自分の声合成ソフト「ボイスター」の携帯版に関する評価と考察

高野哲朗[*1]　渡辺聡[*1]

Evaluation and Consideration of mobile Voistar your-own-voice speech synthesizer

Tetsuro Takano[*1], Satoshi Watanabe[*1]

Abstract – Voistar is a speech synthesizer made from each own voice. It has been provided as a PC application, but this time, we are going to release a mobile version of Voistar. To verify the effectiveness and find the issues, we performed evaluation experiments. As a result, the listeners can understand the voices of mobile Voistar, but we find the users want to input the text more rapidly than we have expected. We considered about more rapid and short-hand input UI.

Keywords: Speech Synthesizer, Mobile Application, Loss of Voice

1.　はじめに

1.1　背景

現在、喉頭摘出や神経性疾患により、本来の自分の声で発話することができなくなる患者は、ディサースリア（発話の実行仮定に関わる障害）だけでも国内で 65～70 万人存在すると推定されている[1]。しかしながら、音声はコミュニケーションを行う上で重要な要素であり、発話機構を失っても音声でコミュニケーションを行いたいという患者は数多く存在する。

喉頭摘出患者においては電気式喉頭や食道発声などの代用音声が広く用いられている。また体が徐々に動かなくなる神経性疾患においては、コンピュータに搭載された音声合成器が用いられることが多い。

いずれも音声を取り戻す方法として有効な手段だが、本来の自分の声とは明らかに異なる音声であるため、コミュニケーションの親密さという点で物足りないという指摘がある。

自分の声ソフト「ボイスター」（以下「ボイスター」）は、この点を解決する音声合成ソフトである[2],[3]。ボイスターの音声合成器は、利用者の音声そのものから作られる。そのため、声色だけでなく、しゃべり方の癖や訛りも再現することができる。

本論文では、ボイスターの携帯版の開発と、その評価について述べる。

1.2　携帯版への要望

ボイスターは従来 Windows 専用ソフト（以下「PC 版」）として提供されてきた（図 1）。Windows は家庭用 PC において大きなシェアを持っており、ユーザーのカバー率を優先して PC 版を開発した。

ところが昨今、ユーザーの利用する端末は PC からスマートフォンに移行しつつある。これは、スマートフォンの記憶容量の増加やクラウド化によって容量制限の問題が解決したこと、タッチパネル操作が一般化してきたこと、スマートフォンアプリケーションの高性能化など、様々な理由によるものである。

これに伴い、多くのユーザーから携帯端末でボイスターを使えないかという問い合わせが増加している。この背景には、上記のスマートフォンの普及に加えて、従来の PC 版ボイスターを使うときに、その携帯性の低さが課題になっていることが推測される。

たとえば、移動しながらの会話や立ち話をするときに、ノート PC を広げることはむずかしい。また在宅時であっても、常に PC を携えていないと発話できないという点は、ソフトウェアの利用に対する心理障壁となり、自分の声としての利用頻度の低下につながる。これらの場面においてスマートフォンでボイスターを利用することができれば、より日常的に自分の声で会話することが可能になる。

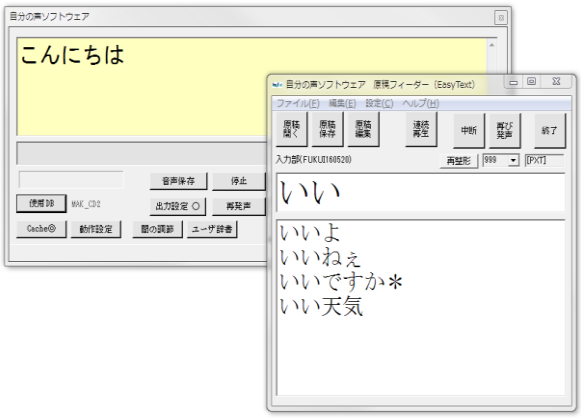

図 1 PC 版ボイスター
Fig 1 Voistar on PC

*1: ヒューマンテクノシステム東京
*1: Human Techno System Tokyo, Co., Ltd.

以上により、携帯版のボイスターの要件は下記のようになる。

1. 聞き手が音声を話者本人のものと認識できる
2. 移動中や立ったままでも会話ができる
3. 入力が面倒と感じない使用感である

我々は携帯版ボイスターを試作し、これらの要件が満たされているかを評価実験で検証した。

1.3 携帯端末の選定

2016 年現在、携帯端末の種類としては、Android、iOS が主流である。また、Windows を搭載した 8 インチ程度の小型タブレット PC もこれに含まれるであろう。

このうち、日本国内でシェアが大きいのは Android および iOS であるため、当面これらの対応で足りると思われる。今回はプロトタイプとして UI/UX の評価も行いたいので、評価予定者の中で利用の多い iOS を評価対象として選定した。

2. 携帯版「ボイスター」について

2.1 ボイスターの作成

個人の音声からボイスターを作成する手順を説明する。これは PC 版と携帯版で共通である。

はじめに利用者の音声を収録する。収録はあらかじめ用意された短文 400〜1000 文より構成される収録原稿の読み上げを基本として行う。

読み上げた音声は、音素単位に波形上の境界情報を付与し、音声合成エンジンから検索可能な形式でデータベース化する。これを音声合成ソフトの一部として組み込み、パッケージ化し、ユーザーの元に提供する。

またこれとは別に、親しい人の名前（家族の呼び名など）や頻繁に使う言葉（あいさつ、方言など）といった音声を録音しておく。これらの音声は、合成音声だと伝わらない発音のニュアンスを伝えるために、録音音声のまま再生する目的で使用される。我々はこれを感情キャッシュ（以下「キャッシュ」）と呼んでいる。キャッシュは同じ言葉を異なるニュアンスで録音することも可能である（「うん（無表情）」と「うん！（元気よく）」など）。

2.2 携帯版の概要

今回試作した携帯版ボイスターは、コミュニケーションをするにあたって必要最低限の機能に絞った。具体的には次の機能を搭載している。

- 文章読み上げ機能
- 入力履歴機能
- キャッシュ再生機能

文章読み上げ機能は、ユーザーの入力した文字列を音声合成器にかけて、生成された合成音を再生する機能である。

入力履歴機能は、ユーザーが入力した文字列をチャット UI で次々に表示していく機能である。これは音声が聞き取れなかったときや合成音声が不明瞭だった場合の救済措置として、話し相手に文字で理解させるための機能である。

キャッシュ再生機能は、収録時に録音したキャッシュを再生する機能である。キャッシュはユーザーがキャッシュの文章を入力するか、キャッシュリストから選択することによって再生される。

2.3 使い方

初回起動時に音声データベースのダウンロードを行う。ユーザーは事前に受け取ったメールに記載された URL から音声データベースを取得することができる。

アプリ起動後の画面（ホーム画面）を図 2 に示す。再生した文字列は吹き出しの中に表示されるため、発言の履歴が残る形になる。

ホーム画面の下部にあるボタンを押下すると、入力画面（図 3）に遷移する。入力画面の上部にあるテキストボックスに文字を入力し、Enter を押下すると、入力画面が閉じ、音声が再生される。画面右下のボタンを押下するとテキストボックスにフォーカスが移る。

画面中央部のリストはキャッシュリストである。キャッシュリストはあらかじめユーザーが設定した優先度に従って並んでいる。ユーザーはテキストボックスにキャッシュ文字列の一部を入力し、絞り込み検索でキャッシュを探すことができる。

図 2 ホーム画面

Fig 2 Home Screen

図 3　入力画面

Fig 3 Input Screen

3.　評価

3.1　評価方法

実験に用いたボイスターの条件は以下の通り。

- 話者：30 代男性（400 文）
- 実行端末：iPod touch 第 5 世代

評価実験は、「1.2 携帯版への要望」で述べたことを考慮し、3 種類の状況下でそれぞれ行った。以下に状況のの詳細を記述する。人数はいずれも話者（携帯版ボイスターの利用者）1 名と口頭で会話する健常な話者数の合算値である。

（状況 1）会議で報告を行う

状況 1 は、もっともアプリを利用しやすい条件を想定したものである。

話者が聞き手に対して、日常業務についての報告を行う。場所は静かな会議室である。話者が自分のペースを守れるよう、聞き手にはあいづち程度の返答のみを行うよう取り決める。

（状況 2）外出先で会話する

状況 2 は、より日常的な会話に近い条件を想定したものである。

話者と聞き手の 2 人は時事の話題に関する会話を行う。場所は室外である。状況 1 とは異なり、聞き手は自分の意見を述べるなど日常的な会話に近い返答を行うよう取り決める。また、安全を確保しつつ歩行しての会話も行う。

（状況 3）複数人の会話

状況 3 は、より複雑なコミュニケーションを想定したものである。

話者と聞き手 2 人の合計 3 人は、時事の話題に関する会話を行う。聞き手は条件 2 と同様、日常的な会話に近い返答を行うこととし、聞き手同士も会話を行うように取り決めた。

3.2　結果

（状況 1）会議で報告を行う

報告の内容を理解することができ、自然なあいづちでコミュニケーションをすることができた。

音質については、カタカナ語などで音声合成のイントネーションが崩れる箇所があったが、それ以外は聞き取りやすいという評価だった。

課題点として、聞き手は常に話者の発話を待機している状態であるため、話者が入力を焦ってしまい打ち込みミスが生じた点があり、打ち込みミスに対してなんらかの救済が欲しいという声があがった。また、現行は入力するまでに二度タップが必要になっているが、これを一度のタップで入力できるようにしてほしいという声もあがった。

（状況 2）外出先で会話する

会話はおおむね成立した。近くを車が通るなどしたときは、音量が負けてしまい聞き返すことがあったが、これは通常の音声による会話でも起こりうることであるため良いものとする。

報告と会話の相違点として、話している途中なのか話し終わったのかの判断を誤り、入力中に次の話題に移ってしまうということがあった。聞き手は実験中にこれに気付き、以後話者の手元を見ながら話を切り出すように変化した。

歩行中の使用に関しては､いわゆる歩きスマホなので､周囲に注意することが難しく使用がこわいという感想があった。

（状況 3）複数人の会話

会話は成立したが、他の 2 人の聞き手が会話をしていて話者が置いて行かれる場面があった。1 対 1 のときの会話と違って 2 人で会話を継続することが可能なので、話者が置き去りになっていることについて気づくのが遅れてしまったものと思われる。

また話者においても、入力と話を追うのを同時に行わなければならないため、文章を考えるのが難しいという意見があった。

4. 結論

4.1 考察

1.2 で述べた要件について検証し、考察する。

1.について各被験者に聴取したところ、イントネーションの違いなどから合成音声であることはわかるものの、声質は本人のようであるという意見であった。従来の PC 版においても同様の意見が得られており、携帯版でその品質が劣化するということはないと言える結果である。

2.について、移動中に文字入力を行うことは周囲の安全性確保が難しく、現段階では行わない方がよい。一方、立ち話や外出先のカフェなどでは、十分コミュニケーションが成り立つことがわかった。より騒々しい場所では、Bluetooth スピーカを利用するか、聞き手にイヤフォンを付けてもらうといった対策が考えられる。

3.について、入力感は携帯端末なだけあって PC を開くような煩わしさは無かった。特に会議室のような PC を使用可能な場所であっても、その場所まで PC を持って行くことが面倒であるという気持ちがあるため、日常使用としての携帯版の有用性が実感できた。一方、携帯端末での入力は通常の音声による発話と比べて格段に時間がかかってしまい、この点がコミュニケーションの障害となることがわかった。この点についてさらに考察する。

どの状況においても、思ったことや感じたことをすぐに声に出して表現したいのに、文字入力が追いつかないことが多々あった。さらに、早く打ち込もうと焦る気持ちから、普段スマートフォンを使い慣れている被験者でも打ち間違いが頻発した。これらの問題は、まずアプリを省入力化するよう UI 設計を考え直す必要がある。また入力を補助するような機能も求められる。ただし、いくら最適化を行っても入力にはある程度の時間がかかるため、聞き手においても、利用者を見て打ち込み中であれば発言を控えるといった配慮が必要である。

また、文字の入力だけではなく、あいづちのように瞬間的に声を出したいという要望にも応える必要がある。1 対 1 の会話では頷きや身振りである程度の表現が可能であるが、複数人だと見過ごされやすい。タイミングを逃すことなく音声であいづちを打つには文字を入力する時間がほとんど無いので、ボタンのような UI が望ましいだろう。

上記が解決されれば、自分の声の魅力もいっそう生きてくると考えられる。

4.2 今後の展望

先述の通り、より少ないステップでスピード感のある入力が行えるよう、アプリを改善する必要がある。

直接的には、文字入力を開始するまでのステップを削減する、または常にキーボードを表示しておくという変更が考えられる。これは今後、ユーザビリティテストを行った上で最良のものを選択する。

状況 1 の報告のように、すでに話すことが決まっている場合は、あらかじめユーザーが作成しておいた原稿を読み上げるような機能が望ましい。これによりユーザーが文字を入力する頻度を大幅に削減できる。また、あらかじめ原稿を再生して音声を確認することができるため、漢字の読み間違いや聞き取りにくい箇所の修正などが可能になる。

一方、状況 2、状況 3 のように、会話の内容が自由であるような場合はあらかじめ原稿を作ることができないため、頻繁に使用する文章を定型文として登録しておく機能が考えられる。これはキャッシュと同等に扱うことができるため、キャッシュリストで検索して選択するという方法で省入力化を実現できる。

あいづちのような瞬間的な反応は、ホーム画面上にボタンとして配置するのが望ましい。ただ、スペースに限りがあるため、ホーム画面を左右にフリックして表示するタブメニューで実現するのが現実的であろう。

また、再生する音声に関して、笑いなどで文字をそのまま合成すると不自然になる場合がある（たとえば「あっはっは」を棒読みしてしまう）。これに対しては、コメディ番組の聴衆の笑い声のように、一般的な効果音を利用することが考えられる。

歩行中の使用は危険であるが、それでも会話したいという気持ちは残るはずである。現状は良い方法は無いが、文字入力や定型文の選択のできるウェアラブル端末が登場すれば、これを利用することが考えられる。

今回は携帯端末を操作できるという前提での開発・検証であったが、今後は神経性疾患で身体が動かなくなった患者向けに、UI のカスタマイズや入力補助スイッチへの対応を行う必要がある。

4.3 まとめ

Voistar の携帯版の評価実験として、携帯端末を用いた音声合成ソフトによるコミュニケーションを行った。PC が利用できない場所や、PC を持ち運ぶのが億劫な場面において利用でき、よりユーザーの生活に密着したツールとなりうることがわかった。

一方、コミュニケーションを行う上で必要とされるスピードに今ひとつ追いつかないという課題が残った。今後、UI や機能の改善とユーザーのインタビューを繰り返し、継続的にアップデートを行っていく。

5. 参考文献

[1] 西尾: ディサースリア臨床標準テキスト, 医歯薬出版株式会社 (2007).

[2] 岩木,他: 喉摘した大学教授による合成音声を使った講義の報告と考察; HIS2008 1334

[3] 岩木,他:リアルタイム操作性を重視した自分の声 VoCA におけるユーザビリティ検討報告; HIS2013 1501D