Vol.2016-AAC-1 No.7
2016/7/30

視覚障害者のための環境文字情報提示システムの検討

岩村 雅一[1,a)]　宮田 武嗣[1]　程 征[1,b)]　田井中 渓志[1,c)]　黄瀬 浩一[1,d)]

概要：本稿では，視覚障害者が環境中に存在する文字情報を利用できるようにする情報提示システムを検討する．文字認識技術を利用して視覚障害者に文字情報を提示する試みは過去にもあるが，いずれも文字が存在する場所，あるいはその範囲がわかっている場合に使用できるように設計されている．視覚に障害がある人の中には失語症のように，文字を指し示すことのできる人もいるが，全盲や弱視の人にとって，文字を自力で発見することは必ずしも容易でないと考えられる．そのため，従来からあるような文字の読み上げ機能や，文字の場所を示す機能に加えて，文字を発見することに工夫する．具体的には，利用者の前方だけでなく，後や横を含めた周囲の文字を小型の全方位カメラを利用して発見する．これにより利用者が自ら文字を発見する必要はなくなるものの，より広範囲をカバーすることが必要になるため，従来からある文字の場所を示す方法が使用できなくなる．そのため，立体音響と頭の向きを実時間で追従するインタフェースを組み合わせることで場所を示すことを検討する．前報 [1] では，オープン型のヘッドホンを使用したシステムを提案したが，オープン型とはいえヘッドホンで耳を塞いでしまうと音に敏感な視覚障害者の貴重な情報源を塞いでしまうのではないかというコメントを受けて，本稿では耳を塞がない骨伝導イヤホンを用いた場合の性能を評価する．

1. はじめに

2010 年の推計では，世界の視覚障害者は 2 億 8500 万人で，そのうち 3900 万人が全盲，それ以外は弱視とされている [2]．健常者は情報の多く (8 割とも言われている) を視覚から得ているが，弱視の人が視覚から得られる情報は限定的であり，視覚に障害がある人は視覚以外の感覚器官から代わりの情報を得ている．本稿では，このような人を支援するインタフェースについて検討する．特に，家の中のような慣れ親しんだ場所ではなく，街中などで利用できるものを考える．

健常者が視覚で得る情報を代わりに伝える手段としては，触覚を利用した点字や，聴覚を利用した音声案内などがある．

点字は，使いこなすことができれば充実した読書生活を送ることができるなどの利点があるものの，その習得には訓練が必要で，特に中途失明者にとっては難しいため，視覚障害者の 10%しか点字を使えない[*1]．また，街中で点字案内板を利用する場合，点字がどこに設置されているのかを把握して，その場所に辿りつかなければいけないという制約がある．

音声案内は，点字を習得していない視覚障害者や健常者でも利用でき，また前述の「点字がどこに設置されているのか」を伝えることもできるという点で優れている．その一方で，音声案内はその性質上，不特定多数に同じ情報を伝えるブロードキャスト型の情報伝達手段であるため，その情報を欲していない人にも否応無しに情報を伝えてしまう．そのため，音声案内をむやみに設置すると騒音になってしまう，あるいは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う問題がある．

この問題を解決する方法の一つとして，音声案内を個人化することが考えられる．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利用者毎に別の情報を伝えることを意味する．このような事を実現しようとすれば，

[1] 大阪府立大学大学院工学研究科，大阪府堺市中区学園町 1 ー 1
Graduate School of Engineering, Osaka Prefecture University, 1-1 Gakuencho, Naka, Sakai, Osaka 599-8531, Japan
[a)] masa@cs.osakafu-u.ac.jp
[b)] zheng386@m.cs.osakafu-u.ac.jp
[c)] tainaka@m.cs.osakafu-u.ac.jp
[d)] kise@cs.osakafu-u.ac.jp

[*1] http://isee-movement.org/

課題 1　各利用者にどのような情報をどのタイミングで伝えるのがいいのか，

課題 2　それらの情報をいかに各利用者に伝送するのか，

などが課題となる．このカテゴリの既存の研究としては，後藤ら杖型デバイス [3] がある．このデバイスは，現在地の案内や目的地までの誘導を声で行うものである．あらかじめ誘導ブロックに RFID タグが埋め込まれており，杖の先端に取り付けられたアンテナで RFID タグを読み取ることで位置データを取得する．このデバイスの利用を考える上での問題は，RFID タグを環境中に設置・維持・管理する必要があることである．

これまで見てきた点字，音声案内，RFID タグはいずれも環境中に何らかの装置を設置するために追加コストが必要であった．それは，健常者が利用しないサービスを視覚障害者のためだけに提供しているからである．したがって，健常者も利用する方法を視覚障害者も利用できれば，装置の設置に要する追加コストを必要としないため，望ましいと考える．

これらの事から，本稿では，環境中の文字情報を利用し，前述の課題 2 を解決することを考える．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．

(1) OTON GLASS[*2]
眼鏡型インタフェースである OTON GLASS 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，文字の場所はわかるものの，文字を読んで文字情報を得ることができない人向けのインタフェースである．

(2) FingerReader [4]
指先に付けたカメラで文字を認識して読み上げる FingerReader が提案されている．このインタフェースを使用するには，文字のある場所を指差す必要があるため，完全に視力を失った人向けというより，弱視の人向けと考えるべきであろう．一度文字行を指差すことができれば，その後は文字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．

(3) Yi らのインタフェースがある [5]
ウェアラブルカメラで撮影した画像から文字情報を探し出し，それを声で読み上げるインタフェースが提案されている．読みたい文字が目の前にあると想定しているため，文字情報の位置を利用者に教える必要はないという立場である．

(4) Goto らのインタフェース [6]
文字情報の位置を利用者に教える必要があるという立場から，Goto らは文字情報を読み上げる際に文字の場所を利用者に伝えるインタフェースを提案している [6]．このインタフェースでは，文字情報の左右の位置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．このインタフェースで使用したカメラは正面しか撮影できないため，利用するには，読みたい文字をカメラの正面に移動する必要がある．しかし，そのためには利用者が文字の場所を把握している必要がある．しかし，文字の場所がわかるのであれば，そもそも文字の位置を利用者に知らせる必要はない点において矛盾がある．

このように，既存手法はどれも帯に短したすきに長しである．

本稿では，利用者は文字の位置がわからないことを想定し，文字情報を伝える際に位置を利用者に教える必要があるという立場を取り，新たなウェアラブルインタフェースを提案する．全方位カメラを用いて利用者の周囲 (全方位) の画像を一度に撮影し，その中から文字を探し出し，文字情報の読み上げの際に文字の方向もわかるようにする．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で文字情報の位置を表すことができるが，提案する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て，特定の方向から文字の読み上げ音声が聞こえることで文字情報の位置を提示する．このような機能を持つインタフェースは，全盲の人のみでなく，弱視の人にも有用と考える．本稿では，立体音響を使用することで，文字の位置がどの程度わかるかを実験により検証する．具体的には，前報 [1] で使用したオープン型ヘッドホンと本稿で新たに導入する骨伝導イヤホンのそれぞれを用いた場合を比較し，性能を評価する．

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．

前節の (1) で述べたように，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．

前節の (2) で述べたように，指先でなぞった文字情報を声で読み上げるものに FingerReader [4] や EyeRing [7] 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

[*2] https://medium.com/@OTONGLASS

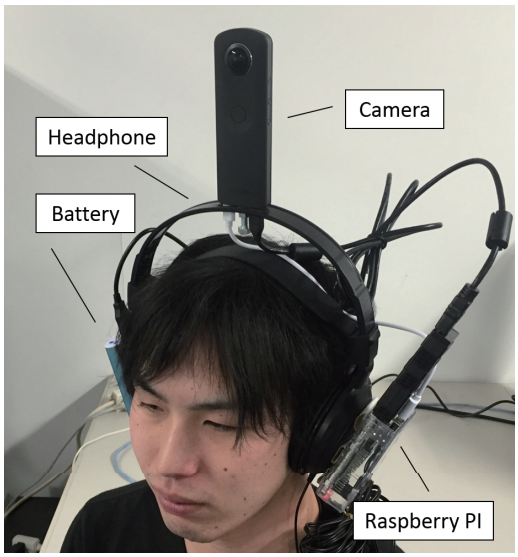

図 1　ヘッドホンを使用した場合のインタフェースの装着図

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．

前節の (3) で述べたように，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，その文字情報を声で教えるものに [8], [9], [10] 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．

前節の (4) で述べたように，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 [6] 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．

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．ここでは，前報 [1] で提案したヘッドホンを使用したシステムについて述べる．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

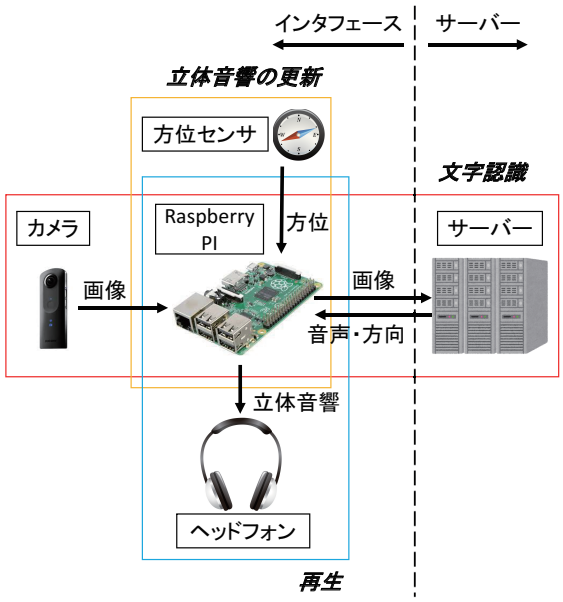

図 2　システムの構成

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．

Raspberry Pi の重さは 45g，RICOH THETA S は約 125g，ヘッドホンは約 230g である．バッテリーの重さが 159g であるため，提案するインタフェースの重さは約 559g 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れられるが，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の約半分程度の重さと言える．

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 [11] 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．

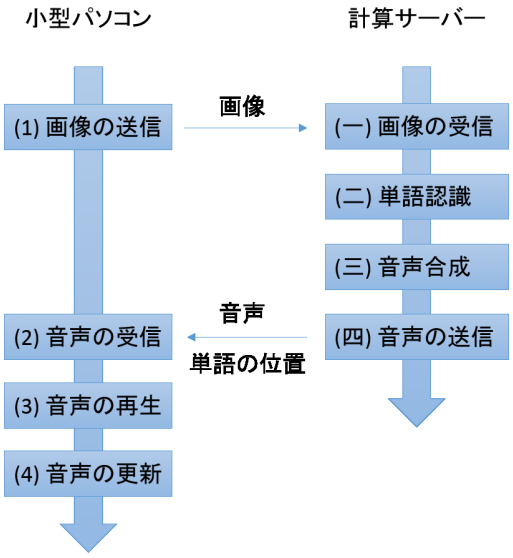

図 3 システムの処理の流れ

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 [12] を用いているため，高速に認識できる．このよう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 [13] 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．実験では，前報 [1] で提案したヘッドホンを使用した場合と，本報で導入する骨伝導イヤホンを使用した場合を試して比較する．骨伝導イヤホンを使用した場合の装着図を図 4 に示す．なお，この写真は評価実験のための装置であるため，全方位カメラは装着していない．

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[*3] 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

[*3] http://www.mmdagent.jp/

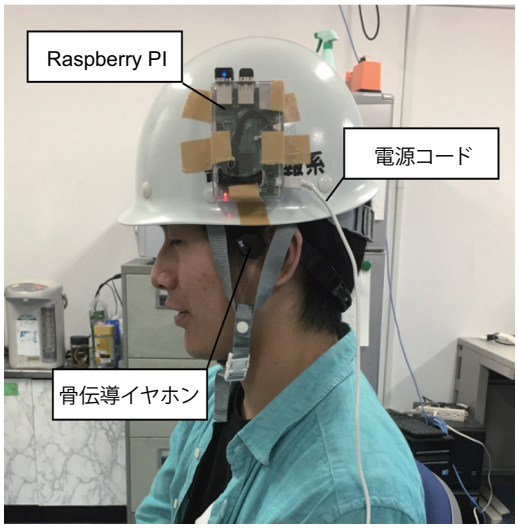

図 4 骨伝導イヤホンを使用した場合の装着図

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．正解率の算出には，角度が完全一致したときのみを正解とする「完全一致」と，角度 1 つ分 (± 30 度) の誤りを許容する「一つ隣の不正解を許容」，音源の左右はの推定は正しいものの，前後を逆に推定した場合の不正解を許容する「前後の不正解を許容」の 3 つの方法を用いた．

ヘッドホンと骨伝導イヤホンを使用した場合の正解率と認識時間をそれぞれ表 1 と表 2 に示す．また，角度毎の詳細データを図 5〜7 と図 8〜10 に示す．まず，ヘッドホンと骨伝導イヤホンを使用した場合の正解率と認識時間を比較すると，概ね同じであるか，ヘッドホンを使用した場合が少し良かった．この原因の一端は，センサの精度ならびに使用方法にあると考える．実験では，方位センサが計測できる 3 軸のうち，2 軸のみを使用して方位を推定していたため，頭の傾きにより，角度の推定に誤差が生じたと考えられる．

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．

角度の特性を見ると，ヘッドホンを使用した場合は音源が左右のどちらかにあるときの精度が最も高く，そこから離れると精度が下がるという傾向が見られた．骨伝導イヤホンを使用した場合は，ヘッドホンを使用した場合に比べて前側の推定精度が悪かった．これは音源を後側に定位させた場合であれば音源位置をうまく推定できるが，前側の場合は全体から聞こえる感じがして，どこから聞こえたかわからないという現象があったためと考えられる．その場合でも，実験 3 のようにビープ恩を鳴らすことで性能は大きく改善できた．

5. まとめと今後の課題

本稿では，視覚障害者が環境中に存在する文字情報を利用できるようにする情報提示システムを検討した．このシステムは，小型の全方位カメラを頭部に装着して，文字検出ならびに認識技術を利用して，利用者の周囲にある文字を発見し，文字の読み上げと文字の位置を提示する．前報 [1] ではオープン型のヘッドホンを使用したシステムを提案したが，本稿では耳を塞がない骨伝導イヤホンを用いて，両者の性能を比較した．その結果，立体音響で文字情報の位置を教える場合には一定の精度が得られたものの，時間がかかるなどの改善が必要な点もあった．今後は方位センサの精度改善などで性能の改善を図るとともに，はじめにで述べた 2 つの課題のうち，本稿では手付かずであった課題 1 の解決に取り組む必要がある．この課題の完全な解決は容易でないが，ユーザインタフェースの工夫により，糸口を見出せればと考えている．

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

参考文献

[1] 宮田武嗣，岩村雅一，黄瀬浩一：立体音響で教える全方位単語感知システム，電子情報通信学会技術研究報告，PRMU2015-194，Vol. 115, No. 517, pp. 179–184 (2016).
[2] Pascolini, D. and Mariotti, S. P.: Global estimates of visual impairment: 2010, *British Journal of Ophthalmol*, Vol. 96, No. 5, pp. 614–618 (2012).
[3] 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 (2003).
[4] Shilkrot, R., Huber, J., Meng Ee, W., Maes, P. and Nanayakkara, S. C.: FingerReader: A Wearable Device to Explore Printed Text on the Go, *Proc. of 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York, NY, USA, ACM, pp. 2363–2372 (2015).
[5] Yi, C. and Tian, Y.: Assistive Text Reading from Complex Background for Blind Persons, *Proc. Camera-Based Document Analysis and Recognition*, Lecture Notes in Computer Science, Vol. 7139, pp. 15–28 (2012).
[6] Goto, H.: Text-to-Speech Reading Assistant Device with Scene Text Locator for the Blind, *Proc. Assistive Technology: From Research to Practice*, pp. 702–707 (2013).
[7] Nanayakkara, S., Shilkrot, R. and Maes, P.: EyeRing: A Finger-worn Assistant, *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, ACM, pp. 1961–1966 (2012).
[8] Yi, C. and Tian, Y.: *Camera-Based Document Analysis*

表 1　ヘッドホン使用時の実験結果

	正解率 [%]	一つ隣の不正解を許容 [%]	前後の不正解を許容 [%]	認識時間 [s]
実験 1	39.3	55.4	68.5	8.0
実験 2	61.3	94.6	61.9	15.9
実験 3	61.3	94.6	63.1	9.9

表 2　骨伝導イヤホン使用時の実験結果

	正解率 [%]	一つ隣の不正解を許容 [%]	前後の不正解を許容 [%]	認識時間 [s]
実験 1	34.2	60.8	57.5	8.4
実験 2	54.2	90.0	56.7	21.9
実験 3	75.0	97.5	75.0	10.5

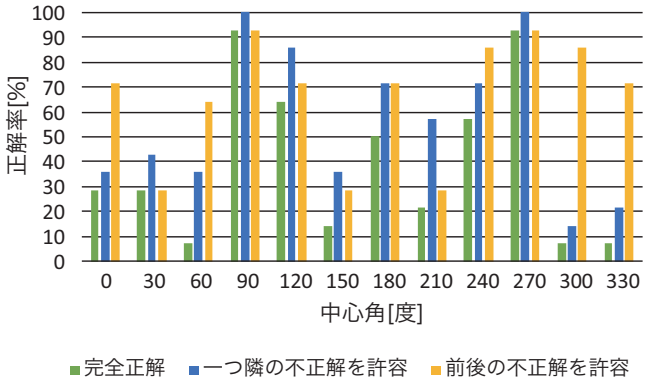

図 5　ヘッドホンを使った場合の角度毎の正解率 (実験 1)

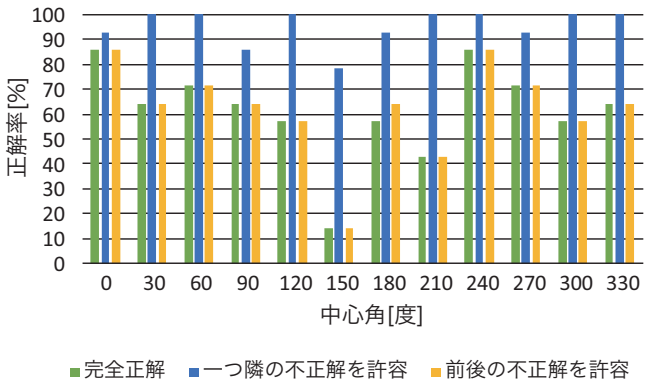

図 6　ヘッドホンを使った場合の角度毎の正解率 (実験 2)

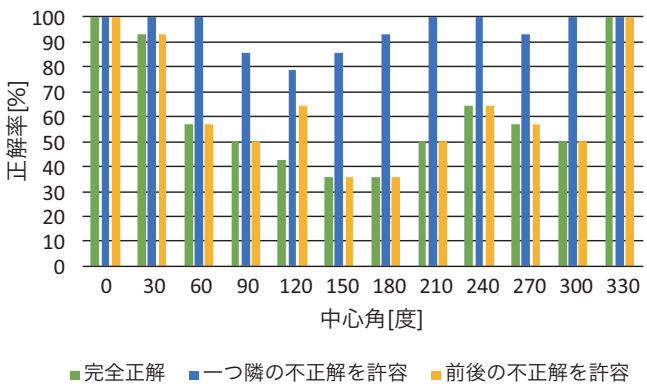

図 7　ヘッドホンを使った場合の角度毎の正解率 (実験 3)

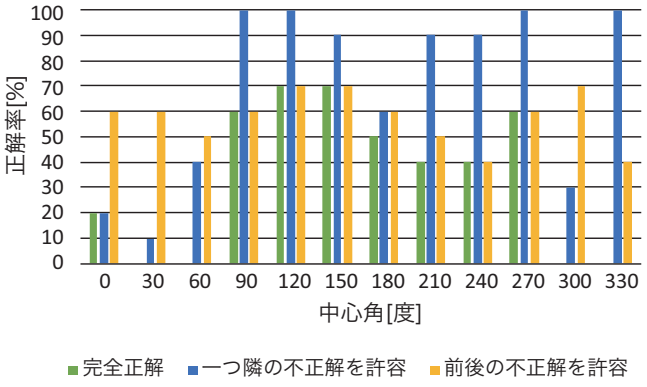

図 8　骨伝導イヤホンを使った場合の角度毎の正解率 (実験 1)

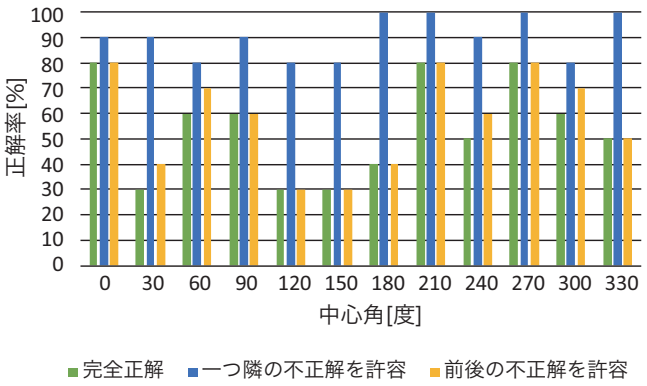

図 9　骨伝導イヤホンを使った場合の角度毎の正解率 (実験 2)

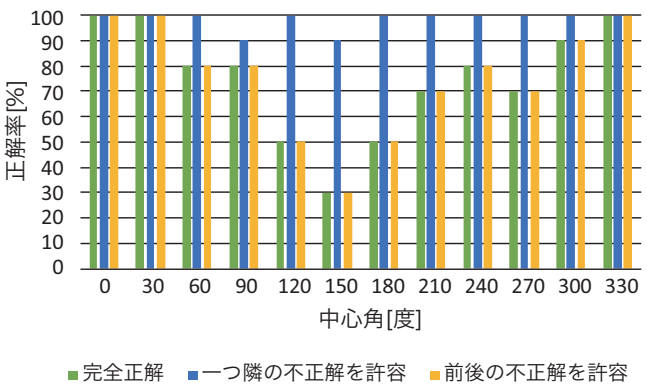

図 10　骨伝導イヤホンを使った場合の角度毎の正解率 (実験 3)

and Recognition: Post-Proc. of 4th International Workshop Camera-Based Document Analysis and Recognition (CBDAR 2011), Lecture Notes in Computer Science, Vol. 7139, chapter Assistive Text Reading from Complex Background for Blind Persons, pp. 15–28, Springer (2012).

[9] Mattar, M. A., Hanson, A. R. and Learned-Miller, E. G.: 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 (online), DOI: 10.1109/CVPR.2005.526 (2005).

[10] Ezaki, N., Bulacu, M. and Schomaker, L.: 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. 683–686 Vol.2 (online), DOI: 10.1109/ICPR.2004.1334351 (2004).

[11] Matsuda, T., Iwamura, M. and Kise, K.: Performance Improvement in Local Feature Based Camera-Captured Character Recognition, *Proc. of 11th IAPR International Workshop on Document Analysis Systems (DAS2014)*, pp. 196–201 (2014).

[12] Iwamura, M., Sato, T. and Kise, K.: 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 (2013).

[13] 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，Vol. I, No. 2-7-6, pp. 343–344 (2010).