TR－SLT－0053

マイクロフォンアレーのパラメータ変化における音声認識率の評価

Evaluation of speech recognition in parameter changes of microphone arrays

鎌本優
Yutaka Kamamoto

発行年月日
2003.9.30

概要

本稿では，音声入力にマイクロフォンアレイを用い，そのパラメータを変化させた場合に，音声認識率がどのように変化するかを評価した．マイクロフォンアレーのパラメータとしては，使用するマイクロフォンの個数，マイクロフォン間の間隔，マイクロフォンに対する音源の角度およびSNRに注目した．音源の到来方向は既知とし，ビームフォーミング手法にはDelay-and-Sumを用いた．BTECテストセット01(510文)に対して，これらの条件のもとでの音声認識率を示した．また，新たなマイクロフォンアレイ手法を開発することを試みた．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

マイクロフォンアレーの
パラメータ変化における音声認識率の評価

Evaluation of speech recognition
in parameter changes of microphone arrays

鎌本 優　(Yutaka KAMAMOTO)

2003年9月30日

目 次

第1章　緒言

1.1　背景

雑音除去技術は様々な工学の分野において需要がある．自動音声認識 (Automatic Speech Recogntion; ASR) においても実環境では雑音や残響により認識率が大幅に低下することから，雑音除去の研究がなされてきている．

マイクロフォンアレーを用いることにより，多チャンネルの音響信号に基づく空間情報が得られ，これらを処理をすることにより雑音除去を行うことができる [1]．マイクロフォンアレーを用いた雑音除去法の中には，空間情報に加え，音声の統計的性質，調音構造，ピッチ，ホルマント構造など，音声特有の情報を積極的に利用する方法も提案されている [2,3]．

現在，擬人化エージェントと ASR を組み合わせて会話をする場合，音声入力の手段として近接マイクロフォンを用いるか，ヘッドセットマイクを用いるためわずらわしい．ここでマイクロフォンアレーを用いれば，画面のエージェントに向かって話し掛けることができ，さらに発話者の方向推定も可能なので，エージェントに発話者の向きに視線を向けることも可能となり，自然なコミュニケーションを交わすことができる．

このように，ASR のためにマイクロフォンアレー手法の技術を向上させることは，大変有用なことである．

1.2 目的

基本的なマイクロフォンアレー手法であるDelay-and-Sumについて，パラメータを変化させた場合に，音声認識率がどのように変化するかをシミュレーションにより調査することを本研究の目的とした．また，その結果をもとに新たな手法を提案することを第2の目的とした．

第2章　実験方法

2.1　検討内容

音声を強調するために，SAFIA [4] や GSC [5] など様々なマイクロフォンアレー手法があるが，ここでは，多くのマイクロフォンアレー手法の基礎となる Delay-and-Sum について調査した．

マイクロフォンアレー手法 Delay-and-Sum を評価するための基準として，

- マイクロフォンの数
- マイクロフォン間の距離
- 音源のマイクロフォンアレーに対する角度
- SNR

に注目し，これらの 4 つの条件をパラメータとして変化させ，音声データを作成した．

作成したデータを用いて Delay-and-Sum 処理後の音声データを音声認識システムに入力し，その認識率により性能評価を行った．

2.2　データの作成

音声認識評価実験を行う準備として，各マイクロフォンアレーの素子で受音される音響信号データを作成した．

音声データとして，ATR の BTEC テストセット 01 を用いた．この評価用データは旅行

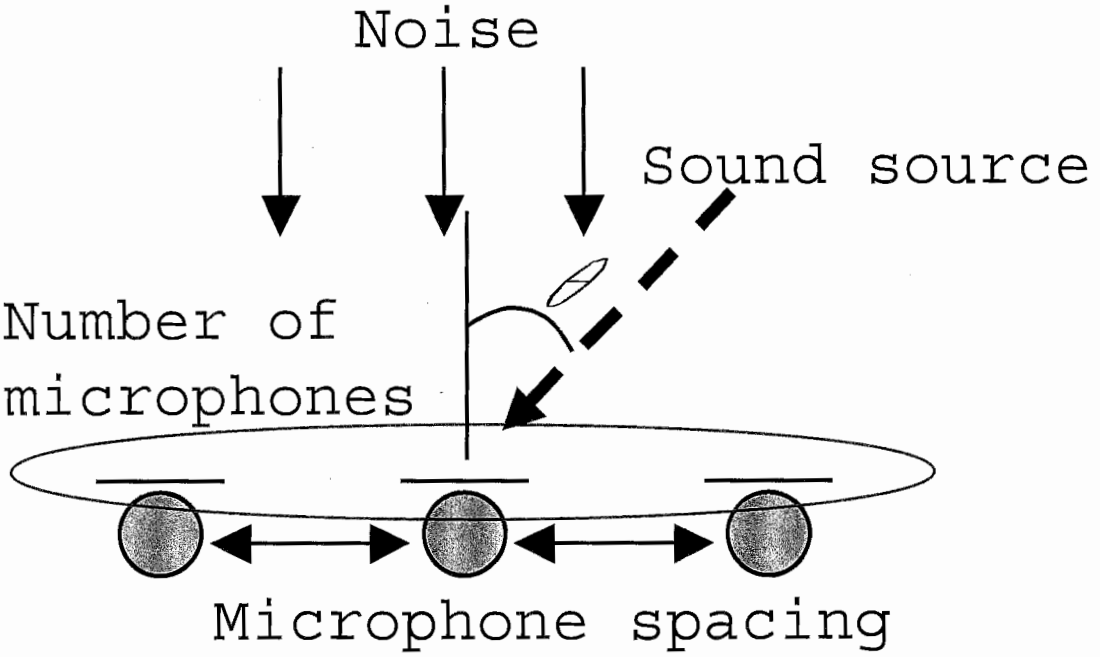

Fig. 2.1 . 4 つの条件

の際に用いられる会話を朗読したもので，全 510 文である．これをマイクロフォンアレーへの入力データとするために MATLAB を用いて加工した．

マイクロフォンの数の分だけ音声データを複製し，それぞれのマイクロフォンにマイクロフォン間隔と角度から計算された時間差を加え，さらにマイクロフォン列に垂直な方向から雑音を加えた．音速は 340m/s とし，雑音は音声の周波数帯域に合わせて，ホワイトノイズに 125Hz から 6000Hz の BPF(チェビシェフ特性) を通したものを用いた．SNR は音声区間の平均振幅から信号のエネルギーを求め，目的の SNR となるように雑音の振幅を変化させた．

パラメータの違いにより，マイクロフォンアレーから得られた多チャンネル信号がどのように変化するかを調査するために，以下のようなパラメータのデータを作成することを試みた．

- マイクロフォンの数：2 個，3 個

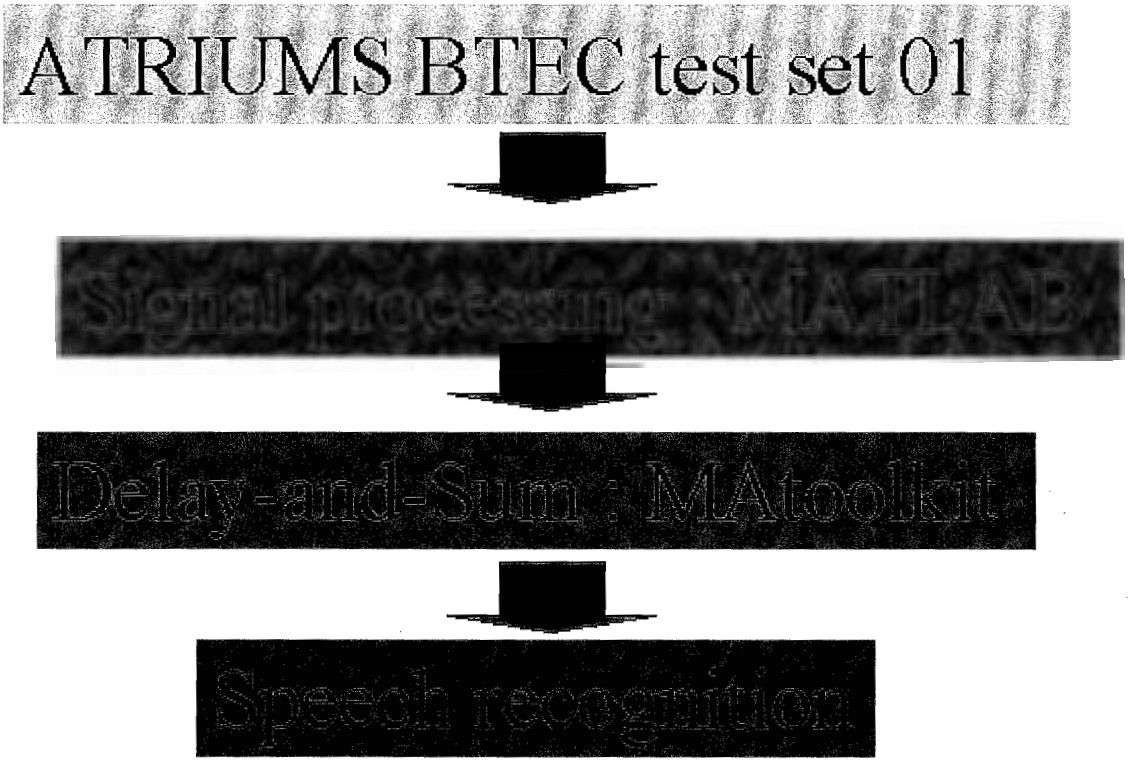

Fig. 2.2 . データ作成の流れ

- マイクロフォン間の距離：5cm，10cm，15cm
- 音源のマイクロフォンアレーに対する角度：−90 度から +90 度まで 5 度間隔で変化させた場合 、
- SNR：20dB から 0dB まで 5dB ずつ雑音を増加させた場合

これらのデータを MAToolkit の Delay-and-Sum のプログラムにより再合成した．

2.3 音声認識実験

音声認識のデコーダには ATRIUMS の EvaluationV1 を用いた．まず，ベースラインの認識率を求めるために，クリーン音声での認識率を求めたところ，単語認識率は約 92% であった．

様々なパラメータでの音声認識率を評価しなければならないため，相対評価により検討することから，認識率とスピードのバランスを考え，雑音環境下の単語認識率が 40% 前後まで下がるような，速度を優先させる設定にした．

第3章　実験結果

3.1　Delay-and-Sum

作成したデータをもとに音声認識による評価実験を行った．各パラメータによる認識率の違いを以下に示していく．

3.1.1　マイクロフォンの数について

マイクロフォンの数が増えると，加算平均の回数を増やすことができるので，理論上はマイクロフォンがN個のときSNRは$\sqrt{N}$倍となる．実験結果も，マイクロフォンの数が増えるほど認識率が上がっているので，数が多いほど良いということが分かった．

3.1.2　マイクロフォン間の距離について

マイクロフォン間隔単独では認識率にそれほど影響をもたらさないことがわかった．しかし，マイクロフォン間隔は次に示す音源方向に影響をもたらすことが分かった．

3.1.3　音源のマイクロフォンアレーに対する角度について

音源の角度によって，SNRの向上度合いが変わることが分かった．また，これはマイクロフォン間隔にも依存していることが分かった．

角度変化に対するSNRの変化と認識率が一致しなかった．SNRが低い角度では認識率が下がるはずであったが，0度付近で急に向上した．

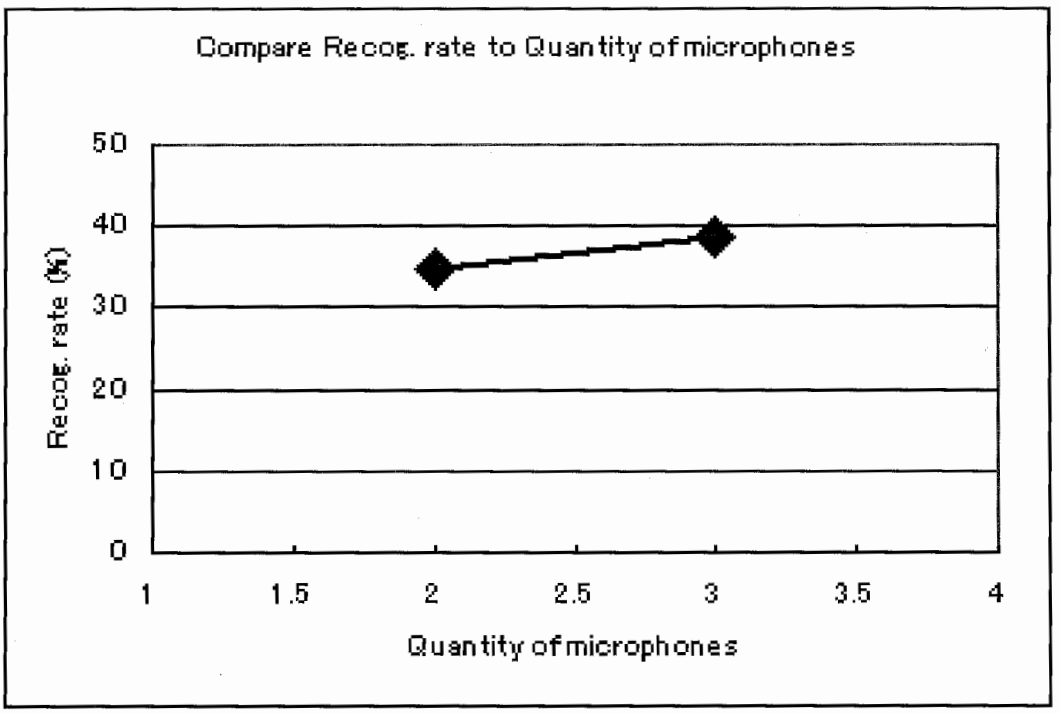

Fig. 3.1．マイク数による認識率の変化 (SNR20dB, マイク間隔 5cm, 音源方向 40^o)

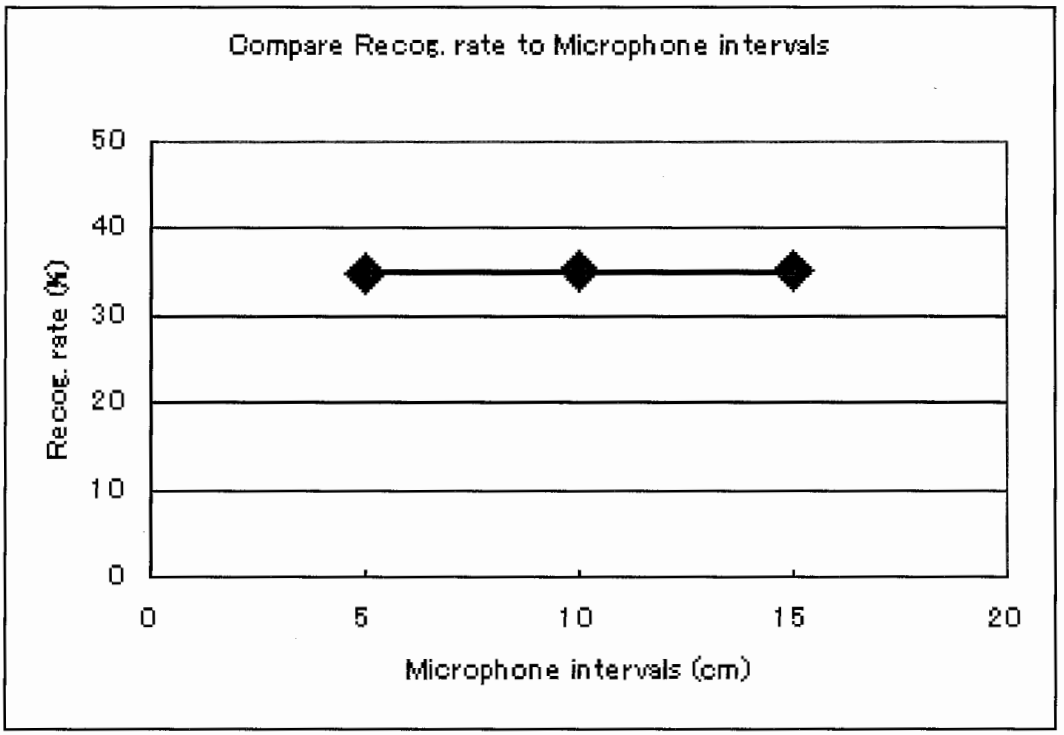

Fig. 3.2 . マイク間隔による認識率の変化 (SNR20dB, 音源方向 40°)

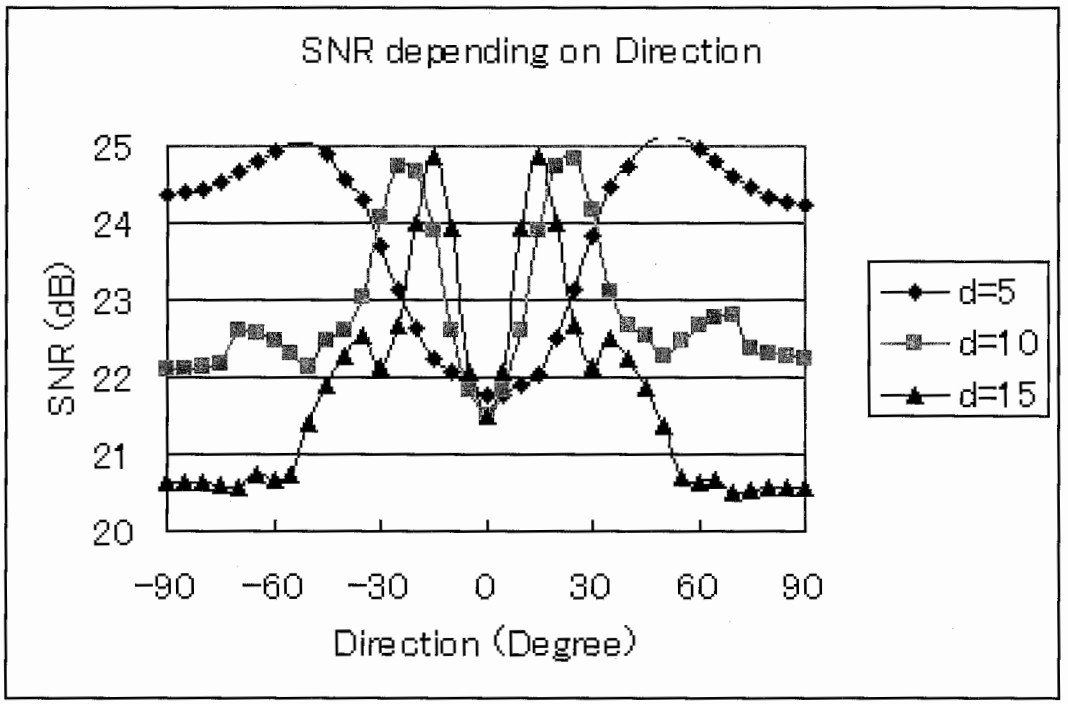

Fig. 3.3 . 角度変化と SNR の関係 (SNR20dB)

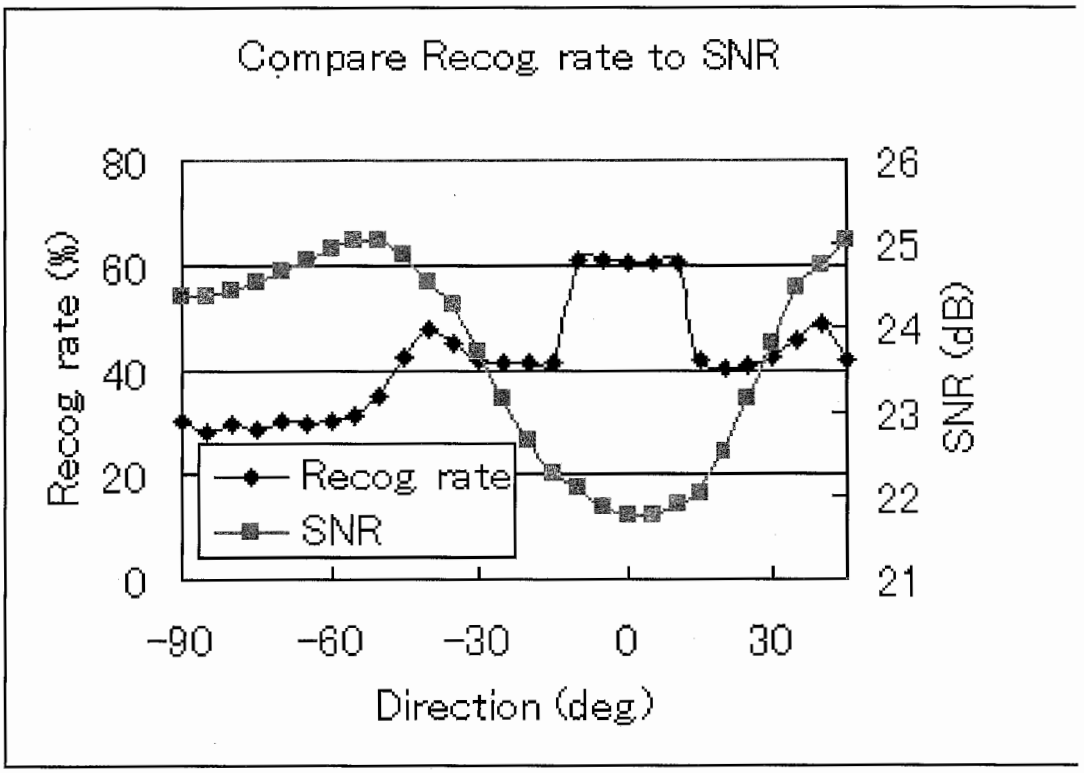

Fig. 3.4 . 角度変化における SNR と認識率の関係 (マイク間隔 5cm,SNR20dB)

3.1.4 SNR について

SNR が高いほど認識率も向上するということが分かった．低 SNR 環境下では認識に時間を用することもわかった．

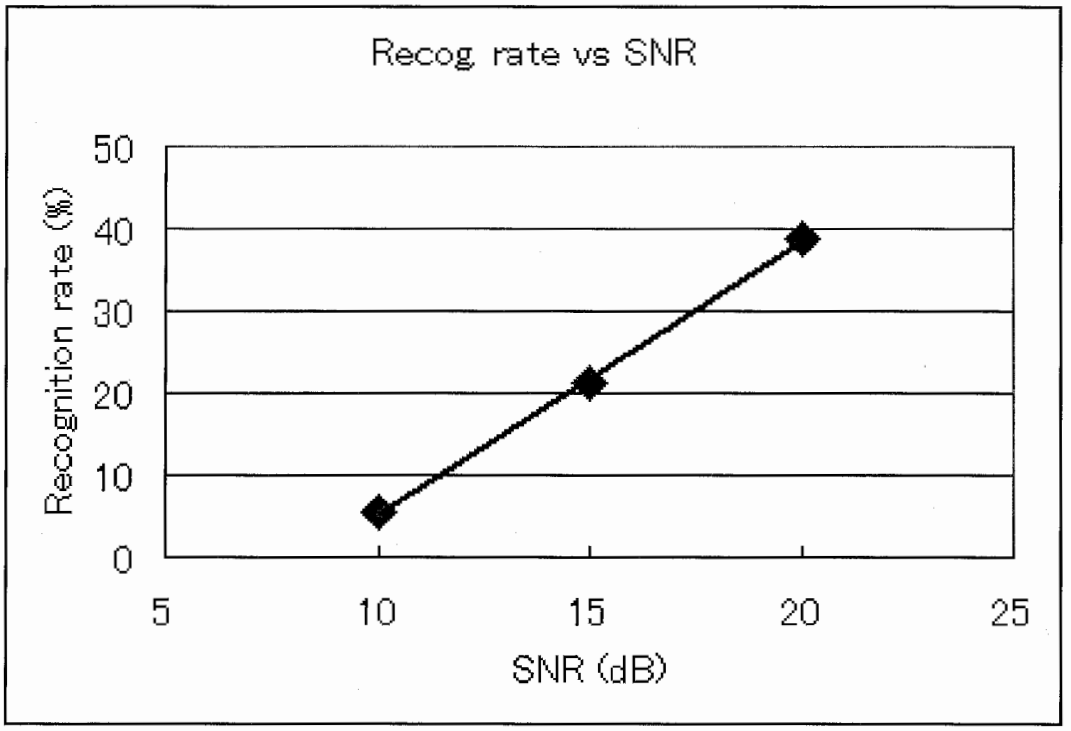

Fig. 3.5 . SNR による認識率の変化 (マイク間隔 5cm, 音源方向 40°)

第4章　考察

4.1　各パラメータによる比較について

結果から，各パラメータと認識率について，以下のようにまとめられた．

- マイクロフォンの数：多いほうが良い
- マイクロフォン間の距離：角度に依存する
- 音源のマイクロフォンアレーに対する角度：間隔と関係し，SNR が変化する
- SNR：高 SNR ほど認識率も高く，計算時間も速い

これらのことから，この 4 つもパラメータを総合的に判断すると，マイクの数を多くすることによって SNR が向上するので認識率が高くなり，また，通常多く使われる音源方向に最適なマイクロフォン間隔にすれば，SNR を向上させることができると考えられる．

Delay-and-Sum はシンプルではあるが，パラメータを変えることにより，音声認識に最適なパラメータが存在するのではないかと考えられる．

第5章　新たな手法の提案

Asymmetrical Delay-and-Sum

5.1　動機

今回の実習ではDelay-and-Sumについての評価を行った．前述の結果や考察からこの方法を基に，新たな手法を作成しようと考えた．

実際，GSCやAMNORなどの適応型フィルタは，予めノイズを入力して，そのノイズを消すようなフィルタ設計を行わなくてはならない．また，フィルタによって音声のスペクトルが変化してしまう危険性もあり，ここでMFCCが変形してしまうと音声認識にとって大きな失敗の原因となってしまう．

そこで，音源の角度とマイクロフォン間隔に注目し，Delay-and-Sumを改良することを試みた．

5.2　原理

4つのマイクロフォンを用意し，その中からマイクロフォン間隔の異なる2つを選択し，1度Delay-and-Sumの処理を行う．さらに処理後の信号を選択し，もう1度Delay-and-Sumの処理を行う．

具体的には，Fig. 5.1. のように，2段階のDelay-and-Sumを行った．数字はチャンネル数である．ここで2度目のDelay-and-Sum処理では，マイクロフォン間隔の異なるものを組み合わせている．これによって，一定のマイクロフォン間隔では角度によってSNRが低下してしまう部分を軽減できると考えた．

さらに，得られた5つの信号を組み合わせて，最大のSNRになる組み合わせのもののみを選択し，出力した．

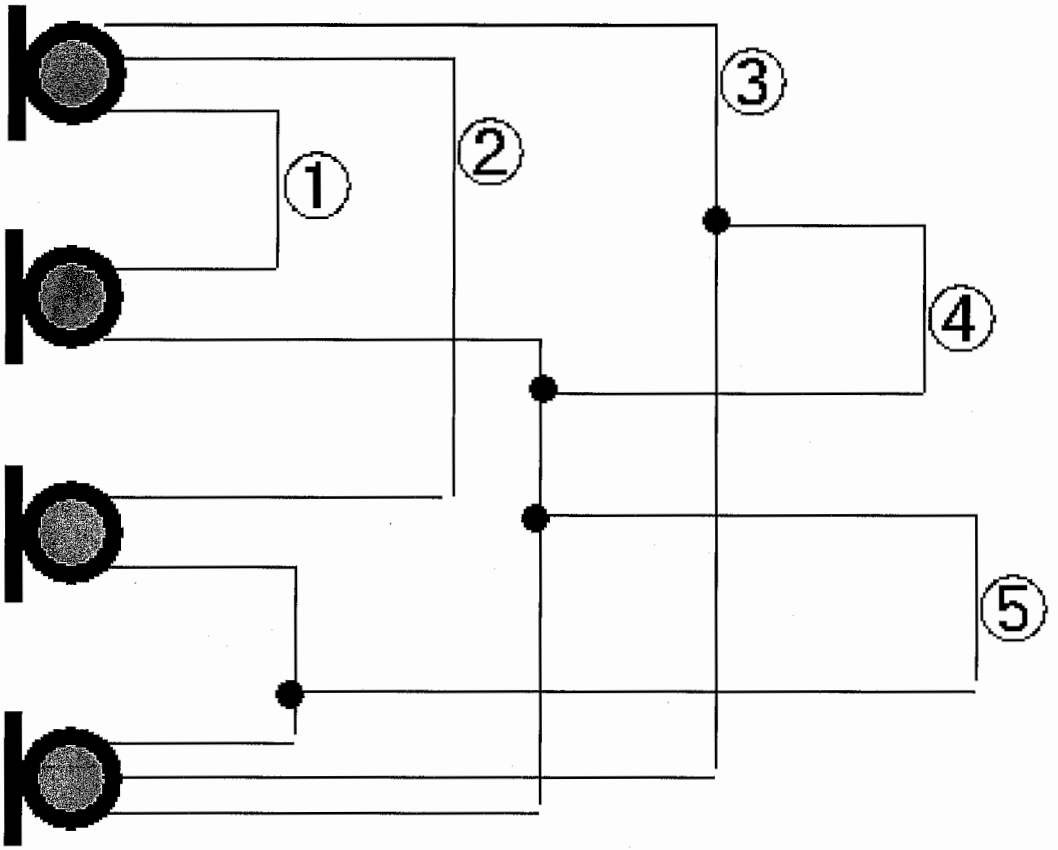

Fig. 5.1 . Asymmetrical Delay-and-Sumの概念図 (数字はチャンネル番号)

5.3 結果

Fig. 5.2. のように，通常のDelay-and-SumよりもSNRを向上させることができた．この図は，各マイクロフォン間隔において，角度を変化させたときの平均値を示している．

そして，音声認識をさせると，Fig. 5.3 のようになった．提案手法により認識率を上げることができた．

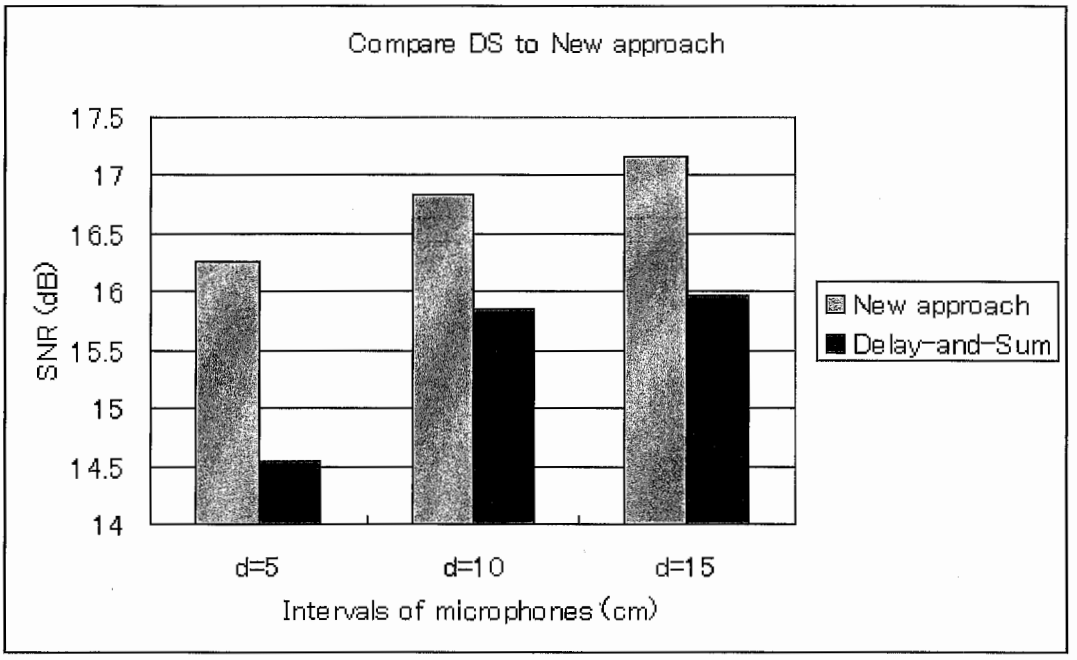

Fig. 5.2 . 新たな手法と通常の Delay-and-Sum の SNR 向上度の比較 (SNR10dB, 音源方向 40°)

5.4 考察

Delay-and-Sum を 2 回作用させるということは，線形変換から，各チャンネルに重みをつけた Delay-and-Sum と同様であった.

このことから，マイクロフォン間隔と音源方向から重み付けを変えることにより，通常の Delay-and-Sum よりも認識率を向上させることができると考えられる.

適応型マイクロフォンアレーでは雑音情報を予め入力したり，周囲の環境のインパルス応答を入力してフィルタを形成させたりするフィードバック方式であるので，計算時間が必要となりリアルタイム性が失われてしまう．しかし，Delay-and-Sum を基にしていれば，原理も単純なのでほぼリアルタイムで音声認識器にデータを送ることができると考えられる.

マイクロフォン間隔を不等間隔に配置するマイクロフォンアレーの研究も行われている

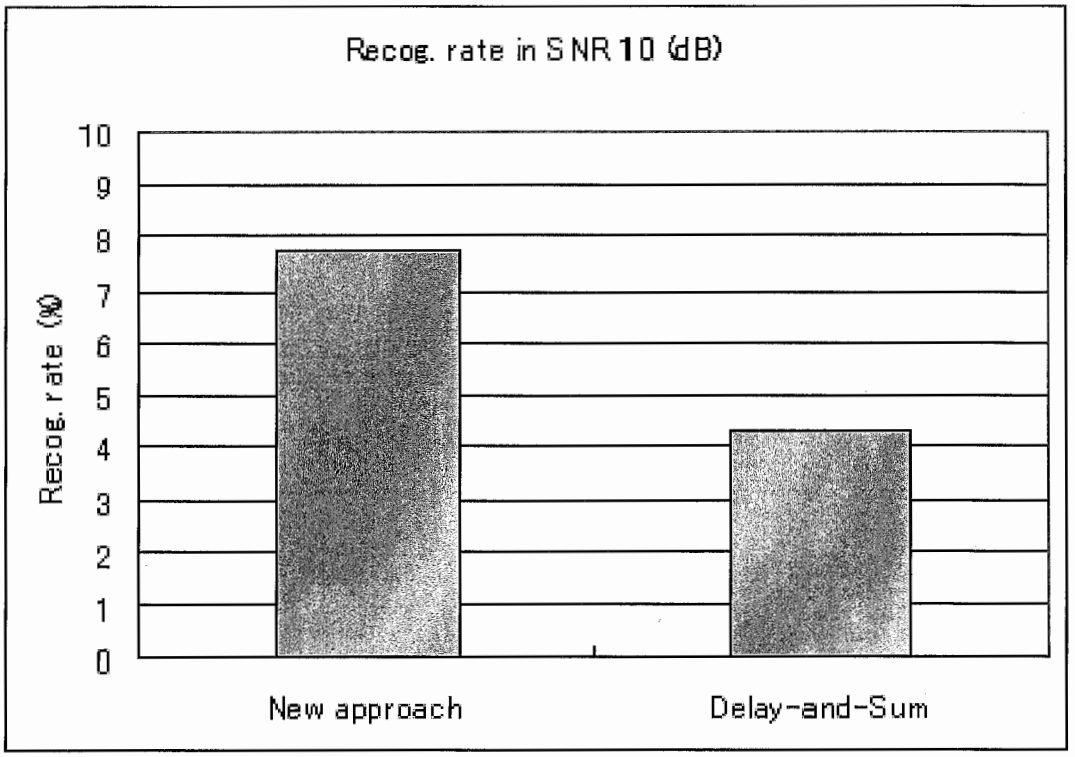

Fig. 5.3 . 新たな手法と通常の Delay-and-Sum の認識率の比較 (SNR10dB, マイク間隔 5cm, 音源方向 40°)

ことから，そのマイクロフォン間隔と音源方向と音源の周波数特性から，各チャンネルのバランスを考え，最適なものを出力するアルゴリズムで実験してみる必要があると考えられる．

また，今回評価できなかった適応型フィルタを用いたタイプのマイクロフォンアレー手法において，パラメータがどのように作用するかも調べなくてはならないと考えられる．

いずれにせよ，音声認識のためのマイクロフォンアレーはまだまだ改良の余地が残されていると考えられる．

第6章　結論

マイクロフォンアレーの手法を評価するための，様々なパラメータにおけるシミュレーションデータを生成するMATLABのM-fileを作成した．

高SNRであることが高認識率であることの必要条件であることから，波形の歪まない高SNRを実現することにより，認識率を高くすることができる可能性があることがわかった．

新たな手法により，通常のDelay-and-Sumよりも認識率を向上させることができた．

謝辞

本実習を進めるにあたり，常に懇切な御指導と御助言を戴きました ATR SLT 第 1 研究室室長 中村哲 博士および同研究室 研修研究員 堀内俊治 氏に心から御礼申し上げます.

また，御指導と御助言を戴きました SLT 第 1 研究室のみなさま，及び，助けていただいた様々な方々にも感謝いたします.

参考文献

[1] 大賀寿郎，山崎芳男，金田豊，“音響システムとディジタル処理，” 電子情報通信学会，(1995)

[2] 水町光徳，“マイクロホン対を用いた雑音除去に関する研究” 北陸先端科学技術大学院大学 博士論文，(2000)

[3] 中村哲, “実音響環境に頑健な音声認識を目指して,” 電子情報通信学会技術報告, SP 2002-12, pp. 31-36, (2002).

[4] 青木真理子, 古家賢一, “騒音下音声強調における空間情報の利用について,” 電子情報通信学会技術報告, SP 2002-11, pp. 23-30, (2002).

[5] 宝珠山治 杉山昭彦, “ブロッキング行列にリーク適応フィルタを用いたロバスト一般化サイドローブキャンセラ,” 電子情報通信学会論文誌 *A*, vol. J79-A, no 9, pp. 1516–1524 (1996).

[6] Joerg Bitzer, et al., “Multi-microphone noise reduction techniques as front-end devices for speech recognition,” *Speech Comunication*, vol. 34, pp. 3–12 (2001).

[7] Michael L. Selzer, et al., “Speech-Recognizer-Based Filter Optimization for Microphone Array Processing,” *IEEE Signal Processing Letters*, vol. 10, no 3, pp. 69–71 (2003).