Vol.2017-AAC-4 No.3
2017/8/26

聴覚障害者のためのARメガネを用いた音声理解

渡辺 大樹[1]　松本 哲也[1]　竹内 義則[2]　工藤 博章[1]　大西 昇[1]

概要：聴覚障害者の主なコミュニケーション方法として，手話や筆談，読唇があるが，店頭での買い物や飲食店での注文の際，手話のできる人がその場にいない時には，困る状況になる．また，筆談であれば手間を要する状況になる．本研究では，聴覚障害者の会話支援を実現するためのシステムを提案する．そのシステムは，音声の到来方向の推定と音声認識の２つの機能から構成した．音声の到来方向と音声認識された単語の表示のための AR メガネとマイクロホンアレーを構成するための複数のマイクを用いてシステムの実装を行った．到来方向の推定と雑音抑制に関して行った予備的な実験の結果について報告する．

キーワード：聴覚障害者，AR メガネ，方向推定，雑音推定，音声認識

Speech understanding system with AR glasses for hearing impaired

Daiki Watanabe[1]　Tetsuya Matsumoto[1]　Yoshinori Takeuchi[2]　Hiroaki Kudo[1]
Noboru Ohnishi[1]

Abstract: As a communication method, the hearing impaired mainly use sign language, writing, lip reading. However, when they will go shopping or order for something at a restaurant, if nobody can use sign language at the place, it becomes an inconvenient situation to communicate. To communicate in writing, it takes time and labor. In this research, we propose a system which is consisted of two functions; estimation of direction of arrival(DOA) of speech and speech recognition in order to realize conversation support of the hearing impaired. We implemented a system which is consisted of AR glasses which display the direction of speech and recognized words and plural microphones to construct a microphone array. We show the results of preliminary experiments in estimation of DOA and noise reduction.

Keywords: Hearing impaired, AR glasses, Direction estimation, Noise estimation, Speech recognition

1. はじめに

聴覚障害者は，会話をする際に主に手話や筆談の形式で行われる．しかし，手話を用いる場合，発信者と受信者の双方が手話を理解している必要がある．現在，手話を習得している人は少なく，手話を用いた会話相手が限られてしまう点，手話ができず対応してもらえない場合がある点などが欠点として挙げられる．一方，筆談を行う場合では，一度紙に書き起こさなければならないために時間が掛かる点，伝えたい全ての情報を短時間で文字に書き起こすことは難しいという点が欠点である．このどちらの場合も必ずサポートをする人が必要となる．このような状況が障壁となり，聴覚障害者にとって会話が必要な場面でも，会話に積極的に参加することが困難，あるいは理解した様子を見せることによる問題がある．

実際に，日常で主に使用するコミュニケーション手段について聴覚障害者 41 名（聴力の自己評価：ろう 32 名，難聴７名，わからない２名）に対して行われたアンケート [1] によると，**表 1** に示すようなアンケート結果（複数回答可）が得られている．発信と受信の双方において全ての属性で手話が最も用いられる手段となっており，受信においては次いで筆談，読話，残存聴力という順番，発信においては発声，筆談という順番になっており，発声によって発

[1] 名古屋大学
Nagoya Uniersity
[2] 大同大学
Daido Uniersity

表 1　日常でのコミュニケーション手段のアンケート結果 [1]

	発信			受信			
	手話	発声	筆談	手話	読話	筆談	残存聴力
ろう	93.8%	50.0%	62.5%	90.6%	50.0%	62.5%	6.3%
難聴	100.0%	71.4%	42.9%	100.0%	71.4%	71.4%	28.6%
わからない	100.0%	50.0%	0.0%	100.0%	50.0%	0.0%	100.0%

信する人が半数を超えている．

また，講義など大勢の人を対象にする場面では，要約筆記を用いた支援 [2] が提供されることがあるが，個人間での会話においてそれと同じ規模でサポートをすることは難しい．

そこで，本研究では，言葉の発信を行うことができる人を対象に，言葉の受信において会話情報支援を行うシステムを提案する．以降では，AR メガネの縁にマイクを複数個取り付け音声を収音し，AR メガネのディスプレイ上に発話方向と，会話内容を提示するシステムでの処理と実験結果を報告する．

2. 関連研究

末光ら [3] による研究では，聴覚障害者の会話情報保障のためにシースルー型メガネを用いた会話情報の字幕化を提案している．透過型ヘッドマウントディスプレイ上に不等間隔に並べられたマイクにより，マイクロホンアレーの設計を行い，インパルス応答を測定し逆畳み込みを行うことで目的音の強調を試みている．等間隔，不等間隔配置のどちらが強い指向性があるかを調べ，不等間隔配置の方が逆畳み込み演算による目的音の強調に適しているとしている．また，「おはよう」や「ありがとう」といった簡単な語を用いて，認識精度の検証を行っており，環境音において信号強調ができたとしている．しかし，インパルス応答を測定し，逆畳み込み処理を行うという処理は，その環境に対応したインパルス応答を測定する必要があるために使用場所に制限がある．

そこで本研究では，環境の変化に対して雑音抑制の面からシステムを提案する．

3. 提案システムの構想

本研究では，音声情報を聴覚障害者に伝える手段として，AR メガネのディスプレイに文字を投影することを考える．AR メガネの AR は Augmented Reality(拡張現実) の略であり，現実世界の光景にメガネを通して，プロジェクターのような画面を映すものである．

AR メガネを用いる理由は，以下の 2 つである．

1. 機器の着用は聴覚障害者のみになるため支援者の負担が軽減される
2. 見ている景色に音声認識の結果を表示することができるため，相手の顔を見ながらコミュニケーションを行うことが可能

これにより，本研究ではスムーズでより自然な会話コミュニケーションの実現を目指している．

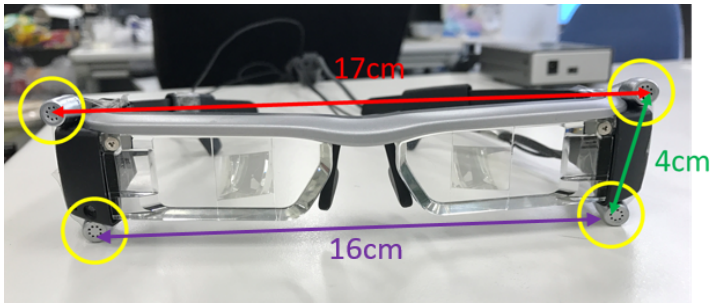

図 1　AR メガネとマイクの配置

図 1 は，本研究で用いる AR メガネとマイクロホンを取り付けた配置について示している．

3.1 システム設計

AR メガネの縁にマイクを取り付け音声の収録を行う．録音は 5 秒間ごとに連続して行い，録音した音声ファイルはネットワーク上の共有サーバーに保存する．保存された音声ファイルを読み込み，音声の到来方向を推定（Matlab で実装）し，次いで音声認識（Julius）を行う．図 2 はシステムの構成と，各部位で行われる処理についてまとめたものである．

使用機器を以下にまとめる．

- AR メガネ
  EPSON 社製スマートグラス MOVERIO BT-200
- 音声同時サンプリング機器
  東京エレクトロンデバイス株式会社製 8 チャンネル音声入力ボード TD-BD-8CSUSB
- マイク
  SONY 社製エレクトレットコンデンサーマイクロホン ECM-CZ10
- 計算機
  音声録音用 PC（Windows 32bit 版），信号処理用 PC（Windows 64bit 版）
- 音声認識器
  Julius version4.4（gmm 版）
- 音声解析ソフト
  Matlab R2015b
- 共有サーバー

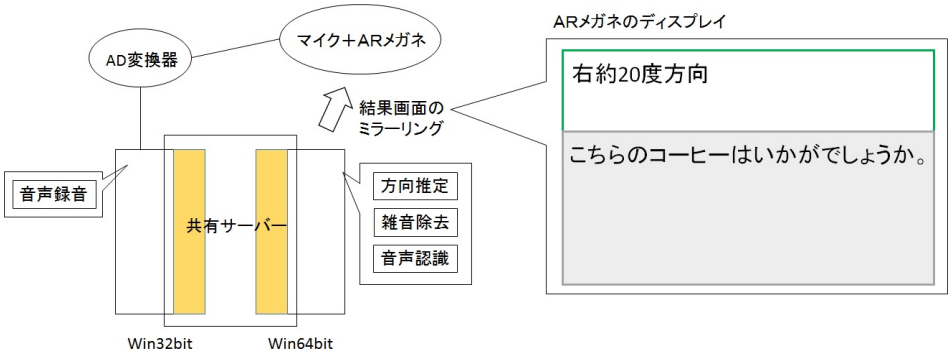

図 2 システムの全体像

4. 目的音の到来方向推定

図 1 に示したマイクロホンアレーの構成上，正面から音声を捉えた場合に，左右のマイクロホンで捉えられた音声の時間差が最小になる．この場合単純に音声を加算するだけで，SN 比の向上が見込まれる．しかし音声の到来方向と頭の向きの間に角度のずれが生じている場合，左右のマイクロホンで伝達関数が異なるために雑音抑圧が難しくなる．利用者は会話相手の方向を認識したら，相手の方向を向いて，音声を正面で捉えることが理想である．このため，音声の到来方向を推定し，提示することを行い頭部を向けることで，それ以後の音声をできるだけ正面で捉えてもらうことが可能となる．

マイクロホンアレーでの方向推定の計算方法を説明する．図 3 は方向推定のモデルを示している．マイクが横に 2 つ配置されている場合，音声が角度 θ の方向から到来したときマイク 1,2 それぞれで捉えた 2 つの音 $x_1(t)$ と $x_2(t)$ では $\frac{d\cos\theta}{c}$（c:音速）だけの時間差が生じる (1)．この時間差を用いて逆関数 (2) により角度を求めることができる．この時間差は相互相関関数 (3) が最大になるときの値として算出する．つまり，2 つの音声の最も相関が高くなる時間差を求めることで，音声の到来方向を求めることができる．

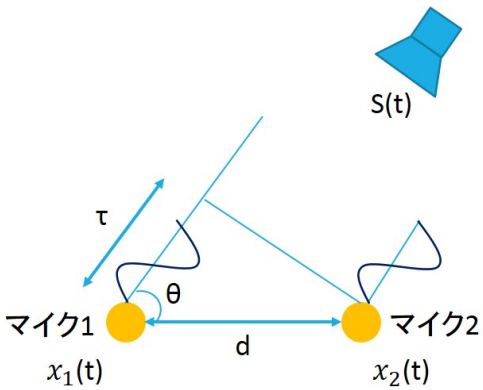

図 3 方向推定モデル

表 2 方向推定結果

入力音声方向	推定方向
左約 45 度	左 49 度
左約 60 度	左 61 度
右約 30 度	右 39 度
左下	左 39 度，下 32 度
左上	左 49 度，上 32 度
右下	右 39 度，下 32 度

$$\tau = \frac{d\cos\theta}{c} \tag{1}$$

$$\theta = \cos^{-1}(\frac{\tau \cdot c}{d}) \tag{2}$$

$$\phi_{1,2}(\tau) = \lim_{\tau\to\infty}\frac{1}{T}\int_0^T x_1(t)\cdot x_2(t+\tau)dt \tag{3}$$

表 2 に方向推定結果の例を示す．サンプリング周波数は 16kHz であり，c の値は $340m/s$ に設定した．使用したマイクは，左右方向の推定には図 1 の上側 2 つのマイク，上下方向の推定には図 1 の左側 2 つのマイクを使用した．

表 2 からわかるように，どの方向もおおよそ正しい方向に推定できている．この時，推定される角度の分解能が 10 度弱であるが，会話を行う距離において人が数度の範囲に多数存在する場面はほとんど無いものと想定できる．また，± 10 度程度の誤差であっても相手を視認することが可能であると考える．推定される角度の精度の向上は，アップサンプリング処理を行うことで可能である．

5. 雑音推定

音声認識結果の向上のためには，雑音抑圧が必要不可欠である．

Julius では，オプションで雑音スペクトルの推定を行うことができる．スペクトルサブトラクションにより，推定した雑音のスペクトルを音声信号から減算することで雑音の抑圧を行う．Julius での雑音推定方法は以下の 2 つがある [4]．

(1) 各入力の最初の数百ミリ秒を雑音区間と仮定してその平均を雑音スペクトルとする方法

(2) あらかじめ付属のツールで雑音スペクトルを推定してファイルに保存しておき，それを読み込む方法

しかし，(1) の場合は，音声の最初のみを用いて雑音を推定するため，雑音が時間的に変化しているときや，音声の最初が雑音ではなかったとき，最初のみに突発的な音が紛れ込んだときなどに十分に対応できず，(2) の場合は，あらかじめ雑音をファイルとして保存しておかなければならないため，手間がかかる且つ環境が変わった場合に対応できないという欠点が存在する．

また，Julius では，信号の振幅とゼロ交差に基づいて発話区間の検出を行っている [5]．これはオンライン入力や長めの音声のファイル入力では，有効であるが，短めの数秒ほどの音声で実行した場合，認識精度が著しく低下する．また，入力の振幅は実行環境（録音ボリューム，マイクと発話者の距離等）に大きく左右され，実行環境ごとに閾値の調整が必要であるとされている．

ここでは，パワースペクトルを用いた雑音推定処理を試みた．処理は以下の通りである．

(1) 処理対象の音声にハイパスフィルタをかける
(2) 音声を t 秒間ごとに区切り，各区間でのパワースペクトルを求める．
(3) 区間 n と $n+1$ のパワースペクトルを比較し，各周波数ビンのパワースペクトルの差の絶対値の平均（M とする）を求める
(4) 上で求めた値が閾値以下であればその区間を雑音区間であるとする

以下に，実際にこのアルゴリズムで雑音区間と発話区間の分離を行った結果を示す．全体が約 5 秒の音声に対して t=0.15 秒で区間を設定した．

図 4 から 7 は連続した 4 つの区間のパワースペクトルである．ここでは，n=4 である．図 4〜5 は雑音部分であり，パワースペクトルに差が見られず，M の値は閾値を下回る．しかし，図 5〜6 にかけては大きく変化しているため，M の値は閾値を上回り，図 5 の中間から図 6 の中間までを発話区間であると判定する．同様に図 6 の中間から図 7 の中間も発話区間となる．発話区間では，区間ごとにパワースペクトルが大きく変化するため，M の値は連続して閾値を上回り，発話区間であると判定される．

図 8 は全区間に対して雑音区間か発話区間かを判定し，もとの音声を 2 つの区間に分離した波形である．中央部が連続した発話区間であり，その両端が雑音区間として切り出されていることがわかる．図 4,5 から，音声の主要な成分の周波数も値が大きくないことがわかる．

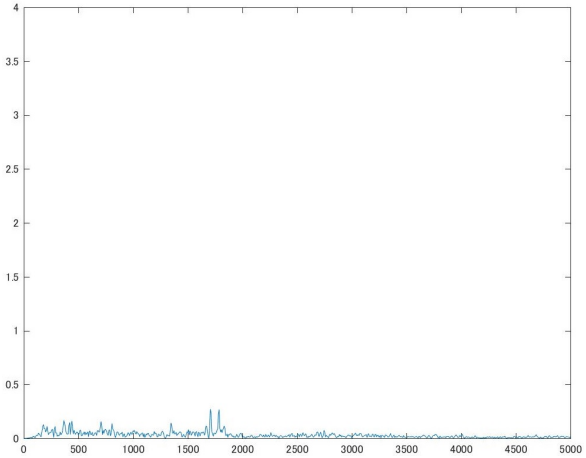

図 4 雑音区間のパワースペクトル（区間 n）

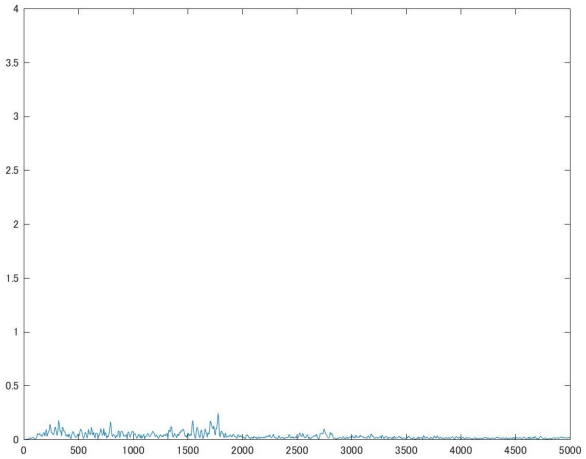

図 5 雑音区間のパワースペクトル（区間 n+1）

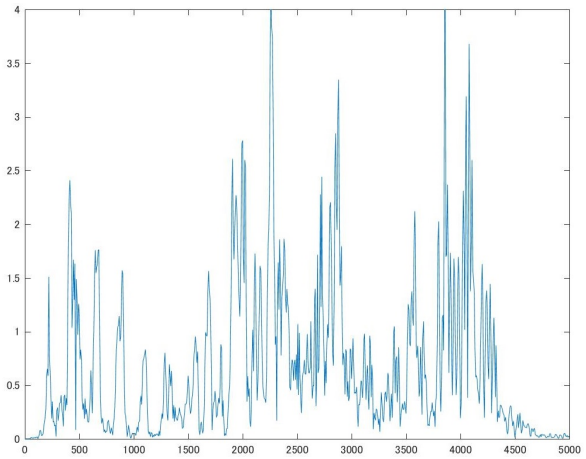

図 6 発話区間のパワースペクトル（区間 n+2）

6. 音声認識

音声認識は Julius のディクテーションキット v4.4 を用いて行う．音響モデルや言語モデルは変更せず，辞書に登録されていない単語は新規登録を行った．

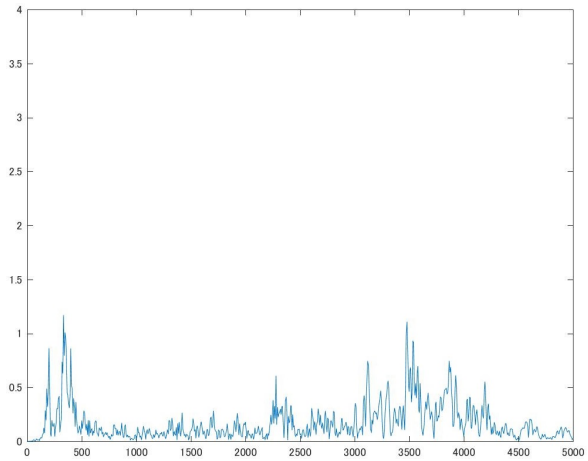

図 7 発話区間のパワースペクトル（区間 n+3）

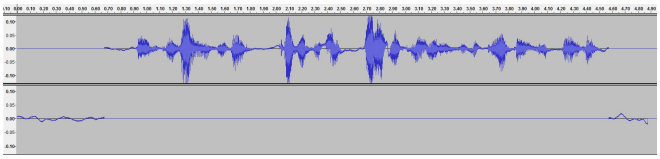

図 8 発話区間と雑音区間を分離した後の波形

発話者は 20 代男性，サンプリング周波数 16kHz とした．AR メガネを床上 1 m の高さに固定し，1 m 離れた位置から，正面，右 30 度，左 30 度の角度から音声を肉声により発話した．雑音として空調やパソコンの待機音が存在した．発話する文章は，[6] から抜粋した以下の 20 の日常会話を用いた．音声は雑音区間を含め，録音開始から録音終了までで 3〜5 秒であった．

(1) 初めまして台湾から参りました．
(2) 日本は初めてなので，まだ何もわかりません．
(3) 僕は南国宮崎の出身です．
(4) 趣味は料理と競馬です．
(5) カラオケが好きなので，皆さんぜひ一緒に行きましょう．
(6) 大学では経済学を専攻していました．
(7) どちらにお住まいですか．
(8) お仕事は，何をなさっていますか．
(9) ご出身はどちらですか．
(10) 天気がいいので，家族でバーベキューをしにいこうと思っています．
(11) 家でＤＶＤを見て過ごそうと思っています．
(12) 昨日受けたテストどうだった．
(13) 来週の火曜日に，テストがあると聞きました．
(14) すみません，銀行口座を作りたいのですが．
(15) ありがとうございます．
(16) こちらの液晶テレビは，いかがでしょうか．
(17) 定価の 40 %引きになっております．
(18) こちらは，新発売のデジカメです．
(19) こちらが咳止めと解熱剤で，こちらは鎮痛剤です．
(20) こちらのコーヒーはいかがでしょうか．

表 3 認識結果（右上のマイク）

	文	品詞	文（割合）	品詞（割合）
正面	14	171	70%	94.0%
右 30 度	14	168	70%	92.3%
左 30 度	10	166	50%	91.2%
全体	20	182		

表 4 認識結果（左上のマイク）

	文	品詞	文（割合）	品詞（割合）
正面	14	173	70%	95.1%
右 30 度	12	167	60%	91.8%
左 30 度	12	168	60%	92.3%
全体	20	182		

文単位と品詞単位を対象に認識精度を算出した．音は同じでも意味の異なる出力の場合は不正解であるとした．品詞分解は，[6] のサイトを用いて行った．

着用者から見て右上に配置したマイク 1 つを用いた場合の認識結果を表 3 に示す．

また，着用者から見て左上に配置したマイク 1 つを用いた場合の認識結果を表 4 に示す．

正面方向が文，品詞共に一番良い結果が得られている．また，図 3 で右側の方が認識率が高いのは，右側のマイクは AR メガネでの反射等の干渉を受けにくいからであると考えられる．左側の場合も同様と考える．

文章ごとに見ると，(1),(5),(7),(8),(15) において全ての場合で，含まれる全ての品詞が正解となった．反対に最も悪いものは (4) であり，「趣味」を「炭」や「シミ」，「競馬」を「電話」や「現場」に誤認識し，総じて半分程の正解数であった．短い文ほど正しく認識できるとは一概には言えず，認識率は発話する単語と人が聞いた場合の聞き取りやすさに依存するものと考えられる．

7. おわりに

聴覚障害者の受信における会話支援システムを提案した．方向推定では概ね良好な結果が得られた．雑音推定では，発話区間と雑音区間の分離を行い，今後は雑音抑圧に応用していこうと考えている．音声認識実験では，定常雑音環境下で 3 方向からの 3〜5 秒の音声に対して 90%以上の精度を示すことができた．今後は，実際の使用を想定した非定常雑音環境下での認識とシステムの全自動化を実現する予定である．

参考文献

[1] 島根陽平，井上清子:聴覚障害者における聾（ろう）と難聴のアイデンティティ，生活科学研究,32,pp:27-35,(2010).
[2] 池田直史ら:音声認識による難入力語の検出を用いた講義

の文字通訳支援システム，信学技報，116(519)，pp:19-24，(2017).
[3] 末光一貴ら:単一チャンネルマイクロフォンアレーによる会話情報の字幕化，HCG,A-4-5,pp:322-326(2016).
[4] The Julius book：第 4 章 フロントエンド処理・特徴量抽出 (online)，入手先 ⟨*https* : *//julius.osdn.jp/juliusbook/ja/desc_feature.html*⟩ (参照 2017.08.01).
[5] The Julius book：第 5 章 音声区間検出・入力棄却 (online)，入手先 ⟨*https* : *//julius.osdn.jp/juliusbook/ja/desc_vad.html*⟩ (参照 2017.08.01).
[6] 生活日本語会話：(online)，入手先 ⟨*http* : *//web.ydu.edu.tw/˜uchiyama/conv/index.html*⟩ (参照 2017.07.14).
[6] 日本語自動品詞分解ツール：(online)，入手先 ⟨*http* : *//tool.konisimple.net/text/hinshi_keitaiso*⟩ (参照 2017.08.01).