音声言語情報処理 39－9
(2001. 12. 20)

雑音DBを用いたモデル適応化HMMのSN比別マルチパスモデルによる雑音下音声認識

伊田 政樹† 中村 哲†

† ATR音声言語コミュニケーション研究所
〒619-0288 京都府相楽郡精華町光台2-2-2

E-mail: †masaki.ida@atr.co.jp

あらまし 音声認識システムを実環境で利用する場合, その認識性能は周囲の環境雑音の混入に大きく影響を受ける. 混入する雑音は多くの場合予測が困難であり, 入力される音声信号と音響モデルの間で不一致が生じ, 認識性能低下の原因となる. このことから, 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. 混入する雑音の問題は, 雑音の種類が未知である問題とSN比が未知である問題の2つに分けて考えることができる. 本稿ではこの問題に対し, 一つ目の雑音の種類が未知である問題に対して既存の雑音データと雑音モデルの適応化によるHMM合成法を用い, 二つ目のSN比が未知である問題に対して複数のSN比に対応した音響モデルを並列に用いる. AURORA2タスクによる評価実験の結果, 1 secの適応データを用いることでSNR = 5dBにおいてベースラインシステムに対して53%の認識性能改善を得た. これは従来法のHMM合成を用いた場合10 secの適応データを用いた場合に匹敵する.

キーワード HMM合成法, 雑音モデル, 非定常雑音, マルチパスモデル

Rapid Model Adaptation with a Prior Noise GMM and Multi-SNR Models for Noisy Speech Recognition

Masaki IDA† and Satoshi NAKAMURA†

† ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288, Japan

E-mail: †masaki.ida@atr.co.jp

Abstract When a speech recognition system is used in a real environment, the recognition performance is affected by surrounding noise. Most additional noises are difficult to predict about kind of noise and SNR, so we cannot avoid the mismatch situation between those of training data and test data. Then we need a method to deal with mismatched noise problems and unknown SNRs. In this paper, we propose an HMM composition-based model adaptation that uses a prior noise data against noise mismatches. We also prepare plural HMMs for several SNRs and select the best model based on acoustic likelihood to deal with the unknown SNRs. Experimental results with AURORA2 task show 53% word accuracy improvement from baseline system with 1 sec real noise data for adaptation. The performance is equivalent to a case with 10 sec real data using the conventional HMM composition method.

Key words HMM composition, noise model, nonstationary noise, multipath model

1. はじめに

音声認識システムを実環境下で使用した場合, 入力音声が歪みを受けることで認識性能が低下する. これらの歪みは伝送経路による歪み (音響特性やマイクロホンの特性) と加法性の雑音の2つに大別できる. 本稿では加法性雑音に対して取り扱う. これらの加法性雑音は一般的に予測困難であり, 多種多様な雑音に対してロバストな音響モデルが求められている.

筆者らは, 雑音の混入に対してロバストな音響モデル構築法としてHMM合成法に着目してきた [1]〜[3]. さまざまな雑音を用いて学習した初期雑音モデルを必要に応じて適応化することで, 雑音の多様性に即座に対応し, SN比ごとに音響モデルの経路を複数個用意してマルチパス化することで, 入力音声のSN比が変化することに対応する. 本稿では, この手法を雑音環境下音声認識の一般的なタスクであるAURORA2データベース [4] により評価した.

2. 多様な雑音下の音声認識における問題点

AURORA2データベースは雑音環境下における音声認識システム評価用データベースである. 詳細は表2. に示す. マイクロホン (伝送経路) の相違を評価するためのC setは今回使用しない. 以下, ベースラインの結果は学習セットのうち雑音を含まない音声データを用いて作成した音響モデル (クリーン HMM) のA set, B setすべての結果の平均を示す.

雑音環境下における音声認識で, もっとも簡単かつ理想的な音響モデルの構築法は入力音声と同じ雑音環境下での学習データを用いて音響モデルを構築する方法である. 以下, 入力音声と同じ雑音環境下の学習データで作成した音響モデルをMatchedモデルと呼ぶ. Matchedモデルの学習には学習セットのうち, 1種類の雑音の混入したサブセットを用い, 評価にはA setのうち対応した雑音が混入した音声データを用いる. これらの平均をMatchedモデルの性能とする. 音響モデル学習データと評価データの混入雑音が異なっている場合の評価として上記Matchedモデルとして作成した音響モデルにA setのうち対応しない雑音の混入した音声データを用いて評価する. HMMの学習・認識はHTKを用いた. 結果を図1に示す. 学習データと入力音声の雑音環境が一致していない場合, 入力音声のSN比の低下に伴って大幅に認識性能が低下している.

表1 AURORA2

タスク: TI-digit (連続数字認識)
サンプリング周波数: 8kHz
16bit PCM / モノラル
training set
雑音: subway, babble, car noise, exhibition hall
SN比: 5dB, 10dB, 15dB, 20dB, clean
全発話数: 8840
test set A
雑音: subway, babble, car noise, exhibition hall
SN比: -5dB, 0dB, 5dB, 10dB, 15dB, 20dB, clean
全発話数: 28028
test set B
雑音: restaurant, street, airport, train station
SN比: -5dB, 0dB, 5dB, 10dB, 15dB, 20dB, clean
全発話数: 28028

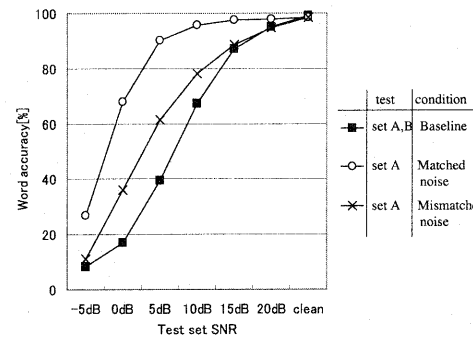

図1 Baseline, Matched modelによる認識結果

一方, 学習データとしてさまざまな雑音を含んだ音声データを用いて音響モデルの学習を行った場合 (Multi-conditionモデル) について検討する. Multi-conditionモデルの学習データにはAURORA2の学習セット全てを用いる. 評価実験として, 学習セットと同じ雑音環境であるA set (雑音条件既知) と学習セットに含まれないB set (雑音条件未知) を用いて比較する. 結果を図2に示す. Multi-conditionモデルを用いた場合には, 雑音環境が未知であるB setの場合, 認識性能はA setに比べて劣る. 雑音環境が既知であるA setの場合は, Matchedモデルを用いた場合に比べて劣化の幅が小さいものの, Multi-conditionモデルの学習には多量の学習データを必要とする問題点を抱えている. 本稿の評価実験においては連続数字認識であるのでデータ量として大きな問題になっていないが, 大語彙連続音声認識システムを考えたと

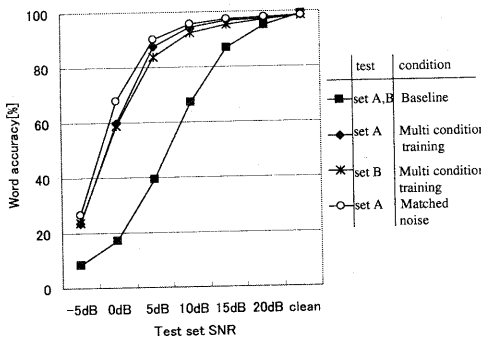

図 2　Multi condition model による認識結果 (set B)

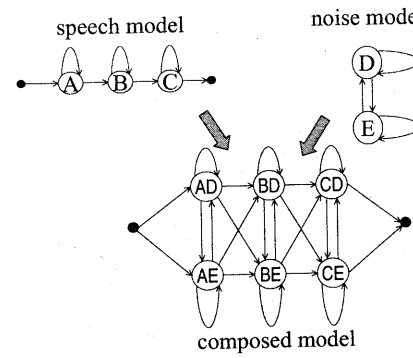

図 3　合成 HMM の構造

き, 非現実的な学習データ量を扱うことになる.

3. 従来法による HMM 合成

従来より音響モデルの環境適応化に用いられてきた手法として MLLR [5] や MAP [6] がある. しかし, これらの方法は適応データとして雑音の混入した音声を必要とするため, 適応データの取得時にユーザに負担をかけるという問題点がある. 適応データとして雑音のみを用いる手法として, 本稿では HMM 合成 [7], [8] に着目する. HMM 合成法は, 事前に clean speech を用いて学習を行った音素の音響モデルと, 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作成する方法である. 本稿では加法性の雑音のみを仮定する. 観測される入力音声のパワースペクトルを Y とし, これを環境雑音のパワースペクトル N と雑音のない clean speech のパワースペクトル 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,

$$Y_{linspc} = S_{linspc} + N_{linspc} \tag{1}$$

一方, 音響モデルは一般的にスペクトルにより特徴抽出されているので,

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

となる. Γ, Γ^{-1} はフーリエ変換およびフーリエ逆変換, 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 3 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, 出力確率分布は各状態において結合される. HMM 合成による雑音環境下の音声認識のブロック図を図 4 に示す.

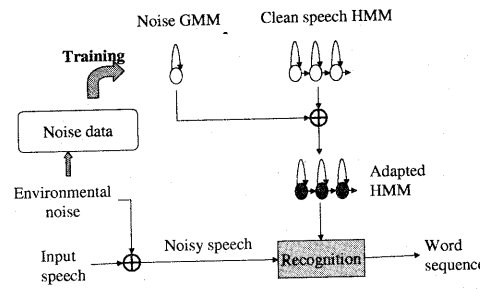

図 4　HMM 合成

4. SN 比別マルチパスモデル

HMM 合成法においては式 (2) に示す通り, 入力音声の SN 比が既知であるという制約がある. この問題の解決のため, 複数の SN 比に対応した適応化 HMM を並列に構築する手法を用いる. この手法は自由発話音声の認識における音響特徴の変動や多様化に対する方法として用いられている [9], [10]. 本手法の概略図を図 5 に示す. 雑音モデルを合成する際に, 入力音声として予測される範囲内のいくつかの SN 比に対応した複数の合成 HMM を得る (図中では SNR = 10, 15, 20 dB). 認識の際, 入力音声の SN 比はわからないので, これら各合成 HMM を 1 つのモデルとして取り扱う. すなわち, 各モデルに複数の SN 比のパスを定義し, デコードする際に最も尤度の高い経路を選択

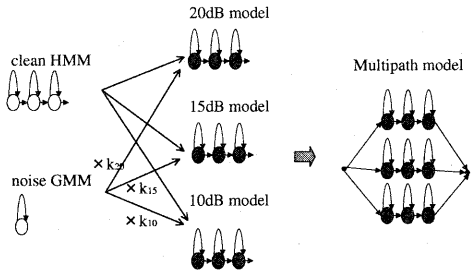

図 5　SN 比別マルチパスモデル

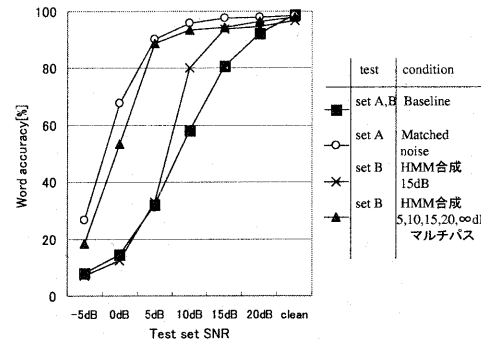

図 6　SN 比別マルチパスモデルによる認識結果 (set B)

させる.

従来法の HMM 合成による音響モデル環境適応化の評価として, AURORA2 の B set を評価データに用いた認識実験を行う. 評価データの各雑音に対して雑音データ 10 sec を用いて雑音モデルの学習を行う. 雑音モデルは 1 状態 8 混合の GMM を用いる. この雑音モデルを用いて以下の 2 つの音響モデルを作成し, 比較する.

- SNR = 15 dB として HMM 合成した音響モデル (HMM 合成 15dB)
- SNR = 5, 10, 15, 20, inf(clean) dB として HMM 合成し, マルチパス化した音響モデル (HMM 合成 マルチパス)

実験結果を図 6 に示す. 比較のため, ベースラインと Matched モデルの結果もあわせて示す.

HMM 合成による適応化によって, SNR = 15 dB 固定の場合 13%の性能向上が見られた. HMM のマルチパス化を用いることで, SNR = 5 dB においてベースラインモデルと比べて 58%高い性能を得た.

5. 雑音 DB とモデル適応化を用いた HMM 合成法とマルチパスモデル

従来の HMM 合成においては, 環境雑音のモデル化に十分な量の雑音データが必要である. 少量の雑音データで環境適応化を行う方法として, 筆者らは雑音 DB とモデル適応化を用いる HMM 合成法を提案してきた [2], [3]. 提案法の概略を図 7 に示す. まず, あらかじめ多様な雑音を含む DB を用いて初期雑音 GMM の学習を行う. また, のちの計算簡単化のため,

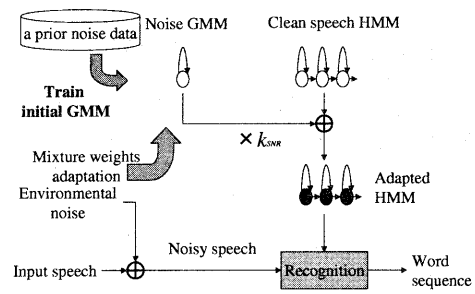

図 7　雑音 DB とモデル適応化を用いた HMM 合成

初期雑音 GMM とクリーン HMM を HMM 合成した初期合成 HMM も準備しておく. 環境適応化の際には, 少量の実雑音データを取得して初期雑音 GMM に混合重み適応化を施し, 適応化雑音 GMM を得る. 適応化には MAP 推定を用いる. 適応化を GMM の混合重み係数に限定しているので, 適応化を行った上で HMM 合成した適応化 HMM と, 初期合成 HMM の間で, 各確率分布の平均や分散が変化することはなく, 環境適応化により変化するのは重み係数のみである. したがって, GMM 適応化で得た重み係数を合成後のモデルに直接反映することで適応化 HMM を得ることができ, 計算量を大きく削減できる. この関係を図 8 に示す.

提案法の HMM 合成による音響モデル適応化の評価として, AURORA2 の B set を評価データに用いた認識実験を行う. 雑音モデルは 1 状態 8 混合の GMM

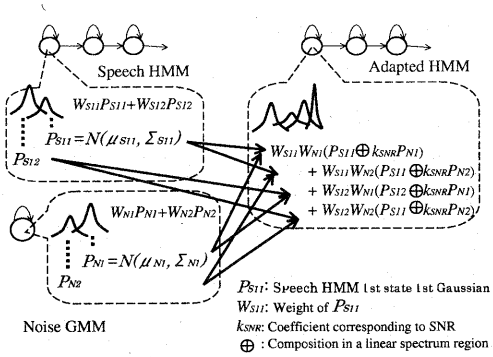

図 8　混合重み適応化の計算

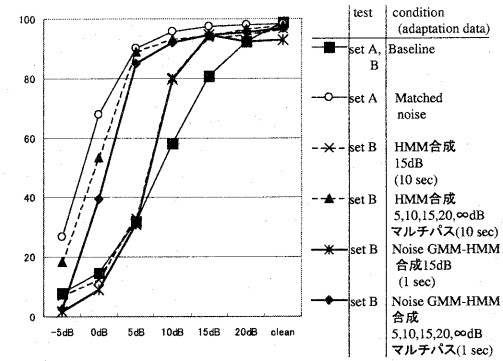

図 9　雑音 GMM-HMM 合成とマルチパスモデルによる認識結果 (set B)

とし, 電子協騒音データベース [11] より 10 sec × 25 種類, 計 250 sec の雑音データを用いて初期雑音モデルの学習を行う. 評価データの各雑音に対して, 雑音データ 1 sec を用いて雑音モデルの適応化を行う. この雑音モデルを用いて,

- SNR = 15 dB として HMM 合成した音響モデル (提案法 15dB)
- SNR = 5, 10, 15, 20, inf(clean) dB として HMM 合成し, マルチパス化した音響モデル (提案法 マルチパス)

を作成した.

実験結果を図 9 および表 2 に示す. 提案法を用いることで, 10 分の 1 の適応データ量で従来法とほぼ同等の認識性能を達成できる. ベースラインモデルに対して, SNR = 15dB 固定の場合において 14%の性能向上が見られた. また, 適応化モデルのマルチパス化により, SNR = 5dB においてベースラインに比べて 53%の性能向上を得た.

6. ま と め

本稿では多様な雑音が混入する音声認識における問題点として, 雑音の種類が変動する問題と SN 比が未知である問題を取り上げた. その解決策として雑音 DB とモデル適応化を用いた HMM 合成法と SN 比別のマルチパスモデルを併用する方法を用いた場合について AURORA2 データを用いて評価実験を行った. 実験の結果, 初期雑音モデルの学習に含まれない雑音環境の評価データ (SNR = 5dB) に対して, 1 sec の実雑音データを適応データとして用いることで 53%の性能向上を得た. これは従来法の HMM 合成を用いた場合においては 10 sec の適応データで得られたものであり, 適応データ量を 10 分の 1 に削減できた.

しかしながら, 10 sec のデータを用いた場合, 従来法の HMM 合成を用いた方がわずかではあるが性能が勝っている現象が見られた. このことから, 雑音環境の定常性をもとに従来法と提案法を選択的に用いる機構について検討を行う必要がある.

謝　　辞

本研究の機会を与えていただきました ATR 音声言語コミュニケーション研究所 山本誠一所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 音声言語コミュニケーション研究所第 1 研究室の研究員諸氏にお礼申し上げます.

文　　献

[1] 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, 2000 年秋季音講論集, 2-5-9, pp. 67–68, 2000 年 9 月

[2] 伊田政樹, 中村哲, 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, 情処研報, 2001-SLP-37-12, pp. 67–72, 2001 年 7 月

[3] 伊田政樹, 中村哲, 雑音 DB とモデル適応化を用いた HMM 合成法における雑音変動耐性の評価 2001 年秋季音講論集, 1-1-17, pp. 33–34, 2000 年 9 月

[4] H. G. Hirsch, D. Pearce, The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions, ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium", 2000 年 9 月

[5] C.J. Leggetter, P.C.Woodland, Maximum Likelihood linear regression for speaker adaptation of continuous density hidden markov models, Com-

表2　雑音 GMM-HMM 合成とマルチパスモデルによる認識結果
(Word accuracy[%])

	Test set A									
	Baseline					Matched model				
	subway	babble	car	exhibition	ave.	subway	babble	car	exhibition	ave.
clean	98.8	99.0	98.8	99.5	99.0	98.3	98.5	98.6	98.3	98.4
20dB	97.0	90.0	96.8	96.2	95.0	98.2	97.9	98.1	97.8	98.0
15dB	92.9	73.4	89.5	91.9	86.9	97.4	97.3	97.9	97.4	97.5
10dB	78.7	49.1	66.2	75.1	67.3	96.0	95.6	96.7	94.9	95.8
5dB	53.4	27.0	33.5	43.5	39.4	92.1	88.3	91.1	89.5	90.2
0dB	27.3	11.7	13.3	16.0	17.1	71.8	61.8	67.1	70.5	67.8
-5dB	12.6	5.0	8.4	7.7	8.4	29.2	26.6	22.0	29.0	26.7
	Test set B									
	Baseline									
	restaurant	street	airport	station	ave.					
clean	98.8	99.0	98.8	99.1	98.9					
20dB	89.2	95.8	90.1	94.4	92.4					
15dB	74.4	88.3	76.9	83.6	80.8					
10dB	52.7	66.7	53.2	59.6	58.1					
5dB	29.6	38.2	30.7	29.7	32.0					
0dB	11.7	18.7	15.8	12.3	14.6					
-5dB	5.0	10.1	8.1	8.5	7.9					
	HMM 合成 15dB (10 sec)					HMM 合成 5,10,15,20,∞dB マルチパス (10 sec)				
	restaurant	street	airport	station	ave.	restaurant	street	airport	station	ave.
clean	93.9	96.2	97.0	99.4	96.6	97.9	96.9	98.8	97.7	97.8
20dB	92.0	95.8	97.0	93.8	94.6	95.7	96.5	97.7	96.5	96.6
15dB	94.9	98.3	89.8	92.4	93.8	92.7	97.8	94.4	92.6	94.4
10dB	77.2	85.7	78.4	79.2	80.1	93.0	94.2	93.8	92.4	93.3
5dB	30.0	35.2	32.4	34.4	33.0	91.9	85.3	89.4	88.8	88.9
0dB	13.2	15.2	10.3	11.6	12.6	50.6	55.6	43.7	64.2	53.4
-5dB	3.0	9.2	7.4	8.5	7.03	20.3	16.2	17.6	19.2	18.3
	Noise GMM-HMM 合成 15dB (1 sec)					Noise GMM-HMM 合成 5,10,15,20,∞ dB マルチパス (1 sec)				
	restaurant	street	airport	station	ave.	restaurant	street	airport	station	ave.
clean	92.4	94.4	91.1	94.2	93.0	96.9	97.2	96.2	97.4	96.9
20dB	91.9	97.2	87.0	93.8	92.5	97.2	95.8	93.2	95.8	95.5
15dB	93.7	97.3	93.7	93.6	94.6	93.8	96.7	92.2	95.4	94.5
10dB	72.5	86.6	78.3	81.6	79.8	89.6	96.6	90.0	92.5	92.1
5dB	25.6	39.3	26.4	32.8	31.0	82.5	87.3	81.5	88.5	84.9
0dB	6.1	16.9	8.5	5.2	9.2	21.9	51.2	37.6	47.1	39.4
-5dB	1.8	1.1	1.1	2.9	1.7	1.3	3.3	2.2	3.3	2.5

puter Speech and Languege, vol. 9, pp. 171-185, 1995

[6] 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

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

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

[9] 奥田浩三, 松井知子, 中村哲, 音節強調発声に頑健な自然発話音声の認識法, 信学技報, SP2000-98, pp. 19–24, 2000 年 12 月

[10] Tetsuya Takiguchi, Satoshi Nakamura, Kiyohiro Shikano, HMM-Separation-Based Speech Recognition for a Distant Moving Speaker, IEEE Trans. Speech Audio Processing, Vol. 9, No. 2, pp.127–140, 2001 年 2 月

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