論　文

MLLRを用いた音響モデルの教師なし環境雑音適応アルゴリズム

山田　実一[†]　馬場　朗[††]　芳澤　伸一[†††]　米良祐一郎[†]
李　晃伸[†]　猿渡　洋[†]　鹿野　清宏[†]

Unsupervised Acoustic Model Adaptation Algorithm Using MLLR in Noisy Environment

Miichi YAMADA[†], Akira BABA[††], Shinichi YOSHIZAWA[†††], Yuuichiro MERA[†], Akinobu LEE[†], Hiroshi SARUWATARI[†], and Kiyohiro SHIKANO[†]

あらまし　MLLRと多数話者データベースを用いたHMM音響モデルの教師なし環境雑音適応アルゴリズムを提案する．対象話者の任意の1文発声と居室雑音を入力として音声データベースから環境適応用雑音重畳データを生成することで，話者に負担をかけずに大量の適応用データを得ることができる．具体的な適応処理は以下の3段階からなる．(1) GMMを用いた話者識別を用いて，入力話者と音響的距離の近い話者をデータベースから選択する．(2)(1)の選択話者の読み上げ音声をデータベースから抽出し，居室雑音を重畳する．(3) その雑音重畳音声を適応サンプルとしてMLLRによる適応を行う．更に，十分統計量と話者距離による教師なし話者適応及びHMM合成法を統合することで，高精度な教師なし統合適応システムを構築する．大語彙連続音声認識において評価した結果，提案手法による適応モデルは環境Matched Modelと同等以上の認識精度を示し，数十サンプルを用いた教師ありMLLRに近い性能が得られた．本適応システムによって認識率はSNRが20 dBの雑音環境下においてmonophoneモデルで48.3%から70.5%に，PTMモデルで60.1%から79.9%に改善された．

キーワード　音響モデル，環境適応，教師なし適応，MLLR，大語彙連続音声認識

1. まえがき

実環境で音声認識を利用するには，話者適応や環境適応などの適応技術が必要不可欠である．音声は話者ごとに音響的特徴が異なるため，一般に不特定話者の音響モデルは，特定話者の音響モデルよりも認識性能が低い．また，雑音が重畳した音声はパワーの低い音声が雑音に隠れるため認識が困難となる．特に，様々な音響的環境でも頑健に動作することのできる実用的な音声認識システムを実現するためには，少量の入力サンプルから高速かつ高精度に適応を行うことのできる適応手法が不可欠である．

雑音環境のための音響モデルの構築手法としては，雑音を学習用音声データベースに重畳して音響モデルを構築する方法が単純であり高い認識性能が得られるが，学習には大量の音声コーパスを必要とする．また，あらゆる環境を想定して各々の環境のための音響モデルをあらかじめ構築しておく必要があり，現実的ではない．

そこで少量のデータに基づいて既存のモデルを適応させる音響モデルの適応技術が盛んに研究されている．この音響モデルの適応手法は大きく分けて教師あり適応と教師なし適応に分けられる．教師あり適応の代表的な手法にはMLLR（Maximum Likelihood Linear Regression）[1]が挙げられ，現在の音声認識システムにおいて広く用いられている．しかし，一般に十分な適応性能を得るには数十文の発声が求められ，またテキストに即した正確な発声が求められるため発声者への負担が大きい．環境適応に関してはHMM合成法[2]についても研究されているが，非線形変換処理

[†] 奈良先端科学技術大学院大学情報科学研究科，生駒市
Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, 630–0101 Japan
[††] 松下電工株式会社東京研究所，東京都
Tokyo Research Laboratory, Matsushita Electric Works, Ltd., Tokyo, 108–8351 Japan
[†††] 松下電器産業株式会社先端技術研究所，京都府
Advanced Technologies Research Laboratories, Matsushita Electric Industrial Co., Ltd., Souraku-gun, Kyoto-fu, 619–0237 Japan

や近似の合成分布は高い精度が得られておらず，大語彙連続音声認識のような大規模なタスクにおいては性能が不十分である．

本論文では，まず雑音環境について，発声者の負担にならずにかつ短時間で適応を行うことを目的として，多数話者音声データベースと MLLR を用いた教師なし環境適応アルゴリズムを提案する．更に，その環境適応アルゴリズムと十分統計量に基づく教師なし話者適応 [6] を組み合わせることで，任意の 1 発話から教師なしの話者適応と環境適応を同時に行う統合的な適応システムを構築する．以下，**2.** で従来手法について述べ，**3.** で教師なし環境適応アルゴリズム，**4.** でこれを用いた統合適応システムを提案する．**5.** では提案手法及びシステムを大語彙連続音声認識で評価し，HMM 合成や，環境 Matched Model，MLLR などの従来手法と比較する．更に，**6.** で十分統計量を用いた雑音適応手法との比較を行う．**7.** で考察をまとめ，**8.** で本論の結論を述べる．

2. 従来の環境適応アルゴリズム

本章では従来の環境適応アルゴリズムの特徴と問題点を述べる．

(a) HMM 合成法

HMM 合成法 [2] は，クリーン環境の音声で学習した HMM と雑音を用いて構築した HMM を合成し，雑音環境の音声 HMM を構築する手法である．音声と雑音は，線形スペクトル領域での加法性が成り立つ性質を利用する．実際には音声認識では特徴量はケプストラム領域で表現されているので，いったん線形スペクトル領域に戻した後，各々を結合し再びケプストラム領域に戻す．ケプストラム領域での音声の確率分布を S_{cep}，雑音の確率分布を N_{cep} とすると，合成分布 O は式 (1) のように表される．C はコサイン変換である．

$$O = \exp(C \cdot (S_{cep})) + \exp(C \cdot (N_{cep})) \qquad (1)$$

HMM 合成法は短時間での適応処理が可能である．しかし，線形スペクトル領域に変換する際に指数変換を行うなど，近似的に雑音に対応した HMM を構築しているため，合成した HMM の精度は高いとはいえない．大語彙連続音声認識などの大規模なタスクにおいては良い性能が得られていない．

(b) Matched Model の作成

Matched Model は，学習用の音声データに対象環境の雑音を重畳して構築したモデルである．この手法は単純で原始的であるが，従来の環境適応において最も効果のある手法である．しかし，大量の学習コーパスからモデルを再構築するのは学習処理に膨大な学習時間を要する．あらかじめ様々な雑音を考慮した Matched Model を作成しておくことも考えられるが，実環境におけるすべての雑音状況を想定することは難しい．このため，適応手法として実際の音声認識システムに用いるのは困難である．

(c) MLLR

MLLR [1] は，初期音響モデルを適応話者の発声データに合致する（ゆう度が大きくなる）ように変換行列を用いて初期モデルのガウス分布の平均値を移動し，かつ，分散値を変更する話者適応アルゴリズムである．初期音響モデルの確率分布の平均と分散の線形変換行列を最ゆう推定する．音韻分布の空間を数十のサブ空間に分け，各々のサブ空間ごとに線形変換を行う．比較的短い時間で適応が行え，かつ効果が大きいので，一般的に広く使用されている．雑音の重畳した音声を用いて MLLR を行うことで，環境適応と話者適応を同時に行うことが可能である．

しかし MLLR は教師あり適応手法であり，適応サンプルの正確なトランスクリプション（音声の書き起し）が与えられることが前提となる．十分な適応を行うためには，指定したトランスクリプションに忠実に従った発声サンプルが数文章から数十文章程度必要であり，すべてを正しく発声するには利用者への負担が大きい．特に駅での発券案内タスクなど，不特定多数にサービスを提供するタスクを考えるとき，このような前提は成立しない．

3. MLLR による教師なし環境適応アルゴリズム

既存の大量の話者の音声データベースを利用して MLLR を用いた環境適応を行うことを考える．すなわち，適応話者の発声を直接用いる代わりに，適応話者に比較的近い音声データ及びそのトランスクリプションをデータベースから大量に抽出し，それらに実環境の雑音を重畳して適応データとして用いる．

本提案手法は教師なしの話者適応手法である．入力音声を音響的距離の近い話者の選択のみに用いるため，入力音声に対してトランスクリプションを必要とせず，任意の発声で適応が行える．

提案する環境適応アルゴリズムを図 1 に示す．あ

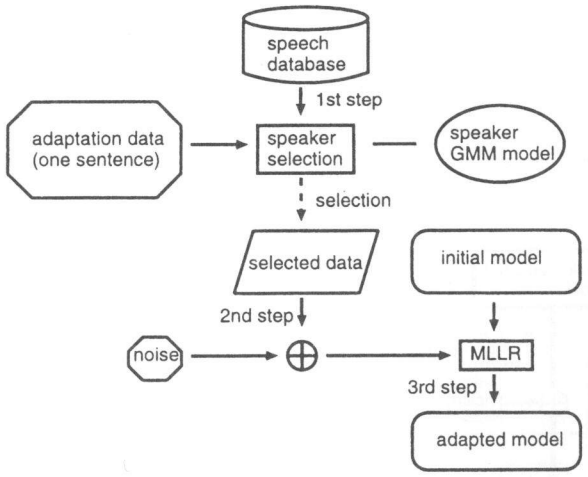


図 1　提案する環境適応アルゴリズム
Fig. 1　Proposed environment adaptation algorithm.

らかじめ，多数話者の音声データベースと，各データベース話者を識別するための話者モデルを用意する．音声データベースはクリーン（雑音が極端に小さい）環境で収録された音声で，トランスクリプション付きであるとする．これには既存の読み上げ音声データベースを用いることができる．話者モデルは音韻独立に構築した GMM（Gaussian Mixture Model）を用いる．これは上記の音声データベース（クリーン環境）から作成する．入力には，雑音環境における任意の 1 文発声を用いる．また，これとは別に環境雑音データを収録する．適応アルゴリズムは 3 ステップからなる．第 1 ステップでは，入力データから特徴抽出した音響特徴量を入力としてデータベース上の全話者の GMM のゆう度を計算し，その上位の話者を，発声者に対して音響的特徴の近い話者集合として選択する．第 2 ステップでは，その選択話者集合の音声データとトランスクリプションをデータベースから抽出し，音声データに対象の環境雑音を重畳する．第 3 ステップでは，第 2 ステップの雑音重畳データを用いて MLLR により音響モデルの環境適応を行う(注1)．

音声データベースを利用することで，少ない任意の入力をもとに大量の環境適応用データを用意することができる．音声データベースにはあらかじめ音声データに対して正確なトランスクリプションが与えられており，MLLR を精度良く実行することができる．また大量の適応データを適応に用いることができるため，適応の収束も早く，比較的短時間で適応が行える．

選択話者の雑音重畳音声データでの MLLR 学習は，環境適応だけでなく話者適応も行っている．更に話者適応の性能を向上させるには，初期音響モデルを改善することも必要となる．特に，MLLR による適応アルゴリズムでは，初期モデルの性能が適応後のモデルの性能に影響を及ぼすことが知られている [4], [5]．次の 4. では，この初期モデルとして，十分統計量と話者選択による教師なし学習 [6] によって，話者適応した音響モデルを用いる．このように話者適応による音響モデルを初期モデルとして用いた環境適応アルゴリズムは，教師なし話者適応アルゴリズムとのある種の統合とも考えられるので，統合システムと呼ぶことにする．

4.　教師なし統合適応システム

初期音響モデルの性能を改善することは，MLLR による環境適応の性能の向上につながることが期待できる．よって，前章で提案した環境適応アルゴリズムの初期音響モデルとして，教師なし話者適応アルゴリズム [6] による話者適応音響モデルを用いる．このように前章で提案した環境適応アルゴリズムを教師なし話者適応と組み合わせることで，教師なしの統合適応システムを構築することができる．

この十分統計量に基づく話者適応手法は，対象話者に音響的特徴が近い話者の十分統計量を用いて対象話者に適応した音響モデルを構築する手法である．この手法は教師なしの適応であり，かつ本適応アルゴリズムと同じく任意の 1 文発声と GMM による話者モデルを用いた教師なし適応であること，また短時間で適応が行えることから，本環境適応アルゴリズムの初期音響モデルとして容易に用いることができる．

本統合適応システムの流れを図 2 に示す．システムは大きく分けて話者選択部，話者適応部，環境適応部の 3 部からなり，話者適応を行った後に環境適応を行う．話者選択部は，話者適応部と環境適応部で共用される．あらかじめ，多数話者の HMM に関する十分統計量を不特定話者の音響モデルより算出し蓄積するとともに，同話者の GMM 話者モデル及び環境適応用音声データベースを準備する．適応処理の具体的な流れは，以下のとおりである．(1) 1 文発声から抽出した音響特徴量と話者モデルを用いてゆう度を計算し，音響的距離の近い話者を選択する．(2) 選択した複数話者の全十分統計量を用いて話者適応した音響モデルを

(注1)：MLLR を用いて環境適応を行う場合は，確率分布の平均だけでなく分散も適応し，かつ繰り返し適応を行うことが極めて効果的であることが確かめられたので本論文の環境適応でも同様の操作を行う [3]．

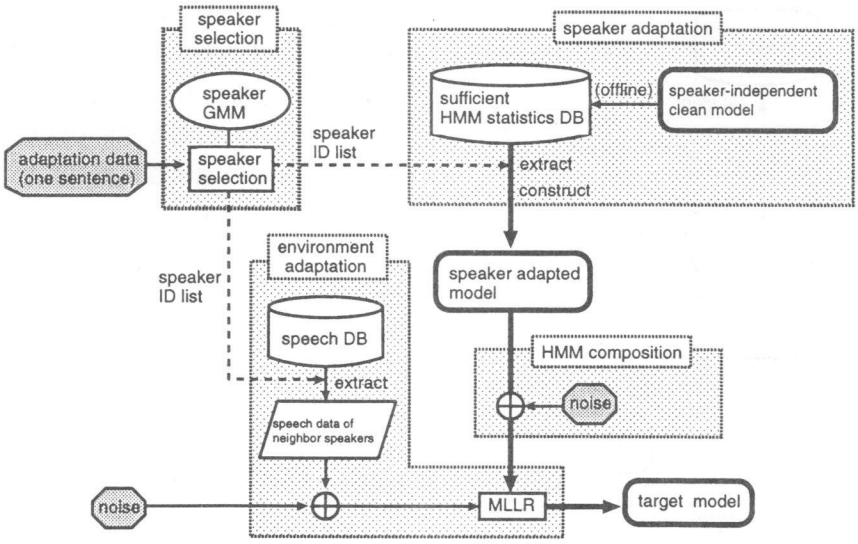


図 2　統合（話者+環境）適応システム
Fig. 2　Integrated adaptation system.

構築する．(3)(1) で選択した話者の音声データを音声データベースから抽出し，雑音を重畳する．(4)(3) のデータを用いて MLLR により (2) の話者適応音響モデルを環境適応する．本統合適応システムは一貫した教師なしの適応システムであり，任意の 1 文発声で話者適応と環境適応を行うことが可能である．

4.1　HMM 合成を用いた初期モデルの改善

環境適応前の初期モデルの性能を押し上げるため，更に話者適応モデルに雑音モデルを HMM 合成 [2] する手法を導入する．話者適応モデルは CMN 処理を行った学習データで構築しているので，あらかじめ学習データからケプストラム平均値を求めておき，それを話者適応モデルに加えた後 HMM 合成を行い，再びケプストラム平均値を差し引く．これによって最終的な認識性能を改善する効果が期待できる．

4.2　入力音声の低パワーの特徴量除去

環境雑音が重畳している入力音声では，雑音の影響でパワーの低い音声部分が隠れて話者識別が難しくなり，話者選択に誤りが生じる傾向が予備実験において確認された．そこで，雑音の影響が小さいパワーの高い部分のみで話者選択を行う手法を用いる．具体的には，入力音声を特徴抽出した後，パワーの低い特徴量を間引いてから話者選択を行う．

5.　大語彙連続音声認識による評価実験

提案した環境雑音適応アルゴリズムの評価を，大語彙連続音声認識で行った．評価する音響モデルは monophone モデルと PTM [10] モデルである．

タスクは日本音響学会の新聞記事読み上げ（JNAS）コーパスの語彙数 2 万語のディクテーションである．適応前の monophone モデルは，「日本語ディクテーション基本ソフトウェア」1999 年度版（IPA'99）[9] の不特定話者 HMM を使用する．PTM モデルは，日本音響学会の新聞記事読み上げ（JNAS）コーパス（男女各 153 名，合計約 45000 文）を学習セットとして不特定話者モデルを構築した．テストセットは，IPA-98-TestSet，つまり，男女各 23 名の合計 200 文の読み上げ音声にそれぞれ 25 dB，20 dB，15 dB の雑音を重畳したものを用いる．適応データは各話者について，JNAS コーパスからテストセット，話者 GMM の学習セット，及び環境適応用データベースに用いるデータセットを除いた中から無作為に 1 文を選び出し，それに雑音を重畳したものを使用する．

言語モデルの単語 N-gram 及び単語辞書は，IPA'99 の毎日新聞記事 75 か月分のものを用いる．認識率は正解率判定ツール [8] を用いて正解文章と自動比較して算出する．音響分析条件は文献 [9] に準ずる．デコーダは Julius [7] を用いた．

5.1　MLLR による教師なし環境適応アルゴリズムの評価

提案した環境適応アルゴリズムを評価した．十分統計量による話者適応のみを行ったモデル（これを提案手法の初期モデルとする）と，提案手法で環境適応を行ったモデルとの性能を比較する．JNAS コーパス約

表 1　実験条件
Table 1　Experimental conditions.

話者数	306（男性 153，女性 153）
話者選択用 GMM	64 混合ガウス分布
選択話者人数 （発声者は除く）	20 （monophone モデル） 40 （PTM モデル）
十分統計量話者数	306
環境適応用選択データ数	20，60，100，200 文 （選択話者ごとに 1，3，5，10 文）

45000 文のうち，話者モデル及び十分統計量に用いたのは約 42000 文（話者ごとに約 140 文）で，残りの約 3000 文（話者ごとに約 10 文）を環境適応部の音声データベースとして使用する．重畳する雑音として，180 秒の居室雑音（主に計算機雑音と空調のファン音）を用いる．実験条件を表 1 に示す．話者選択用 GMM はクリーン環境の音声データで構築した．なお，話者選択を行う際，評価対象話者の GMM は選択対象から除外する．

実験結果を図 3 に示す．結果の詳細は付録 **1.** を参照されたい．なお話者選択用 GMM はクリーン音声で作成されているが，実環境を想定して雑音を含む音声で話者選択を行った場合（図中の「Select-noisy」），及び話者選択を雑音重畳前のクリーンな適応データで行った場合の実験値（図中の「Select-clean」）を比較する．また適応はテストセットの話者ごとに 1 文のみで行った．

提案した環境適応手法により，話者適応のみの場合（図中の before）から monophone モデルでは約 8～25%，PTM では約 7～25%の認識率の向上が見られた．SNR が低いほど，認識率の改善幅が大きい傾向が見られた．また，選択データ数を増やすにつれて認識率が向上したが，比較的少ない選択データ数でも高い性能が得られることが示された．雑音重畳音声による話者選択（Select-noisy）は，どの SNR においてもクリーン音声による話者選択（Select-clean）と同等の性能が得られた．ただし精度の差は SNR が低くなるにつれて若干広がる傾向が見られた．

5.2　統合適応システムの性能評価及び改善

MLLR による教師なし環境適応アルゴリズムと話者適応を統合した統合適応システムの適応性能を，統合適応前の不特定話者モデル，更に HMM 合成法 [2] 及び環境 Matched Model と比較した．HMM 合成法，環境 Matched Model に使用する初期モデル及び雑音データはともに **5.1** と同じものを使用した．HMM 合成法による適応音響モデルは，1 状態 1 混合の雑音 HMM と初期モデルを HMM 合成したものである．環境 Matched Model は，初期モデルを雑音を重畳した JNAS コーパス全体で EM アルゴリズムによる学習を行ったものである．統合適応システムにおける選択データ数は 200 文を用いた．また提案手法については，**4.1** の HMM 合成による初期モデルの改善を施さない場合についても調べた．

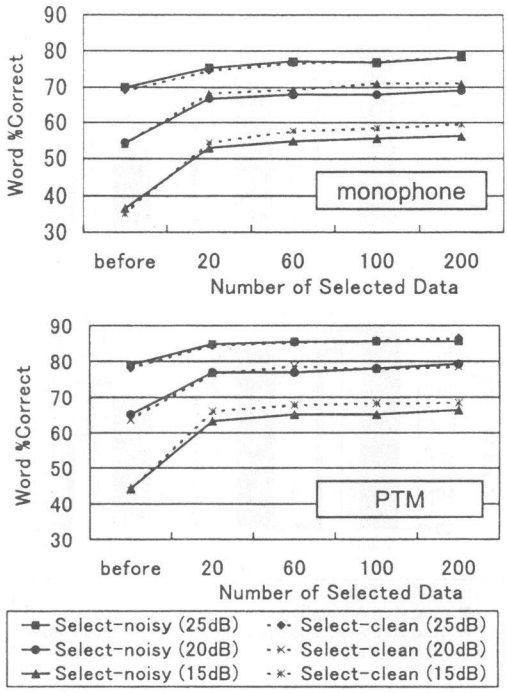


図 3　提案した環境適応アルゴリズムの効果
Fig. 3　Effect of proposed environment adaptation algorithm.

結果を図 4 に示す．詳しい結果は付録 **2.** を参照されたい．提案システムで適応した音響モデルは，HMM 合成法によるモデルよりも monophone で約 12～15%，PTM で約 7～12%高い認識率を示し，環境 Matched Model と同程度かそれ以上の性能が得られた．また環境適応の前に雑音モデルを HMM 合成することで，初期モデルが改善され，更なる認識率の向上が得られることが確認できた．

5.3　MLLR による教師あり適応との比較

提案手法を MLLR による教師あり適応と比較した．MLLR による教師あり適応の初期音響モデルとして不特定話者音響モデルを用いた．教師あり適応学習用音声としては，居室雑音を重畳した発声者自身の音声データを用いた．この適応学習用音声データは，評価

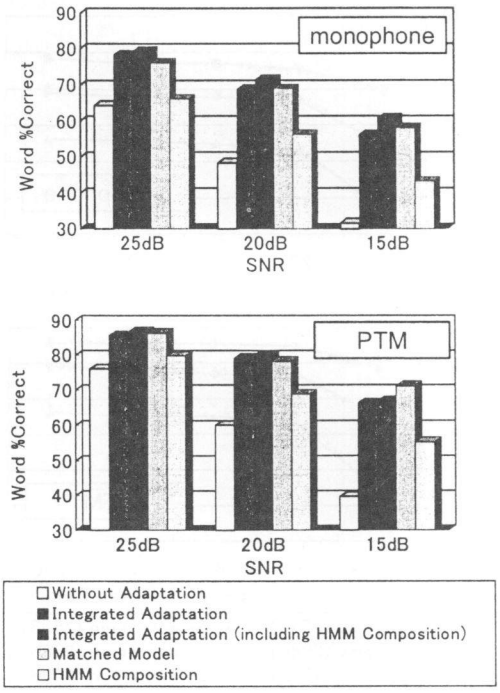


図 4 統合適応システムの精度比較
Fig. 4 Comparison of adaptation systems.

表 2 MLLR による適応実験の条件
Table 2 Condition of MLLR experiment.

音声データベース	JNAS コーパス（音素バランス文）
雑音データ	居室雑音（180 秒）
適応データ	話者ごとに 10 文と 50 文に雑音重畳
SNR	20 dB

表 3 MLLR との比較結果
Table 3 Comparison with MLLR.

monophone モデル

	適応手法	適応文数	単語正解率
教師あり	MLLR	10	72.0%
		50	76.9%
教師なし	統合適応システム（選択データ数 200）	1	71.2%

PTM モデル

	適応手法	適応文数	単語正解率
教師あり	MLLR	10	80.3%
		50	84.9%
教師なし	統合適応システム（選択データ数 200）	1	78.5%

用音声や GMM の作成に用いた音声データとは別のものである．実験条件を表 2 に，認識実験結果を表 3 に示す．統合適応システムによる適応音響モデルは，MLLR による適応モデルに及ばないものの，1 文の教師なし適応データで MLLR に迫る性能が得られることがわかった．

表 4 環境適応時間の比較
Table 4 Comparison of adaptation time.

適応手法	文章数	適応時間
Matched Model	約 45000	24 h
HMM 合成法	なし	30 sec
提案法 （選択データ数 20）	1	30 sec
（選択データ数 60）		100 sec
（選択データ数 100）		160 sec
（選択データ数 200）		370 sec

（model: monophone, 提案法は MLLR を行った時間）

5.4 処理時間の比較

monophone モデルにおけるおおよその環境適応のための処理時間の比較を表 4 に示す．提案法は Matched Model に比べてはるかに速く，HMM 合成法と同程度の高速な適応処理が行えることが確認された．なお PTM モデルの処理時間は monophone モデルの約 4 倍であった．このことから，適応にかかる処理時間がモデルのもつガウス分布数にほぼ比例していると考えられる．

6. 雑音重畳音声の十分統計量を用いた適応法との比較

本提案手法と異なるアプローチの一つとして，十分統計量と話者距離による教師なし話者適応 [6] の枠組みで環境適応も行うことが考えられる．すなわち，話者ごとに，想定される雑音ごとの重畳音声データをも含んだ十分統計量を準備しておいて，それらの中から入力に近い雑音と話者を同時に選択する．これによって話者適応のみならず環境適応をも一括して行うことができる．本章ではこの手法と提案手法を比較する．話者選択は，雑音重畳音声による GMM 話者モデルを用いて行う．この手法は図 2 の統合適応システムにおいて，話者選択部と話者適応部（環境適応としても使用）を使用し，環境適応部を削除したものと考えられる．実験条件を表 5 に示す．GMM 及び十分統計量は，clean 及び 2 種類の SNR で構築した．話者選択部においては，まず，ゆう度が高い上位 100 個の雑音重畳話者を選択し，その中で一番多かった SNR を対象の SNR として決定する．その後その SNR の話者データの中からゆう度上位話者を適応用話者として選択する．

monophone モデルでの実験結果を表 6 に示す．適応データの SNR と話者選択部において決定された SNR のミスマッチは見られなかった．雑音環境下においては適応前と比べ約 30%の認識率の向上が見られ，

表5　十分統計量を用いた適応実験の条件
Table 5　Condition of noise adaptation using sufficient statistics.

雑音データ	居室雑音（180秒）
SNR	20 dB，10 dB，(clean)
音響モデル	monophone（16混合ガウス分布）
話者数	306（男性153，女性153）
GMM	64混合ガウス分布
話者GMM数	306 × 3
十分統計量数	306 × 3
選択話者人数	20（発声者は除く）
適応データ	1文章に雑音重畳
評価データ	IPA-98-TestSet（46人，200文）に雑音重畳

表6　十分統計量を用いた適応による単語正解率
Table 6　Word %correct of noise adaptation by sufficient statistics method.

	SNR		
	clean	20 dB	10 dB
適応前	83.2	48.1	13.7
適応後	85.8	74.4	48.3
統合適応システム	*	71.2	*

(model: monophone)

提案手法を上回る性能を示した．しかし，この適応法はあらかじめあらゆる雑音を想定して十分統計量と話者モデルを用意しなければならないという問題点があり，雑音の種類とSNRに対応した莫大な十分統計量と話者モデルを必要とするので実際の使用は難しい．一方提案手法は環境に動的に対応するためより少ないデータベースですみ，効率が良いといえる．

7. 考　　察

5.1において，SNRが下がるにつれて（特に15 dB）選択cleanと選択noisyの認識率の差が若干広がった．これは選択noisyにおいて話者選択がうまく行われていないことが考えられる．しかし実際には選択された話者を比較してみると，選択cleanのゆう度上位話者が選択noisyに現れていない場合が数例であり，認識率の差はそれほど大きくなかったことから，雑音環境下においても話者選択はほぼ成功していると考えられる．これは，**4.2**で述べたパワーの低い音響特徴量をカットした効果が現れていると考えられる．

Matched Modelは性能は良いが，1回の適応に対して非常に多くの学習データを必要とし，また学習時間が非常にかかるという難点がある．HMM合成法は適応時間は非常に短いが性能があまり良くないので，大語彙連続音声認識において高い性能は望めない．教師あり適応であるMLLRは適応時間が比較的短く，性能も良い．しかし，数十文を正確に発声しなければならず，ユーザに大変な負担が強いられる．提案法は，適応時間が短く，適応データも1文と短く，更にこの適応データは話者の音響的特徴の近い話者を選択するためのものなので，正確に発声する必要もなく，発声者の負担も極めて小さい．また，25 dBの雑音環境下で86%（PTM）近い認識率が得られた．これは，HMM合成法よりも高く，Matched Modelと同程度以上の性能であった．

8. むすび

本論文では，音声データベースとMLLRを用いて任意の1文発声から教師なしの環境適応を行うアルゴリズムを提案した．対象話者と音響的特徴が近い話者の音声をデータベースから大量に抽出し，環境雑音を重畳してMLLRを行うことにより，発声者へ負担をかけずに比較的短時間で高精度な環境適応が可能である．更に，本手法に十分統計量と話者距離を用いた教師なし話者適応とHMM合成法を統合した，教師なし統合適応システムを提案した．本システムはHMM合成法による環境適応よりも高い適応性能を示し，短い適応処理時間でMatched Modelと同程度以上の認識性能が得られた．また，教師あり適応で数十文の決められた文発声を必要とする通常のMLLRと比較して，本システムは任意の1文のみの入力でMLLRに迫る認識率が得られた．提案手法による音響適応モデルを大語彙連続音声認識によって評価したところ，認識率はSNRが20 dBの雑音環境下で，monophoneモデルにおいては48.3%から70.5%，PTMモデルで60.1%から79.9%に改善された．

今後の課題としては，話者選択部，初期モデルの更なる検討を行う．また，データベースの中に音響的特徴の近い話者が存在しない場合の環境適応についても検討を行う．

謝辞　本研究は，NEDO（新エネルギー・産業技術総合開発機構）の援助を受けて行われた．御協力頂いた関係各位に感謝致します．

文　　献

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, pp.171–185, 1995.

[2] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech composition of hidden Markov models," Proc. Eurospeech93, vol.3, pp.1031–1034, Sept.

1993.
[3] P.C. Woodland, D. Pye, and M.J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," Proc. ICSLP, pp.1133–1136, 1996.
[4] M. Padmanabhan, L.R. Bahal, D. Nahamoo, and M.A. Picheny, "Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition system," Proc. ICASSP95, pp.701–704, 1995.
[5] Y. Gao, M. Padmanabhan, and M. Picheny, "Speaker adaptation based on pre-clustering training speakers," Proc. Eurospeech97, pp.2091–2094, 1997.
[6] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, and K. Shikano, "Unsupervised speaker adaptation based on the sufficient HMM statistics of selected speakers," Proc. ICASSP2001, May.2001.
[7] 李 晃伸，河原達也，堂下修司，"単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識，" 信学論 (D-II)，vol.J82-D-II, no.1, pp.1–9, Jan. 1999.
[8] 山本俊一郎，伊藤克亘，鹿野清宏，中村 哲，"ディクテーションにおける日本語の特性を考慮した単語正解率判定ツール，" 音響学会講演論文集，pp.155–156, March 1999.
[9] 河原達也，李 晃伸，小林哲則，武田一哉，峯松信明，嵯峨山茂樹，伊藤克亘，伊藤彰則，山本幹雄，山田 篤，宇津呂武仁，鹿野清宏，"日本語ディクテーション基本ソフトウェア（99 年度版）の性能評価，" 情処学研報，2000-SLP-31-2, 2000.
[10] 李 晃伸，河原達也，武田一哉，鹿野清宏，"Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識，" 信学論 (D-II)，vol.J83-D-II, no.12, pp.2517–2525, Dec. 2000.

付　　録

1. 選択データ数と単語正解率

提案した環境適応アルゴリズムにおける選択データ数と単語正解率の一覧を表 A·1 に示す．

2. 種々の環境適応手法による単語認識率

種々の環境適応手法ごとの単語正解率と単語正解精度を表 A·2 に数字で示す．

表 A·1　選択データ数に対する単語正解率の詳細
Table A·1　Word %correct of various number of selected data.

SNR=25 dB
適応前（選択 clean）：68.9%(monophone)，78.3%(PTM)
適応前（選択 noisy）：70.0%(monophone)，79.2%(PTM)

適応データ数	20	60	100	200
選択 clean（monophone）	74.6	76.8	77.2	78.6
選択 noisy（monophone）	75.4	77.2	76.9	78.4
選択 clean（PTM）	84.5	85.4	85.8	86.6
選択 noisy（PTM）	84.9	85.5	85.6	85.7

SNR=20 dB
適応前（選択 clean）：54.1%(monophone)，63.2%(PTM)
適応前（選択 noisy）：54.5%(monophone)，65.2%(PTM)

適応データ数	20	60	100	200
選択 clean（monophone）	68.1	69.2	71.1	71.2
選択 noisy（monophone）	66.7	67.9	67.9	68.9
選択 clean（PTM）	76.7	78.7	77.8	78.5
選択 noisy（PTM）	77.0	77.0	78.1	79.2

SNR=15 dB
適応前（選択 clean）：35.0%(monophone)，43.6%(PTM)
適応前（選択 noisy）：36.6%(monophone)，44.3%(PTM)

適応データ数	20	60	100	200
選択 clean（monophone）	54.5	57.8	58.5	59.5
選択 noisy（monophone）	53.0	55.0	55.7	56.3
選択 clean（PTM）	66.0	67.8	68.2	68.4
選択 noisy（PTM）	63.3	65.2	65.2	66.4

表 A·2　種々の環境適応手法ごとの認識率の詳細
Table A·2　Recognition rate of various adaptation methods.

monophone モデル

Algorithm/SNR	25 dB	20 dB	15 dB
適応なし	64.3[62.3]	48.3[46.0]	31.5[30.0]
HMM 合成	66.1[63.9]	56.2[54.0]	43.1[40.3]
Matched Model	76.1[74.1]	68.9[67.1]	58.0[55.3]
統合適応	78.4[76.4]	68.8[67.0]	56.3[54.2]
統合適応（含 HMM 合成）	78.2[76.6]	70.5[68.6]	58.0[55.5]

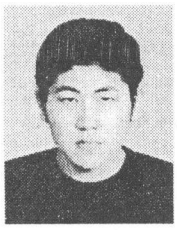
(Word %correct [Word accuracy])

PTM モデル

Algorithm/SNR	25 dB	20 dB	15 dB
適応なし	76.0[74.1]	60.1[57.2]	39.7[37.7]
HMM 合成	79.7[77.4]	68.9[66.1]	55.2[51.9]
Matched Model	86.3[84.6]	78.2[76.2]	71.2[69.3]
統合適応	85.7[84.1]	79.2[77.5]	66.4[63.4]
統合適応（含 HMM 合成）	86.7[85.1]	79.9[77.9]	67.0[64.2]

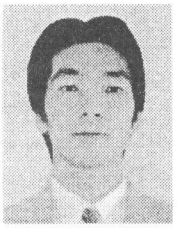
(Word %correct [Word accuracy])

（平成 13 年 6 月 4 日受付，9 月 14 日再受付）



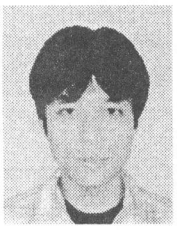
山田　実一　（学生員）

平 11 会津大・コンピュータ理工卒．平 13 奈良先端科技大学院大情報科学研究科博士前期課程了．同年（株）アドバンスト・メディアに勤務．現在音声認識の研究開発に従事．



馬場　朗　（正員）

平 6 九工大・工・電気卒．平 8 九大大学院・総合理工学研究科博士前期課程了．同年松下電工（株）入社．平 12～13（財）イメージ情報科学研究所に出向．現在音声認識の研究開発に従事．日本音響学会会員．



芳澤　伸一　（正員）

1994 名大・工・電気卒．1996 同大大学院修士課程了．1999 同大大学院博士課程了．在学中，ニューラルネットワークによるクラスタリングに関する研究に従事．1999 松下電器産業株式会社入社．2000～2001（財）イメージ情報科学研究所に出向，現在に至る．工博．主として音声・音情報処理の研究に従事．電気学会，計測自動制御学会，日本音響学会各会員．

米良祐一郎

平 12 阪大・基礎工・情報科学卒．同年奈良先端科技大学院大情報科学研究科博士前期課程入学．音声認識に関する研究に従事．日本音響学会会員．

李　晃伸　（正員）

平 8 京大・工・情報卒．平 10 同大大学院修士課程了．平 12 同博士後期課程了．同年より奈良先端科学技術大学院大学情報科学研究科助手．主として音声認識・理解の研究に従事．博士（情報学）．情報処理学会，日本音響学会各会員．

猿渡　洋　（正員）

平 3 名大・工・電気卒．平 5 同大大学院修士課程了．平 12 同大学院博士課程了．工博．平 5 セコム（株）入社．セコム IS 研究所音声情報処理研究室において，超音波アレー信号処理に関する研究に従事．平 12 奈良先端科学技術大学院大学助教授．音響アレー信号処理，ブラインド処理，音場再生などに関する研究に従事．平 13 本会論文賞受賞．日本音響学会，IEEE 各会員．

鹿野　清宏　（正員）

昭 45 名大・工・電気卒．昭 47 同大大学院修士課程了．同年電電公社武蔵野電気通信研究所入所．昭 59～61 カーネギーメロン大客員研究員．昭 61～平 2ATR 自動翻訳電話研究所音声情報処理研究室長．平 4NTT ヒューマンインタフェース研究所主席研究員．平 6 より奈良先端科学技術大学院大学情報科学研究科教授．音情報処理学講座を担当．工博．主として音声・音情報処理の研究及び研究指導に従事．昭 50 本会米沢賞．平 3IEEE SP 1990 Senior Award，平 6 日本音響学会技術開発賞．平 12 情報処理学会山下記念研究賞．平 13 VR 学会論文賞．IEEE，ISCA，情報処理学会，音響学会，VR 学会各会員．