

論文

インタビュー音声の文字起こしにおける ミライスピーカーの活用

大谷 紀子 岡部 大介

聴力が低下した高齢者はテレビの音量を上げる傾向にある。家族からの指摘により音量を抑えると、内容を十分聞き取れないままテレビを視聴することになる。テレビ視聴時間の長い高齢者にとってテレビ音声の聞き取りにくさは重要な課題である。本課題の解決策のひとつとして、音量を上げずに聞き取りやすさを向上させる「ミライスピーカー」が注目されている。本製品は高齢者や同居家族から高い評価を得ており、空港や銀行などの施設でも活用されているが、販売元の株式会社サウンドファンは新たな分野への展開を模索している。本研究では、ミライスピーカーの新たな活用分野の開拓を目的として、インタビュー音声の文字起こしにおける活用可能性を検討する。インタビューで収録した音声を再生しながら確認や修正を行いつつ文字起こしを行う場面を想定し、ミライスピーカー、高音質で高価なスピーカー、およびパソコン内蔵スピーカーにおける自動文字起こしの認識精度を比較する。

キーワード：ミライスピーカー インタビュー音声 文字起こし 認識精度 誤認識

1. はじめに

聴力は加齢に伴って低下し、日常生活におけるさまざまな音声が聞き取りにくくなる^{[1][2]}。聴力が低下した高齢者は、テレビやラジオの内容を聞き取るために音量を上げる傾向があるが、大幅に音量を上げると周囲の人にストレスを与える可能性がある。家族から「うるさい」と言われた高齢者が、内容を十分聞き取れないままテレビを視聴する状況が生じることもある。

高齢者の多くはテレビに依存した生活を送っており、ニュースや情報番組を通じて情報を収集するだけでなく、ドラマの視聴によって予想通りの展開に安心感を得たり、自己の状況に当てはめて教訓を得たり、自己の人生を振り返ったりしている^[3]。総務省情報通信政策研究所が実施した「令和5年度情報通信メディアの利用時間と情報行動に関する調査」^[4]によると、60代のテレビ視聴時間は平日で257分、休日で307.6分である。また、年齢が高いほどテレビ視聴時間が長くなる傾向が示されていることから、より高齢の人のテレビ視聴時間はさらに長いと推測される。以上より、高齢者にとってテレビの音声の聞き取りにくさは重要な課題といえる。

本課題の解決策のひとつとして、音量を上げなくてもテレビの音声の聞き取りやすくなる「ミライスピーカー」が注目されている。聞こえに困難を抱える高齢者や同居

家族から高い評価を得ており、空港や銀行、公共機関などでの呼出し用スピーカーとしても活用されている。ミライスピーカーの販売元である株式会社サウンドファンは、ミライスピーカーのターゲットをさらに拡大することを目指し、新たな分野への展開を模索している。

質的心理学やフィールド認知科学、ライフヒストリーの社会学などの分野では、インタビューデータの分析を通じて新たな知見が得られることがある。インタビューデータを分析するにあたり、音声データの文字起こしが不可欠である。かつては、人間が音声を繰り返し聞きながら手作業で文字に起こしていたため、文字起こし作業には多大な労力が必要であった。しかし、近年では音声データを入力することで自動的に文字起こしをするソフトウェアやサービスが提供されており、文字起こし作業にかかる労力は大幅に軽減されている。一方で、音声を再生しながら確認や修正を行いつつ文字起こしをする場合、音声データ入力による自動文字起こしは必ずしも有効とはいえない。

本研究では、ミライスピーカーの新たな活用分野の開拓を目的として、インタビュー音声の文字起こしにおける利用可能性を検討する。インタビューで収録した音声を再生しながら確認や修正を行いつつ文字起こしを行う場面を想定し、ミライスピーカー、高音質で高価なスピーカー、およびパソコン内蔵スピーカーにおける自動文字起こしの認識精度を比較する。

2. ミライスピーカー

ミライスピーカーは株式会社サウンドファンが製造、

OTANI Noriko

東京都市大学メディア情報学部情報システム学科教授

OKABE Daisuke

東京都市大学メディア情報学部社会メディア学科教授

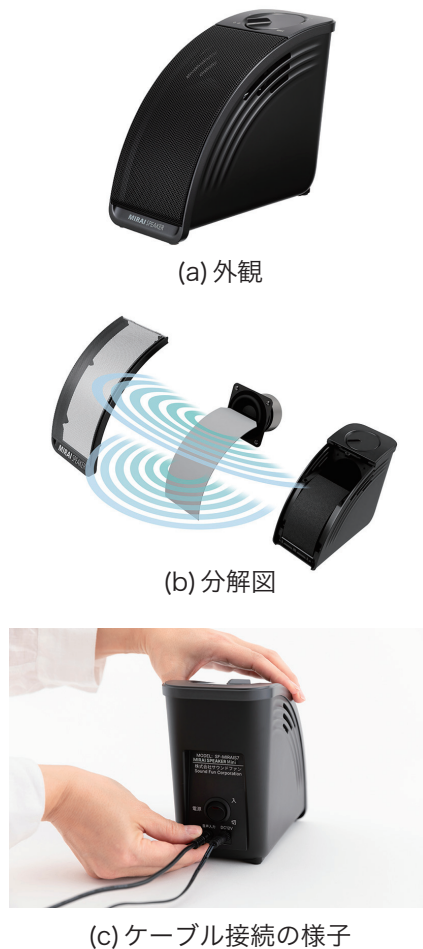


図1 ミライスピーカー・ミニ

販売しているスピーカーであり、音質よりも言葉の聞き取りやすさを重視して開発された。現在、モノラルタイプのミライスピーカー・ミニと、ステレオタイプのミライスピーカー・ステレオが販売されている。YouTubeのサウンドファン公式チャンネルで公開されている動画¹では、ミライスピーカーについて以下のように説明されている。

ミライスピーカーとはテレビの言葉をくっきりと聞こえやすい音に変換するスピーカー。
聞こえにくいテレビの言葉をくっきりとクリアに変換する
だから今まで聞き取りにくかったテレビの言葉が
音量を上げなくてもはっきりと聞き取りやすくなる
聞こえにくい方には聞こえやすくご家族にはうるさくない
そして音の広がりも通常のテレビに比べて部屋全体に、
遠くまで。
だから離れた場所でも聞こえやすい

¹「55秒で説明！ミライスピーカーとは？」<https://www.youtube.com/watch?v=3173Zg9b8Ic>

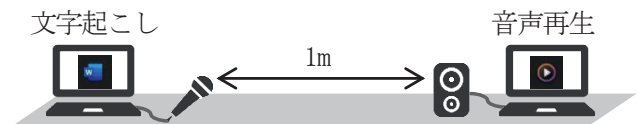


図2 測定環境

ミライスピーカー・ミニの外観と分解図をそれぞれ図1(a), 図1(b)に示す。従来のスピーカーの振動板は円錐形であるが、ミライスピーカーの振動板は平板を湾曲させた形状となっており、テレビやラジオの音声をミライスピーカーから出力すると、音量を上げなくても言葉をより明確に聞き取ることができる。また、「曲面サウンドの原理」により、ミライスピーカーから出力された音は特定の方向に偏ることなく空間全体に均一に広がるため、テレビの視聴位置の自由度が高い。

さらに、図1(c)に示すように、3.5mm オーディオケーブルを用いてテレビのヘッドホン端子とミライスピーカーのAUX IN端子を接続し、電源ケーブルをコンセントに差し込むだけで使用できる。言葉の聞き取りやすさや設置の容易さに加え、税込み20,000円未満という手頃な価格も相まって、聞こえに困難を抱える高齢者や同居家族から高い評価を得ている。

3. 自動文字起こしの認識精度の測定

3.1 方法

岡部らの研究^[5]で実施した2回のインタビューの冒頭15分の音声を用いて、自動文字起こしにおける認識精度を測定した。以降、2回分のインタビュー音声をそれぞれ音声A、音声Bと呼ぶ。両者とも男性1名を対象とするインタビュー音声で、主な質問者は男性2名と女性1名の計3名である。音声データだけでなく、人手で文字起こしをしたテキストデータも存在する。

測定環境の模式図を図2に示す。スピーカーを有するパソコンとマイクを接続したパソコンを1m離して配置し、前者で再生したインタビュー音声を後者で收音して自動文字起こしをした。パソコンのOSはいずれもWindowsであり、收音にはFIFINE製USBコンデンサマイク、文字起こしにはMicrosoft Wordのディクテーション機能を用いた。新たに高価な機材やソフトウェアを購入したり、追加のサービスを契約したりすることなく、容易に文字起こし作業ができる環境での検証が重要と考えての選定である。音声出力には、以下の3種類のスピーカーを使用した。

- ①ミライスピーカー・ミニ
 - ②Bose製のスピーカーシステム Companion 20 multimedia speaker system
 - ③Panasonic製パソコンCF-QV内蔵スピーカー
- ②はミライスピーカー・ミニよりも高価で音質が良いも

の、③は外付けスピーカーを用意せずとも使用できる手軽なものとしての比較対象である。②を使用する際には、①と同様にオーディオケーブルと電源ケーブルの接続が必要になるほか、左右のスピーカーを接続し、右のスピーカーにコントローラを接続する必要がある。以降では、各スピーカーをそれぞれMirai, Bose, PCと表記する。

3種類のスピーカーを用いて得られたテキストデータと、人手で作成されたテキストデータを日本語形態素解析システムMeCabにより単語に分割した。得られたテキストデータのうち、前者を評価単語列、後者を正解単語列とし、スピーカーごとに単語単位の認識精度 Acc を算出した。 Acc は、正解単語列の単語数 N 、削除誤り(Deletion error) D 、置換誤り(Substitution error) S 、挿入誤り(Insertion error) I を用いて、式(1)により求められる。

$$Acc = \frac{N-D-S-I}{N} \quad (1)$$

ここで、削除誤り D は正解単語列に含まれるが評価単語列には含まれない単語の数、置換誤り S は正解単語列に含まれる単語が評価単語列で別の単語に置換されている数、挿入誤り I は正解単語列に含まれず評価単語列に含まれる単語の数を表す。

評価単語列に含まれる単語が認識されなかったり、含まれていない単語が挿入されたりしたのではなく、誤認識によって語数が変化した場合には、正解単語列における単語数を基準として誤り単語数をカウントする。例えば、「人間描いた」が「人間がいた」と誤認識された場合には、「描い」という1語が「が」と「い」の2語に対応するが、「描い」を1語の誤認識とみなし、置換誤り1としてカウントする。同様に、「絵の」が「でも」と誤認識された場合、「絵」と「の」の2語が「でも」という1語に対応するが、「絵」と「の」の2語が誤認識されたとみなし、置換誤り2としてカウントする。また、かな漢字変換や表記の違いは無視し、読みの正誤で判断する。

3. 2 結果

3. 2. 1 認識精度

音声A, Bを対象としたときの3種類のスピーカーの認識精度を表1に示す。いずれの音声においても、認識精度はMiraiとBoseがPCを大幅に上回る結果となった。誤認識の内訳をみると、PCでは削除誤りが非常に多く、認識に必要な音声を十分に出力できていないといえる。MiraiとBoseを比較すると、認識精度はわずかながらMiraiがBoseを上回ったが、音声A, Bで誤認識の内訳の傾向が異なる。削除誤りについては、単純にMiraiが認識できた単語をBoseが認識できなかったのではなく、いずれか一方のみが認識できた単語、両者が認識できた

単語、両者とも認識できなかった単語が混在している。詳細については3.2.3節にて述べる。

次に、発話内容の把握に影響を及ぼさない語に着目する。ここでは、「えっと」、「あの」、「まあ」など言いよどみや思考の整理のための間投詞、「うんうん」、「ふーん」など相槌や理解の伝達のための間投詞、および「かき、髪型は」の「かき」など言い間違いに相当する部分を内容把握に影響しない語とする。内容把握に影響しない語の認識精度を表2、表1の誤りから表2の誤りを除外した認識精度を表3に示す。いずれの音声およびスピーカーにおいても、内容把握に影響しない語の認識精度は低く、結果として表3の認識精度は表1より高くなっている。内容把握に必要な他の語と比べて発声音量が小さかったり、発音が明瞭でなかったりすることに起因すると考えられる。また、相槌のための間投詞は、他の話者の発話の途中に入ることが多い点も認識精度を低下させる原因の一つといえる。MiraiとBoseの認識精度はPCよりも高いという傾向が表1～3に共通してみられる。しかし、MiraiとBoseの各誤り数と認識精度には有意な差はみられないものの、表2と表1, 3では異なる傾向が示されている。具体的には、表2ではMiraiのほうがBoseよりも削除誤りが多いが、表3では関係が逆転している。また、挿入誤りは表2ではいずれのスピーカーにおいても発生

表1 全単語の認識精度

	音声 A			音声 B		
	Mirai	Bose	PC	Mirai	Bose	PC
N	2522			2279		
D	221	273	1007	101	117	577
S	193	209	204	99	99	174
I	8	13	21	11	19	31
Acc	0.833	0.804	0.511	0.907	0.897	0.657

表2 内容把握に影響しない語の認識精度

	音声 A			音声 B		
	Mirai	Bose	PC	Mirai	Bose	PC
N	182			67		
D	85	84	117	25	22	38
S	7	8	6	2	2	7
I	0	0	0	0	0	0
Acc	0.495	0.495	0.324	0.597	0.642	0.328

表3 内容把握に影響しない語を除外した認識精度

	音声 A			音声 B		
	Mirai	Bose	PC	Mirai	Bose	PC
N	2340			2212		
D	136	189	890	76	95	539
S	186	201	198	97	97	167
I	8	13	21	11	19	31
Acc	0.859	0.828	0.526	0.917	0.905	0.667

していないが、表3ではMiraiのほうがBoseよりも少ない。以上の結果は、MiraiがBoseよりも発話内容把握に影響しない語を省略して書き起こす傾向にある可能性を示唆している。

3. 2. 2 人名と専門用語の認識

音声A、Bには、以下に示すような5つの人名と3つのAI関連専門用語が【】内の回数だけ出現する。

- (1) インタビューの姓+敬称「さん」【5回】
- (2) インタビューのペンネーム+敬称「さん」【9回】
- (3) インタビューの名【1回】
- (4) キャラクタの名+敬称「ちゃん/さん/無」【7回】
- (5) 研究協力者の姓「わかつき」+敬称「君」【2回】
- (6) StableDiffusion 【9回】
- (7) AnyRoLA 【2回】
- (8) chatGPT 【3回】

(4)の敬称は「ちゃん」が3回、「さん」が2回で、残り2回は敬称がつけられていない。(1)はMiraiとBoseで5回中3回が置換誤りとなり、うち1回は同じ語に誤認識された。(2)は両スピーカーで9回とも置換誤りとなり、うち3回は同じ語に誤認識された。一方、(3)と(4)は両スピーカーですべて正しく認識された。(1)は統計的に少ない姓、(2)は一般的な人名にはないペンネームであるが、(3)は統計的に多い名前、(4)は発音が平易で、人のニックネームやマスコットキャラクタの名前としても頻繁に使用される名前であることが、認識精度の違いにつながったといえる。

(5)～(8)のMiraiとBoseによる認識結果を表3に示す。ここで、○は正しい認識、×は削除誤りを表す。表3全体における両スピーカーの認識精度は同じであるが、Miraiには削除誤りがなく、置換誤りについても読みが大きくかけ離れた誤りはみられなかった。今回自動文字起こしに用いたディクテーション機能には、特定の誤認識を回避するための学習機能は提供されておらず、一般的に使われる語のみが認識可能な状況であった。以上の点を考慮すると、未知の言葉であっても何らかの意味をもつ可能性のある音声について、MiraiはBoseよりも言葉を表す音声として出力する傾向があると考えられる。

3. 2. 3 スピーカーごとの誤認識内容の比較

MiraiとBoseにおける削除誤りには、文の途中の助詞が認識されなかった削除誤り、および文末における1～2音が認識されなかった削除誤りが多く確認された。両者の例を表5および表6に示す。表には、Boseでは正しく認識されたがMiraiでは削除誤りとなった例、Miraiでは正しく認識されたがBoseでは削除誤りとなった例、MiraiとBoseの両方で削除誤りとなった例に分類して記載している。

表4 人名と専門用語の認識結果

正解	Mirai	Bose
わかつき君	分かつ気分 ○	わか杉文 かずき君
StableDiffusion	テーブルフュージョン テーブルディフュージョン ステイブルフュージョン ○ ○ ○ ○ ○	センブリフュージョン ○ ○ × ○ ○ ○
AnyRoLA	にローラー 何色だ	にローラー ×
chatGPT	ちゃんとGPT ○ ○	ちゃんとGPT ○ ○

表5 文の途中の助詞が欠落した削除誤りの例

スピーカー	正解	誤認識
Miraiのみ	これはアニメの 金魚を入れた	これアニメの 金魚入れた
Boseのみ	言語モデルは変換して 女性を描く 影響を受けた イラストを全部入れた そういうのをたくさん	言語モデル変換して 女性描く 影響受けた イラスト全部入れた そういうのたくさん
MiraiとBose	それはでも 結局は人間の 画像を新たに 学習させた 顔を構造的リアルに それを色濃く てるなと思う かなとは思います 絵としてはかわいい こだわられてなくて	それでも 結局人間の 画像新たに 学習させた 顔を構造的リアルに それ色濃く てるなと思う かなとは思います 絵としてはかわいい こだわられなくて

主語を示す係助詞や目的語を示す格助詞、共感や確認を表す終助詞などが認識されなくても、同じ意味内容を伝えることは可能である。しかし、削除誤りによって、網掛け部分のように伝わる意味やニュアンスが変わる場合もある。文中の助詞の欠落については、MiraiとBoseの両方で同様に意味が変わる削除誤りが発生していた。一方で、意味が変わるような文末における削除誤りはBoseに多くみられた。文中の助詞や文末の1～2音は他の語より小さい音量で発声される傾向があるため、表5および表6に示されるような削除誤りは発生しやすいと考えられる。しかし、Boseでのみ発生する文末での削除誤りが多く見受けられることから、Miraiでは意味内容が変わるような削除誤りが少なく、発話内容を正しく把握するために必要な語が適切に書き起こされていると考えられる。

次に、インタビュー音声特有の話し言葉に着目する。

表6 文末における削除誤りの例

スピーカー	正解	誤認識
Miraiのみ	なるほどね	なるほど
Boseのみ	言うんですかね ありますね おもしろいな ですけど っていうのは 入れて 感じかもしれない	言うんです あります おもしろい です っていう 入れ 感じかもしれ
MiraiとBose	良かったですね 知らないね ないですね 感じるかなあ 発信してる 運営してる	良かったです 知らない ないです 感じるか 発信して 運営して

表7 話し言葉が書き言葉として認識された例

スピーカー	正解	誤認識
Miraiのみ	あるんですかね 書いてる あんま	あるのですかね 書いている あんまり
Boseのみ	構図なんで	構図なので
MiraiとBose	そうっすね いいんすよ ないっすね ないすけど なんすけど なってる 通り過ぎてる かもしんない っていう やっぱ 頭ん中	そうですね いいんですよ ないですね ないすけど なんすけど なっている 通り過ぎている かもしれない という やっぱ 頭の中

話し言葉の認識に関しては、発話通りに認識された部分と、話し言葉が書き言葉として認識された部分の両方がみられた。後者の例を表7に示す。MiraiとBoseの両方で同様の誤認識が多くみられたが、意味内容が変わるような誤認識はなかった。

MiraiとBoseで異なる語に誤認識された例を表8、その他の誤認識の例を表9に示す。読みが大きく異なり、認識の理由が理解できないような誤認識はみられないが、スピーカーの違いによらず、意味は大きく変わっている。

4. おわりに

本研究では、ミライスピーカーの新たな活用分野を見出すため、インタビュー音声の文字起こしにおける認識精度を測定した。高音質で高価なBOSE製スピーカーおよびパソコン内蔵スピーカーと比較した結果、ミライスピーカーの認識精度はパソコン内蔵スピーカーを大きく上回り、BOSE製スピーカーと同等、あるいはわずかに優れていることが確認された。インタビューデータを分析する研究者が認識精度80～90%をどのように評価するかについては、今後の調査が必要である。しかし、価格や設置の容易さを考慮すると、ミライスピーカーの採用には十分な利点があるといえる。

表8 MiraiとBoseで異なる誤認識をした例

正解	Miraiの誤認識	Boseの誤認識
画風の 自動で アニメとか萌え系の 福笑い 福笑い 感じかなさっきの 学習量の アニメ特化の アート系の絵 この絵には 脳の	爆風の 実際 アニメとか防衛系の 服あれ 少ない 感じかな先 終了の アニメとかの パンといけない この世には ものを	ガフの 実際に アニメとかも理系の 福原 服はない 感じ川崎 主要な アニメどっかの あのいけない このようには 南野

表9 その他の誤認識の例

スピーカー	正解	誤認識
Miraiのみ	今回 一回 煽りに 商業漫画家なんで 絵として 解釈で 動画で 生成に 二段に 構図の 書き方とか 少しいじりながら 海外 アートっぽく	公開 いっぱい 青に 所有漫画なんで 円として 会社で とかで 生産に にらに 鴻巣の 株価とか 少し一時半から 彼が アドっぽく
Boseのみ	平面的な 作家さん的には 構図が AIは プロンプトを アニメの絵で 置いたり 習ったり バストアップ 出した結果 絶対 押せば 触らない	天然的な サッカーのときには 坊主が 変愛は ポンプを アニメのうえで 降ったり なかったり バックアップ だしだけか ずっと お世話 変わらない
MiraiとBose	できてる最中 細部まで 画風ですけど 〇〇さんの絵感 目は 学習した 立体視して アニメ絵の 荒かったり パッと見では 人間描いた 配置してる シンクロしやすい 取り込んでる 書くうえで むしろ ラフっぽく	できてる最初 最後まで 歌風ですけど 〇〇さんの印鑑は レイは 握手した 立体して アニメの あったり パッと見では 人間がいた 愛してる 進行しやすい 取り組んでる 革命で 虫を lovepop

今回の測定結果から、ミライスピーカーでは内容把握に影響を及ぼさない語が認識されにくい可能性が示唆された。インタビューデータの分析において、内容把握に影響しない語を分析対象に含めるか否かは分析の目的によって異なる。研究者が目的に応じて適切なスピーカーを選定できるよう、文中の助詞や文末の1～2音の削除誤

り、話し言葉の認識傾向を含めた詳細な調査が求められる。

ミラリスピーカーの開発過程では、聞こえに困難を抱える日本人高齢者を対象として言葉の聞き取りやすさを評価し、最適な振動板の形状を決定したため、日本語に特化した特性をもつと考えられる。角田^[6]は、人の感情的な声、虫の音、鳥や動物の鳴き声、風・波・雨などの自然音を識別する脳の部位が、日本語話者と他言語話者で異なることを実験により明らかにした。ミラリスピーカーが日本語に特化している場合、自然音などの言葉以外の音声の聞こえ方も日本語話者向けになっていると考えられ、インタビュー音声に含まれる言語以外の音声が認識精度に影響を及ぼす可能性がある。今後は、日本語以外の言語における有効性を検証するとともに、言語以外の音声を対象にした調査も必要である。

謝辞

本研究は東京都市大学メディア情報学部の研究助成を受けて実施されました。また、測定に使用したミラリスピーカー・ミニは株式会社サウンドファンより貸与を受けました。ここに深く感謝申し上げます。

参考文献

- [1] 太田有美, “加齢性難聴の病態と対処法”, 日本老年医学会雑誌, Vol.57, No.4, pp.397-404, 2020.
- [2] 立木孝, 笹森史朗, 南吉昇, 一戸孝七, 村井和夫, 村井盛子, 河嶋寛, “日本人聴力の加齢変化の研究”, AUDIOLOGY JAPAN, Vol.45, No.3, pp.241-250, 2021.
- [3] 大坪寛子, 国広陽子, “高齢者にとってのテレビ: 記憶の中のテレビと現在のテレビ視聴”, メディア・コミュニケーション: 慶応義塾大学メディア・コミュニケーション研究所紀要, No.62, pp.107-119, 2012.
- [4] 総務省情報通信政策研究所, “令和5年度情報通信メディアの利用時間と情報行動に関する調査報告書”, 2024.
https://www.soumu.go.jp/main_content/000976455.pdf
- [5] 岡部大介, 平綿素望, 若月玲之, 大谷紀子, “画像生成AIと対峙した専門家の「作風」にかかる語りの分析”, 認知科学, Vol.32, No.1, 2025.
- [6] 角田忠信, “日本語人の脳”, 言叢社, 2016.