音声言語情報処理 15－7
（1997. 2. 7）

マイクロフォンアレイを用いたCSCW向け音声認識モジュール
-アクティブマイクロフォン-

遠藤 隆　長屋 茂喜　中沢 正幸　古川 清　岡 隆一

新情報処理開発機構つくば研究センタ

〒305 茨城県つくば市竹園1-6-1つくば三井ビル13F
Tel:0298-53-1687 E-Mail:enchan@trc.rwcp.or.jp

あらまし　本報告では，画像認識結果から得られた話者の位置情報を用いて，マイクロフォンアレイの指向性を動的に制御し，話者からマイクまでの伝達特性を推定することで，多数の話者が同時に発声した場合やマイクから離れた位置から発声した場合にも認識性能の低下が少ないCSCW向け音声認識モジュールを提案する．連続音声からのワードスポッティングによる評価を行ない，(1)同時に2人の話者が発声している環境において認識率が30%改善，(2)160cm離れた位置で発声された場合に15%の認識率の改善，を確認した．

キーワード　マイクロフォンアレイ，音声認識，マルチモーダル，ジェスチャ認識

Speech Recognition Module for CSCW using a Microphone Array
- Active Microphone -

Takashi ENDO Shigeki NAGAYA Masayuki NAKAZAWA Kiyoshi FURUKAWA Ryuichi OKA

Tsukabu Research Center
Real World Computing Partnership

Tsukuba Mitsui Building 13F,
1-6-1 Takezono Tsukuba-shi, Ibaraki 305
Tel: 0298-53-1687 E-Mail:enchan@trc.rwcp.or.jp

Abstract - This report proposes a recognition module for use in CSCW that suffers little degradation in recognition performance even when more than one person speaks at the same time and they speak at a distance from a microphone. This is accomplished by controlling directionality using a microphone array and estimating transmission characteristics from speakers to microphones. On the basis of evaluation performed by word spotting from continuous speech, it has been found that this module raises the recognition rate by (1) 30% in an environment where two people are speaking at the same time, and (2) by 15% when people speak at a distance of 160 cm from a microphone.

key words Microphone array, Speech Recognition, Multiple modalities, Gesture Recognition

1. はじめに

我々が普段行っている会議では，アイコンタクトや表情といった形態のコミュニケーションが一つの重要な要素になっている．そこで，音声・ジェスチャーを入力インタフェースとすることで，通常の会議と同様，ユーザがより自然な形でコミュニケーションを取りながら協調作業を行うことができる，新たな形態のCSCWシステム，Open Cooperative Work Space(OCoWS)[Nagaya1996]を我々は提案している(図1参照)．OCoWSで用いる音声インタフェースは，離れた位置からの声も入力可能で，ユーザはマイクの位置を意識する必要のない，自然な音声インタフェースを目標としている．

このような音声インタフェースでは音声認識の性能が低下するという大きな問題が生ずる．これは，(1) ユーザのすぐ側にマイクがある場合に比べ目的話者以外の音声や雑音の影響が非常に大きくなる，(2) 壁からの反射等によって音声波形に歪み(伝達歪み)が生ずるため認識モジュール内のリファレンスと合わなくなる，の2 つの問題があるためである．

問題(1)のような目的話者音声以外の妨害音を抑制するために，我々はマイクロフォンアレイ[金田1995]を用いることにした．マイクロフォンアレイの音声認識への有効性に関しては，奈良先端大学の青木らが報告している[青木1996]．このとき目的話者位置の推定には，音声による推定手法[陶山1996]ではなく，より簡単な画像認識による推定手法[長屋1994]を採用した．問題(2)の伝達歪みについては，リファレンスに対して歪み適応を施すことにした．

本報告では，OCoWSを構成する音声認識モジュールとして，マイクロフォンアレイによる妨害音の抑制と，線形予測分析からの伝達特性の推定による歪み適応化により，複数話者が同時に発声する環境下やマイクロフォンから話者までの距離が大きい場合でも認識率の低下が少ないアクティブマイクロフォンを提案する．

2. アクティブマイクロフォン

アクティブマイクロフォンは，(1) 画像認識による目的話者位置の推定，(2)目的話者音声を選択的に抽出するマイクロフォンアレイ，(3)目的話者からマイクまでの伝達歪みを推定し，認識系を適応化する処理の3 つから構成され，次のように動作する．

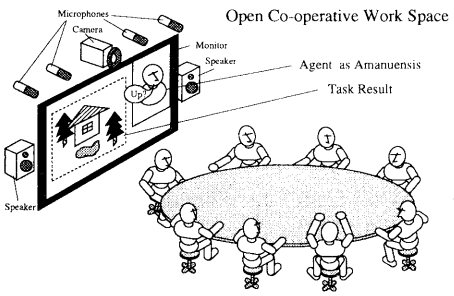

図1 Open Co-operative Work Space(OCoWS)

まず，複数話者の中から挙手などの特定の動作を行った者を目的話者として認識し，その位置を推定する．次に，マイクロフォンアレイとデジタルフィルタにより，マイクの指向性を動的に制御して目的話者に向け，目的話者以外の妨害音を抑制するとともに目的話者音声のS / N 比を改善する．最後に，あらかじめ推定した伝達歪みを畳み込んで適応させた認識用音声テンプレートを用いて音声認識処理を行う．

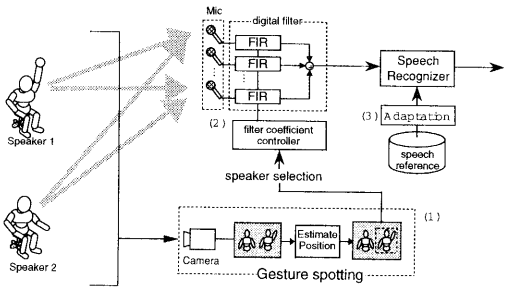

図2 アクティブマイクロフォン

2.1 ジェスチャ認識

既に我々は，複数の人物がいる映像から人物領域の切り出しや大きさの正規化といった操作なしに，人数・位置・ジェスチャの種別をリアルタイムで検出する手法を提案している[長屋1994]．図3にこのジェスチャ認識手法による検出結果を示す．画像中のテキストは話者毎に検出された重心位置・ジェスチャの種別を示している．システムは，あらかじめ設定された特定ジェスチャを行うユーザを見つけるとその位置情報をマイクロフォンアレイ制御部(2)に送る．

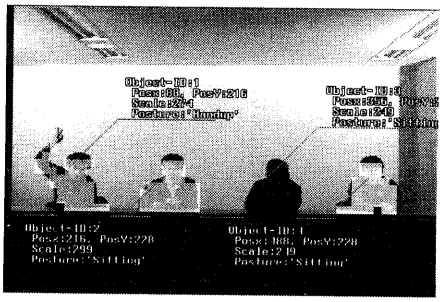
図3 画像認識による目的話者の位置推定例

2.2 目的話者音声の抽出

目的話者以外により生ずる妨害音を抑制するための指向性制御は，FIR フィルタを接続したマイクロフォンアレイにより実現できる．このFIRフィルタの係数推定にはA

MNOR[金田,1992]アルゴリズムを用いる.デジタルフィルタ部のシステム構成を図4に示す.

AMNORアルゴリズムでは,目的音源の方向からの信号に対してはゲイン1.0のディレイフィルタ,目的音源方向以外からの信号に対してはゲイン0　のフィルタとして働くよう係数を適応的に推定する.今回はRLS法[辻井,1995]により係数推定を行う.

目的音源方向の歪みと目的音源以外の信号の抑制の能力はトレードオフの関係にあり,目的音源方向の歪みを小さくすると,目的音源方向以外からの信号の抑制能力が低下する.フィルタの特性は白色雑音のゲインAの値で制御される.Aの値が大きい場合は,目的音源方向の信号に対して歪みが少なくなる方向に係数が誘導され,遅延和アレーに近づくように動作する.

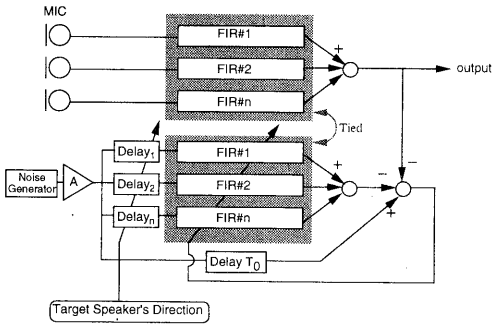

図4 デジタルフィルタ部

2.3 伝達歪みへの適応

図5に本システムの伝達歪みのモデルを示す.

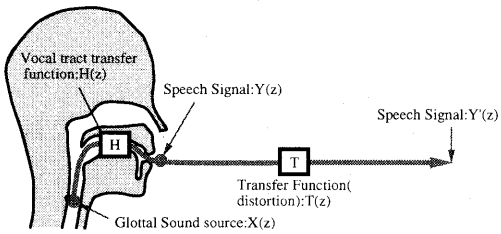

図5 歪みモデル

人間の音声は,母音の場合声帯の音源波形$X(z)$と声道伝達関数$H(z)$　との積としてモデル化することができる.口から発する音声を$Y(z)$とすると,

$$Y(z)=H(z)X(z) \tag{1}$$

となる.母音の場合声道伝達関数$H(z)$は,全極形のシステムとしてモデル化され,(2)式で表される.

$$H(z)=\frac{1}{1+\sum_{i=1}^{p}\alpha_i z^{-i}} \tag{2}$$

上式のα_iは線形予測分析により求めることができる.

また,マイクロフォンに到達する音声$Y'(z)$は,口からマイクロフォンまでの伝達特性を$T(z)$とするならば,(1)式との積として表される.

$$Y'(z)=T(z)H(z)X(z) \tag{3}$$

ここで$T(z)$が全極形システムとしてモデル化できるならば,$T(z)H(z)$もまた全極型となる.

$$T(z)H(z)=\frac{1}{1+\sum_{j=1}^{q}\beta_j z^{-j}} \tag{4}$$

上式のβ_jはマイクロフォンの入力信号$Y'(z)$から線形予測分析によって求めることができる.

(4)式を(2)式で割ることにより,$T(z)$が求められる.

$$T(z)=\frac{T(z)H(z)}{H(z)}=\frac{1+\sum_{i=1}^{p}\alpha_i z^{-i}}{1+\sum_{j=1}^{q}\beta_j z^{-j}}=1+\sum_{k=1}^{\infty}h_k z^{-k} \tag{5}$$

(5)式に示されるインパルス応答をh_kの値が十分小さくなる次数で打ち切ることで,FIRフィルタを構成し,音声認識系のリファレンス音声に畳み込むことで,伝達歪みに対する適応化処理が可能となる.

3. 評価実験

2.1節で述べたようにジェスチャ認識によって良好に目的話者の位置を求める手法は既に実現している.そこで今回は既にユーザの位置情報が得られているものとして,目的話者音声の抽出及び伝達歪みへの適応について評価実験を行った.具体的には,連続発声音声に対して,2.2,2.3節で述べた処理を適用した後,連続DPによるワードスポッティングを行い,その認識率を調べた.実験内容を以下の二通りである.

(1) 目的話者音声の抽出
- 2人の話者が同時発声している状況で一方の話者の音声を認識.

(2) 伝達歪み適応
- マイクから離れた位置の話者音声の認識.

このときの認識系の分析条件を表1に示す.

表1 分析条件

項目	内容
サンプリング	16[KHz],16[bit]
フレームシフト	7.5[ms]
フレーム長	12.5[ms]
特徴量	ぼかしベクトル場[岡,1986]

3.1 目的話者音声の抽出

2 人の男声話者が同時に発声している状況で，目的話者の音声を分離し，認識する実験を行った．

3.1.1 実験条件

2 人の話者の位置関係とマイクの配置を図6 に示す．実験では，図6中の話者Person1を目的話者として認識を行った．目的話者抽出の純粋な性能を調べるために，あらかじめ独立に収録した2 人の男声音声を計算機上で合成したものを実験試料として用いた．話者Person1の位置を既知として与えた．FIR フィルタのタップ数は，最も離れているマイク同士の距離を音速で割った値にサンプリング周波数を乗じた値以上になるように選択した．マイクの本数が2本，4本の場合に30タップ，8本の場合に66タップに設定した．フィルタ係数の推定には，話者Person2のみが発声している音声区間0.5秒を用いた．なお，実験はすべてオフラインで行った．

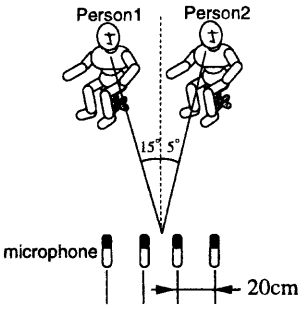

図6 話者分離実験条件

3.1.2 実験結果

マイクロフォンアレイを構成するマイクの本数とS/N 比の関係を図7 に，単語スポッティングの認識率を図8 に示す．マイクの本数に比例してS/N比が向上し，マイクを8本用いた場合，マイク1本の場合と比較して27[dB]のS/N比を得た．認識率も30%→60%と改善されたが，マイクの本数を2本より増やしてた場合，S/N比が向上し人間が聴取した場合は明確に聞き取りやすくなっているのが確認できたが，認識系の認識率の向上は確認できなかった．

マイクの本数ごとの指向性の分布状態を図7 に示す．マイクの本数が増える毎に指向性のビーム幅が狭くなることが確認できる．直線上に並んだ遅延和アレーでは，アレー長の拡大と共にビーム幅が狭くなる[金田1995]．AMNORも，目的話者方向の周波数歪みを抑える場合は遅延和アレーに性質が近づき，同様なビーム幅特性を示しているものと予想される．

なお，処理速度は実時間の約2倍であった．

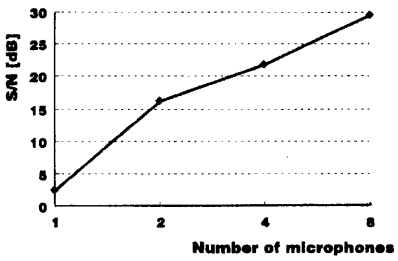

図7 マイクロフォンの数とS/N比

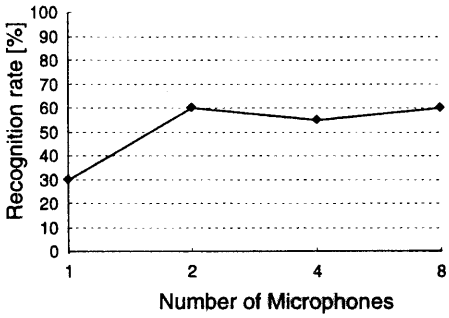

図8 認識率

3.2 伝達歪み適応

マイクと話者の距離を大きくした際に生ずる認識性能の低下が，伝達歪み適応によってどの程度改善されるかを評価する実験を行った．

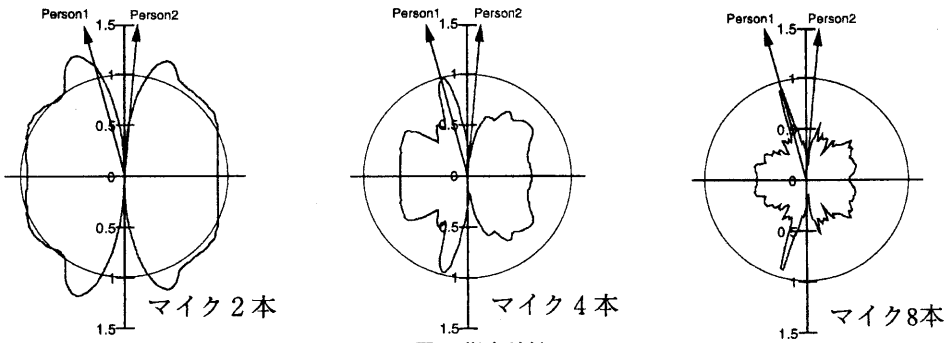

図7 指向特性

3.2.1 実験条件

話者とマイクロフォンの位置関係を図9に示す.実験条件を揃えるため,あらかじめ防音室で収録した音声をスピーカーで呈示した.リファレンス音声は特定話者を用いた.認識実験は以下の3条件で行った.

(条件1) 通常(ordinary)
図9中のマイク-Mic1で収録した音声を認識.

(条件2) 目的話者音声の抽出 (directivity control)
図9中のマイク2本を用いてマイクロフォンアレイを構成し,話者方向に指向性を制御して目的話者のみから収録した音声を認識.

(条件3) (条件2) +伝達歪み適応(adaptation)
(条件2) の音声について,さらにその伝達特性を推定し,認識系のリファレンス音声に畳み込んで認識.伝達特性は,2 秒分の音声から線形予測分析した結果により推定.

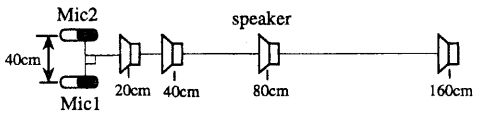

図9 話者とマイクロフォンの位置関係

3.2.2 実験結果

認識結果を図10に示す. (条件1)では,スピーカーとマイクとが離れるにつれ,認識率が劣化している.(条件2)の指向性を制御した認識結果では,S/N比が1～2dBほど改善され,距離による認識率の低下が抑えらることが確認された. (条件3)の指向性制御に伝達歪みの適応を組み合わせた結果では,実験したすべての距離において認識率の改善が認められ,伝達歪み適応の有効性が確認された.

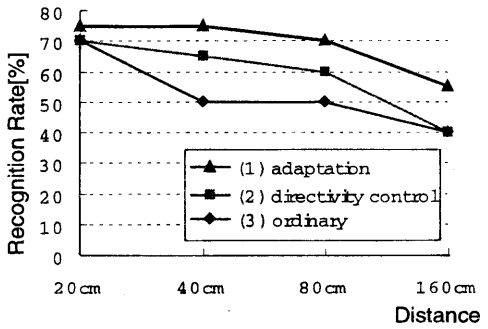

図10 認識率

図11にスピーカ・マイク間が160cmの場合の伝達特性の推定結果を示す.インパルス応答が7 0 サンプル程で収束しており,FIRフィルタとなっていることが確認できる.

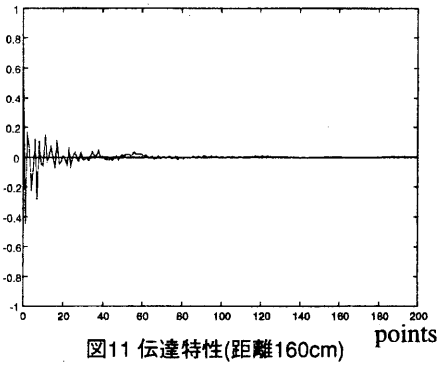

図11 伝達特性(距離160cm)

4. 考察

目的話者音声の抽出実験において,アクティブマイクロフォンを使わない場合でも30%の認識率が得られていたが,その内容を調べると,息継ぎなどのために話者2がちょうど発話していない区間であった.2本以上のマイクを使った場合は,話者2 が発声している区間の単語も認識可能となり,認識率の向上が図られていた.FIRフィルタによって話者2の音声が抑制された信号を聴取すると,話者2 の音声は摩擦音や破裂音などがめだつ,ささやき声のような音声になっていた.8本のマイクロフォンを用いて目的話者音声を抽出を行った際の目的話者方向と話者2の方向の周波数特性を図12に示す.話者2の方向の周波数特性は,低域のゲインが低く,高域に向かって高くなってゆくという,ちょうど人間の音声の周波数特性を上下に反転したような特性を示している.これは,FIRフィルタの係数を推定する際に,話者2の話者の音声を抑制するように適応推定した結果であり,その特性によって話者2の音声がささやき声のような音質になると説明される.

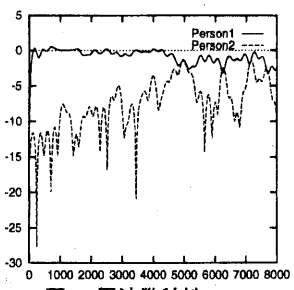

図12 周波数特性

伝達歪み適応では,特定話者リファンレンス音声を用いて行ったため,良好な結果が得られたが,不特定話者のデータへ適用する場合,話者適応も考慮に入れなければならず,今回の手法がそのまま適用可能かどうかは検討を要する.

今回の実験では，FIRフィルタ畳み込み処理に実時間の約2倍かかっているが，現在DSP上に実時間で動作するシステムをインプリメント中で，まもなく稼働予定である．また，FIRフィルタの適応推定処理のon/off制御については人手により行ったが，将来は口唇画像の認識による発話の有無の推定結果から自動することを検討している．

5. まとめ

本報告ではCSCW向け音声認識モジュールであるアクティブマイクロフォンについて提案した．アクティブマイクロフォンは，マイク位置に対してある程度自由な位置からユーザの発声を可能にするとともに，複数ユーザが同時に発声する場合でも個々の音声を認識する．アクティブマイクロフォンの構成要素の内，マイクロフォンアレイによる目的話者音声の抽出と，伝達歪みの推定による音声リファレンスの適応化処理について，その動作原理を説明した．また，評価実験を行い，(1)同時に2人の話者が発声している環境において認識率が30%改善すること，(2)160cm離れた位置で発声された場合に15%の認識率の改善することを確認し，本手法の効果を示した．

謝辞 本研究に日頃支援いただく，新情報処理開発機構島田潤一所長に感謝いたします．また，熱心に議論いただく情報統合研究室の皆様に深く感謝いたします．

参考文献

[Stefik1988]Stefik, Bobrow, Kahn, Lanning, Suchman : "WYSIWIS Revised : Early Experience with Multiuser Interface", ACM Trans. on Office Information Systems,5(2), pp.147-167 (1988).

[Harrison1990]Harrison, Minneman, Stults, Weber : "Video, A Design Medium. Report No.", SIGCHI Bulletin (1990).

[Nagaya1996]Nagaya, Endo, Itoh, Kiyama, Oka : "A Proposal of Novel Information Integration Architecture - Open Cooperative Work Space", Proc. of the International Conference on Multisensor Fusion and Integration for Intelligent Systems, pp.425-432 (1996 Washinton.D.C).

[金田1992]金田："アダプティブマイクロホンアレー"，信学論誌，B-II, Vol. J75-BII, No.11, pp. 742-748 (1992-11).

[金田1995]金田："マイクロホンアレーによる指向性制御"，音響学会誌，Vol. 51, No. 5, pp. 390-394 (1995-5).

[陶山1996]陶山，高橋，岩倉："2段階のデータ選別による複数音源定位"，信学論誌，A, Vol. J79-A, No. 6, pp.1127-1137 (1996-6).

[辻井1995]辻井："適応信号処理"，昭晃堂 (1995).

[青木1996]青木，山田，滝口，中村，鹿野："マイクロフォンアレーとHMM合成による実環境における音声認識実験"，音講論集，2-Q-2 (1996-9).

[長屋1994]長屋，関，岡："多重解像度特徴によるジェスチャ認識"，信学技報，PRU95-99 (1995-7).

[岡1989]岡："音素片候補からなるフレーム記号列に対する文章スポッティング"，音講論集，3-6-7, pp.83-84(1989-3).

[木山1993]木山，伊藤，岡："連続構造化法を用いた文意スポッティング"，音講論集，2-7-7, pp.83-84 (1993-10).

[岡1986] 岡："スペクトルベクトル場とスペクトルの音声認識における有効性比較について"，信学論誌，D, J69-D, No.11, pp.1704 (1986-11).

[伊藤1995]伊藤，木山，関，小島，張，岡："同時複数話者の音声およびジェスチャのリアルタイム統合理解におけるNovel Interface System"，情報処理学会，音声言語情報処理研究会資料，7-3 (1995-7).