音声言語情報処理 35－6
（2001．2．2）

環境雑音適応アルゴリズムの大語彙連続音声認識による評価

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＊

＊ 奈良先端科学技術大学院大学 情報科学研究科
＊＊ イメージ情報科学研究所

あらまし　本報告では、MLLRを用いた音響モデルの環境雑音適応アルゴリズムを提案し、大語彙連続音声認識により環境適応音響モデルの性能を評価する。提案手法は発声者の任意の1文発声と居室雑音に基づく本人の声を使わない教師なし適応であり、次の3段階からなる。(1) GMMによる話者モデルを使って音響的特徴が近い話者を選択する。(2) 充足統計量による話者適応を行なって教師なし話者適応クリーンモデルを作成する。(3) 選択された話者の雑音重畳音声データを用いてMLLRで適応する。この適応音響モデルとHMM合成による音響モデル、不特定話者雑音重畳モデル、さらに教師あり適応との比較を行なった。本適応アルゴリズムによるモデルで大語彙連続音声認識実験を行なったところ、25dBの雑音環境下においてPTMで86.6%の単語認識率が得られ、不特定話者HMM合成モデルより約10%、不特定話者EM学習モデルと同程度の認識率を得ることができ、環境適応アルゴリズムの有効性を示すことができた。

キーワード　音響モデル, 環境適応, MLLR, 大語彙連続音声認識

Performance of Environment Adaptation Algorithms in Large Vocabulary Continuous Speech Recognition

Miichi YAMADA＊　Akira BABA＊＊
Shinichi YOSHIZAWA＊＊　Yuichiro MERA＊
Akinobu LEE＊　Hiroshi SARUWATARI＊　Kiyohiro SHIKANO＊

＊ Graduate School of Information Science, Nara Institute of Science and Technology
＊＊ Laboratories of Image Information Science and Technology

Abstract　This paper proposes an acoustic model adaptation algorithm using MLLR in noisy environments, and describes the large vocabulary continuous speech recognition (LVCSR) performance in noisy environments. This algorithm is an unsupervised adaptation one. This proposed method is based on (1) selecting speakers which is close to an utterance speaker, (2) performing the speaker adaptation with speaker sufficient HMM statistics, and (3) performing the MLLR noisy adaptation with noisy speech data of the speakers which are selected in (1). This adapted acoustic model is compared in LVCSR with HMM composition acoustic models with noise HMM models, EM trained acoustic models with noise-added speech database, and MLLR supervised adaptation acoustic models. The word recognition rate is 86.6% with the PTM model, which is about 10% higher than the HMM composition model and about the same as the EM trained acoustic models. These results show the effectiveness of the proposed noisy environment adaptation algorithm.

Key words　Acoustic Model, Environment Adaptation, MLLR, LVCSR

1 はじめに

実環境で音声認識を利用するには話者適応と環境適応は必要不可欠であり、また非常に難しい課題である。音声は話者ごとによって音響的特徴が違うので不特定話者の音響モデルにおける認識率は特定話者の音響モデルよりも悪い。また雑音の入った音声では音声パワーの低い発話は雑音に隠れて認識は非常に難しくなる。一般的に雑音環境における音響モデルの適応にはその環境での雑音を音声データベースに重畳して音響モデルを学習したモデルが高い認識率をもたらすことがわかっているが、大量の音声コーパスと多くの時間を必要とする。またこのような音響モデルは特定の環境で構築されるので環境の変化に対してその都度音響モデルを構築しなければならない。よって実際の使用は難しい。

そこで時間とデータ量の少ない適応が必要となり今までさまざまな研究がなされてきた。これまでの適応法においては話者適応に関しては発話者本人が決められた文章を正確に発声を行なう教師ありの適応がほとんどである。教師あり適応では、比較的多くの発声文章を必要とするので、特にお年寄りや体の不自由な方々にとっては大変な労力である。環境適応には HMM 合成法 [1][2] などがあるが連続音声認識においては満足な結果が得られていない。そこで、話者と環境の適応を両方行なうシステムを開発中で、話者適応には芳澤らが提案した充足統計量を用いた教師なし適応法 [3] を用いた。今回、環境適応には話者適応で用いられる MLLR(Maximum Likelihood Linear Regression)[4] を用いた。MLLR は話者適応と環境適応の両方を行なうことができ、今回は充足統計量を用いた方法と併用し、教師なしの適応として MLLR を使用する環境適応アルゴリズムを提案した。さらに、従来の環境適応法である HMM 合成法や雑音重畳モデルと比較した。

2 環境適応アルゴリズム

我々が提案している適応システムを図 1 に示す。前もって、話者毎の HMM(Hidden Markov Model) に関する充足統計量を不特定話者の音響モデルより算出し蓄積する。さらに、話者ごとの 1 状態の GMM(Gaussian Mixture Model) を話者モデルとして作成しておく。適応システムは 3 段階のステップからなる。第 1 ステップでは発話文から抽出した音響特徴量を入力とし、話者毎に作られた GMM による話者モデルで尤度計算し、音響的特徴の近い話者を選択する。第 2 ステップでは選択された複数の話者の充足統計量を用いて発声話者に適応した音響モデルを構築する。第 3 ステップでは選択された複数の話者の雑音重畳音声データを用いて、MLLR により話者適応音響モデルからの環境適応を行ない話者・環境適応の音響モデルを構築する。第 1～第 2 ステップの詳しいアルゴリズムは芳澤らの話者適応法 [3] を参照されたい。

ここで、第 3 ステップの詳細を述べる。第 3 ステップでは話者適応されたモデルに対して環境雑音適応を行なう。手法としては第 1 ステップで選択された話者の雑音重畳データを使い、MLLR[4] によって適応する。つまり話者適応によって選ばれた話者と同じ話者で適応を行なう。あらかじめ用意しておいた環境雑音適応用のデータベースから選択された話者の音声データを使用し、その音声データに入力音声の非発話区間の雑音データを重畳する。この雑音重畳音声データを用いて、話者適応音響モデルから MLLR で環境雑音適応をおこなう。従来、MLLR は話者適応に用いられる手法であるが、雑音重畳音声データを用いることにより環境適応も行なうことが出来る。また MLLR を使用することで話者適応も同時に行なうことが可能である[1]。

この適応アルゴリズムは一般的に行なわれている発話者本人のデータを使う教師ありの適応ではなく、発話者に近い話者のデータで適応していることにより教師なしの適応であることが言える。さらに充足統計量や音声データベースを用意することで発話者は適応用に 1 文章のみ発声すれば良く、数多くの文章を発声する必要もなく発話者の負担が非常に軽くなる。

表 1: 音声分析条件

サンプリング周波数	16 kHz
分析フレーム長	25 msec
フレームシフト長	10 msec
特徴パラメータ	MFCC(12 次元) +Δ MFCC(12 次元) +Δパワー

3 連続音声認識実験

本報告における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エンジン JULIUS[5] を用いた連続音声認識で行なった。また認識率は正解率判定ツール [6] を用いて正解文章と自動比

[1]MLLR による話者適応では、平均値のみの適応を 1 回繰り返すだけで十分であるが、環境適応においては、平均値・分散ともに適応し、かつ、2 回以上繰り返す必要がある。

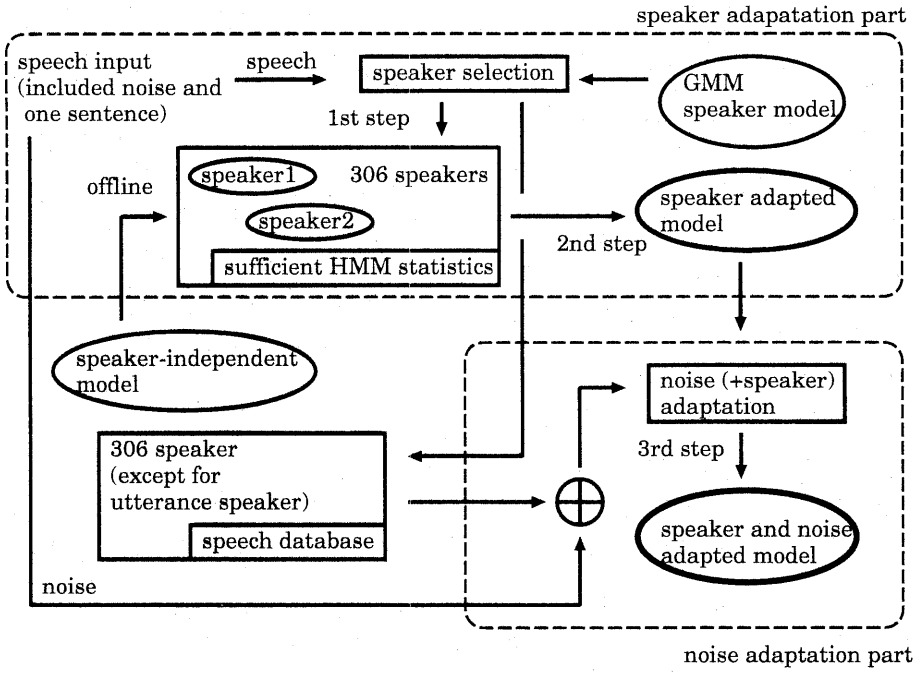

図 1: 話者＋環境適応システム

較して算出した。音声分析条件を表 1 に載せる。分析時、文章ごとに CMN(Cepstrum Mean Normalization) 処理を施している。音響モデルとして不特定話者の HMM を用い、簡素な単一音素のモデル（43 音素モデル）であるモノフォンモデルと、トライフォン間で分布の共有を行なう PTM(Phonetically Tied-Mixture)[8] モデルの 2 種類を用いた。言語モデルは日本語ディクテーションソフトウェア [7] に含まれている毎日新聞記事 75 カ月分から作られた語彙数 2 万語の単語 N-gram モデルを使用した。評価文として、46 人の評価用話者による 200 文章で認識実験を行なった。なお評価用データおよび適応用データには居室雑音を重畳したものを用いて実験を行なった。

3.1 雑音重畳音声の充足統計量を用いた適応とその問題点

本報告の適応アルゴリズムを用いる前に、雑音重畳音声の充足統計量を用いたアルゴリズムの性能について述べる。この適応アルゴリズムでは、雑音重畳音声データから充足統計量を算出し、話者モデルも雑音重畳音声による GMM を用いて話者選択を行なう。なお、適応システムは第 2 ステップまでである。実験条件を表 2 に示す。実験条件は、GMM および充足統計量は clean および 2 種類の SNR(Signal-to-Noise Ratio) である。居室雑音は主に計算機雑音と空調のファン音である。話者選択部においては、まず尤度上位 100 名の中で一番多い SNR を見つけ、その SNR の尤度上位話者を適応用話者として選択した。音響モデルはモノフォンを使用した。

実験結果を表 3 に示す。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本システムの有効性が見られた。しかしこの適応法に関する問題点は、あらゆる雑音を想定して充足統計量と話者モデルを用意しなければならないことである。よって、雑音の種類と SNR に対応した莫大な充足統計量と話者モデルを必要とするので実際の使用は不可能である。本報告の適応システムではこの結果を目標の一つとしている。

3.2 適応実験と評価

今回提案した環境雑音適応アルゴリズムの評価を行なった。音声データ、雑音データを表 4 に示す。基本的に表 2 と同じである。2 種類の話者選択に基づく実験を行なった。1 つめは第 1 ステップにおいて入力音声に clean データを使って話者選択を行なった後、雑音重畳データで環境雑音適応を

表 2: 実験条件（雑音重畳音声の充足統計量）

音声データ	JNAS 音声（約 45000 文）
雑音データ	居室雑音
SNR	20dB, 10dB, clean
音響モデル	モノフォン (16 混合)
話者 GMM(64 混合)	306 話者× 3
充足統計量	306 話者× 3
選択話者人数	20 人（発声者は除く）
入力音声	1 文章
評価データ	46 人,200 文章

表 3: 雑音重畳音声の充足統計量おける適応におけるモノフォンの単語認識率 (%)

評価 SNR	clean	20dB	10dB
適応前	83.2	48.1	13.7
適応後	85.8	74.4	48.3

行なった（この実験を clean-noisy と呼ぶ）。2 つめは雑音を重畳した入力音声で話者選択を行ない、選択話者の充足統計量および環境適応用データで音響モデルを適応させた（この実験を noisy-noisy と呼ぶ）。比較として本人の文章（10 文、50 文）で MLLR を行なった教師あり適応音響モデルと、HMM 合成法で構築した不特定話者モデル、さらに雑音重畳音声で学習した不特定話者モデルも評価した。

2 つめの話者選択実験において、入力データにおいてパワーの低い MFCC パラメータをカットした。これは雑音によって話者識別が難しくなることを防ぐためである。

3.2.1 評価 1：環境適応文章数による性能比較

始めに環境適応の文章数と単語認識率の関係を比較した。実験結果を表 5 に示す。適応前に比べ、モノフォンでは 17～27%、PTM では 9～29%認識率が向上し、適応文章数が増えるにつれて認識率が若干上がった。また、SNR が 15dB に下がると clean-clean と clean-noisy の差が少し広がった。

3.2.2 評価 2：HMM 合成モデル、雑音重畳モデルとの比較

次に本適応アルゴリズムと他のアルゴリズムとの比較をした。比較したアルゴリズムは clean な音響 HMM と雑音 HMM を合成する HMM 合成法（不特定話者）[1][2]、雑音重畳音声を用いて EM(Expectation-Maximization) アルゴリズムによる音響モデルの学習（これを雑音重畳学習と呼ぶ）行なう方法（不特定話者）、そして本手法（200 文章）である。その結果を表 6 及び図 2 にまとめた。HMM 合成法よりもモノフォンで約 15%、PTM で約 10%認識率が高い。雑音重畳学習学習と比べるとほぼ同程度の認識率が得られた。

表 4: 実験条件（環境雑音適応）

雑音データ	居室雑音
SNR	25dB, 20dB, 15dB
音響モデル	モノフォン（16 混合） PTM（状態数 2500 , 64 混合）
話者 GMM(64 混合)	306 話者 (clean)
充足統計量	306 話者 (clean)
選択話者人数 (発声者は除く)	20 人（モノフォン） 40 人（PTM）
入力音声	1 文章
環境適応データ	各選択話者ごとに 1,3,5,10 文

3.2.3 評価 3：教師あり適応との比較

次に適応話者本人の発話文章を使って MLLR の教師あり適応を行なった結果と本提案手法（教師なし適応）を比較した。教師あり適応における実験条件を表 7 に示す。その他の条件（初期モデル、評価文）は同じである。実験結果を本手法の結果と共に表 8 に示す。教師あり適応が教師なし適応より認識率が数%高かった。また、表 3 の雑音重畳充足統計量の 20dB のモノフォンの結果を表 8 に示しておく。

3.2.4 評価 4：適応時間の比較

最後にモノフォンモデルにおけるおおよその適応時間の比較を表 9 に示す。提案法は HMM 合成法と同じかそれより若干時間がかかるが雑音重畳学習に比べてはるかに速い（PTM に関してはモノフォンの約 4 倍であった）。

4 考察

評価 1 において SNR が 15dB のとき、clean-noisy と noisy-noisy の認識率の差が少し広がった。これは noisy-noisy において話者選択がうまく行なわれていないことが考えられる。実際に選択された話者を clean-noisy と比較して見てみると入力が男性の話者であるにもかかわらず、女性の話者が尤度上位に入っていることがたまに見受けら

表 5: 適応文章数と単語認識率 (%)

SNR=25dB

適応前: 64.3 %(モノフォン), 76.0 %(PTM)

適応文章数	20	60	100	200
clean-noisy(モノフォン)	74.6	76.8	77.2	78.6
noisy-noisy(モノフォン)	75.4	77.2	76.9	78.4
clean-noisy(PTM)	84.5	85.4	85.8	86.6
noisy-noisy(PTM)	84.9	85.5	85.6	85.7

SNR=20dB

適応前: 48.3 %(モノフォン), 60.1 %(PTM)

適応文章数	20	60	100	200
clean-noisy(モノフォン)	68.1	69.2	71.1	71.2
noisy-noisy(モノフォン)	66.7	67.9	67.9	68.9
clean-noisy(PTM)	76.7	78.7	77.8	78.5
noisy-noisy(PTM)	77.0	77.0	78.1	79.2

SNR=15dB

適応前: 31.5 %(モノフォン), 39.7 %(PTM)

適応文章数	20	60	100	200
clean-noisy(モノフォン)	54.5	57.8	58.5	59.5
noisy-noisy(モノフォン)	53.0	55.0	55.7	56.3
clean-noisy(PTM)	66.0	67.8	68.2	68.4
noisy-noisy(PTM)	63.3	65.2	65.2	66.4

れた。しかしパワーの低い音響パラメータをカットしているので、話者 GMM がクリーンモデルであるにもかかわらず認識率の差はそれほど大きくなかった。このことから雑音環境下での話者選択はほぼ成功していると思われる。認識率の差を埋めるならば話者選択部においてさらなる改善が必要であると思われる。

EM アルゴリズムによる雑音重畳学習は認識率が良いが一回の適応に対して学習データが非常に多く、学習時間が非常にかかり不便である。HMM 合成法は適応時間が非常に速いが認識率が悪く、大語彙連続音声認識においての使用は厳しい。教師ありの適応は適応時間が速く、認識率も良いが多くの文章を正確に発声しなければならず、お年寄りや声の不自由な人たちにとっては大変な負担がかかる。本手法では適応時間が速く適応するための入力文章も 1 文と短く、さらにこの入力文章は話者の音響的特徴から近い話者を選択するためなので、正確に発話する必要はないので労力をほとんど必要としない。また、認識率はオフィス環境に近い 25dB で 86%(PTM) 近い認識率が得られ、HMM 合成法よりも高く、雑音重畳学習と同程度であることが確認された。

表 6: 適応アルゴリズムによる単語認識率 [単語正解精度] (%) の比較

モノフォン

アルゴリズム/SNR	25dB	20dB	15dB
適応前	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]
雑音重畳学習	76.1[74.1]	68.9[67.1]	58.0[55.3]
提案法 (clean-noisy)	78.6[76.8]	71.2[69.5]	59.5[57.6]
提案法 (noisy-noisy)	78.4[76.4]	68.8[67.0]	56.3[54.2]

PTM

アルゴリズム/SNR	25dB	20dB	15dB
適応前	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]
雑音重畳学習	86.3[84.6]	78.2[76.2]	71.2[69.3]
提案法 (clean-noisy)	86.6[84.9]	78.5[76.5]	68.4[65.9]
提案法 (noisy-noisy)	85.7[84.1]	79.2[77.5]	66.4[63.4]

表 7: 実験条件 (教師あり適応)

文章数	各話者ごとに 10 文章、50 文章
SNR	20dB
音響モデル	モノフォン、PTM

5 まとめ

MLLR を使った教師なし環境適応アルゴリズムの評価を行なった。発話者の負担をより少なくするため、入力文章と話者 GMM より特徴の近い話者を選択し話者の充足統計量によってモデルの再構築を行なった後、選択話者の雑音重畳データを用いて MLLR による環境適応を行なった。モノフォンモデルは 25dB で 78.6%、20dB で 71.2%、15dB で 59.5%、PTM モデルにおいてはそれぞれ 86.6%、78.5%、68.4%の認識率が得られた。これは、HMM 合成法より高い認識率で、かつ、雑音重畳学習と同程度の認識率であり、本手法の有効性が確認できた。

今後は考察で述べた話者選択における入力文章と話者モデルについてさらなる検討をする。また HMM 合成による雑音重畳モデルからの適応なども考え、環境適応前の初期モデルについて検討する。

謝辞

本報告は NEDO(New Energy and Industrial Technology Department Organization, 新エネルギー・産業技術総合開発機構) の援助を受けて行なわれた。御協力いただいた関係各位に感謝いたします。

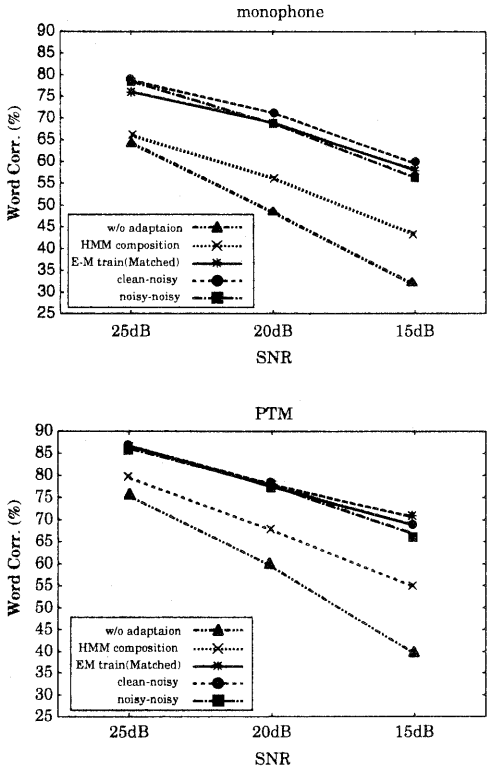

図 2: 各種アルゴリズムにおける単語認識率の比較

表 8: 教師なし適応（提案法）と教師あり適応との単語認識率 (%) の比較

モノフォン, 20dB

教師あり	10 文章	72.0
	50 文章	76.9
教師なし (充足統計量)		74.4
教師なし (提案法)	20 文章	68.1
	60 文章	69.2
	100 文章	71.1
	200 文章	71.2

PTM, 20dB

教師あり	10 文章	80.3
	50 文章	84.9
教師なし (提案法)	20 文章	76.7
	60 文章	78.7
	100 文章	77.8
	200 文章	78.5

表 9: 適応時間

適応手法	文章数	適応時間
雑音重畳学習	45000	24 h
HMM 合成法	なし	30 sec
提案法	20	30 sec
提案法	60	100 sec
提案法	100	160 sec
提案法	200	370 sec

参考文献

[1] M.J.F.Gales and S.J.Young, "Am improved approach to the hidden Markov model decomposition of speech and noise", Proc.ICASSP, pp.233-236, 1992.

[2] Frank Martin, Kiyohiro Shikano, Yasuhiro Minami, "Recognition of Noisy Speech Composition of Hidden Markov Models", Eurospeech93, 33.3, pp.1031-1034, 1993-9.

[3] 芳澤伸一, 馬場朗, 松波加奈子, 米良祐一郎, 山田実一, 鹿野清宏, " 充足統計量と話者距離を用いた音韻モデルの教師なし学習", 情報処理学会研究報告, SLP34-15,pp.83-88,2000.

[4] C.J.Leggetter, 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.

[5] 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", 電子情報通信学会論文誌, J82-D-II No.1, pp.1-9, 1999.

[6] 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, 音学講論, pp. 155-156, Mar.1999.

[7] 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", 情報処理学会研究報告, 99-SLP-31-2, 2000.

[8] 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", 電子情報通信学会論文誌,J83-D-II No.12, pp.2517-2525, 2000.