社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

立体音響で教える全方位単語感知システム

宮田　武嗣†　岩村　雅一†　黄瀬　浩一†

† 大阪府立大学大学院工学研究科
〒 599–8531 大阪府堺市中区学園町 1–1
E-mail: miyata@m.cs.osakafu-u.ac.jp, {masa,kise}@cs.osakafu-u.ac.jp

あらまし　私たちは身の回りにある文字情報を見て必要な情報を取り入れるが，視覚障害者はそれができない．そのような人が視覚の代わりに触覚や聴覚を使って情報を取り入れられるような補助ツールとして点字や音声案内がある．しかし点字は設置している箇所に辿り着けなければ利用できないことや，音声案内はむやみに設置すると健常者にとっては騒音となる問題がある．それに対してこれまで様々なウェアラブルインタフェースが提案されてきたが，どこに文字情報があるか分からなければ使用できないことや，正面以外の方向にある文字情報を認識できないといった問題がある．そこで全方位から文字情報を探し出し，何という文字情報がどこにあるか声で教えるウェアラブルインタフェースを提案する．このインタフェースでは，実際に文字情報が位置する方向から声が聞こえるように，立体音響を用いて文字情報を読み上げる．本稿では，文字情報の位置を教えるために立体音響を用いることが適切がどうか確かめるべく，評価実験を行った．

キーワード　視覚障害者，ウェアラブルインタフェース，全方位，立体音響

1. はじめに

私たちには五つの感覚器官があるが，人間が外界から得る情報の 8 割は視覚からもたらされると言われている．しかし視覚に障害を持ち，目から情報を取り入れられない人が世界に約 3900 万人いる [1] とされており，そのような人は生活する上で不便が多いと考えられる．そのような不便を減らすために，目から取り入れる情報を他の感覚器官から取り入れることができれば，大きな助けとなる．そのような補助ツールとして街中の点字や音声案内がある．点字や音声案内は，視覚の代わりに触覚や聴覚を使って情報を伝えるものであるが，点字は設置箇所に辿りつけなければ利用できないことや，凹凸パターンを覚えなくては情報を読み取れないといった問題がある．それに比べると，音声案内は設置箇所に辿りつくことや内容を理解することは容易である．しかし，音声案内をむやみに設置すると健常者にとっては騒音となる可能性や，音声案内により得られる情報は利用者がその時に求めている情報と必ずしも一致しないといった問題がある．それを解決するためには，音声案内を個人化することが一つの解決策になる．

音声案内の個人化とは，スピーカーで多数の人に同じ情報を伝えるのではなく，例えばヘッドホンを使って個人毎に別の情報を伝えることを意味する．その場合，どういった情報を使用者に伝えるか，その情報をいかにヘッドホンを制御しているデバイスに送るが次に問題になる．後藤らは現在地の案内や目的地までの誘導を声でする杖型デバイスを提案している [2]．この杖の先端には RFID タグを読み取るアンテナが取り付けられており，あらかじめ誘導ブロックに埋め込まれた RFID タグから位置データを取得する．しかし，RFID タグの設置コストの関係から大量に設置できないことが最大の問題である．したがって，使用者に伝える情報は設置コストのかからない方法で環境から取得できる方法が望ましい．

このような観点から，文字情報は最も有力な情報源となり得る．私たちの身の回りには文字情報が沢山あることや，健常者も利用することから特段のコストを掛けずとも情報が更新し続けられることがその利点である．そのため，文字情報を利用した装着型のインタフェースは既にいくつか提案されている．眼鏡型インタフェースである OTON GLASS(注1) は使用者の視線付近の文字情報を声で読み上げる．これは失語症や弱視，外国人のような，文字情報自体は読めないが，読みたい文字情報の場所が分かる人向けのインタフェースである．同様に，指先でなぞった文字情報を声で読み上げるものに FingerReader [3] がある．FingerReader には文の行を正しくなぞれるように振動によって誘導する機能があるが，どこに文字情報があるか分からなければ使用できないのは OTON GLASS と同じである．文字情報の位置が分からなくても使用できるものに，ウェアラブルカメラで撮影した正面の画像から文字情報を探し出し，それを声で読み上げる Yi らのインタフェース [4] がある．しかし，このインタフェースは何という文字情報があるか教えるだけで，その文字情報がどこにあるかは教えない．それを解決したものに，何という文字情報がどこにあるか声で教える Goto らのインタフェース [5] がある．このインタフェースでは，文字情報

(注1)：https://medium.com/@OTONGLASS/what-is-oton-glass-f62ed8318dd0#.73zkz8azi

の左右の位置の違いは左右のヘッドホンの音量の違いで表し，上下の位置は声のトーンで表す．しかし，このインタフェースで使用しているカメラは正面しか撮影できないため，正面にある文字情報のみを対象としている．得たい情報が正面ではなく側面にあった場合，健常者は周りの環境からそれを予測し振り向くことができるが，視覚障害者は周りの環境を得ることができず，どこに文字情報があるか予想できない．したがって，正面だけの撮影では得たい情報を逃してしまうと考えられる．

そこで本稿では，全方位から文字情報を探し出し，何という文字情報がどこにあるか声で教えるウェアラブルインタフェースを提案する．このインタフェースでは全方位カメラを使用することで，正面だけではなく一度に全方位の画像を撮影できる．そして，文字認識技術を用いて全方位画像に含まれる文字を認識し，その結果を声で読み上げる．正面にある文字情報を対象とする Goto らのインタフェースは音量や声のトーンの違いだけで文字情報の位置を表すことができるが，このインタフェースは全方位にある文字情報を対象とするため，それだけでは全方位のどこに文字情報が位置するのか完全には分からない．そこで，音声を再生する際に立体音響を用いることで，実空間での文字情報の位置を再現する．これにより，実際に文字情報が位置する方向から音声が聞こえるように，文字情報を読み上げることができる．この装着型のインタフェースは，生活する上で常時装着することを想定する．そのため，このインタフェースに求められる要件として長時間装着していても不快に感じないことが挙げられる．このインタフェースに取り付けられている全方位カメラは軽量であるため，提案するインタフェースの重さは一般的なバイク用ヘルメットの重さとあまり変わらない．そのため，このインタフェースを長時間装着していても，不快に感じられないことが期待される．

2. 関 連 研 究

本節では，視覚障害者の補助を目的とした関連研究について述べる．文字情報を読む行動を声で補助するものに OTON GLASS がある．OTON GLASS は二つのカメラを持ち，アイカメラはユーザーの視線を取得し，シーンカメラは利用者が読みたい文字情報を撮影する．これは文字情報を読むのに時間がかかったり，読み間違える人には便利であるが，全く見えない人は使用できない．指先でなぞった文字情報を声で読み上げるものに FingerReader [3] や EyeRing [6] がある．これらを利用することで，これまで点字のある特別な書籍でしかできなかった読書を一般の印刷物でもできるようになる．視覚障害者にとって，文章を正確になぞることや文の改行を判断することは難しいが，FingerReader は文の行を検出し，4 つの振動モーターを使って指先に移動方向を教える．EyeRing には通貨を教えたり，物体の色を教えるといった機能もある．これらはどこに文字情報があるか分からない場合には使用できないといった問題がある．ウェアラブルカメラで撮影した正面の画像から文字情報を探し出し，その文字情報を声で教えるものに [4, 7, 8] がある．これらは，どこに文字情報があるか正確に分かっていなくても使用できるが，何という文字情報があるかを教えるだけであり，その文字情報がどこにあるかは教えない．我々の研究に最も似たものとして，何という文字情報がどこにあるか声で教える Goto らのインタフェース [5] がある．このシステムでは，文字情報の左右の位置の違いは左右のヘッドホンの音量の違いで表し，上下の位置は声のトーンで表す．そして，骨伝導ヘッドホンを用いることで外界の音を遮断せずに聴覚で情報を受け取ることができる．このシステムで使用しているカメラは正面しか撮影できないため，使用者は正面にある文字情報しか知ることができない．正面だけの撮影では取り入れたい情報を逃してしまうのは前述の通りである．

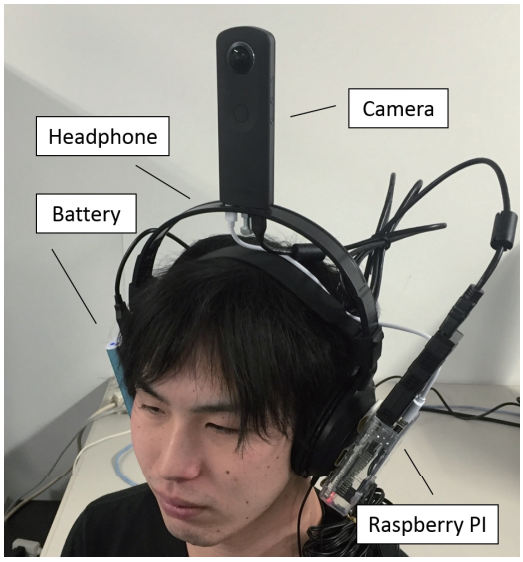


図 1　インタフェースの装着図

3. 提案システム

提案システムは，全方位文字認識と立体音響を組み合わせることで，使用者の周囲にある文字情報が何という文字かそしてどの方向にあるのかを声で教える．インタフェースの装着例を図 1 に示す．

3.1 インタフェースの構成

インタフェースは，図 2 で示すように主に 4 つの部品で構成される．一つ目は，全方位カメラの RICOH THETA S である．本インタフェースでは，このカメラを HDMI キャプチャボードに通してウェブカメラとして使用する．このカメラには二つの魚眼レンズが前後に搭載されており，二つの方向を同時に撮影する．この時，得られる画像の解像度は 1920 × 1080 であり，この中に二つの魚眼画像が含まれる．そして，二つの魚眼画像を繋ぎ合わせることで，全方位の画像が得られる．二つ目は，小型パソコンの Raspberry Pi 2 ModelB である．Raspberry Pi はカメラの映像を取得し，計算サーバーへ転送する．そして，サーバーから音声ファイルを受け取り，ヘッドホンで再生する．また，Raspberry Pi には HDMI キャプチャボード (FEBON168) と無線 LAN 子機 (BUFFALO WLI-UC-GNME) を取り付けている．HDMI キャプチャボードを使うことによって，USB 接続で RICOH THETA S に接続するときと比べて，高解像度な画像を取得することができる．また，無線 LAN 子機はサーバー

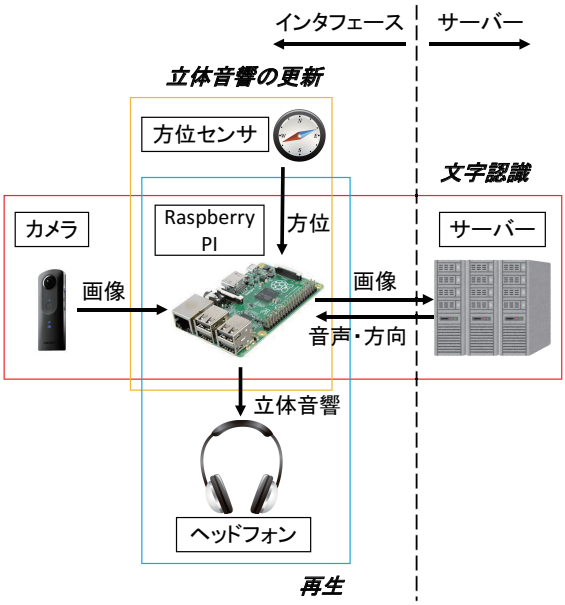


図 2　システムの構成

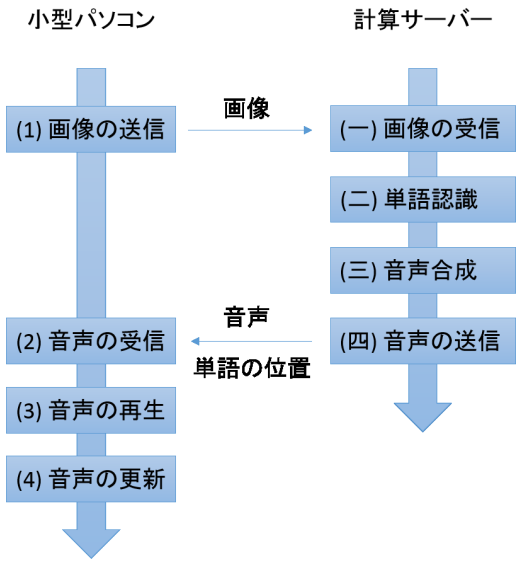


図 3　システムの処理の流れ

と接続するために使用する．Raspberry Pi の CPU は ARM Cortex-A7 900Mhz，メモリは 1GB である．そして三つ目は，ヘッドホン (audio-technica ATH-TAD300) である．外出中にこのインタフェースを装着しても周りの音が聞こえるようにオープン型のヘッドホンを使用する．そして四つ目は，頭部の回転を検出するために使用する方位センサ (HMC5883L) である．その他に，Raspberry Pi の電源用として使用するモバイルバッテリーがある．モバイルバッテリーの容量は 5600mA であり，これはシステムを半日稼働させるのに十分な容量である．

RICOH THETA S の重さは約 125g，スマートフォンの重さは約 125g，ヘッドホンの重さは約 467g である．そのため，提案するインタフェースの重さは約 717g である．頭部に装着するものとしてバイク用ヘルメットが挙げられるが，一般的なバイク用ヘルメットの重さは 1.0kg を超えており，提案インタフェースはバイク用ヘルメットよりも軽いと言える．

3.2　処理の流れ

提案システムの処理の流れを図 3 に示す．図中の (1) や (一) といった数字は，3.1 節や 3.2 節の数字とそれぞれ対応する．提案システムの処理は，小型パソコンと計算サーバーに分けられる．主に，小型パソコンはインタフェースの処理を，計算サーバーは文字認識をする．

3.3　小型パソコン

小型パソコンでは，次のように処理が行われる．(1) カメラの画像を取得し，それを計算サーバーへ送信する．この時，計算サーバーへ送信する画像は魚眼画像である．魚眼画像は歪みが生じているため，それに含まれる文字も歪んでしまう．そのため，魚眼画像のまま文字認識をすると，認識精度が著しく低下すると考えられる．したがって，魚眼画像を他の図法に変換する必要があるが，この処理は計算量が多い．そこで，この処理を計算サーバーに委託することで，処理の高速化を図る．(2) 計算サーバーから認識結果を受け取る．計算サーバーから受け取るものは，認識結果の音声ファイルと単語の位置する方向である．(3) 音声を再生する．音声を聞くだけで単語の位置が分かるようにするため，立体音響を用いる．立体音響は立体空間での音環境を再現する技術である．この技術では，立体空間での音源と聞き手の相対位置によって，音量や耳に到着するまでの時間などを変えることにより，音の変化を作り出す．提案システムでは使用者を原点にし，実世界で単語が位置する方向に音源を配置する．これにより，実際に単語が位置する方向から，単語を声で読み上げることができる．提案システムでは使用者と単語との距離を測ることができないため，使用者と単語との距離は常に一定値とした．(4) 音声を顔の向きの変化に応じて更新する．提案するインタフェースには 3 地軸磁気センサを取り付けており，このセンサの値を使うことで頭部の向きの変化を検出して，頭部の向きが変われば，それに応じて音声を更新する．これにより，使用者が振り返ったとしても実際に文字情報が位置する方向から声が聞こえるというように，音声をリアルタイムに更新できる．文字情報の方向が分かりにくい場合でも，頭の向きを変えて，声の聞こえ方の違いを感じ取ることで，より明確に文字情報の位置を知ることができる．

3.4　計算サーバー

計算サーバーでは，次のように処理が行われる．(一) 小型パソコンから 2 つの魚眼画像を受け取り，それを正距円筒図法に変換する．正距円筒図法は，地球投影法の一つで，緯線と経線が直交かつ等間隔になるように変換する方法である．(二) 松田らの手法 [9] で単語を認識する．松田らの手法は文字認識手法であるが，データベースに単語画像を登録することによって単語認識として使用する．初めに，(一) で変換された画像から局所特徴量を抽出する．次に，抽出された局所特徴量をあらかじめデータベースに登録されている局所特徴量とマッチングする．そして，マッチングされた局所特徴量の配置を使って単語を認識する．松田らの手法では，局所特徴量のマッチングに近似最近傍探索 [10] を用いているため，高速に認識できる．このよう

表 1　正解率と認識時間

	正解率 [%]	認識時間 [s]
実験 1	39.3	8.0
実験 2	61.3	15.9
実験 3	61.3	9.9

に高速な文字認識手法を用いることで，このインタフェースは認識結果の高頻度な更新ができる．(三) Open JTalk [11] を用いて音声ファイルを生成する．(四) 小型パソコンへ音声ファイルと単語の位置を送信する．

4.　提案システムの評価

4.1　評 価 実 験

本実験では，文字情報の位置を教えるために立体音響を用いることが有効であるかを確かめる．提案システムでは，単語が位置する方向から声が聞こえるように立体空間に音源を設置する．設置箇所は，水平方向は中心角を 30 度ずつに区切った 12 方向，鉛直方向は仰角を 0 度と固定し，計 12 種類である．使用者と音源の距離は常に一定とした．再生する声は Mei(注2) で音声合成した約 2 秒の“大阪府立大学”である．実験の手順として，初めに 12 種類の声の聞こえ方の違いを確認してもらうために，0 度～330 度の音声を順に再生する操作を 2 回繰り返した．そして次に音源を 12 種類の中からランダムに設定し，どの方向から聞こえるか回答をもらった．このとき，音声は回答をもらうまで再生し続け，音声の再生開始から回答までの時間を認識にかかる時間とした．同様に，12 種類全ての方向に対してこの操作を行った．実験 1 では，被験者の頭の向きを固定し，立体音響による声を聞くだけで単語が位置する方向が分かるか検証した．実験 2 では，被験者は自由に頭の向き変えて，声の聞こえ方の違いを感じ取ることでより明確に単語が位置する方向が分かるか検証した．実験 3 では，実験 2 に加えて，単語を正面にしたらビープ音を鳴らすことでより明確に単語が位置する方向が分かるか検証した．正解率と認識時間を表 4.1 に示す．実験 1 と実験 2 を比較すると，声の聞こえ方の違いを感じ取ることでより明確に単語が位置する方向が分かったと言える．認識時間が増えたのは頭の向きを変えて音の違いを確かめるのに時間を要したと考えられる．実験 2 と実験 3 を比較すると，単語を正面にしたらビープ音を鳴らすことでより素早く単語が位置する方向が分かったと言える．

4.2　アンケート

提案インタフェースの有効性を示すため，アンケートを実施した．被験者は 20 代の男女 14 人である．アンケートは選択形式で 8 題設問し，それぞれ 5 つの選択肢から選んでもらった．質問内容とその選択肢を以下に示す．

Q1　このインタフェースを長時間装着すると不快か

Q1 選択肢　(1) 全く不快ではない (2) 不快ではない (3) どちらでもない (4) 不快 (5) 非常に不快

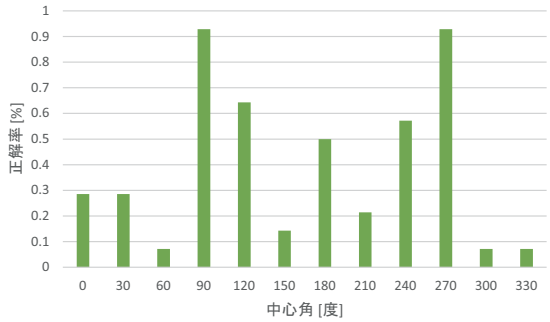


図 4　角度毎の正解率 (実験 1)

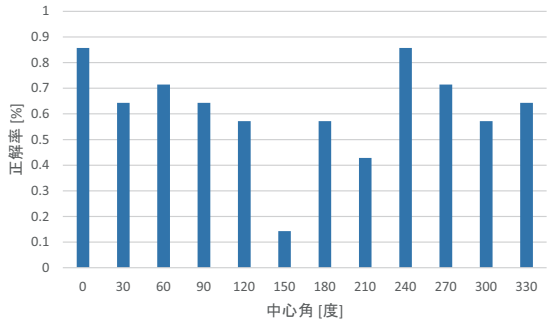


図 5　角度毎の正解率 (実験 2)

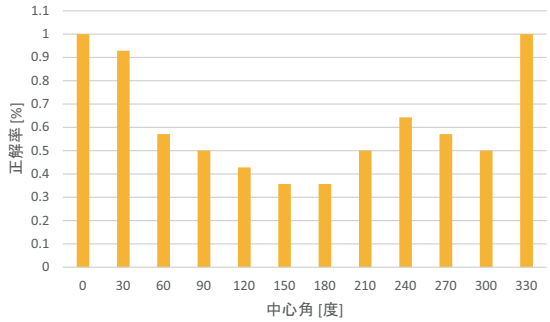


図 6　角度毎の正解率 (実験 3)

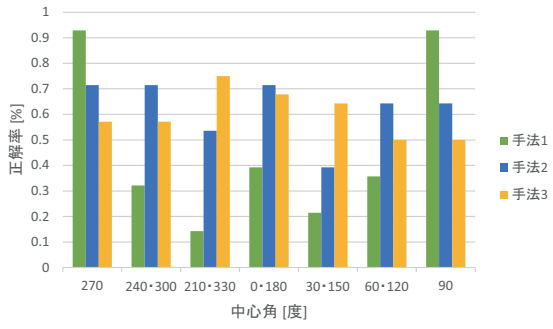


図 7　正面と背面の差を除いた場合の正答率

Q2　このインタフェースは頭部に装着するには重いか

Q2 選択肢　(1) 全く重くない (2) 重くない (3) どちらでもない (4) 重い (5) 非常に重い

(注2)：http://www.mmdagent.jp/

Q3　このインタフェースを装着して外出するのは恥ずかしいか
Q3 選択肢　(1) 全く恥ずかしくない (2) 恥ずかしくない (3) どちらでもない (4) 恥ずかしい (5) 非常に恥ずかしい

Q4　このインタフェースを装着して外出すると危険か
Q4 選択肢　(1) 全く危険ではない (2) 危険ではない (3) どちらでもない (4) 危険 (5) 非常に危険

Q5　このインタフェースを装着すると周りの音が聞こえなくなるか
Q5 選択肢　(1) 全く変わらない (2) 変わらない (3) どちらでもない (4) 聞こえない (5) 全く聞こえない

Q6　このインタフェースを装着すると動き辛いか
Q6 選択肢　(1) 全く動き辛くない (2) 動き辛くない (3) どちらでもない (4) 動き辛い (5) 非常に動き辛い

Q7　単語の位置は分かりやすいか (実験 1)
Q7 選択肢　(1) 非常に分かりやすい (2) 分かりやすい (3) どちらでもない (4) 分かりづらい (5) 非常に分かりづらい

Q8　単語の位置は分かりやすいか (実験 2)
Q8 選択肢　(1) 非常に分かりやすい (2) 分かりやすい (3) どちらでもない (4) 分かりづらい (5) 非常に分かりづらい

Q9　単語の位置は分かりやすいか (実験 3)
Q9 選択肢　(1) 非常に分かりやすい (2) 分かりやすい (3) どちらでもない (4) 分かりづらい (5) 非常に分かりづらい

アンケート結果を図 8 に示す.

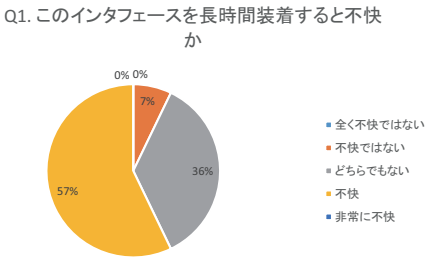
5. まとめと今後の課題

本稿では，視覚障害者の補助を目的としたウェアラブルインタフェースを提案した．このインタフェースは全方位カメラで周囲を撮影し，その中に含まれている文字を認識する．そして立体音響を使って，文字情報が位置する方向からその文字情報を読み上げる．それにより視覚障害者でも何という文字情報がどの方向にあるか知ることができる．検証実験の結果，立体音響で文字情報の位置を教えるには精度が不十分であった．今後の課題として，正面と背後といった立体音響では区別しづらい方向の音を識別できるようにするために，それらを区別する合図を導入することが考えられる．またアンケートの結果，提案インタフェースを長時間装着することは不快であることが分かった．その原因としてインタフェースが重いこと挙げられるため，スマートフォンを用いるなどインタフェースの小型化をする必要がある．

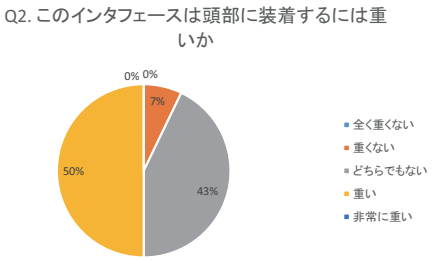
謝辞　本研究は，JST CREST ならびに JSPS 科研費 25240028 の補助による．

文　献

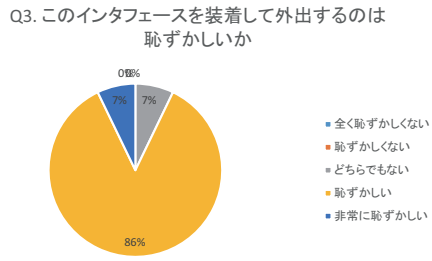
[1] S.P.M. Donatella Pascolini, “Global estimates of visual impairment: 2010,” British Journal of Ophthalmol, vol.96, no.5, pp.614–618, 2012.
[2] 後藤浩一，松原　広，深澤紀子，水上直樹，“駅環境における携帯端末を用いた視覚障害者向け情報提供システム,” 情報処理学会論文誌，vol.44，no.12，pp.3256–3268，Dec. 2003.
[3] R. Shilkrot, J. Huber, W. Meng Ee, P. Maes, and S.C. Nanayakkara, “FingerReader: A wearable device to explore printed text on the go,” Proc. of 33rd Annual ACM Conference on Human Factors in Computing Systems, pp.2363–2372, 2015.
[4] C. Yi and Y. Tian, “Camera-based document analysis and recognition: Post-proc. of 4th international workshop camera-based document analysis and recognition (cbdar 2011),” vol.7139, chapter Assistive Text Reading from Complex Background for Blind Persons, pp.15–28, Lecture Notes in Computer Science, Springer, 2012.
[5] H. Goto, “Text-to-speech reading assistant device with scene text locator for the blind,” Proc. Assistive Technology: From Research to Practice, pp.702–707, 2013.
[6] S. Nanayakkara, R. Shilkrot, and P. Maes, “Eyering: A finger-worn assistant,” CHI ’12 Extended Abstracts on Human Factors in Computing Systems, pp.1961–1966, 2012.
[7] M.A. Mattar, A.R. Hanson, and E.G. Learned-Miller, “Sign classification using local and meta-features,” IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2005, San Diego, CA, USA, 21-23 September, 2005, p.26, 2005.
[8] N. Ezaki, M. Bulacu, and L. Schomaker, “Text detection from natural scene images: towards a system for visually impaired persons,” Proc. of 17th International Conference on Pattern Recognition (ICPR 2004), vol.2, pp.683–686Vol.2, Aug. 2004.
[9] T. Matsuda, M. Iwamura, and K. Kise, “Performance improvement in local feature based camera-captured character recognition,” Proc. of 11th IAPR International Workshop on Document Analysis Systems (DAS2014), pp.196–201, April 2014.
[10] M. Iwamura, T. Sato, and K. Kise, “What is the most efficient way to select nearest neighbor candidates for fast approximate nearest neighbor search?,” Proc. 14th International Conference on Computer Vision (ICCV 2013), pp.3535–3542, Dec. 2013.
[11] 大浦圭一郎，酒向慎司，徳田恵一，“日本語テキスト音声合成システム Open JTalk,” 日本音響学会春季講論集，第 I 巻，pp.343–344，March 2010.



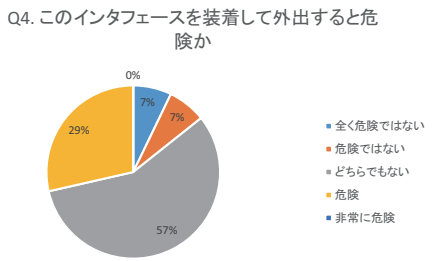
(a) Q1 結果



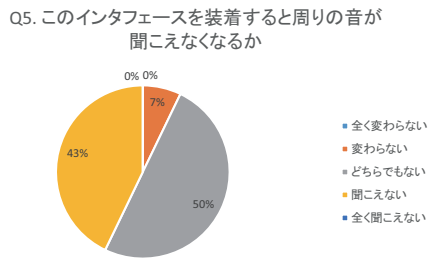
(b) Q2 結果



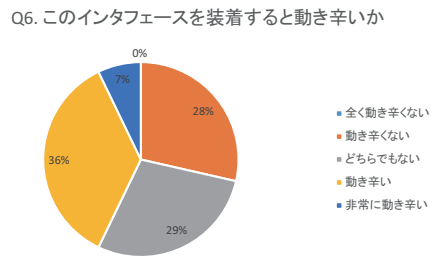
(c) Q3 結果



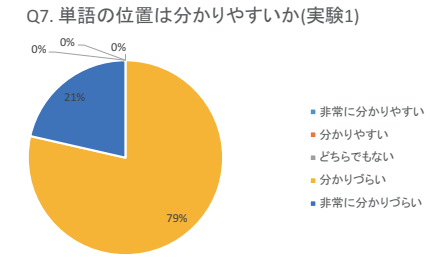
(d) Q4 結果



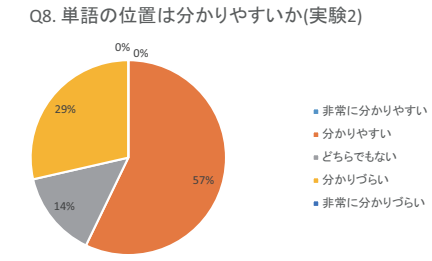
(e) Q5 結果



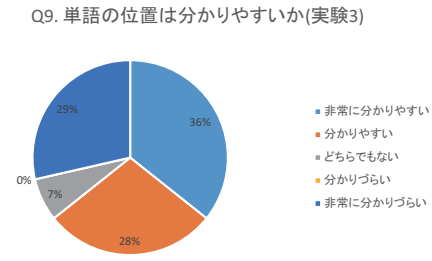
(f) Q6 結果



(g) Q7 結果



(h) Q8 結果



(i) Q9 結果

図 8　アンケート回答結果