音声言語情報処理 17－9
（1997. 7. 18）

おもろい音声サービスって何やろ!?
— 音声サービスに求められる課題について —

天白成一* 橋本雅行* 傍島康雄* 小川 均**

*株式会社アルカディア
**立命館大学

1 はじめに

昨今の音声合成、音声認識等の技術は、実用レベルに到達しつつあると考えられている。しかしながら、現状においては、技術の成熟度と比較して、期待されたほどには利用されてはいない。確かに、音声合成や音声認識のソフトウェアは、パソコンにバンドルされて出荷されており、利用者が増えているようではあるが、ビジネスとして成立するには至っていない。これらのバンドルの目的は、音声技術の利用を訴えるのではなく、むしろ販促品としての位置付けである。

先に嵯峨山 [1, 2] がメーリングリスト等で調査したときにもいろいろな意見が出ていたが、現状では、市場動向を大きく変化させるようなブレークスルーが見当たらないのが本音であろう。この状況は、もう少しさかのぼれば、中津による市場動向の報告 [3] にも指摘されており、ここ 10 年間に大きな変化は認められない。

音声言語システムの開発をF1のフォーミュラー・カーの開発に例えて、岡田が「いいエンジンを持っていても、それに見合うシャシーの供給が受けられなければ、本来の性能を引き出すことができない。その意味で、総合技術、あるいはシステムとしてのバランス感覚も重要だ。」と述べている [4]。音声認識や音声合成のいわゆる究極の目標の達成を待っていては、いつまでたっても音声サービスを提供することなどできない。むしろ、現状の技術で何か新しい音声サービスをするという観点で、研究開発を眺め直す必要があるように思う。また、精緻な氷細工を造るのに、長刀を振るうようなアンバランスも不要である。ここ数年で必要とされる技術課題を明確化することも重要である。

2 おもろい音声サービス

メディアとしての音声の特徴は、手軽、リアルタイム、非文法的、韻律 (アクセント、イントネーション、リズム)、モダリティ(感情、個人性) の存在がある。しかしながら、音声を他のメディア (文字情報、図形情報) と同様の目的や同様の機能実現のために使用すると、例えば、文章の伝達、3次元物理関係の定義などに使用すると、音声のメディアとしての特徴が損なわれることになり、利用する気を喪失させる。したがって、これらの特徴を生かせる応用に限定するか、他のメディアと合わせて利用することにより音声が重要な役割を果たすようにする必要がある。本稿では、これら音声のメディアとしての特徴を生かしたおもろい音声サービスの例を考えてみる。

ちょっと賢い留守番電話

「はい」「もしもし」「こちはら○○です。」この各フレーズを個別に録音しておいて、各フレーズの間を少しづつ間を空けて、相手が、喋りだすのを待つ。ちょっとした音声パワーの検出だけで開発できるのがミソ。でも、発話を促すには、間をどのように制御すれば良いかも分かるやもしれない。

TV ゲームに応用しよう

TV ゲームに音声認識を利用した場合、例えば、音声テトリス [4] では、右・左・回れなどを音声でコントロールする。流行りのアクションゲームでは、「昇龍拳!!」と叫べば、技を蹴り出す。ただし、必殺技の名称を音声認識するのではなくて、単に音声パワーの強度や時間変化を利用する。もちろん、音声パワーに対応して、相手へのダメージが異なるようにしておく。速さとスピードの変化をどのようにつけるか、裏技本が出版されることは間違いない。

うなずきマシーン

自動販売機が定型文を喋ると、酒に酔った中年サラリーマンが、語りかけるという話がある。そんな中年サラリーマンの不満は、自動販売機が、返事をしないことだそうだ。そこで、音声で語りかけると、とにかく、何か返事をしてくれる機械を開発しよう。いつも上司にがみがみ言われている人にとっては、ストレスを発散するに良いかもしれない。あるいは、独り暮らしの老人向けにも良いかもしれない。

つぶやきマシーン

何を喋っているかは、良く分からないが、とにかく楽しそうに、話し続ける。これは、テレビのCMでお馴染みのシャベリタランティーノと同様、とにかく喋る機械である。チェス・ゲームで人間とコンピュータの戦いが話題になるから、シャベリタランティーノとつぶやきマシーンとの一騎打ちもさぞかし話題をさらうであろう。

漫才の研究

うなずきマシーンとつぶやきマシーンが完成したら、漫才の研究をしよう。如何にしてボケるか? 如何にしてツッコむか? これは、難しい研究である。絶妙のタイミングで、人間とのコミュニケーションを行い、笑いを誘う姿は、さぞかし見物であろう。まずは、二丁目劇場で修業してから、なんば花月を目指そう。

エンターテイメントを目指そう

先頃、MIDIで制御による歌う音声合成器が発売された。近い将来には、歌う音声合成器がユーミンや小室ファミリーのバックコーラスに使われるかもしれない。そのためには、高品質音声合成技術(??)が必要になりますね。さて、カラオケでの音痴を補正するために、半音程度のズレであれば、現状の技術でも実時間で音程を変換することができるでしょう。これがあれば、多少の無理な高い音を含んだ曲でもちゃんと歌えますよ。

ルパンは生きている

アニメのルパン三世の声でお馴染みの山田康雄さんが亡くなられて、撮影中の映画を完成させるために、ものまねの栗田貫一さんが主人公のルパンの声の代役を演じて話題になった。しかし、熱烈なファンは、先のルパンの声と違うと評価は厳しい。これは、ものまねでは、喋り口調を真似ているのであって、決して、その者の声を真似ているのではないからである。そこで、この課題を解決するために、栗田貫一さんに演じてもらった後、音声変換でさらに、本物のルパンの声に近づけよう。ただし、決して山田康雄さんを復活させるのではないからお間違えなく。

3 まとめ

我々は、もしかして前提のない研究開発を強いられているのではないだろうか? 人工知能で良く知られたチューリングテストの音声版を考えてみる。

「もしも、肉声に限りなく近い合成音声を作りだす技術ができたとする。」

「もしも、全ての音声を音声認識できる技術ができたとする。」

このような仮説を設けた上で、果たしてどのような音声サービスが利用者に望まれるのであろう。例え肉声に限りなく近い合成音声を得られたとしても、文字情報を全て音声情報に変える必要はまったくない。規則合成を英語では Text-To-Speech と言うが、書き言葉と喋り言葉は、大幅に異なる。書き下された文章を音声で聞くのは、極めて冗長である。もし、電子化された文字データのない世界で、音声合成の研究をしたら、どのようなアプローチになるのであろう。Text-To-Speech から Text を取れば、To-Speech だけが残る。合成音声を生成する部分だけを使った音声合成技術のサービスへの利用は考えられないのであろうか? どのみち、日本語解析の部分では、固有名詞の読み間違いの問題を回避することは到底できない。事実、人間だって読み間違う。あるいは、読めない言葉も少なからず存在する。音声認識においても事情は、同じことで全てを音声で機械に入力する必要はない。テンキーの方が、優れた入力デバイスであるという局面が少なからず存在する。このような観点からマルチモーダルな環境での音声の役割について、検討する必要はある。もちろん、新田[5]が指摘するように、音声だけではサービスはできない。また、ここで陥ってはいけない罠は、音声が主たる役割を果たすという妄想に捕らわれてはならないことである。音声は、あくまで補助的な情報伝達手段である。T.P.O. をわきまえて、アプリケーションやサービスにおける音声の果たす役割を、音声の特徴を生かすように熟慮することがまず第一である。

参考文献

[1] 嵯峨山 茂樹: "なぜ音声認識は使われないか・どうすれば使われるか?", 情報処理学会研究報告, 94-SLP-1, Vol. 94, No. 40, pp.23-30, May 1994.

[2] 嵯峨山 茂樹: "音声認識技術実用への課題", 情報処理学会誌, Vol. 36, No. 11, pp.1047-1053, Nov 1995.

[3] 中津良平: "音声認識・合成技術の市場動向", 日本音響学会誌, Vol. 48, No. 1, pp.60-65, Jan 1992.

[4] 岡田美智男: "聞き耳をたてるコンピュータ", 竹内郁雄編「A I 奇想曲」, NTT 出版, 1992.

[5] 新田恒雄: "GUI からマルチモーダル UI(MUI) に向けて", 情報処理学会誌, Vol. 36, No. 11, pp.1039-1045, Nov 1995.