両方が必要であり、そのような企業は、世界でGoogleくらいしかなかったからだ、と思っています。

つまり、Googleの音声認識技術のキモは、「すごい技術の発明」―― というよりは、むしろ、毎日集まってくる、宇宙規模の膨大な音声情報を、"理論上"ではなく、"現実"に処理することのできるシステムを、"理論上"ではなく、"現実"に作ってみせた ―― という点にある、と私は読み取りました。

ベイズ理論の概念と似ているN-gramモデル

では、最後に、N-gramモデルについて、私なりの理解で簡単に説明を試みます。

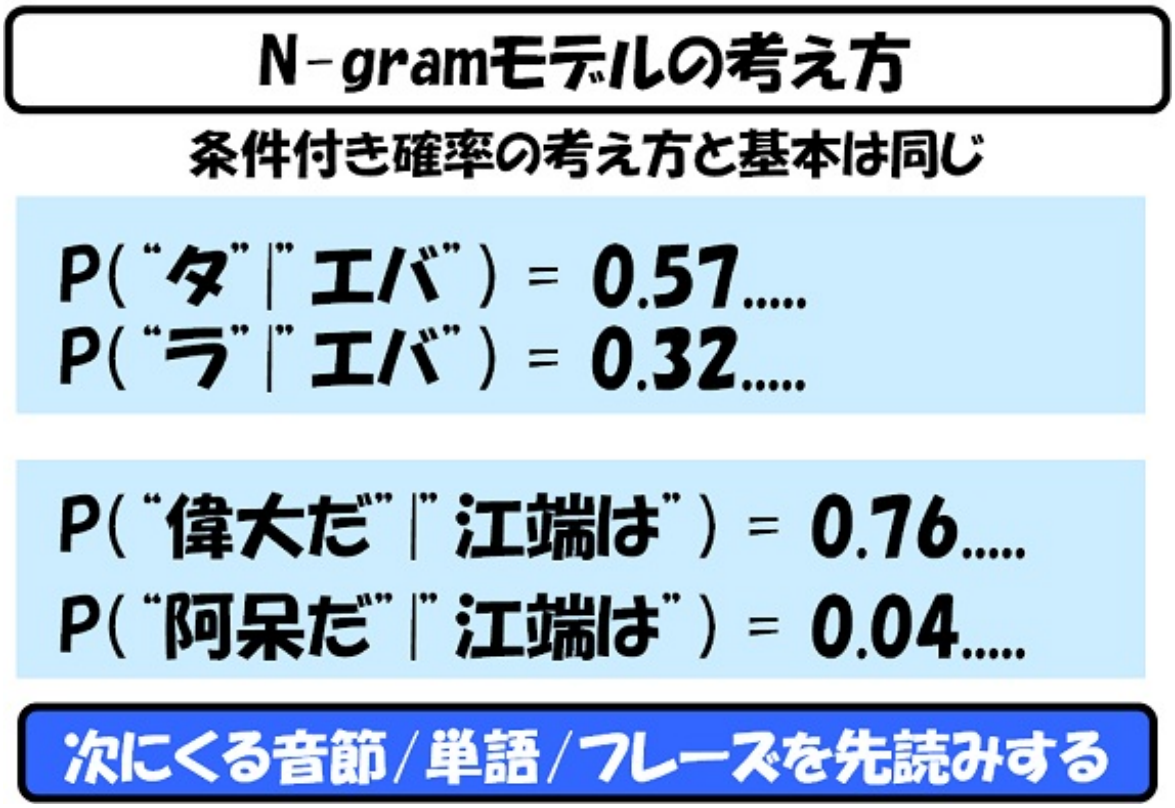
以前、この連載で、ベイズ推論のお話をしました（「困惑する人工知能 ～1秒間の演算の説明に100年かかる!?」）。まずは、この内容を思い出してください（まあ、思い出さなくても良いですが）。



N-gramモデルとは、ぶっちゃけ「条件付き確率」のことです。ある事象の発生確率が、その直前までに既に発生してしまった事象によって、コロコロと変化する確率のことです。

この例では、パートナー（男性）の浮気の確率が、他の条件によって（タンスの中からパンティが発見されたという事実で）変化することを示す、「"パンティ発見"条件付き"浮気"確率」を表しています。

N-gramモデルの考え方も、概念的には同じです（多分）。



"エバ"と出てきたら、次が「タ」となり"エバタ"となる確率と、"エバ"と出てきたら、次が「ラ」となり、"エバラ"となる確率を計算して、高いものからリストアップしておきます。こうしておけば、"エバ"の後の音節が不明瞭でも、力づくで音声認識を押し進めることができます。

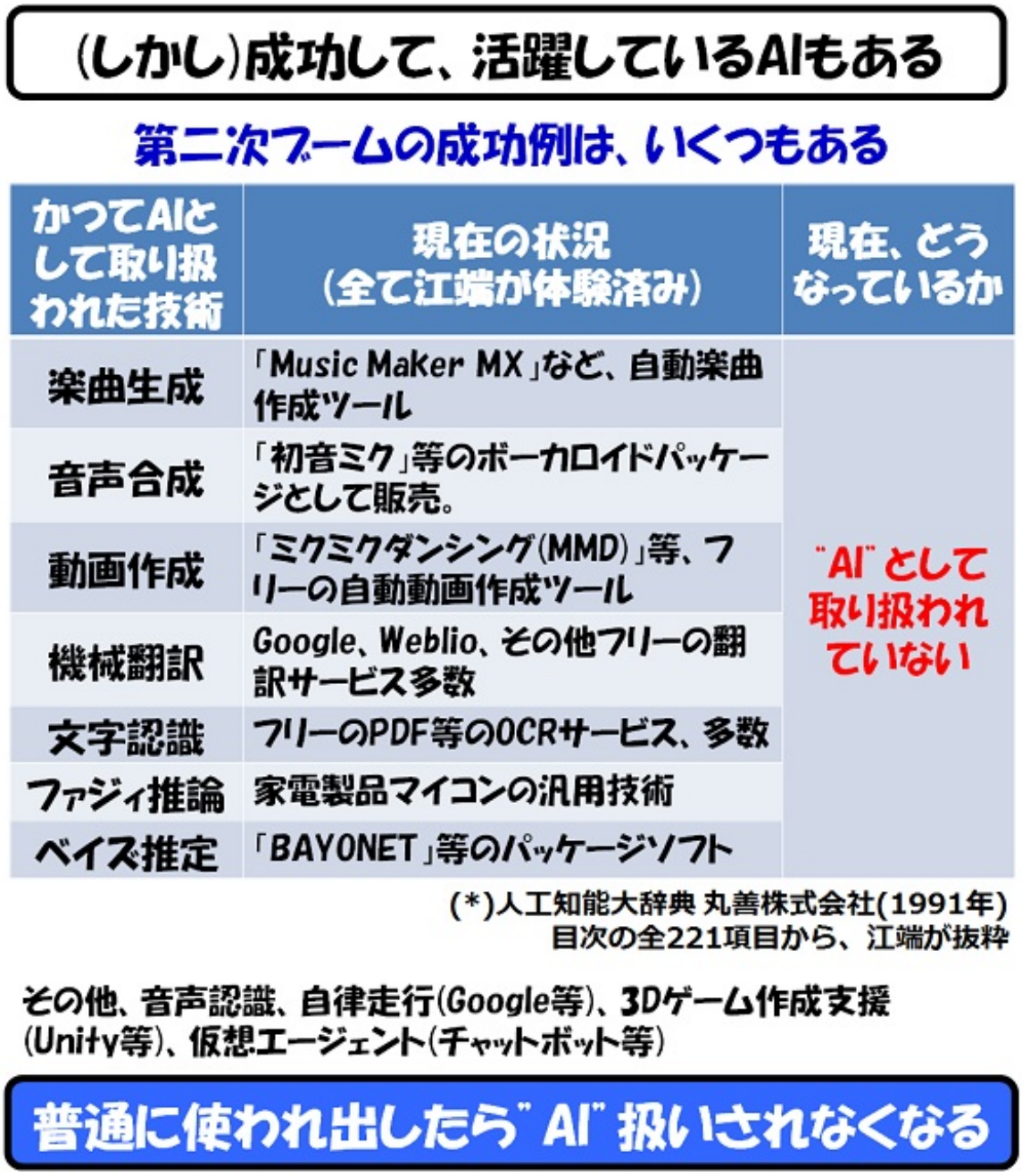
また、文節単位でも、"江端は"がくれば、普通に「偉大だ」が確率的に上位にくるはずです。"江端は"の後に「阿呆だ」が登場する確率が下位に沈むのは間違いありません。なぜなら「江端は偉大だ」は、もはや、慣用句といってもいいレベルにあるからです（うそです）。

“実用化されたAI”に対する世間の評価

音声認識技術も、希望と絶望を（少なくとも一巡）繰り返した技術ではありますが、他のAI技術と違い、（特定IT企業のみが有する膨大なコンピュータパワーを使い倒してはいるものの）ほぼ私たちが望むレベルに達したという、稀有（けう）な、そして、幸運な技術と言えそうです。

このことは、もっと世間に評価されても良いはずです。しかし、冒頭のトレンドデータからは、この音声認識技術が、大騒ぎされているという傾向は読み取れません。

私は、ここに、以前私が提示した、AI技術に対する世間の評価の姿勢が見られるような気がするのです。



「陰湿な人工知能 ～「ハズレ」の中から「マシな奴」を選ぶ」より抜粋

　つまり ――

ひとたび、実用化されてしまえば、AI技術として認識されなくなる

ということであり、私は、今でもこの現実に腹立たしい思いをしています。

　誰であれ、技術の恩恵を受けたなら、その技術の開発者に最大級の敬意と賛辞を贈るべきだと思うのです。

　特に、わが国は、エンジニアの成果に対する敬意と賛辞（と金銭的報酬）に、著しく欠けていると思うのですよ*)。

＊）ぜひ、ご一読ください。著者のブログより「「名誉」のフリーライド（ただ乗り）

ねえ、そう思いますよね。エンジニアの皆さん。

それでは、今回のコラムの内容をまとめてみたいと思います。

【1】冒頭にて、米国赴任時に、航空会社の自動電話応答システムで、一方的に無視されたという、忌々しい（いまいましい）江端の黒歴史と、今年のゴールデンウイークに、Googleの音声認識技術に「腰を抜かす」程に驚いた、というお話をしました。

【2】音声認識技術は、1999年PC用音声認識ソフトの大失敗で、冬の時代に突入し、10年の月日を経た後、一気に実用化に成功して今日に至った ―― という、江端の音声認識技術の歴史観をご紹介しました。

【3】「音声」は、（1）周波数単位に分解できる音の波（音波）の集合体であって（2）人間ごとに、別々の楽器から出てくる音のようなものであることを説明し、最低最悪の情報伝達信号であることを明らかにしました。

【4】音声認識技術とは、別々の楽器から出てくるどんな音であっても、そこから、文字に変換しなければならないという、とてつもなく難しい技術であることを、その1つの技術であるMFCC法を例として説明しました。

【5】しかし、「音声認識技術」は、恐しく面倒で地道な努力を続ける割に、その努力に比例してその認識率が一向に向上する訳でもない、という、研究者やエンジニアにとっては、苦行とも言える作業が必要となることを説明しました。

【6】Googleの音声認識技術のキモが「音声は聞き取れなくて当然」という開き直りにあり、「それならば、聞き取れない音節を『先読み』してやれば良い」という逆転の発想にある、という江端の理解を明らかにしました。

【7】最後に、「音声認識技術の実用化を、もっと多くの人が絶賛し称賛すべきであり」「技術の恩恵を受けたなら、その技術の開発者に最大級の敬意と賛辞を贈るべきだ」という江端の熱い思いを語りました。

「音声認識技術」について、私たちが真っ先にすべきこと

冒頭で、私は「英語のコミュニケーションの相手は、「人間」ではなく「機械」になる」という仮説を申し上げましたが、実際のところもっとエゲつない時代がくるかもしれません。

自分で計算しておいてなんですが、日本の人口問題は、気が遠くなりそうなほど絶望的に深刻な問題です（「“電力大余剰時代”は来るのか（前編）～人口予測を基に考える～」）。そんでもって、人口問題は、高齢社会問題、税収の問題、その他の問題と強く関連しています。

現時点で、私が考えている現実的な解は2つです。(1)機械化と、(2)移民受け入れ条件の緩和です。

特に上記(2)に関しては、日本の労働力を担ってくれる外国の人を大量に受けいれるしかないと思っています。特に、高齢化社会における医療、介護現場では、本当にこの手段しか残っていないじゃないかと思っています。

しかし、日本は、外国人労働者をなかなか受け入れません。

一例ですが、例えば、日本で「看護師をやりたい」というアジアの人は、結構な数いるようなのですが、この受け入れ数は、ものすごく少ないと聞いています。大きな理由の1つが「日本語」です。資格試験も日本語なら、実地試験も日本語で、しかも、そのトライアルの期限が非常に短い(たしか3年だったかな?)。

しかし、これでは、最初から「受け入れを拒否」しているようなものです。特に語学に関しては、日本人であれば、(英語教育で)その難しさをよく理解しているはずです。

一方、医療は、人命をあずかる仕事です。円滑なコミュニケーションが前提でなければ、命にかかわる問題が起こることも、確かなのですが、とはいえ「日本語が流暢(りゅうちょう)に使えることが前提」などと、ぜいたくを言っていられる時間は、そんなに残されていません。

ならば、資格試験を「英語」で統一してしまえば良いのです。フィリピン、シンガポール、インド、マレーシアからの人ならオールクリア。台湾、香港でも英語を使える人は多いそうです。

しかし、そうなると、自然、病院でのオフィシャルランゲージが「英語」とならざるを得ません。日本人の医師や看護師、その他の業務に関しても、英語によるコミュニケーションが徹底されることになるかもしれません。

まあ、彼らはその資質がある(のかな?)としても、最大の問題は、

患者も「英語が使えること」が前提となることです。

英語で自分の病状を説明しなければならず、正しく説明できない場合は、自分の命が危機にさらされることになります。

つまり、「英語が使える/使えない」によって、国内の介護・医療の品質について、差別を受ける時代がやってくる、ということです。

「グローバル化」なるものは、実は外側(海外)に向いているのではなく、本当は、内側(国内)に向いているのではないか ―― 最近の私は、そんなことばかり考えています。

はっきり言って、私たち「英語に愛されない日本人」の多くにとって、「英語しか通用しない病院に入院させられること」は、「絶滅収容所に収監されること」と同義……とまでは言いませんが、「英語は嫌い」だけど「介護サービスは受けたい」の両方を成り立たたせることが難しい時代になる可能性は十分にあると思います。

しかし、発想を逆転させれば、カーナビでもゲームでもメールでも、そしてスマホの音声入力でもない、音声認識技術の新しいキラーアプリケーション ―― 医療・介護通訳 ―― の肥沃な市場が、地平線のかなたまで広がっているということです*)。

*)江端の試算では、2053年に65歳以上の人口は35%に至ります(参考記事(外部媒体に移行します))。

もちろん"音声認識技術"だけでは足りず、併せて"自動同時翻訳技術"も推し進める必要もあるでしょう。

そして、そのために、最も大切なことは「その新しい希望に満ちた老後のパラダイムを作ることができるのは、一体どこの誰か?」ということを考えることです。

はっきり言って、ハードウェアの設計も、ソフトウェアのコーディングも、音声デバイスの開発もできないような政治家や行政庁のトップごときの意向を、忖度(そんたく)なんぞしている場合じゃありません。

皆さんが真っ先に行わなければならないこと ―― それは、

研究員やエンジニアを、それらの研究開発に向わせるように仕向け、そして、彼らを、これ以上もないほど、リスペクトして、チヤホヤして、お世辞と追従の限りを尽して、いい気分にさせ、最大級の待遇と報酬を与え続けること ―― 。

これ以外に、私たち日本人の未来はありません(断言)。

⇒「Over the AI ――AIの向こう側に」⇒連載バックナンバー

Profile

江端智一（えばた ともいち）

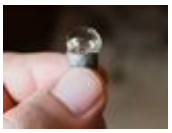
日本の大手総合電機メーカーの主任研究員。1991年に入社。「サンマとサバ」を2種類のセンサーだけで判別するという電子レンジの食品自動判別アルゴリズムの発明を皮切りに、エンジン制御からネットワーク監視、無線ネットワーク、屋内GPS、鉄道システムまで幅広い分野の研究開発に携わる。

意外な視点から繰り出される特許発明には定評が高く、特許権に関して強いこだわりを持つ。特に熾烈（しれつ）を極めた海外特許庁との戦いにおいて、審査官を交代させるまで戦い抜いて特許査定を奪取した話は、今なお伝説として「本人」が語り継いでいる。共同研究のために赴任した米国での2年間の生活では、会話の1割の単語だけを拾って残りの9割を推測し、相手の言っている内容を理解しないで会話を強行するという希少な能力を獲得し、凱旋帰国。

私生活においては、辛辣（しんらつ）な切り口で語られるエッセイをWebサイト「こぼれネット」で発表し続け、カルト的なファンから圧倒的な支持を得ている。また週末には、LANを敷設するために自宅の庭に穴を掘り、侵入検知センサーを設置し、24時間体制のホームセキュリティシステムを構築することを趣味としている。このシステムは現在も拡張を続けており、その完成形態は「本人」も知らない。

本連載の内容は、個人の意見および見解であり、所属する組織を代表したものではありません。

関連記事



「AIは最も破壊的な技術領域に」Gartnerが分析

Gartner（ガートナー）は、「先進テクノロジーのハイプサイクル：2017年」を発表した。この中で、今後10年にわたってデジタルビジネスを推進する3つのメガトレンドを明らかにした。

外交する人工知能 〜 理想的な国境を、超空間の中に作る

今回取り上げる人工知能技術は、「サポートベクターマシン（SVM）」です。サポートベクターマシンがどんな技術なのかは、国境問題を使って考えると実に分かりやすくなります。そこで、「江端がお隣の半島に亡命した場合、“北”と“南”のどちらの国民になるのか」という想定の下、サポートベクターマシンを解説してみます。

正体不明のチップを解析して見えた、“オールChina”の時代

中国メーカーのドローンには、パッケージのロゴが削り取られた、一見すると“正体不明”のチップが搭載されていることも少なくない。だがパッケージだけに頼らず、チップを開封してみると、これまで見えなかったことも見えてくる。中国DJIのドローン「Phantom 4」に搭載されているチップを開封して見えてきたのは、“オールChina”の時代だった。

攻撃的ホームセキュリティ ーー メイドが侵入者を迎え撃つ

さて、今回はいよいよEtherCATでホームセキュリティシステムの構築を開始します。「江端さんのDIY奮闘記」の名に恥じない作業風景をご覧いただきましょう。

誰も望んでいない“グローバル化”、それでもエンジニアが海外に送り込まれる理由とは？

今回は実践編（プレゼンテーション[後編]）です。前編ではプレゼンの“表向き”の戦略を紹介しましたが、後編では、プレゼンにおける、もっとドロドロした“オトナの事情”に絡む事項、すなわち“裏向き”の戦略についてお話します。裏向きの戦略とは、ひと言で言うなら「空気を読む」こと。ではなぜ、それが大事になってくるのでしょうか。その答えは、グローバル化について、ある大胆な仮説を立てれば見えてきます。

“技術の芽吹き”には相当の時間と覚悟が必要だ

今回は、200年以上の開発の歴史を持つ燃料電池にまつわるエピソードを紹介したい。このエピソードから見えてくる教訓は、1つの技術が商用ベースで十分に実用化されるまでには、非常に長い時間がかかるということだ。