音声言語情報処理 37−12
（2001. 7. 14）

HMM 合成を用いた雑音環境下音声認識における環境音 GMM の適応化

伊田政樹　　中村哲
ATR 音声言語通信研究所
〒 619-0288 京都府相楽郡精華町光台 2-2-2
E-mail: masaki.ida@slt.atr.co.jp

概要　実環境下で音声認識システムを使用する場合, 周囲の環境に依存した雑音がマイクロホンから混入することは避けられない. 混入する雑音の多くは予測することが難しく, 変動する雑音の混入に対してロバストな音響モデルが求められている. 本稿ではこの問題に対し, 雑音データベースにより構築した環境モデルの適応化を組み込んだ HMM 合成法を提案する. 従来法においては, 環境音のモデルの生成に使用環境の実雑音を用いた学習を行っている. しかし, 実用上の制約から取得できる実雑音のデータ量は限られているので, 少量のデータから得られる環境音モデルは変動に対して弱いという問題がある. そこで, 初期環境音モデルを雑音データベースを用いて用意しておき, 少量の実雑音データで適応化を行う方法を提案する. 評価実験より, 環境音モデル生成に要する実雑音データ量を 20% に削減することができ, 同時に雑音変動に対するロバスト性も確認できた.
キーワード　HMM 合成法, 環境適応化, 環境音モデル, 非定常雑音

Speech Recognition in Noisy Environments Using HMM Composition and Noise GMM Adaptation

Masaki IDA　　Satoshi NAKAMURA
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan
E-mail: masaki.ida@slt.atr.co.jp

abstract　When using a speech recognition system in a real environment, noise from the surrounding affects recognition performance. Most additional noises are difficult to predict, so we need a robust acoustic model for nonstationary noises. In this paper, we suggest a new HMM composition method with noise GMM adaptation. In the conventional HMM composition, the noise model is trained with real noise data. But as the amount of training noise data will always insufficient to match any real noise, such a noise HMM will be of limited use. In a proposed method, we prepare an initial noise GMM based on a noise database, then a noise GMM adaptation with a small amount of real noise data. Experimental results show two improvements, the amount of noise data necessary for adaptation is reduced to 20%, and HMM is robust against nonstationary noise
keywords　HMM composition, environmental adaptation, noise model, nonstationary noise

1　はじめに

実環境で音声認識システムを使用した場合, マイクロホンに周囲の環境に依存した雑音が混入することは避けられない. 混入する雑音を予測することは困難であることが多く, 変動する雑音の混入に対してロバストな音響モデルが求められている.

従来から音響モデルの環境適応化についてさまざまな研究が行われてきた. 大別して, 適応データに雑音の混入した音声を用いる方法と, 適応データに雑音データのみを用いる方法がある. 前者の方法として MLLR [1] や MAP [2], あるいはこれらを応用した方式が用いられてきた. これに対し, HMM 合成 [3][4] やヤコビ適応 [5] などの方法は後者の方法である. 適応データを取得する容易さの観点で後者のほうが有利である. HMM

合成法は事前に環境雑音のモデル化を実雑音データを用いて行う．したがって，環境が変化した場合，その都度環境のモデル化を行う必要がある．ヤコビ適応は環境の変動による音響モデルの変動を線形近似で表すことで短時間での適応化処理を実現している．しかし，Taylor 展開の 1 次項で近似しているため大きな変動に対して近似誤差が無視できなくなる問題点がある．

筆者らは，あらかじめ複数の環境音を環境音モデル学習データとして与えた HMM 合成法を提案し，混入する環境音が未知の場合における性能評価を行った [6]．本稿では，あらかじめ複数の環境音を元に作成した環境音 GMM を適応化することで，短時間のデータを用いて変動に対してロバストな環境音 GMM を生成し，これを HMM 合成して音響モデルを環境適応化する方法について提案する．

2 HMM 合成

HMM 合成法は，事前に clean speech を用いて学習を行った音韻の音響モデルと，環境雑音のモデルとを合成することで，モデル化された環境雑音に適応した音響モデルを生成する方法である．

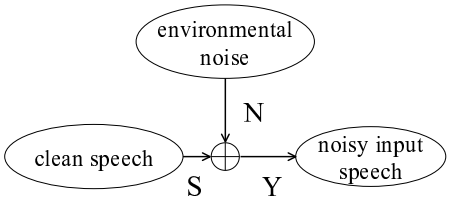

図 1: 仮定する音声信号の観測系

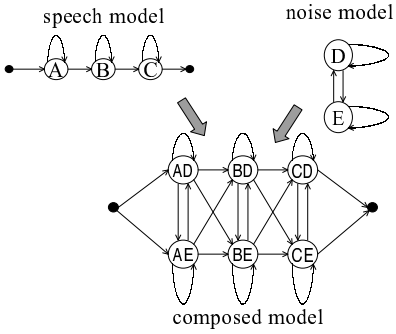

図 2: 合成 HMM の構造

本稿では入力音声の観測系を図 1に示すモデルと仮定する．観測される入力音声を Y とし，これを環境雑音 N と雑音のない clean speech S で表す．環境雑音の加法性は線形スペクトル領域において成立し，

$$Y_{linspc} = S_{linspc} + N_{linspc} \quad (1)$$

一方，音響モデルは一般的に対数ケプストラムにより特徴抽出されているので，

$$Y_{logcep} = \Gamma^{-1} \log[\exp\{\Gamma(S_{logcep})\} + k \exp\{\Gamma(N_{logcep})\}] \quad (2)$$

となる．Γ はフーリエ変換，k は SN 比に応じて決定する係数である．式 (2) を HMM に適応した場合，合成 HMM の構造は図 2に示すように各 HMM の直積で表される．遷移確率は対応する遷移確率の積で求められる．出力確率分布は各状態において結合される．音韻 HMM，環境音 HMM の各出力確率が正規分布で表現されている場合，合成された出力確率はその和で与えられる．

3 従来法 HMM 合成法の問題点

3.1 実騒音データ量と認識性能

図 3に示す通り，従来法の HMM 合成は使用環境の実騒音データを用いて環境音モデルの学習を行っており，性能は実騒音データの量に依存する．本節では予備実験として実騒音データ量と認識性能の関係について調べる．

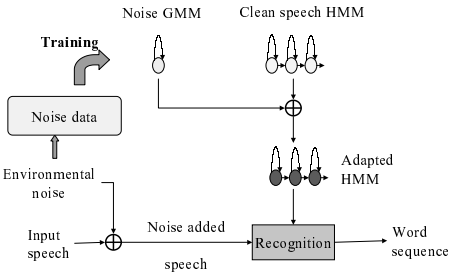

図 3: HMM 合成による雑音環境下の音声認識 (従来法)

実験条件を表 1に，実験結果を図 4に示す．図中の clean speech - clean HMM は clean speech HMM を用いて雑音の混入していない音声を認識した場合の性能，clean HMM は clean speech HMM を用いて展示会場雑音の混入した音声を認識した場合の性能，exhibi-HMM はあらかじめ展示会場雑音を重畳した音声データで作成した HMM を用いた場合の認識性能である．実験結果の図より，学習データ量が少ない場合，十分にある場合に比べて音声認識性能 (Word Accuracy) が劣化し，その傾向は混合数が大きいほど顕著に表れることがわかる．また，データ量に関しては，この評価環

表 1: 実験条件

音声データ	:	サンプリング周波数 16kHz 16bit PCM monoral
特徴ベクトル	:	$12Cep+\Delta Cep+\Delta pow$ (pow は合成時のみ使用) フレーム周期 20msec ハミング窓 フレームシフト 10msec CMN なし
Clean speech HMM	:	性別依存 / 話者非依存 triphone 1 音素あたり 4 ～ 5 状態 全 1400 状態の HMnet ATR 音声データベース 計 3619 話者を用いて学習
Noise GMM	:	1 状態 {1,2,4,8}mixture {100, 10, 5, 3, 1}sec の電子協展示会場雑音 (ブース内) で学習
評価データ / タスク	:	ATR 旅行対話評価セット 42 対話 連続単語音声認識 電子協展示会場雑音 (ブース内) を SNR=15dB になるよう重畳

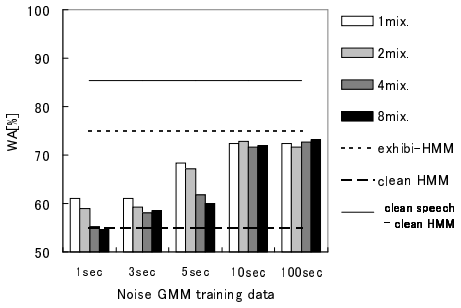

図 4: 環境音 GMM 学習データ量と認識性能

境を GMM で表すためには 10sec の実雑音データが必要であるといえる. しかしながら, 実際の使用環境として変動する雑音環境を考えた場合, 雑音環境に変化がみられるたびに 10sec の実雑音データを取得することは難しく, 適応化に必要な実雑音データ量を削減する必要がある.

3.2 雑音変動に対するロバスト性

変動する雑音環境を考えた場合, 適応化ののち認識までの間に環境雑音が変動する場合や発話中に変動する場合についても考慮しなければならない. 先ほどの 10sec の展示会場雑音に対し作成した 2 混合の環境音 GMM を HMM 合成した音響モデル の環境変動に対するロバスト性を調べるため, 以下の 3 つの評価データによる認識性能を比較する. 評価データは, ATR 音声 DB の旅行対話評価セットに以下の雑音を SNR = 15dB となるよう重畳したものである.

Exhibi 電子協展示会場ブース内雑音
(環境音 GMM 学習データと一致)

Comp 電子協計算機室雑音 (未知雑音)

Nonstationary 開始～ 0.5sec に展示会場雑音, 以降に計算機室雑音 (雑音環境が途中で変動)

評価実験の結果を図 5に示す. 未知雑音が混入した場合 (Comp), 混入雑音が変動した場合 (Nonstationary) いずれの場合も大幅に認識性能が低下している.

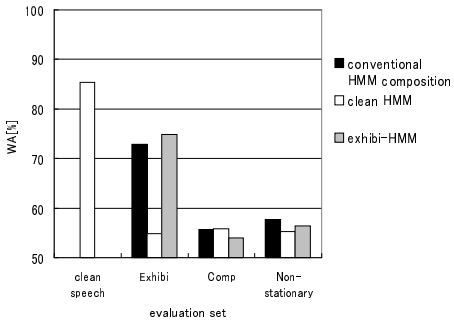

図 5: 従来法 HMM 合成の環境変動に対するロバスト性

4 雑音データベースを用いた HMM 合成と環境音 GMM 適応化

前節で従来法には, 十分な実雑音データ量が必要であることと, 環境変動に対するロバスト性が低いことの 2 つの問題点が存在することを示した. この問題点に対処するため, 本節では環境音モデルの生成に雑音データベースと環境音モデルの適応化を用いた方法を提案する. あらかじめ雑音データベースから用意した初期環境音モデルの適応化を行うことで, 環境音モデルを実雑音データだけから学習する従来法に比べて実雑音データの量を減らすことを考える. また, 環境音モデルが雑音データベースに含まれる雑音の特性をあわせ持つことで環境変動に対するロバスト性を与える. 提案手法の概略を図 6に示す.

あらかじめ雑音データベースのさまざまな雑音を用いた初期環境音 GMM を学習する. この初期環境音 GMM と clean speech HMM を合成した初期合成 HMM も準備しておく. 適応の際には, この初期環境音 GMM に, 少量の使用環境の実雑音データを用いて混合重み係数の適応化を施す. 適応化には MAP 推定を用いる. 適応化を重み係数に限定することにより, 初期合成 HMM の重み係数に適応化環境音 GMM の重み係数を反映するだけで適応化 HMM を得ることがで

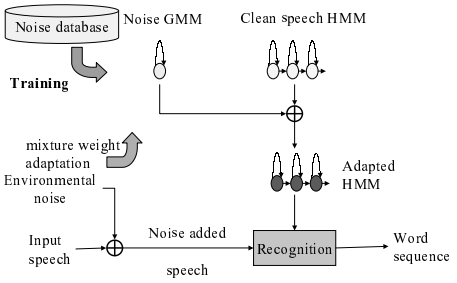

図 6: 雑音データベースを用いた HMM 合成

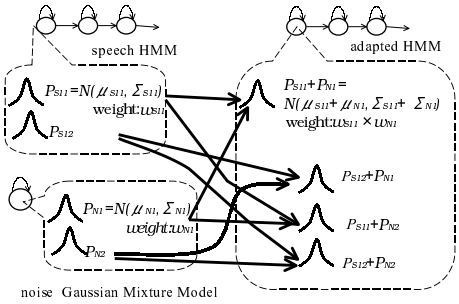

図 7: GMM 重み係数の適応化

きる. このことを図 7に示す. 図中, speech HMM と環境音 GMM を合成する場合を考える. speech HMM の第 1 状態の出力確率分布が 2 混合の混合正規分布

$$w_{S11}P_{S11} + w_{S12}P_{S12}$$

で与えられ, 環境音 GMM も同様に

$$w_{N1}P_{N1} + w_{N2}P_{N2}$$

で与えられるとする. w はそれぞれ混合重み係数である. $P_{S11}, P_{S12}, P_{N1}, P_{N2}$ はそれぞれ正規分布であり, 平均 μ と分散 Σ で表される.

$$P_{S11} = N(\mu_{S11}, \Sigma_{S11})$$
$$P_{S12} = N(\mu_{S12}, \Sigma_{S12})$$
$$P_{N1} = N(\mu_{N1}, \Sigma_{N1})$$
$$P_{N2} = N(\mu_{N2}, \Sigma_{N2})$$

このとき合成後の HMM の第 1 状態の出力確率分布は次式で表される.

$$\begin{aligned} & w_{S11}w_{N1}(P_{S11} \oplus P_{N1}) \\ + \quad & w_{S12}w_{N1}(P_{S12} \oplus P_{N1}) \\ + \quad & w_{S11}w_{N2}(P_{S11} \oplus P_{N2}) \\ + \quad & w_{S12}w_{N2}(P_{S12} \oplus P_{N2}) \end{aligned} \quad (3)$$

正規分布の和は, 平均・分散それぞれの和として表される. たとえば,

$$\begin{aligned} P_{S11} \oplus P_{N1} &= N(\mu_{S11}, \Sigma_{S11}) + N(\mu_{N1}, \Sigma_{N1}) \\ &= N(\mu_{S11} + \mu_{N1}, \Sigma_{S11} + \Sigma_{N1}) \end{aligned} \quad (4)$$

環境音 GMM の混合重み適応化を行った場合, w_{N1}, w_{N2} のみが適応化されるので, 合成後のモデルにおいても式 (3) の各項の確率分布の平均や分散は変化しない. したがって, 適応化された GMM の重みを反映することで適応化 HMM を得ることができ, 適応化の都度, HMM 合成を行う必要はない. これに対し従来法では, 各項の確率分布が環境音モデルを学習するたびに変わるため, その都度式 (2) にしたがった計算が必要になる.

5 評価実験

5.1 初期 GMM 学習雑音データの構成と認識性能

ここでは, まず初期環境音 GMM を構築するために用いる雑音データの種類や量の関係を調べるため, 初期環境音 GMM を適応化なしで合成した場合の合成 HMM の連続単語音声認識性能評価実験を行う. 特徴ベクトル, Clean speech HMM および評価データに関する条件は表 1と同一である. 環境音 GMM の構成は 1 状態 8 混合とする. 初期環境音 GMM の学習データセットは電子協騒音データベース [7] を用いる. 電子協騒音データベースは全 17 種類の騒音データが収録されており, 評価データに重畳した雑音 (展示会場ブース内雑音) および, それに類似した雑音 (展示会場通路雑音) を除く 15 種類の雑音データを用いる. 雑音の種類 (異なり数) と量 (時間) により表 2に示す,A1 ～ A5 の 5 種類について調べる.

結果を図 8に示す. 図中 conventional は従来法により評価環境の雑音データだけを用いて 2 混合の環境音 GMM の学習を行った場合で, 図 4に示したものと同一である.

学習セットに含まれる雑音の種類が多い A3 セットが最も高い性能を示す. 学習データの総量が同じ場合 (A2 と A4,A1 と A5 の比較) においては, 多種の雑音を含んでいる方 (A2,A1) が良い性能を示す. 以上より, 雑音の種類が多いほど, またデータ量が多いほど性能が良いことがわかる.

次に学習セットに評価データと同じ雑音や類似した雑音データが含まれる場合について調べる. A1 ～ A5 セットに類似雑音 (展示会場通路雑音) を加えたものを B セット, 評価雑音 (展示会場ブース内雑音) を加えたものを C セットとする. 詳細を表 3に示す.

表 2: 初期環境音 GMM 学習データセット

A1 set	2sec × 15 種	計 30sec
A2 set	4sec × 15 種	計 60sec
A3 set	10sec × 15 種	計 150sec
A4 set	10sec × 6 種	計 60sec
A5 set	10sec × 3 種	計 30sec

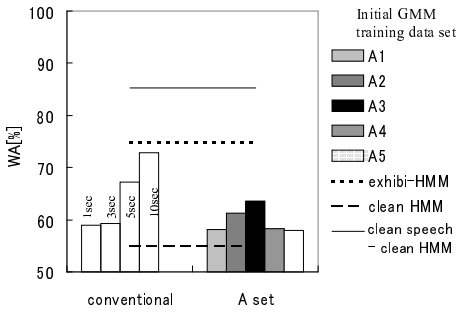

図 8: 雑音 DB による初期環境音 GMM を用いた合成 HMM の認識性能

図 9に示す通り, 類似雑音を含む場合 (B セット) や評価雑音を含む場合 (C セット) はそれ以外の雑音が少ない場合 (B5,C5 セット) に高い性能を示す. すなわち, 環境音 GMM を構成する要素のうち評価雑音環境に相当する要素の重みが大きいほうが認識性能が向上している. このことからさまざまな雑音データを用いて学習した環境音 GMM(A3,B3,C3 セット) に重み適応を施すことで, 少量の適応データを用いて評価雑音環境に近づけることで, 認識性能の向上が可能になると考えられる.

5.2 環境音 GMM の重み適応化による性能向上

環境音 GMM の重み適応化による認識性能について評価を行った. 評価は初期 GMM として A3 セット,B3

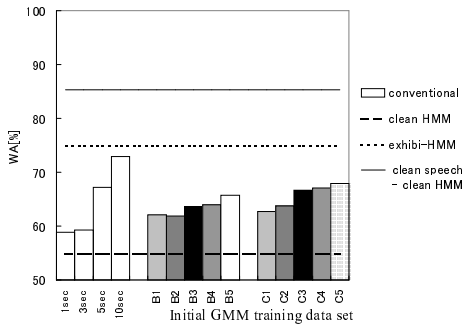

図 9: 雑音 DB に評価雑音や類似雑音が含まれる場合の合成 HMM の認識性能

表 3: 評価環境と同一あるいは類似データを含む学習データセット

B1 set	A1 + 類似雑音 2sec	計 16 種 32sec
B2 set	A2 + 類似雑音 4sec	計 16 種 64sec
B3 set	A3 + 類似雑音 10sec	計 16 種 160sec
B4 set	A4 + 類似雑音 10sec	計 7 種 70sec
B5 set	A5 + 類似雑音 10sec	計 4 種 40sec
C1 set	A1 + 評価雑音 2sec	計 16 種 32sec
C2 set	A2 + 評価雑音 4sec	計 16 種 64sec
C3 set	A3 + 評価雑音 10sec	計 16 種 160sec
C4 set	A4 + 評価雑音 10sec	計 7 種 70sec
C5 set	A5 + 評価雑音 10sec	計 4 種 40sec

セット,C3 セットを用いて学習したものについて行った. 重み適応に用いる適応データとして, 評価環境 (展示会場ブース内雑音) データを 0.5sec, 1sec, 3sec 与えた場合について比較する.

実験結果を図 10に示す. conventional および without Adapt. は図 8および図 9に示したものと同一である. A3 セットに適応化を行うことで初期環境音 GMM 学習データセットに評価環境が含まれない場合においても認識性能が向上する.GMM 適応データ量が 1sec の場合で, 従来法の 5sec の実雑音データを環境音 GMM の学習に用いた場合に匹敵する.

学習データセットに評価環境や類似環境が含まれている場合 (B3,C3) の認識性能の向上と比較しても, 含まない場合 (A3) の性能向上は遜色なく, 提案手法が有効に機能していると考えられる.

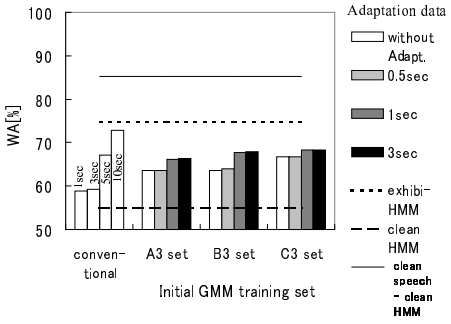

図 10: 環境音 GMM 適応化を行った合成 HMM の認識性能

5.3 雑音環境変動に対するロバスト性

前節の結果, 環境音モデルの適応化を行うことで, 性能向上と引きかえに環境変動に対するロバスト性が損なわれている恐れがある. そこで本節では, 提案手法に

よって適応化を行った音響モデルの, 雑音環境変動に対するロバスト性を調べる. 評価対象の音声データと音響モデルは以下の通りである.

評価データ

ATR 音声 DB の旅行対話評価セットに以下の雑音を SNR=15dB となるよう重畳したもの

Exhibi 電子協展示会場ブース内雑音 (環境音 GMM 学習データまたは適応データと一致)

Comp 電子協計算機室雑音 (提案法の初期環境音 GMM 学習データに含まれる)

NOISEX NOISEX-92 DB よりエンジンルーム雑音 (未知雑音)

Nonstationary 開始〜 0.5sec に展示会場雑音, 以降に計算機室雑音 (雑音環境が途中で変動)

音響モデル

conventional 10sec 従来法 HMM 合成モデル (10sec の展示会場雑音で学習した 2 混合環境音 GMM を合成)

conventional 5sec 従来法 HMM 合成モデル (5sec の展示会場雑音で学習した 2 混合環境音 GMM を合成)

A3 + GMM adapt 提案法 (A3 セットによる 8 混合初期環境音 GMM を 1sec の展示会場雑音で適応化)

実験結果を図 11に示す. clean speech HMM を用いて雑音の混入した音声を認識した場合の性能, およびあらかじめ展示会場雑音を重畳した音声データで作成した HMM を用いた場合の認識性能 (exhibi-HMM) も合わせて示す.

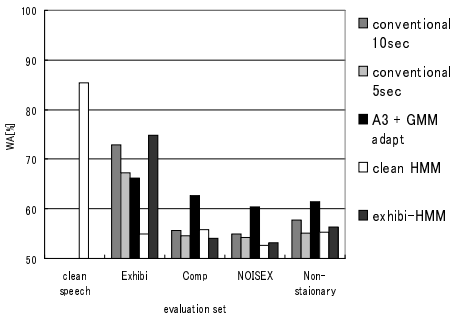

図 11: 提案法の環境雑音変動に対するロバスト性

図 11に示す通り, 従来法による HMM 合成では学習していない環境に対して弱く, clean speech HMM を用いた場合並みに認識性能が低下しているのに対し, 提案法では認識性能の低下が大幅に抑制されている. 従来法では環境音 GMM に実雑音データだけの情報しか含んでいないのに対し, 提案法では初期環境音 GMM 学習データとしてさまざまな音の情報を含んでいることが有効に働いていると考えられる.

6 まとめ

本稿では, 従来法の HMM 合成について適応のための実雑音データ量と雑音変動に対するロバスト性の 2 つの問題点を取り上げ, その解決策として雑音データベースを用いた HMM 合成と, 環境音 GMM 適応化を用いる方法を提案した. 実雑音データ量の問題に関しては, 従来法で 5sec の実雑音データを用いて実現した性能を 1sec の実雑音データを適応化に用いることで実現した. 雑音変動の問題に関しては, 初期環境音 GMM の学習セットに含まれる雑音と含まれない雑音の両面について, その効果を確認した.

今後の課題としては, 本稿の提案においては SN 比が一定である条件を課しており, SN 比の変動に対するロバスト性について検討を行うことが挙げられる. また, 環境音 GMM の適応範囲を分布の平均や分散に拡大し, 性能向上をはかることと, 発話と発話の間の無音区間や, 発話内の無音区間を利用した逐次適応への応用を考えている.

謝辞 本研究の機会を与えていただきました ATR 音声言語通信研究所 山本誠一社長に感謝いたします. また, 本研究を進めるに際し有益なご助言をいただきました ATR 音声言語通信研究所第 1 研究室の研究員諸氏にお礼申し上げます.

参考文献

[1] C.J. Leggetter, P.C.Woodland, "Maximum Liklihood linear regression for speaker adaptation of continuous density hidden markov models", Computer Speech and Languege, vol. 9, pp. 171-185, 1995

[2] J.L. Gauvian, C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", Trans SAP, vol. 2, No. 2, IEEE, pp. 291-298, Apr. 1994

[3] M.J.F. Gales, S.J. Young, "HMM Recognition in Noise Using Parallel Model Combination", Proc. of EUROSPEECH, pp. 837-840, Sep. 1993

[4] F. Martin, K. Shikano, Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models", Proc. of EUROSPEECH, pp. 1031-1034, Sep. 1993

[5] S. Sagayama, Y. Yamaguchi, S. Takahashi, J. Takahashi, "Jacobian Approach to Fast Acoustic Model Adaptation", Proc. of ICASSP, pp. 835-838, 1997

[6] 伊田政樹, 松井知子, 中村哲, "HMM 合成による環境音重畳音声の認識", 音講論集, 2-5-9, 2000 年 9 月

[7] 電子協騒音データベース, http://it.jeita.or.jp /jhistory/committee/humanmed/speech/noisedbj.html