Vol.2010-SLP-83 No.10
2010/10/29

それで結局 認識率はどれくらいなんですか？ ― 音声認識性能の虚実 ―

平沢 純一[†] 村上 久幸[†] 田中 幸[†] 木伏 祐治[†]

音声認識エンジンを用いて音声認識 HMI の製品・サービスを開発するカスタマは，品質（認識精度）について何をどのようにリクエストしてくるのか？を紹介しながら，音声認識エンジンのベンダはどのような技術サポートを提供することが可能なのか？を議論する．議論のポイントは，「カスタマからの期待」と「現在の音声認識技術で可能な技術サポートの現実レベル」との間に存在する溝を，どのようにして埋めたらよいのか？である．

How Customers Request Recognition Accuracy

Jun-ichi Hirasawa[†] Hisayuki Murakami[†]
Miyuki Tanaka[†] and Yuji Kibuse[†]

This article describes how customers developing products and services with speech HMI would question and request for speech recognition engine vendors in terms of speech recognition accuracy requirement. It also discusses what types of support speech vendors can provide in response to their requests. The key issue here is to fill the gaps between customers' expectations and what speech vendors could actually provide.

1. はじめに

実用化はされている

音声認識 HMI はどうしたら実用化され，普及するのか？が議論されるようになってからどれくらい経つだろうか．確かに音声認識 HMI をめぐる技術開発はまだまだ課題が山積みだ．工学的アプローチのごく基本であるはずの「性能評価」や「品質保証」についてすら，音声認識 HMI の分野では方法論が十分に確立されているとは言い難い面があるのも，情けないながら正直な現実である．

その一方で，実は「既に実用化はなされている」と考えるのも現実と言ってよいのではないだろうか．筆者のまったくの個人的な印象ではあるが，日本（語）の音声認識 HMI 技術のコミュニティには他国・他地域にはない，不思議な“自虐的悲観論”のような空気がある．また実際に多くの日本市場の開発顧客（カスタマ）に接していると「アンチ音声認識」派に多く出会うのも事実である．これが何らかの歴史的経緯によるものなのか，日本独自の文化的な要因から来るのか，それはそれで興味深い話題ではある．しかし，悲観的であろうが，アンチであろうが，その原因が何であれ，もはや我々は「実用化は既に立派に始まっている」という段階に来ていると考えてもよいのではないか．

製品・サービス開発の現場で起きていること

たとえ音声認識 HMI に技術的に未熟な面があろうと，製品やサービスを開発するカスタマは現状で実現可能な範囲で製品化を行う．そこでは「どうしたら実用化されるのか？」などという段階の議論は存在しない．あるのはただ「カスタマの期待に応えられるか，応えられないか」，応えられないなら「代わりにカスタマに何をどう提供すれば納得してもらえるのか」という格闘だけである．

カスタマは酔狂で製品やサービスを出すのではない．許される費用の制約のもとで，いかに魅力的な製品やサービスを世に問うか？の真剣勝負である．その中で殊勝にも「音声認識 HMI 技術」に興味を持ってくれるのである．だからと言って無条件のシンパになってもらえている訳でもない．そこでは，音声認識 HMI 技術が彼らカスタマのお眼鏡に適うだけの代物かどうか厳しく吟味される．

その過程では当然カスタマから「認識率（性能）はいくつか？」と問われる．音声認識に特有の事情や難しさ（次章以降で紹介する）を説明し，話はそれほど単純ではないことをカスタマに伝えても，カスタマからは「それで結局，認識率はどれくらい

[†] ニュアンス コミュニケーションズ ジャパン株式会社
Nuance Communications Japan K.K.

なんですか？」と問われることからは逃げられない．仮に運良く実際の製品・サービスの開発プロジェクトの開始まで漕ぎ着けられたとしても,その後にはお馴染みの「認識しません」が待っていることもしばしばである．

本稿では音声認識 HMI 技術の品質（認識性能，認識精度）に焦点を当て，実際のプロジェクトでカスタマからどのような要求や質問が寄せられるかを述べながら，それらに対して，音声認識エンジンのベンダはどのような技術サポートを提供することが可能なのか？を議論する．以下，2 章でプロジェクトとして契約が成立する以前の pre-sales フェーズでの取り組み，3 章で実際の開発プロジェクトとして認識性能はどのように扱われるか，4 章ではプロジェクトを終了した後の保証（warranty）フェーズで生じる事態について述べる．

2. プロジェクト契約前 ― pre-sales フェーズ ―

本章では，実際の開発プロジェクトとして契約が成立する前の段階で，カスタマからはどのような質問や要求が寄せられ，音声ベンダ側からはどんな情報提供や技術サポートが可能か？について述べる．最大のポイントは，認識精度が認識エンジンに固有なひとつの値ではなく，タスク設定（ドメイン，話者，環境など）に大きく依存するという事情をどこまでカスタマと共有できるか？である．

認識精度のタスク依存性

音声認識エンジンの性能を大きく左右するのは，それぞれのエンジンが内蔵する音響モデルである.しかしながら,音響モデルだけですべての性能が決まる訳でもない．同じ認識エンジン（音響モデル）を用いても，単に「ハイ/イイエ」の 2 語を認識させるだけなのか，1 から 1,000 までの自然数を認識させるのか，桁数任意の電話番号のような連続数字を認識させるのか，に応じて，当然ながら認識性能は変化する．認識性能に影響を与えるのは認識タスクのドメインや仕様だけではない．評価に用いる音声データはどんな話者がどのように発声したものなのか，音声の収録環境はどうか．実にさまざまな要因の帰結として，ひとつの「認識率」が算出されるのである．

したがって，カスタマは「仕様」と「評価用のテストセット音声データ（評価データ)」を規定して初めて「認識率」を算出することができる，という言い方が最も正しい．正確さを追求するなら,「認識精度はいくつですか？」というカスタマからの質問には,「それはお客様（の設定）次第です」との回答になってしまう．

もちろん，これで「ああ，そうですか．了解です」などと納得するカスタマはいない.「それはわかったけど．でも大体どのくらいなの？」と引き下がらないのは当然のことだ．認識率が未算出のカスタマのタスク設定に関して，認識精度の予測値を語れる技術があればよいが，まずは「認識性能はさまざまなタスク設定に大きく依存するのです」という事情を繰り返し説明して理解を得るのが第一歩だ．

SDK 提供による試用

もちろんボーっと手をこまねいている必要はない．まだ認識率を算出していないのなら，カスタマのタスク設定で評価計測してみればよいだけのことだ．通常，音声認識エンジンベンダは PC 上で動作する SDK などをカスタマに積極的に提供する．タスク仕様を定め，評価用の音声データを揃えれば，カスタマの最終製品での認識性能を近似的に求めることが可能だ．ここでのポイントは，ライブ発声（その都度マイクに向かって発声）したのでは，たとえ同一話者が同一内容を発声したとしても公平な性能比較ができないという事情をカスタマと共有することだ．そのためには評価用の音声データをファイルで収録・保存することが必須となる．

「仕様を定め」「音声データを揃え」れば認識精度を算出できる．しかし，そもそも「音声認識 HMI を導入しようかどうしようか」と考えている段階のカスタマにここまでの意欲と余力があることは稀である．検討段階でそこまで終えられるくらいのカスタマなら，そもそも「認識精度はどのくらいですか？」などと質問してくることはないだろう．自ら算出できないからこそ，質問してくるのだから．

実績や導入事例の紹介

ならばあとは「実物」を試してもらうしかない．すでに市場導入されている実際の製品があればそれを試してもらうのが早道だ．しかしすでに導入されている製品ということは「既存タスク」であることを意味する．カスタマが何らかの新しいタスク仕様を企画しているケースでは実例を参考にするのも限界がある．また同じタスク仕様であったとしても，誰がアプリ開発の主体なのか？の違いに応じて，できあがる製品が達成する性能が異なってくることは十分にありえる．実際の製品は認識エンジンだけでできあがっている訳ではないのだから．

結局のところ，pre-sales フェーズで「認識率はいくつですか？」と問われても，音声認識エンジンベンダは「具体的な認識率を提示する」ことはできず，ましてや「認識率を保証する」ことなどできないのが現実である．だからと言ってエンジンベンダが「認識率なんて事前に出せる訳ないじゃないか！」とキレてしまったらそこまでである．何とかカスタマが把握したい性能の参考になるような情報を，手を替え品を替え模索しながら，可能な限りの手段を尽くすのである．騙し売りでもよいのなら「我が社の認識エンジンは業界最高の 99.5%を実現しました」などと吹聴すればよいだけだ．しかし，カスタマに真摯に対応しようとするのなら，這いつくばうようにカスタマとの誠実なやりとりを続けるほかない．

3. 開発プロジェクト

誰だって認識精度は高い方がよいに決まっている．本章では達成可能な最善の認識精度の実現のために，実際の音声認識 HMI 製品・サービスの開発プロジェクトの中で行われている，性能（認識精度）に関わる開発アイテムの例を紹介する．

UI コンサルテーションの実施

実際の音声認識 HMI 製品・サービスでの体感性能を下げる，もっともありがちな原因は「語彙外発声（未知語, Out Of Vocabulary; OOV）」であろう．認識性能の評価実験を行う場合にも，通常，評価用のテストセットに敢えて意図的に語彙外発声を含めておくことは考えにくいので，語彙外発声のリスクはできるだけ開発段階の中で潰しておくことが求められる．同時に，エンドユーザに語彙外発声をされにくい仕様設計にしておくことは，のちのちのエンドユーザ満足，開発カスタマ満足に直結する．

音声対話システムが複雑になってくると，緻密に設計しているつもりでも，システムからのガイダンス/プロンプトと，その後の音声入力を待ち受けるグラマ/辞書の「語彙」が対応しなくなってしまうことがある．また言語学者（Speech Scientist）の知見から，ガイダンスの言い回し（wording）とグラマ側に登録しておく言い換え語のバリエーションとの関係に対して，有効な助言が可能なケースもある．さらに単一の認識語彙に対しても複数の読み・発音（phonetic transcriptions）のバリエーションを登録しておくことで少しでも認識精度を向上させられるケースもある．このように，音声ベンダによるコンサルテーションの実施は，認識精度の向上に関して重要な役割を持っている．

実音声データの収録

前章でも述べたように，性能評価には収録された音声データのファイルを用いる．のちのち製品・サービスの性能評価を適切に行なおうとするのであれば，音声データの収録は必須の開発アイテムである．ポイントは，可能な限り製品・サービスの実際の使用場面に近づけた収録をすることである．実際に想定されるユーザ層と同じ構成で収録の話者（被験者）を用意したい．収録場所（発声環境）も，実際に製品・サービスが使われる環境にできるだけ似ている方がよい．要は，実際のアプリケーションで音声認識エンジンに入力される音声とできるだけ同じ音声を収録しておきたいのである．

しかしながら容易に想像されるように，実音声データの収録は非常にコストの高い開発作業アイテムである．まして，実使用環境に近づけようとすれば尚更である．そこで少しでもコストを低減させるために，様々なノウハウが試みられている．もちろんシミュレーションなどを多用した場合，理想的な評価用の実音声データの収録条件からは少なからず乖離してしまうという現実はいかんともしがたい．あとは収録コストと評価データの厳密性とのトレードオフを鑑みながら，カスタマにとって「納得のいく評価データ」を準備するしかない．

いずれにせよ，「評価用の音声データを用意することもなく，認識性能の話をするなんてナンセンスだ」という感覚だけはカスタマと共有しておきたい．

認識エンジンパラメータのチューニング

認識精度の低下（誤認識）を招く要因はいろいろあるが，もっとも大きな要因のひとつに「音声区間（の始端・終端）検出の失敗」がある．認識すべき対象となる音声の区間が適切に取得できていなければ話にならない．そこで音声区間検出のミスを最小限にするため，認識エンジンのパラメータをチューニングすることが行われる．

「エンジンをチューニングする」とは，音声認識エンジンに馴染みの薄いカスタマには魅惑的に響く危険な用語であり，「チューニング」さえすればあたかも低かった認識精度がみるみると改善していくかのような過度の期待を抱かせてしまう危険がある．音声区間検出のパラメータは「厳しすぎず・緩すぎず」の最適なバランスが求められているに過ぎず，チューニングとはその最適値を探し当てるだけの作業である．探し当てた結果，「デフォルト値が最適でした」などということも当然起こりうる．

また，カスタマの中には「性能評価試験の結果，このコマンド（単語）の認識率が悪かったので，このコマンドの認識率を上げてください」というような「個別リクエスト」を寄せるケースがある．もちろん，そのコマンドが製品の中で非常に重要なのでこのようなリクエストをしてくるのだ，という背景はよくわかる．シーンごとやコマンドごとに，製品としての受け入れ基準となる音声認識率を定義するカスタマもよく見かける．これも当然のことだ．

しかし特定の語にバイアスを掛ける作為を施すという処置は，「それまで正しく認識していた語の認識を逆に悪くしてしまうかもしれない」という副作用とセットで検討されなければならない．結局，全体のバランスの中で判断するしかないのだ．あちらも立てつつ，こちらも立てる，ということが難しいのが現状の音声認識技術なのだ．

エンジン側で打てる手はすべて打った．それでもなおカスタマに「誤認識を起こすので何とかしてほしい」と求められたら，あとは原点に帰って「入力された音声データを真摯に眺める」しかない．そこから言えるアドバイスをカスタマに伝える．案外，最後の改善施策は，（認識エンジン内ではない）カスタマ機器側での音声入力系の音質改善を検討することだったりする例も多い．

Vol.2010-SLP-83 No.10
2010/10/29

4. プロジェクト終了後 ― 保証（warranty）フェーズ ―

何とかプロジェクトの終了まで漕ぎ着け，実際の製品やサービスが世に出る．とりあえずはメデタシメデタシであるが，本当に辛いのはここからだ．実稼働の後に寄せられる質問やリクエスト（もはやクレームと呼び名が変わる）は，開発の現場から寄せられるものではなく，リアルなエンドユーザからのものであることも多く，それだけにカスタマ側でも扱いが重くなることが多い．その一方で，実稼動しているサービスや製品に関わる内容であるため，「24 時間稼働しているサービスを止めてはならない」など，許される改修の手法にも制限があることが多く，厄介である．

正直に言えば，このフェーズに来て突然に何か魔法のような対応方法が存在する訳はないのである．現状の技術レベルで可能な手は，pre-sales フェーズ，開発プロジェクトの中ですべて投入しているものである．せめてできることと言えば，前章でも述べたように「実際の音声データを見て，誤認識の原因に関する分析結果をアドバイスする」くらいである．それとて誤認識はひとつの要因だけから生じるとは限らないので，明快な原因分析を語れるとは限らない．

せめて，保証フェーズに入っての「クレームをレポートする際のフォーマット（プロトコル）」を事前に定めておきたい．ポイントは「認識エンジンベンダー側でも現象が再現できるようにレポートしてもらう」点である．音声認識エンジンの場合，「当該の音声ファイル」を共有して，現象を再現可能にしておくことだ．

決して笑い話ではなく実際に起こるケースだが，カスタマから「とにかくエンドユーザさんが怒っている．『認識しないから何とかしろ』とのことだ」というだけのレポートがエンジンベンダに上がってくる．もしかしたら，そのシーンでは認識対象となっていない単語を懸命に繰り返していたのかもしれない．それとも興奮し過ぎで発話音量が大きすぎたのかもしれない．まさかとは思うものの，機器側で音声入力を取得し損ねていて音声認識エンジンには何も入力されていなかったのかもしれない．カスタマとの間で正しい現象の切り分けをするためにも「再現可能な体制」を事前に申し合わせておくことが重要となる．

5. おわりに

誤解を恐れず正直に言えば，現状の音声認識エンジンではどう逆立ちしても認識できない発話は存在する．認識できないものはできない．しかしその一方で「打てる筈の手を打っていないがために認識できていない」ケースが多いことも事実だ．本来であれば認識できるはずの発話が，エンジン性能以外の理由で認識できないことになるのだ．そもそも要求仕様の段階で無理があったのに仕様の策定段階で問題として発覚しそこねたのかもしれない．あるいは，待ち受けるグラマや語彙の設計や実装に不十分なところがあったのかもしれない．あるいは，シナリオ（対話フロー）のわずかな工夫があれば，問題にならなかったことかもしれない．とにかく発覚がプロジェクトの後半になればなるほど不幸は増す．エンドユーザは不満を覚える．カスタマは苛立つ．音声ベンダは当惑する．みんなの「こんなハズじゃなかった」を起こさないためには，カスタマと初めて出会った，あの打ち合わせでの「今度，音声認識を導入してみようと思うのですが，認識率はどのくらいなんですか？」と問われた瞬間から，勝負は始まっている．

謝辞　本発表の機会を与えてくださった旭化成株式会社 庄境 誠氏に感謝いたします．