

発話位置依存 CMN とマルチマイクロフォンアレイ処理の併用 による遠隔発話の音声認識

王 龍標[†] 北岡 教英[†] 中川 聖一[†]

† 豊橋技術科学大学 情報工学系

〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †{xiaowang,kitaoka,nakagawa}@slp.ics.tut.ac.jp

あらまし 遠隔環境において、伝送歪みは音声認識の性能を大きく劣化させる。本稿では、発話位置依存ケプストラム平均正規化とマルチマイクロフォンアレイ処理の併用による遠隔発話の音声認識を提案する。各マイクロフォンによる認識結果によって認識単語候補に投票し、最多得票の単語を最終の結果として選択する（投票法），もしくは各マイクロフォンによる単語尤度を同一単語毎に加算して尤度最大の単語を最終の結果として識別する（最大連合尤度法）。さらに、投票法あるいは最大連合尤度法と遅延和ビームフォーミングを統合するマルチマイクロフォンアレイを提案する。マルチマイクロフォンアレイ処理前に、各チャンネルの入力を位置依存 CMN により補正し、音声認識を行う。シミュレーション環境と実環境において孤立単語認識実験を行った。実環境において、提案した発話位置依存 CMN とマルチマイクロフォン処理の併用手法では、従来の発声毎 CMN に基づく遅延和ビームフォーミング処理より 3.2% (50.0% の相対エラー減少率) の改善を達成することができた。

キーワード 遠隔発話音声認識、位置依存ケプストラム平均正規化、マルチマイクロフォン処理、マイクロフォンアレイ

Robust Distant Speech Recognition by Combining Multiple Microphone-array Processing with Position Dependent CMN

Longbiao WANG[†], Norihide KITAOKA[†], and Seiichi NAKAGAWA[†]

† Department of Information and computer Sciences, Toyohashi University of Technology

1-1, Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan

E-mail: †{xiaowang,kitaoka,nakagawa}@slp.ics.tut.ac.jp

Abstract In a distant environment, channel distortion may drastically degrade speech recognition performances. In this paper, we propose a robust distant speech recognition by combining multiple microphone-array processing with position dependent Cepstral Mean Normalization (CMN). In this paper, the maximum voting method (that is, *Voting method*) or the maximum summation likelihood (that is, *Maximum-summation-likelihood method*) of whole channels were used to obtain the final result. Furthermore, we combined the delay-and-sum beamforming with *Voting method* or *Maximum-summation-likelihood method*, which was called multiple microphone-array processing. Before multiple microphone-array processing, the system compensated the input features by proposed position dependent CMN and performed speech recognition for each channel. We conducted our experiments in both simulated and real environments. The proposed combinational method improved the 100 isolated word recognition performance remarkably in both situations. We achieved 3.2% improvement (50.0% relative error reduction rate) than the delay-and-sum beamforming with conventional CMN in a real environment.

Key words distant speech recognition, position dependent CMN, multiple microphone processing, microphone array

1. はじめに

現在、音声インターフェースが急速に普及してきているが、近接マイクロфонを用いたものが主である。より多くのアプリケーションにおいて自然で使い勝手のよい音声インターフェースを実現するためには、実環境下のハンズフリー音声認識[1][2]の技術が必要となる。近年、実環境におけるハンズフリー音声処理/認識に関する研究が行われている。文献[3]では、高残響環境下における反射音を利用したマルチビームフォーミング法を提案し、その実環境における有効性は音声認識実験などを通じて実証された。文献[4]では、出力音声を強調するかわりに、正確な仮説の尤度を最大化する特徴の系列を生成するマイクロфонアレー処理方法が提案された。文献[5]では、マイクロфонアレーを用いてハンズフリー連続数字の認識実験を行っている。

しかしながら、遠隔環境下で、さまざまな距離からの音声による音声を認識することは直接音の減衰や反射音の重複により性能低下を招く。マイクロфонと音源の距離が1m~2mになると大きく認識精度は低下する[6]。これらを含む性能低下の原因は実環境とトレーニング環境のミスマッチであるといえる。このミスマッチを減少するために、入力音声の特徴量を補正するのは重要な方法である。特に、伝達特性のミスマッチを補正する簡易で効果的な解決策として、CMN(Cepstral Mean Normalization)[7]がよく使用されている。CMNは簡単に実現できるために、現在多くのシステムに使われている。しかし、一般的なCMNでは、認識する音声の入力が終了した後に補正パラメータを推定するため、実時間処理向きではない[7]。時間同期型の認識アルゴリズムは本来入力と並行して認識処理が進められるが、一般的なCMNを併用すると並行に処理が行えなくなる。この問題に対し、リアルタイムに認識を行うために、過去に認識した音声から補正パラメータを事前に算出しておき、音声を補正する方法が提案されている[8]。この方法はreal-time CMNと呼ばれている。しかしながら、遠隔環境で異なる発話位置からの伝達特性はそれぞれ異なるため、このreal-time CMNでは話者移動によって発生する伝達特性の急速な変化に追従できず、遠隔ハンズフリー音声認識におけるミスマッチの補正には不十分である。

本研究では、話者位置によって事前推定されたパラメータを用いたPosition Dependent CMN(位置依存CMNと呼ぶ)による頑健な遠隔音声認識手法および話者認識手法を提案する。部屋をいくつかの区域に分割しておき、事前に区域の中心位置からマイクロфонまでの伝達特性(位置依存の補正パラメータ)を計測しておく。本システムは、四つのマイクロфонペアの間の到着時間差によって音源位置を推定する[10][11][12]。文献[10][11][12]では、話者位置同定結果は話者がどの区域(60cm×60cm)で発声したのかを推定できることを示した。そして、推定した発話位置によって事前に計測した補正パラメータを選択し、CMNによって音声を補正して遠隔発話の音声認識を行う。

また、遠隔環境下では、各マイクロфонの音声信号はCMNによる正規化誤差、マイクロфонの位置、音源とマイクロфонとの距離、発声方向及びマイクロfonの品質などによって影響される。単一のマイクロfon音声処理では、これらの状況に十分に対応できず、音声認識性能の低下を招く。そこで、新しいマルチマイクロfon音声処理を提案する。音声認識システムでは、各マイクロfonに収録した音声信号を独立に認識する。各マイクロfonによる認識結果によって認識単語候補に投票し、最多得票の単語を最終の結果として選択する(投票法)、もしくは各マイクロfonによる単語尤度を同一単語毎に加算して尤度最大の単語を最終の結果として識別する(最大連合尤度法)。複数のマイクロfonでは、異なるマイクロfonアレイによって形成するビームは実験環境、マイクロfonペアの距離、マイクロfonの品質などによって違う指向性音声を形成する。従って、複数のマイクロfonアレイと提案した投票法と最大連合尤度法を統合し、マルチマイクロfonアレイ音声処理手法を提案する。

2. 位置依存ケプストラム平均正規化

2.1 一般的なCMNとreal-time CMN

実環境で音声認識を用いる際に考慮すべき雑音として、加算性雑音のほかに乗算性雑音がある。これは音声の伝達特性への影響のことであり、マイクロfonなどの受音系の違いや話者とマイクロfonの距離の違いなどの要因によって生じる。音声認識において学習環境とテスト環境における各ケプストラム係数の平均の差を補正するケプストラム平均正規化(Cepstral Mean Normalization: CMN)[7]は簡易で効果的な伝達特性正規化手法の一つである。この方法は伝達チャンネルと録音装置に起因する歪みを補正することができる。

音声信号 s に伝達特性 h の影響が加わった場合を考える。このとき観測音声信号 x は

$$x = s \otimes h \quad (1)$$

となる。この音声をケプストラム領域に変換すると、

$$C^x = C^s + C^h \quad (2)$$

となる。従って、認識する音声のケプストラム C_t は、乗算性雑音(伝達特性もしくはチャンネル歪み)はケプストラム領域においてテスト環境とトレーニング環境との伝達特性の差を表すパラメータ ΔC を用いて

$$\bar{C}_t = C_t - \Delta C \quad (t = 0, \dots, T) \quad (3)$$

と補正される。

一般的なCMNにおいて、補正パラメータ ΔC は

$$\Delta C = \bar{C}_t - \bar{C}_{train} \quad (4)$$

として求められる。ここで、 \bar{C}_t と \bar{C}_{train} はそれぞれ認識する音声のケプストラムの平均とトレーニング音声のケプストラムの平均である。従って、従来のケプストラム平均正規化手法で

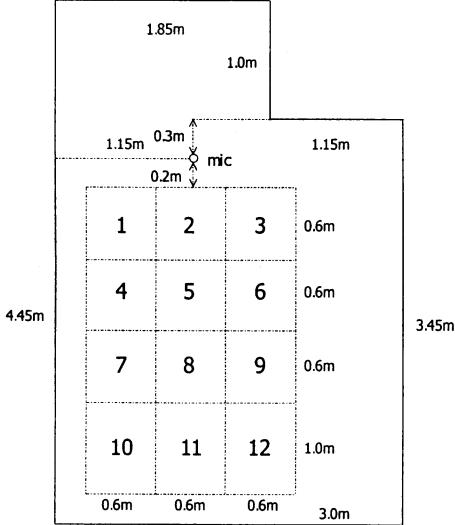


図 1 部屋の構造図（部屋の大きさ：横 = 3m 縦 = 3.45m 高さ = 2.6m）

は、認識する音声の入力が終了し、補正パラメータを算出してからしか認識が開始できない。時間同期型の認識アルゴリズムは入力と並行して認識処理が進められるが、一般的な CMN を併用すると並行に処理が行えなくなり、実時間処理ができない。一般的なケプストラム正規化手法のもう一つの問題は、認識する発話が短い場合ケプストラム平均が正確に推定できることが挙げられる。

文献[8]では、チャンネル歪みが急激に変化することのないものであるとの考えに基づいて、次のように改良を行った。この方法では、過去に認識したいくつかの発話から補正パラメータを事前に算出しておく。新しい補正パラメータは

$$\Delta C = \bar{C}_{environment} - \bar{C}_{train} \quad (5)$$

になる。ここで、 $\bar{C}_{environment}$ は特定の環境下で発声された発話のケプストラム平均である。本稿ではこの方法を従来の CMN と区別し、real-time CMN と呼ぶ。文献 [8] では、補正パラメータは過去に認識した発話によって事前に計算される。即ち、 n 番目の発話に対する補正パラメータを

$$\Delta C^{(n)} = (1 - \alpha) \Delta C^{(n-1)} - \alpha \times (\bar{C}_{train} - \overline{C^{(n-1)}}) \quad (6)$$

として求める。ここで、 $\Delta C^{(n)}$ と $\Delta C^{(n-1)}$ はそれぞれ $n - 1$ 番目と n 番目の発話に対する補正パラメータ、 $\overline{C^{(n-1)}}$ は $n - 1$ 番目の入力音声のケプストラム平均である。

Real-time CMNにより、入力音声終了を待たずに認識処理が開始できる。さらに、real-time CMNの補正パラメータは十分な発話のケプストラム係数から推定されるため、一般的なCMNより伝達特性歪みをより正確に補正することができる。

2.2 話者位置同定の real-time CMN への導入

CMN、特に real-time CMN は音声認識システムの性能の改

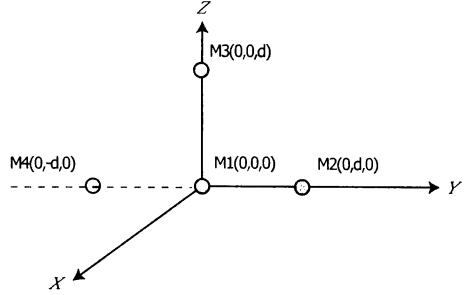


図2 マイクロフォンの配置 (d=20cm)

善に有効である。しかしながら、遠隔環境下では、話者とマイクロフォンの距離やそれによる部屋の反響などのために、異なる発話位置からの伝達特性は非常に違ってくる。従って、話者位置が変化する場合、伝達特性歪みが急激に変化するために、real-time CMN では十分に対応できない。

そこで本稿では、話者位置同定を real-time CMN と併用する位置依存ケプストラム平均正規化 (Position Dependent CMN: PDCMN) を用いる [11] [12] [13]。位置依存ケプストラム平均正規化の補正パラメータは

$$\Delta C = \bar{C}_{position} - \bar{C}_{train} \quad (7)$$

となる。ここで、 $\bar{C}_{position}$ はある特定の位置で発声された発話のケプストラム平均である。本稿では、部屋は図1のように12の区域に分割し、各区域の中央から発声された音声を用いてケプストラム平均 $\bar{C}_{position}$ を計測する。

3. マルチマイクロフォン音声処理手法

本研究の位置依存 CMN では話者位置推定に図 2 のように 4 つのマイクロфонを用いることを想定している。しかし、話者と各マイクロfon間の伝達特性やマイクロfonの個体差により、各マイクロfonでの受音を認識した結果は異なる。ここで、それを利用した音声認識性能の改善を試みる [14] [15]。

3.1 投票法

4つのマイクロフォンによる認識結果に従って投票し、最多得票の単語を最終の結果として選択する。この方法を投票法と呼ぶ。投票法は次のように定義する。

$$\hat{W} = \arg \max_{W_R} \sum_{i=1}^{\#channel} I(W_i, W_R), \quad (8)$$

$$I(W_i, W_R) = \begin{cases} 1 & \text{if } (W_i = W_R) \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

ここで、 W_i は i 番目のマイクロフォンの認識単語であり、 $I(W_i, W_R)$ はインジケータを示す。 $\#channel$ はマイクロフォン数である。最多得票の結果が 2 種類以上あれば、話者に最も近いマイクロフォンの結果を認識結果として出力する。

3.2 最大連合尤度法

各マイクロフォンによる単語尤度を 4 つのマイクロフォンにわたって同一単語毎に加算して尤度最大の単語を最終の結果とする。この方法は最大連合尤度法と呼ぶ。最大連合尤度法は次のように定義する。

$$\hat{W} = \arg \max_{W_R} \sum_{i=1}^{\# \text{channel}} L_{W_R}(i), \quad (10)$$

ここで、 $L_{W_R}(i)$ は i 番目のチャンネルの単語 W_R の尤度である。

3.3 マルチマイクロフォンアレイ処理

遠隔環境下で、マイクロフォンアレイを用いて目的信号の方の指向性を形成することにより、高音質に目的音声を受音する遅延和アレイ [16] がよく利用されている。Flanagan らは遅延和アレイの性能を改善する方法として、低次の反射音にも超指向性を形成して有効利用するマルチビームフォーミング [17] を提案している。しかし Flanagan らのマルチビームフォーミングは、部屋の形状や音源の位置などの情報は既知であるという前提条件が必要であった。また、Flanagan らの方法では、複数のビームを形成するのに、一つのアレイしか用いないため、実環境下で十分に対応できない。一方、複数のマイクロフォンでは、異なるマイクロフォンアレイにより形成するビームは実験環境、マイクロフォンペアの距離、マイクロフォンの品質、測定誤差などによって違う指向性音声を形成する。これらの異なる指向性音声による認識結果を投票法あるいは最大連合尤度法により統合する方法を提案する。これをマルチマイクロフォンアレイ法と呼ぶ。そこで、図 2 の四つのマイクロフォンによってアレイ 1 (マイク 1, マイク 2, マイク 3), アレイ 2 (マイク 1, マイク 2, マイク 4), アレイ 3 (マイク 1, マイク 3, マイク 4), アレイ 4 (マイク 2, マイク 3, マイク 4), アレイ 5 (マイク 1, マイク 2, マイク 3, マイク 4) の 5 つのアレイを独立に構成し、これらの結果を投票法または最大連合尤度法で統合する。

4. 実験

4.1 実験条件

提案手法を不特定話者小語彙 (100 単語) 孤立単語認識により評価した。実験は $3.45m \times 3m \times 2.6m$ の部屋で行った。特に障害物のない環境 (シミュレーション環境) と様々な物で実環境を再現した環境 (実環境) を用意した。シミュレーション環境では、部屋の中で、マイクロフォンと音源以外に何にも置いてない。実環境では、図 3 のように、白板やテーブルや椅子やテレビなど物は部屋の中に置いた (ゼミナール部屋)。部屋は図 1 のように 12 (3×4) 個の矩形の区域に分割した。各区域の中心位置からマイクロフォンまでの伝達特性 (位置依存の補正パラメータ) をスピーカーから再生した音声のケプストラム平均を求める事により事前に計測しておいた。区域の幅の $60cm$ の半分以内の誤差で音源位置を精度良く同定するのは可能である [11] [12]。本実験では、発話位置に対応する区域



図 3 実験を行うゼミナール部屋 (実環境)

は正確に推定できると仮定した。

テストデータは東北大・松下単語データベースの 20 話者の発話を用いた。発話内容は各話者 200 個の孤立単語であり、最初の 100 単語はテストデータとして使用し、残りの 100 単語は式 (7) におけるケプストラム平均 $\bar{C}_{\text{position}}$ を推定するために使用した。すべての発話は各区域の中心位置においてスピーカーから再生することにより各位置において発声した発話をシミュレートした。サンプリング周波数は $12kHz$ である。フレーム長は $21.3ms$ 、フレーム周期は $8ms$ 、256 点数のハミング窓を用いた。116 個の 4 状態の left-to-right 音節 HMM [18] [19] を音響モデルとして使用した。各状態の出力分布は全共分散行列をもつ多次元正規分布の 4 混合である。HMM のパラメータは、ATR 連続音声テキスト 503 文の 8 話者による読み上げ音声および ATR 音韻バランス 216 単語の 10 話者による読み上げ音声から切り出した音節により音節毎に学習し、日本音響学会音声データベース (30 話者) および新聞読み上げコーパス JNAS (125 話者) を用いて MAP 推定による連結学習を行った。学習データのケプストラム平均は JNAS コーパスの学習データを用いて求めた。特徴量は 10 次元の MFCC とそれらの Δ , $\Delta\Delta$ 係数および対数パワーの Δ , $\Delta\Delta$ 係数から構成される、計 32 次元である。

4.2 シングルマイクロフォンによる認識実験

遠隔環境においてスピーカーから再生された孤立単語の認識実験を行った。

シミュレーション環境と実環境下における認識結果は表 1 に示す。ここで、全ての結果は 4 つの独立マイクロフォンから得られた認識結果の平均値である。表 1 では、提案手法である Position Dependent CMN ('PDCMN') をベースライン ('CMN なし')、一般的な CMN ('発声毎 CMN')、代表的な区域 (ここでは区域 5) のケプストラム平均を用いた CMN ('区域 5 の CMN') および Position Independent CMN ('PICMN') と比較している。発声毎 CMN では各発話から求めたケプストラム平均に基づいて補正パラメータを算出した。'区域 5 の CMN' は部屋全体の中心位置のみの音声から事前に求めたケプストラム平均を補正パラメータとして全て

表 1 シングルマイクロфонによる認識結果 (4つ独立マイクロfonによる得られた結果の平均値: %).

区域	シミュレーション環境 (empty room)					実環境 (seminar room)				
	CMN なし	発声每 CMN	区域 5 の CMN	PICMN	PDCMN	CMN なし	発声每 CMN	区域 5 の CMN	PICMN	PDCMN
1	90.1	94.5	96.2	96.3	96.6	86.3	92.8	94.2	95.0	95.7
2	94.7	96.4	97.7	97.7	97.7	95.4	95.7	97.4	97.7	97.4
3	90.2	94.0	96.0	96.2	96.2	94.3	94.6	96.8	97.1	96.8
4	88.5	93.3	94.9	95.0	95.2	87.4	92.9	93.1	94.6	95.6
5	91.5	94.6	96.7	96.7	96.7	92.1	93.8	96.3	96.0	96.3
6	87.2	92.8	95.0	95.2	95.4	90.9	93.2	95.2	96.1	95.9
7	86.6	93.2	94.3	94.6	95.4	89.0	92.4	94.3	94.8	95.7
8	86.8	92.8	94.4	94.5	94.6	91.4	91.4	93.8	94.1	94.7
9	87.3	91.6	94.0	94.3	94.4	92.3	93.1	95.9	96.4	96.0
10	82.3	90.7	92.6	92.7	93.3	84.9	90.0	90.5	91.8	93.5
11	85.7	90.8	93.4	93.5	93.7	86.9	90.9	91.7	93.2	94.1
12	84.7	89.5	92.5	92.7	92.8	85.9	89.8	90.9	93.3	93.3
Ave.	88.0	92.9	94.8	95.0	95.2	89.7	92.5	94.2	94.9	95.4

の区域の音声を補正する方法である。PICMN では、すべての 12 区域の補正パラメータを平均して算出し、位置に関わらず同一の補正パラメータを用いてケプストラム係数を補正した。なお、PDCMN と PICMN では、20 人に対してそれぞれ共通な補正值 (ΔC) を用いた。

まず、CMN を用いない場合、認識率はマイクロfonからの距離が大きくなるにつれて認識性能が大きく劣化していることがわかる。部屋の反響などの影響により伝達特性が大きく変化し、認識に影響していることが確認できる。発声毎 CMN を用いた場合、効果はあるが十分ではない。本実験で用いたテストデータの発話長が短く、音韻的なバランスが発話内でとれていなかったため、補正值を求めるためのケプストラム平均の信頼性が低い。したがって、伝達特性の補正効果と音韻バランスの悪さが相殺される形となった。もちろん、長い発話でテストすればあまり問題とならないと考えられるが、実用を考えた場合にコマンド等短い発話を用いられることは多いと考えられ、問題になる。一方、PICMN や「区域 5 の CMN」、PDCMN では、事前に十分な量の発話から補正值を求めておくため、この問題が生じない。すべての区域においてベースラインや発声毎 CMN の性能を上回った。

シミュレーション環境下で、PDCMN は「区域 5 の CMN」あるいは PICMN よりわずかの改善であったが、実環境（ゼミナール部屋）では、PDCMN は「区域 5 の CMN」あるいは PICMN より音声認識率が有意に改善できた。実環境下で、提案した PDCMN は CMN なしより 55.3%、発声毎 CMN より 38.7%、「区域 5 の CMN」より 20.7%、PICMN より 9.8% の相対エラー減少率を達成することができた。実環境下の結果により、音源とマイクロfonの距離は遠いほど、その改善が大きい。シミュレーション環境下では、部屋の中で障害物がないので、異なる区域からマイクロfonまでの伝達特性の差は大きくないと考えられる。一方、実環境（ゼミナール部屋）では、様々な物の反射の影響により伝達特性の位置に依存性が大きくなり、異なる区域からマイクロfonまでの伝達特性の差はシ

ミュレーション環境よりかなり大きいため、提案した位置依存 CMN (PDCMN) は他の方法より有効な改善ができた。

4.3 マルチマイクロfonによる認識実験

シミュレーション環境と実環境下での、シングルマイクロfonと複数マイクロfonによる結果を表 2 と表 3 に示す。投票法 (Voting method: VM) または最大連合尤度法 (Maximum-summation-likelihood method: MSLM) いずれもはシングルマイクロfonより頑健な認識結果が得られた。実環境下で、PDCMN を用いる場合、最大連合尤度法はシングルマイクロfonより 21.6% の相対エラー減少率を達成した。PDCMN とマルチマイクロfonアレイ（遅延和アレイ & 最大連合尤度法）の併用手法は PDCMN と遅延和アレイの併用よりも 11.1%、発声毎 CMN と遅延和アレイの併用（即ち、従来手法）よりも 50.0% の相対エラー減少率を達成した。最大連合尤度法は投票法より若干良いが、この原因は各チャンネルの尤度の加算により全てのチャンネルの潜在的な信頼性を利用できるためと考えられる。シミュレーション環境下では、シングルマイクロfonと同じ傾向があり、PDCMN と PICMN の差はあまり大きくなかった。一方、実環境下で、マルチマイクロfonにより PDCMN は PICMN より大きく改善ができた。最大連合尤度法と遅延和アレイを統合する場合、PDCMN は PICMN よりも 11.1% の相対エラーを削減した。

5. まとめ

本稿では、投票法及び最大連合尤度法という新しいマルチマイクロfon音声処理による頑健な遠隔発話の音声認識手法を提案した。さらに、投票法または最大連合尤度法と位置依存ケプストラム平均正規化との併用によるマルチマイクロfonアレイ法を提案した。シミュレーション環境と実環境において不特定話者小語彙（100 単語）孤立単語認識により評価した。実環境で、位置依存ケプストラム平均正規化は一般的な CMN や位置独立ケプストラム平均正規化より認識率を大幅に改善できただ。また、提案したマルチマイクロfonアレイ音声処理手法

表 2 シングルマイクロフォンと複数マイクロフォンによる音声認識結果の比較 (シミュレーション環境: %)

	single micro- phone	multiple microphones								
		VM	MSLM	beamforming					VM + beamforming	MSLM + beamforming
				array 1	array 2	array 3	array 4	array 5		
CMN なし	88.0	89.7	89.2	90.6	91.2	90.8	90.9	91.4	91.6	91.6
発声毎 CMN	92.9	94.2	94.1	93.7	94.6	93.8	94.0	94.4	94.7	94.8
PICMN	95.0	96.0	96.0	95.7	96.1	95.8	96.0	96.1	96.4	96.4
PDCMN	95.2	96.0	96.1	95.9	96.3	96.0	96.2	96.2	96.5	96.5

表 3 シングルマイクロフォンと複数マイクロフォンによる音声認識結果の比較 (実環境: %)

	single micro- phone	multiple microphones								
		VM	MSLM	beamforming					VM + beamforming	MSLM + beamforming
				array 1	array 2	array 3	array 4	array 5		
CMN なし	89.7	91.3	91.6	90.9	91.3	91.3	91.0	91.4	91.9	91.9
発声毎 CMN	92.5	94.2	94.5	93.5	93.3	93.3	93.4	93.6	94.1	94.2
PICMN	94.9	96.0	96.2	95.7	96.1	95.8	96.0	96.1	96.3	96.4
PDCMN	95.4	96.4	96.6	96.2	96.3	96.2	96.1	96.4	96.7	96.8

はシングルマイクロフォン音声処理手法あるいは単一の遅延和アレイより良い認識性能が得られた。PDCMN とマルチマイクロフォンアレイ（遅延和アレイ & 最大連合尤度法）の併用手法は発声毎 CMN と遅延和アレイの併用よりも 50.0% の相対エラーを削減した。

文 献

- [1] B.H. Juang, F.K. Soong, "Hands-free telecommunications", Proceedings of the Workshop on Hands-Free Speech Communication (HSC-2001), pp. 5-10, 2001.
- [2] T. Takiguchi, S. Nakamura, K. Shikano, "HMM-Separation-Based Speech Recognition for a Distant Moving Speaker", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 2, pp. 127-140, Feb. 2001.
- [3] 西浦 敬信, 中村 哲, 鹿野 清宏, “反射音を利用したマルチビームフォーミングによる音声認識”, 電子情報通信学会論文誌, Vol. J83-D-II, No. 11, pp. 2198-2205, Nov. 2000.
- [4] M.L. Seltzer, B. Raj, R.M. Stern, "Likelihood-Maximizing Beamforming for Robust Hands-free Speech Recognition" IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 5, pp. 489-498, Sep. 2004.
- [5] M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani, "Experiments of hands-free connected digit recognition using a microphone array", Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 490-497, 1997.
- [6] 中村 哲, “外に強い音声認識を目指して”, 日本音響学会誌, Vol. 57, No. 10, pp. 662-667, 2001.
- [7] S. Furui, "Cepstral analysis technique for automatic speaker verification", J. Acoust. Soc. Amer., Vol. 55, pp. 1204-1312, Jun. 1974.
- [8] N. Kitaoka, I. Akahori, S. Nakagawa, "Speech recognition under noisy environments using spectral subtraction with smoothing of time direction and real-time cepstral mean normalization", Proceedings of the Workshop on Hands-Free Speech Communication (HSC-2001), pp. 159-162, 2001.
- [9] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay", IEEE Trans. Acoust. Speech Signal Process., ASSP24(4):320-327, Aug. 1976.
- [10] L. Wang, N. Kitaoka and S. Nakagawa, "Distant speech recognition based on position dependent cepstral mean normalization", Proceedings of the 6th IASTED International Conference on Signal and Image Processing (SIP-2004) pp. 249-254, 2004.
- [11] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speech recognition based on position dependent CMN", Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH2004-ICSLP), pp.2409-2052, 2004.
- [12] 王 龍標, 北岡 教英, 中川 聖一, “発話位置依存ケプストラム平均正規化による遠隔発話の音声認識”, 電子情報通信学会技術研究報告, 2004-SLP-51, pp. 25-30, May. 2004.
- [13] 王 龍標, 北岡 教英, 中川 聖一, “発話位置依存ケプストラム平均正規化による遠隔発話の音声認識と話者認識”, 電子情報通信学会技術研究報告, 2004-SP-79, pp. 47-52, Nov. 2004.
- [14] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speech recognition based on position dependent CMN Using a Novel Multiple Microphone Processing Technique", Proc. of EUROSPEECH-2005, pp. 2661-2664, 2005.
- [15] 王 龍標, 北岡 教英, 中川 聖一, “発話位置依存 CMN とマルチマイクロフォン処理による遠隔発話音声認識”, 日本音響学会講演論文集, 2-7-13, pp.83-84 (2005.9)
- [16] J. Flanagan, J. Johnston, R. Zahn and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms", J. Acoust. Soc. Amer., Vol. 78, pp. 1508-1518, June, 1985.
- [17] J. Flanagan, A. Surendran and E. Jan, "Spatially selective sound capture for speech and audio processing", Speech Communication, Vol. 13, pp. 207-222, 1993.
- [18] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition.", Proc. International Workshop on Automatic Speech Recognition and Understanding, pp.393-396, 1999.
- [19] 中川 聖一, 花井 建豪, 山本 一公, 峯松 信明, “HMMに基づく音声認識のための音節モデルと triphone モデルの比較”, 電子情報通信学会論文誌, Vol.J83-D II, No.6, pp.1412-1421, Jun. 2000.