ヒューマンインタフェース 97−28
音声言語情報処理 40−28
（２００２．２．２）

スペクトルサブトラクションを用いた十分統計量に基づく環境・話者適応

山出 慎吾 * 松浪 加奈子 * 馬場 朗 ** 芳澤 伸一 ***
李 晃伸 * 猿渡 洋 * 鹿野 清宏 *

* 奈良先端科学技術大学院大学 情報科学研究科
** 松下電工株式会社 東京研究所 *** 松下電器産業株式会社 先端技術研究所

あらまし 実環境での音声認識では，環境適応や話者適応といった音韻モデルの適応技術が必要となる．本研究では，スペクトルサブトラクションを用いた十分統計量に基づく環境・話者適応を行う．十分統計量を用いた環境・話者適応アルゴリズムでは，任意の 1 発声文と数十秒の雑音データのみを入力として教師なし学習を行う．適応手順は，以下の 3 段階からなる．(1) 雑音データをデータベースに重畳し十分統計量を計算する．(2) 任意の 1 発声文に対し GMM 話者モデルを用いて音声データベースから音響的特徴が近い話者集合を選択する．(3) 選択された話者集合の十分統計量を用い適応モデルを生成する．さらに発声文および雑音データを重畳した音声データベース全体にスペクトルサブトラクションを施すことで，SNR を改善し高精度な適応を行う．20dB の SNR における認識実験では，提案手法により適応した音韻モデルの認識率は，雑音データを重畳した音声データから EM アルゴリズムを用いて作成された matched モデルに比べ，PTM で約 76%から約 82%に改善された．さらに，この適応モデルを初期モデルとして MLLR による適応を行った場合，単純に MLLR を行うよりも高い認識精度が得られることを示す．また，雑音レベルの変動の認識率への影響についても述べる．

キーワード スペクトルサブトラクション，十分統計量，環境話者適応，教師なし適応，音韻モデル

Unsupervised Speaker and Noise Adaptation based on HMM Sufficient Statistics with Spectral Subtraction

Shingo YAMADE* **Kanako MATSUNAMI***
Akira BABA** **Shinichi YOSHIZAWA*****
Akinobu LEE* **Hiroshi SARUWATARI*** **Kiyohiro SHIKANO***

* **Graduate School of Information Science, Nara Institute of Science and Technology**
** **Matsushita Electric Works, LTD** *** **Matsushita Electric Industrial Co., LTD**

Abstract Noise and speaker adaptation techniques are essential to realize robust speech recognition in real environments. In this paper, we propose an efficient unsupervised noise and speaker adaptation method based on HMM sufficient statistics with spectral subtraction. Our adaptation method basically consists of three steps. (1) Noise data of the target environment is added to the training speech database to calculate noise-added HMM sufficient statistics. (2) Given a target speaker, the speakers that are acoustically close to the target speaker are selected from the database, using speaker GMMs. (3) The resulting adaptation model is obtained by constructing models from the HMM sufficient statistics of the selected speakers. Furthermore, we apply spectral subtraction to suppress noise signal and improve the SNR to both input utterances and the training database. Recognition experiments show that our adapted PTM can achieve about 82% word accuracy on 20dB SNR condition with only one utterance and environment noise data, which is about 6% higher than the environment-matched models trained by E-M algorithm. We also examined robustness of the adapted models in various SNR conditions. Integration with the supervised MLLR method is also investigated.

Key words Spectral Subtraction, HMM Sufficient Statistics, Noise and Speaker Adaptation, Unsupervised Adaptation, Acoustic Model

1　はじめに

実環境における大語彙連続音声認識においては，周囲の環境における雑音や発声者の特徴を考慮する必要がある．しかし，雑音には非常に多くの種類が存在し，それらすべての雑音を網羅するデータをあらかじめ持つことは不可能である．認識システムがおかれている環境の雑音に適応するには，その環境の雑音自身を用いることが望ましい．少ない発声文章数や運用の場での環境雑音をもとに，その環境および話者に音韻モデルを適応させることが必要である．広く用いられている MLLR(Maximum Likelihood Linear Regression)[1] に基づく教師あり環境・話者適応では，比較的高精度な適応モデルが得られるが，適応には数十文を必要とする．よって音韻モデルを適応するには認識を行う前に数十文章発声することが要求され，ユーザに大きな負担をかける．

本論文では，スペクトルサブトラクション [2] を用いた十分統計量に基づく環境・話者適応のアルゴリズムを提案する．十分統計量を用いた環境・話者適応アルゴリズムは，任意の 1 発話文と数十秒の雑音データから教師なし学習を行うことが可能である．音韻モデル学習用の clean 音声データベースに対して環境雑音を重畳するとともに，その重畳後の音声データから十分統計量を算出し，入力話者に対する類似話者を用いて適応を行う．十分統計量を用意することにより，話者に適応した音韻モデルを任意の 1 文から高速に学習できる．さらに，発声文および雑音データを重畳した音声データベースに対しスペクトルサブトラクションを行うことで，SNR が改善された発声文および十分統計量を算出することが可能となり，高精度な適応が可能となる．

また，実際の環境においては雑音レベルが常に変動すると考えられる．したがって，本研究では，適応後のモデルの雑音レベルの変動に対する頑健性についても評価する．さらに，高精度な適応を目指して，提案手法で適応したモデルを元に MLLR を行う方法についても評価を行う．

以下，2 章では音韻モデルを適応する際に用いる十分統計量に基づく環境・話者適応アルゴリズムについて述べ，3 章ではスペクトルサブトラクションによる雑音抑圧方法について述べる．4 章では提案するスペクトルサブトラクションを用いた十分統計量に基づく環境・話者適応について述べる．5 章では提案手法を用いて適応した音韻モデルの評価実験結果を示し，雑音レベルの変動に対する頑健性に対する実験結果について述べる．6 章では提案手法を用いて適応した音韻モデルを初期モデルとして MLLR ををを用いて適応した場合の認識性能を報告する．7 章ではまとめを述べる．

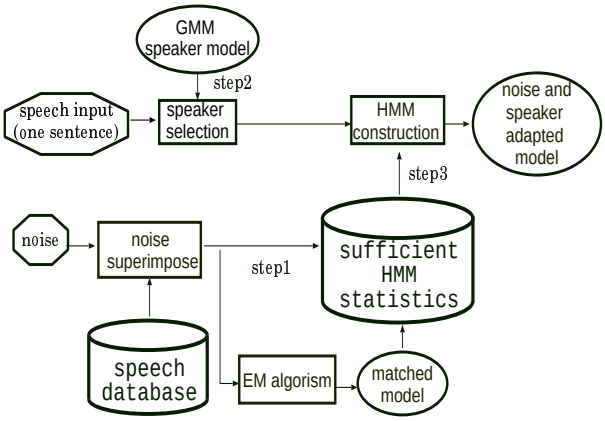

図 1: 環境・話者適応アルゴリズム

2　十分統計量に基づく環境適応および話者適応アルゴリズム

十分統計量に基づく環境・話者適応アルゴリズムの概念図を図 1 に示す．入力としては，任意の 1 発声文と数十秒の雑音データのみでよい．提案手法の特徴として，適応に用いる発話文が任意の 1 文章であることから，入力話者にかかる負担が少ない．また，認識システム運用の場の環境雑音データを用いて適応を行うので，あらゆる環境雑音に適応することが可能となる．適応手順は以下の通りである．

1. 音声データベースの全話者について雑音重畳データから十分統計量を計算して保存する．
2. GMM 話者モデルを用いてテスト話者に音響的特徴が近い話者を N 人選択する．
3. 選択された N 人の話者の十分統計量を用いて適応モデルを再構築する．

以上で環境適応および話者適応された音韻モデルが作成される．

提案手法では，環境適応が環境雑音データを重畳した音声データベースを用いて，十分統計量を作成することで行われるため，いったん環境雑音データを用いて十分統計量を作成しておくと，その環境に適応した話者適応モデルを作成するには十分統計量を用いた適応を行うだけで十分であり，話者の交替にも高速に適応することができる．

以下，適応アルゴリズムの各工程における詳細を述べる．

2.1　十分統計量の計算

十分統計量とは，データベースの統計的性質を表すのに十分な統計量のことである．HMM における十分統計量とは，各正規分布の平均，分散，

EM カウントのことであり [3]，この十分統計量を用いて，高速に音韻モデルを学習する．本研究では，音声データベースとして JNAS[4] を用いる．これは Clean 環境で収録された新聞記事読み上げ音声データであり，これに雑音を重畳して十分統計量を計算する．これに入力雑音を重畳して十分統計量を計算する．十分統計量は，EM アルゴリズムにより不特定話者モデルから 1 回学習を行うことにより作成される．

2.2 話者選択

話者選択では任意の 1 発声文と GMM(Gaussian Mixture Model) 話者モデルを用い，テスト話者に音響的特徴が近い話者を音声データベースから選択する．GMM は混合正規分布で表現した話者ごとの確率モデルである．また各話者はそれぞれ 1 状態 64 混合で表現されている．厳密には話者選択用 GMM を入力発声の環境ごとに作成する必要があるが，環境ごとに GMM を作成することは非効率である．そこで入力音声の中で雑音レベルの低い部分のみを用いて話者選択を行うことで，clean データで作成した GMM をそのまま用いることとする [5]．具体的には，入力音声の全フレームの中でパワーが低い半分のフレームを削除した特徴量ファイルを入力ファイルとして選択を行う．音響的特徴が近い話者を選択する際の距離尺度としては尤度を用い，尤度上位 N 人を選択話者として選択する．

2.3 適応モデルの再構築

2.2 節で選択された N 人の話者の十分統計量を用いて，環境および話者に適応したモデルを再構築する．

3 スペクトルサブトラクションによる雑音抑圧

スペクトルサブトラクション [2] は，雑音が定常であることを利用して，非音声区間の信号より雑音の特徴量を推定しておき，雑音混じりの音声の特徴量から雑音を取り除き，元の音声信号を推定する信号処理である．時刻 t の雑音混じりの信号を $y(t)$，真の音声信号を $s(t)$，雑音信号を $n(t)$ とすると，観測される雑音混じりの信号 $y(k)$ は，次のように表される．

$$y(t) = s(t) + n(t)$$

ここで，窓の位置を m で表した短時間分析による両辺のフーリエ変換をとると次の式が得られる．

$$Y(f,m) = S(f,m) + N(f,m)$$

$Y(f,m)$，$S(f,m)$，$N(f,m)$ は，周波数 f の複素スペクトルを表す．振幅スペクトルは入力 y から推定雑音を減算したものとし，位相は入力信号のものを使用する．この方法により推定される信号は次のように表される．

$$\hat{S}(f,m) = \left||Y(f,m)|^2 - \alpha E_m\left[|N(f,m)|^2\right]\right|^{1/2} \cdot e^{j \arg\left(Y(f,m)\right)}$$

$E_m\left[|N(f,m)|^2\right]$ は非音声区間で推定した雑音信号，α は減算調節パラメータである．また，推定された音声の振幅スペクトルが負である時の問題を回避するため，次に示す式のように任意の係数 $A(0 < A < 1)$ を用いてフロアリング処理を行う．

$$\hat{S}(f,m) = Y(f,m) \cdot A$$
$$\left(\text{ただし } |Y(f,m)|^2 - \alpha E_m\left[|N(f,m)|^2\right] < 0\right)$$

4 スペクトルサブトラクションを用いた環境話者適応技術

ここでスペクトルサブトラクション [2] と 2 章で述べた十分統計量に基づく環境・話者適応アルゴリズムを統合した概念図を，図 2 に示す．提案法では，スペクトルサブトラクションを発話文章と雑音データを重畳した音声データベースに対して行う．スペクトルサブトラクションを行うとフロアリングなどの処理を行っているため，入力音声波形が歪んでしまい clean な音韻モデルでの認識では，十分な性能が発揮されない．そこで入力に対してのみスペクトルサブトラクションを行うのではなく，音韻モデルにおいて波形の歪みを考慮するため音声データベースにもスペクトルサブトラクションを行い，スペクトルサブトラクションに適応した音韻モデルを使用する．具体的には，再構成する際の初期モデルとして，スペクトルサブトラクション後の音声データベースから EM アルゴリズムに基づいて作成した音韻モデルを用いる．また，スペクトルサブトラクションを行った後の音声データベースから十分統計量を計算して保存しておく．スペクトルサブトラクションを用いた場合の適応手順は以下の通りである．

1. 音声データベースにおける全話者の雑音重畳データに対し，スペクトルサブトラクションを行った後，十分統計量を計算して保存する．
2. 発話文に対しスペクトルサブトラクションを行い，GMM 話者モデルを用いてテスト話者に音響的特徴が近い話者を N 人選択する．
3. 選択された N 人の話者の十分統計量を用いて適応モデルを再構築する．

以上により環境・話者適応された音韻モデルを構築する．

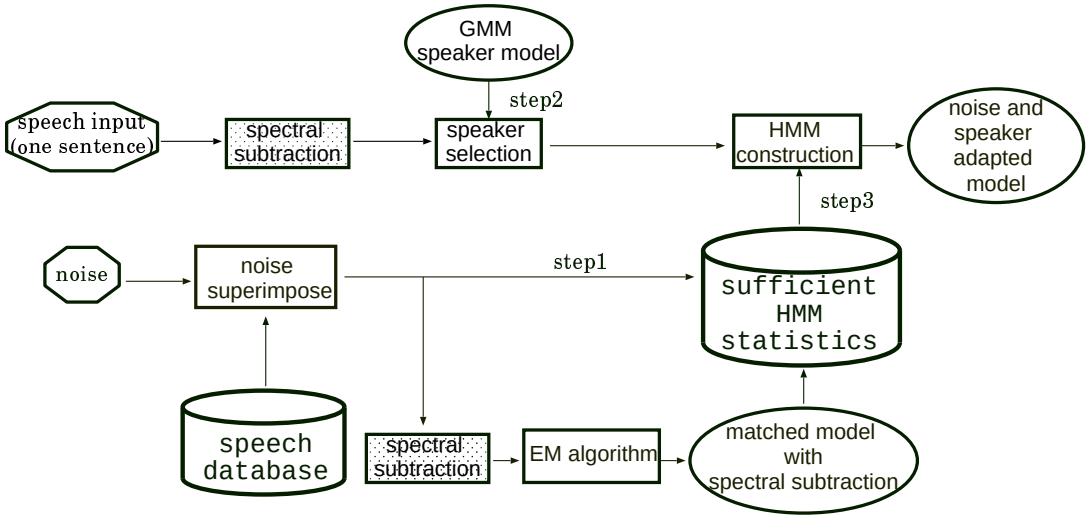

図 2: スペクトルサブトラクションを用いた環境・話者適応アルゴリズム

5 認識実験

音声認識におけるスペクトルサブトラクション [2] の有効性，および提案したスペクトルサブトラクションを用いた十分統計量に基づく環境・話者適応の評価実験を行った．オフィス環境の居室雑音を対象とした適応実験を行い，提案手法をスペクトルサブトラクションを用いない場合，通常の環境 matched モデル (EM アルゴリズムで学習したモデル) と比較する．

5.1 実験条件

音声データベースは JNAS[4] の音声データベースを用いる．JNAS データベースは，男性 153 人，女性話者 153 人の計 306 人の話者で構成されており，各話者ごとに 50 文の音素バランス文と約 100 文の新聞読み上げ文を持つ．今回の実験では 306 人中男性 130 人，女性 130 人の計 260 人を学習用話者とし，残る男性 23 人，女性 23 人の計 46 人を評価用話者として用いた．

学習データ数は，各話者につき音素バランス文 50 文，新聞読み上げ文 100 文である．評価用データとしては各話者につき 4，5 文章，計 200 文章である．サンプリング条件は 16kHz，16bit，特徴量は窓シフト長 10ms で分析した 12 次元の MFCC(Mel-frequency cepstrum coefficient) とデルタ MFCC，デルタパワーを用いる．言語モデルは語彙数 20k の新聞記事から構築した 3-gram を使用し，デコーダは Julius を用いる．音韻モデルは 43 音素 3 状態 16 混合のモノフォンモデルと 2000 状態 64 混合の PTM(phonetic tied mixture model)[6] を使用する．また，音韻モデルは不特定話者モデルである．

スペクトルサブトラクション [2] に用いる減算調節パラメータ α は 2.0，フロアリング係数は 0.5 とし、選択話者数はモノフォンモデルの場合では 20 人，PTM の場合では 40 人とする．これは，Clean 音韻モデルでの予備実験において最適であったパラメータを用いている．

話者選択で用いた GMM 話者モデルは一人あたり 1 状態 64 混合で構成される．GMM 話者モデルは各話者につき約 150 文章の音声データ (雑音重畳なし) から構築した．環境適応対象は居室雑音下での発声であり，マイク収録した 3 分間の居室雑音を評価用データに重畳したものに対してスペクトルサブトラクションを行い新しい評価用データとする．SNR は 15dB，20dB，25dB とする．

5.2 実験結果

モノフォンモデルと PTM でのそれぞれのスペクトルサブトラクション [2] の効果を表す結果を，図 3，図 4 に示す．グラフにおいて縦軸は単語正解精度 (%)，横軸は入力雑音の SNR(dB) である．baseline は Clean モデル (適応なし，不特定話者モデル) での認識率，SS は入力に対してスペクトルサブトラクションを行い Clean モデルでの認識を行った認識率，matched は雑音を重畳した音声データベースより EM アルゴリズムを用いて作成した不特定話者音韻モデルの認識率，SS matched は雑音を重畳した音声データベースに対し，スペクトルサブトラクションを行い EM アルゴリズムを用いて作成した，スペクトルサブトラクションに適応した音韻モデルの認識率を示している．また clean-clean は Clean モデルを用いて Clean データを認識した結果である．

実験結果より，Clean モデルでもスペクトルサ

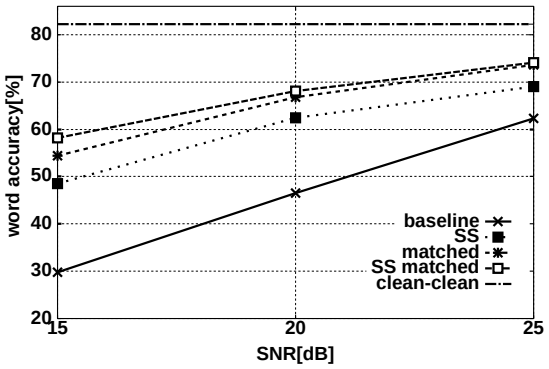

図 3: スペクトルサブトラクションの効果 (monophone)

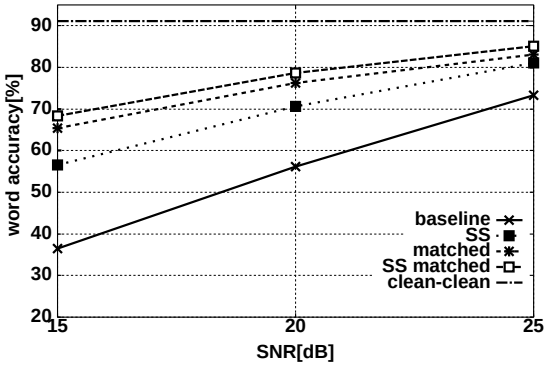

図 4: スペクトルサブトラクションの効果 (PTM)

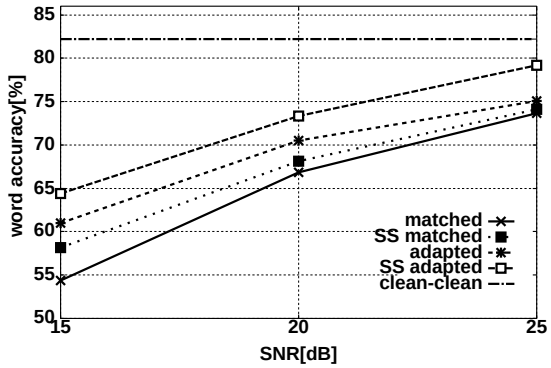

図 5: 環境・話者適応の効果 (monophone)

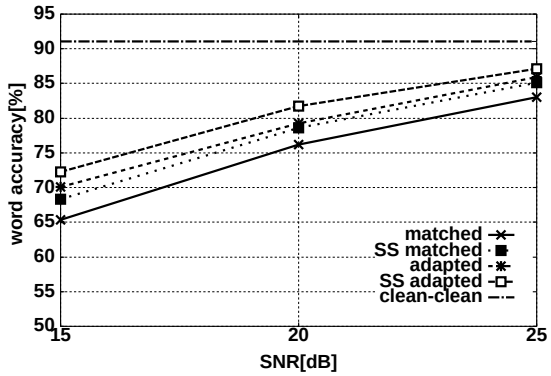

図 6: 環境・話者適応の効果 (PTM)

ブトラクションを行うことで baseline の性能を大きく上回り，モノフォンモデルで 15dB の場合では約 20%，25dB では，約 6%認識率が改善した．また PTM でも同様の結果が得られ，15dB の場合では約 20%，25dB では，約 8%認識率が改善した．しかしながら，モノフォンモデル，PTM 共に matched の性能を越えることはできなかった．原因として，スペクトルサブトラクションを行うとフロアリングなどの処理を行っているため，音声波形に歪みが生じ Clean モデルでの認識では，その歪みの部分の影響が大きいためであると考えられる．

そこで，雑音データを重畳した音声データベースに対してスペクトルサブトラクションを行い，波形の歪みを考慮した SS matched モデルで実験を行った．SS matched では，波形の歪みを含むスペクトルサブトラクションに適応したモデルであるので，認識精度が matched よりモノフォンモデルで 15dB の場合では約 4%，PTM で 15dB の場合で約 3%向上した．このことにより雑音環境での発声におけるスペクトルサブトラクションの有効性が示せた．

続いて提案手法を用いたモノフォンモデルと PTM でのそれぞれの環境・話者適応後モデルの認識率を，図 5，図 6 に示す．グラフにおいて縦軸は単語正解精度 (%)，横軸は SNR(dB) である．adapted はスペクトルサブトラクションを用いずに環境・話者適応した音韻モデルでの認識率，SS adapted は，スペクトルサブトラクションを用いて環境・話者適応した音韻モデルでの認識率となっている．比較対象として，前述の実験結果の clean-clean，matched および SS matched を示す．

実験結果より，matched と adapted および SS adapted を比較すると，モノフォンモデルで 15dB の場合で，matched に比べ SS adapted の方が約 10%高い認識率が得られた．また PTM においては 15dB の場合で matched に比べ SS adapted の方が約 13%高い認識率が得られた．このことにより提案法では，matched モデルよりも adapted モデルで高い認識率が得られ，さらにスペクトルサブトラクションを組み合わせた SS adapted モデルで，より高い認識率を得られることがわかる．また雑音レベルが高いほど，改善幅が高くなっている．モノフォンモデル，PTM のそれぞれの音韻モデルにおいてほぼ同様の改善傾向が見られた．

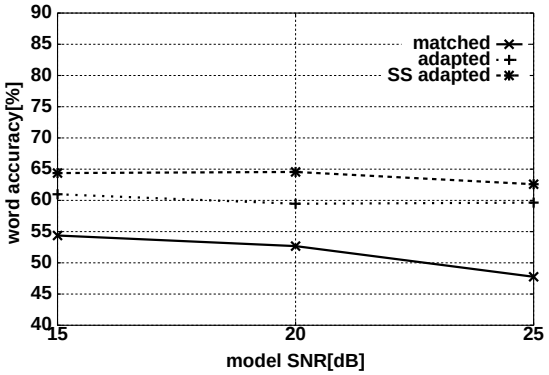

図 7: 異なる SNR のモデルにおける認識性能 (monophone:15dB SNR テスト音声入力)

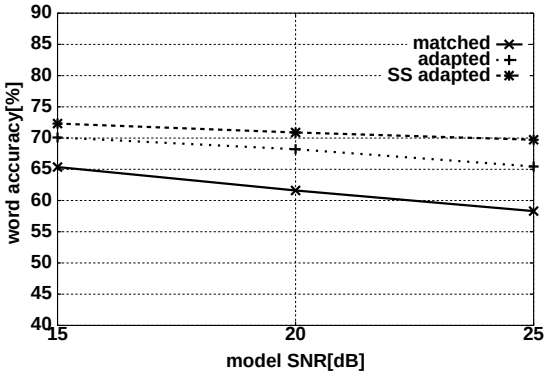

図 8: 異なる SNR のモデルにおける認識性能 (PTM:15dB SNR テスト音声入力)

5.3 異なる SNR に対する頑健性

SNR はユーザおよび環境によってさまざまに変動する．そのため，評価用データの SNR と異なる SNR の適応モデルを用いても，ある程度の認識性能が確保できることが求められる．そこで異なる SNR に対する音韻モデルの頑健性を評価するため，5.2 節と同様の条件のもと，評価用データの SNR と異なる SNR の音韻モデルを用いた認識実験を行った．音韻モデルとしては，前述した adapted モデルと SS adapted モデル，matched モデルを用いる．SNR は音韻モデルについては 15dB，20dB，25dB，評価用データについては 15dB である．モノフォンモデルと PTM でのそれぞれの認識実験結果を図 7，図 8 に示す．グラフにおいて縦軸が単語正解精度 (%)，横軸が音韻モデルの SNR(dB) である．また，matcthed と SS adapted の縦軸が単語正解精度 (%)，横軸が評価用データの SNR，10dB，15dB，20dB，25dB，30dB の場合の PTM における認識実験結果を図 9 に示す．詳しい実験

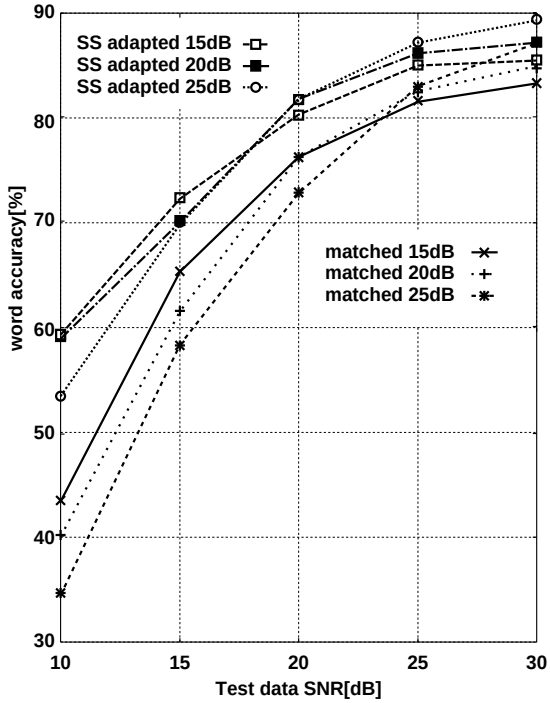

図 9: 異なる SNR の評価データにおける認識性能 (PTM)

結果は付録 A，付録 C に示す．

以上の結果により，matched モデルに比べ SS adapted モデルは，図 7，図 8 においてよりグラフが平坦になってることがわかり，また図 9 において各 SNR でのモデル間の認識率がより接近していることがわかる．よって提案手法を用いることで SNR の変動にも頑健で，一度環境適応を行えばある程度環境に適応できることがわかる．

6 MLLR の初期モデルとしての有効性

現在最も一般的な教師ありの話者適応・環境適応法は，MLLR(Maximum Likelihood Linear Regression)[1] である．MLLR は適応データに対する尤度を最大化するように線形行列を求め，その行列を用いて HMM の平均，分散ベクトルを変換することにより音韻モデルを適応させるアルゴリズムである．MLLR は教師あり適応であるので，あらかじめ決められた発話内容を確実に読み上げる必要がある．個人に特化した音声認識システムのように教示が可能なアプリケーションにおいては，さらにこの MLLR を行うことでさらなる精度の向上が期待できる．また MLLR の観点からは，その適応モデルの認識率は初期モデルに依存することが知られている [3, 7, 8].

提案法で適応した音韻モデルが，MLLR の初

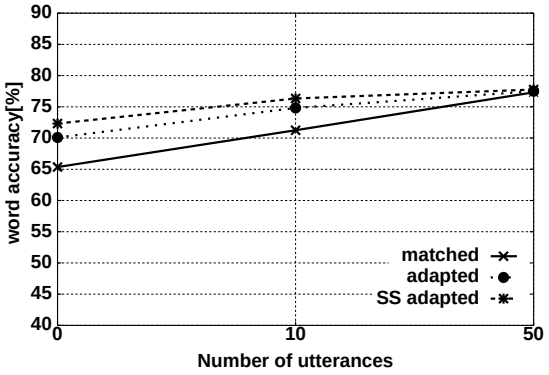

図 10: MLLR の初期モデルとしての有効性の評価 (15dB SNR PTM，平均・分散，繰返し 3 回)

期モデルとして有効であるかを検証するため，SS adapted，adapted，matched を初期モデルとしてその平均値と分散に対して，繰返し 3 回の MLLR を行い，適応したモデルの評価を行った．MLLR で適応する際に用いる適応文章は，各話者につき JNAS データベース上の各話者の全発話文から評価用文章を除いた 10 文，50 文を使用した．

PTM での 15dB での認識実験結果を図 10 に示す．グラフの縦軸は単語正解精度 (%)，横軸は適応文章数である．SS adapted，adapted は，SNR が 15dB のデータについて，提案手法で適応したモデルを初期モデルとしたときの実験結果である．同様に matched は matched モデルを初期モデルとしたときの実験結果である．詳しい実験結果は，付録 B，付録 D に示す．adapted，SS adapted と matched を比較すると MLLR 適応前の認識率がそれぞれ 70.1%，72.3%，65.4%であったのに対し，10 文章で MLLR 適応した結果は 74.8%，75.2%，71.3%となっており，matched に比べて SS adapted の方が約 4%高い認識率を示した．50 文章で MLLR を行った結果は，どのモデルも飽和状態にあるとみられ，ほぼ同程度の認識率を示した．

また，PTM で 15dB の場合において，平均値のみに対して 1 回の MLLR を行った結果および前述した平均・分散に対して 3 回の MLLR を行った結果を表 1 に示す．matched に比べ adapted，SS adapted の方が 10 文章，50 文章と共に高い認識率を示し，提案手法の有効性が示せた．

これらにより，提案法で適応した音韻モデルは，特に少ない文章の基での MLLR の初期モデルとしても有効であることがわかり，MLLR を行うことでさらに精度が向上することが確かめられた．

表 1: MLLR の初期モデルとしての有効性の評価 (15dB SNR PTM:平均 (繰返し 1 回), 平均・分散 (繰返し 3 回))

		Word accuracy[%]		
	model	before	10 utterances	50 utterances
平均	matched	65.4	69.2	71.0
	adapted	70.1	72.4	74.2
	SS adapted	72.3	73.3	76.7
	model	before	10 utterances	50 utterances
平均 分散	matched	65.4	71.3	77.3
	adapted	70.1	74.8	77.5
	SS adapted	72.3	75.2	77.7

7 おわりに

本論文では，スペクトルサブトラクションを用いた十分統計量に基づく環境・話者適応手法を提案した．提案法ではある任意の 1 発声文と数十秒の環境雑音を用いて適応を行うことができる．さらに，スペクトルサブトラクションを施すことで音韻モデルの適応精度を向上させることができる．また，評価用データと SNR が異なる音韻モデルを用いてもある程度の認識性能を確保できることを示した．最後に現在最も一般的な適応法である MLLR の初期モデルとしての有効性を示した．

参考文献

[1] C.J. Leggetter, C. Woodland: “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” Computer Speech and Langage, Vol.9, pp.701-704, 1995

[2] S.F. Boll: “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Transaction on Acoustics,Speech, and Signal Processing, ASSP-33, vol.27, pp.113-120,1979

[3] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, K. Shikano: “Adaptation Based on Sufficient HMM Statistics of Selected Speakers,” Proceedings of ICASSP, pp.341-344, 2001

[4] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi: “JNAS : Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research,” The Journal of the Acoustical Society of Japan(E), Vol.20, pp.199-206, 1999

[5] M. Yamada, A. Baba, S. Yoshizawa, Y. Mera, A. Lee, H. Saruwatari, K. Shikano: “Unsupervised Noisy Environment Adaptation Algorithm Using MLLR and Speaker Selection,” Proceedings of EuroSpeech, pp.869-872, 2001

[6] A. Lee, T. Kawahara, K. Takeda, K. Shikano: “A New Phonetic Tied Mixture Model for Efficient Decoding,” Proceedings of ICASSP, pp.1269-1272, 2000

[7] Y. Gao, M. Padmanabhan, M. Picheny: “Speaker Adaptation Based on Pre-clustering Training Speakers,” Proceedings of EuroSpeech, pp.2091-2094, 1999

[8] M. Padmanabhan, L.R. Bahal, D. Nahamoo, M.A. Picheny: “Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulay Speech Recognition System,” Proceeding of ICASSP, pp.701-704, 1995

付録 A：異なる SNR における認識性能 (monophone)

		matched model		
	SNR	15dB	20dB	25dB
Test data	10dB	36.3	31.8	27.3
	15dB	54.4	52.7	47.8
	20dB	66.0	66.8	63.5
	25dB	72.1	73.2	73.6
	30dB	71.8	74.6	77.7
		SS matched model		
	SNR	15dB	20dB	25dB
Test data	10dB	43.5	42.5	39.7
	15dB	58.2	57.7	55.7
	20dB	66.9	68.1	67.0
	25dB	70.9	72.6	74.1
	30dB	71.4	74.4	77.2
		adapted model		
	SNR	15dB	20dB	25dB
Test data	10dB	41.7	38.1	38.9
	15dB	61.0	59.4	59.6
	20dB	69.4	70.5	70.5
	25dB	74.0	76.0	75.1
	30dB	76.7	77.9	80.0
		SS adapted model		
	SNR	15dB	20dB	25dB
Test data	10dB	50.7	49.8	45.6
	15dB	64.4	64.6	62.6
	20dB	71.5	73.3	73.1
	25dB	74.1	77.0	79.2
	30dB	75.3	78.8	82.7

付録 B：MLLR の初期モデルとしての有効性の評価 (monophone，平均・分散，繰返し 3 回)

	matched model		
SNR	before	10 utterances	50 utterances
15dB	54.4	64.5	70.5
20dB	66.8	75.5	80.4
25dB	73.6	80.3	83.7
	adapted model		
SNR	before	10 utterances	50 utterances
15dB	61.0	66.2	70.7
20dB	70.5	76.1	80.8
25dB	75.1	80.3	83.3
	SS adapted model		
SNR	before	10 utterances	50 utterances
15dB	64.4	67.7	71.7
20dB	73.3	77.5	80.3
25dB	79.2	81.0	84.3

付録 C：異なる SNR における認識性能 (PTM)

		matched model		
	SNR	15dB	20dB	25dB
Test data	10dB	43.4	40.1	34.5
	15dB	65.4	61.6	58.3
	20dB	76.2	76.2	72.8
	25dB	81.5	82.6	83.0
	30dB	83.2	85.0	87.1
		SS matched model		
	SNR	15dB	20dB	25dB
Test data	10dB	53.8	53.1	48.3
	15dB	68.3	66.5	66.0
	20dB	76.9	78.6	77.7
	25dB	80.8	83.6	85.1
	30dB	82.3	85.2	87.2
		adapted model		
	SNR	15dB	20dB	25dB
Test data	10dB	51.4	45.8	40.3
	15dB	70.1	68.2	65.5
	20dB	79.7	79.3	77.0
	25dB	84.5	85.9	85.9
	30dB	85.6	87.7	89.2
		SS adapted model		
	SNR	15dB	20dB	25dB
Test data	10dB	59.3	58.9	53.4
	15dB	72.3	70.9	69.7
	20dB	80.4	81.8	81.7
	25dB	85.0	86.2	87.2
	30dB	85.5	87.2	89.4

付録 D：MLLR の初期モデルとしての有効性の評価 (PTM，平均・分散，繰返し 3 回)

	matched model		
SNR	before	10 utterances	50 utterances
15dB	65.4	71.3	77.3
20dB	76.2	82.0	86.7
25dB	83.0	88.2	91.0
	adapted model		
SNR	before	10 utterances	50 utterances
15dB	70.1	74.8	77.5
20dB	79.3	83.5	85.8
25dB	85.9	89.5	90.5
	SS adapted model		
SNR	before	10 utterances	50 utterances
15dB	72.3	75.2	77.7
20dB	81.8	84.2	86.3
25dB	87.2	89.2	89.2