TR-S-0006

展示会場の実データを用いた音声認識性能評価実験

An Evaluation of Speech Recognition Performance in Real Exhibition Environments

西川 憲一郎 Ken'ichiro Nishikawa
伊田 政樹 Masaki Ida

2000.9.8

要旨

音声認識システムを実環境下で動作させる場合、周囲の雑音が入力音声に混入することによる認識性能の低下が避けられない。この対雑音ロバスト性について種々の手法が研究されているが、一般的にシミュレーション実験による評価が多用されており、実環境において有効に機能するか検証する必要がある。本稿では、その一手法である HMM 合成について検討する。実験は展示会場において実際に収録した音声データを用いて行った。

目次

1　はじめに

近年の計算機技術の発展により音声認識技術は実用段階に達した。実際に使用する環境においては周囲に雑音が存在し、認識性能低下の原因となる。そのため、雑音の混入した音声入力に対しても安定した認識手法が求められ、様々な手法が研究されている。しかし、これらの研究においては一般的に実験室内あるいは計算機上のシミュレーション評価が多用されており、実環境においても同様に有効であるか検証する必要がある。本稿では、雑音混入対策の一つとしてHMM合成を用いた音響モデルの適応化について検討する。実験は実際の展示会場において収録した雑音データを用いてHMM合成を行い、同じ会場で収録した音声を用いて評価した。

2　HMMによる統計的音声認識

本実験を行うに当たり、使用した音声認識システムに関する理論について述べる。

2．1　音声認識とは

入力音声パターンをIフレームの時系列$X = x_1, x_2, \cdots, x_I$とし、それに対し最もよくマッチングする単語列$W = w_1, w_2, \cdots, w_N$を見つけ出すことである。

すなわち、

$$P(W \mid X) = P(w_1, w_2, \cdots, w_N \mid x_1, x_2, \cdots, x_I)$$

を最大にする単語列Wを見つけ出す問題となる。

ここでP(W|X)は、

$$P(W \mid X) = \frac{P(X \mid W) \cdot P(W)}{P(X)}$$

と書ける。P(X)は入力パターン自身の生起確率であるので、単語列によらない。

よって、音声認識の問題は、P(X|W)・P(W)を最大化するWを求める問題となる。

P(W)は単語列の事前確率であり、入力パターンXとは無関係な確率である。

この単語列の確率は、

$$P(W) = P(w_1, w_2, \cdots, w_N) = P(w_1) \prod_{i=2}^{N} P(w_i \mid w_{i-1} \cdots w_1)$$

として与えられ、（統計的）言語モデルと呼ばれる。

以上より、

$$P(X \mid W) = P(x_1, x_2, \cdots, x_I \mid w_1, w_2, \cdots, w_N)$$

$$= \sum_{1, <I_2 < \cdots < I_{31}} P(x_1, x_2, \cdots, x_{I_1} \mid w_1) P(x_{I_1+1}, x_{I_1+2}, \cdots, x_{I_2} \mid w_2)$$

となり、

任意の音声パターンの部分系列$x_i x_{i+1} \cdots x_j$と単語w_kに対して、

$P(x_i x_{i+1} \cdots x_j \mid w_k)$が計算できれば原理的に計算できることになる。

この$p(x_i \cdots x_j \mid w_k)$の計算にHMM（隠れマルコフモデル）を用いる。

HMMによる音声認識においては音声単位として音韻をとる。いったん音韻のモデルが構築できれば、音韻モデルをつなぎあわせて、任意の単語や単語列や文章を構成することができる。

音声波形は、時間とともに、音韻とともに変化している。そして音声波形全体は、非定常的な信号であるが、局所的には比較的定常な信号であると見ることができる。HMMは、このような定常な信号をつなぎあわせて、非定常な信号を表現するのに適した統計的信号源モデルである。

HMMは、遷移する状態の集まりとして表され、確率としては、状態の遷移の確率を表す遷移確率と、状態が遷移する時に観測ベクトルの確率を出力する出力確率とからなる。出力確率は、音声のHMMによるモデル化において、スペクトル距離尺度に対応している。これらにより、任意の音韻列の出力される確率、すなわち、もっともらしさを算出する。

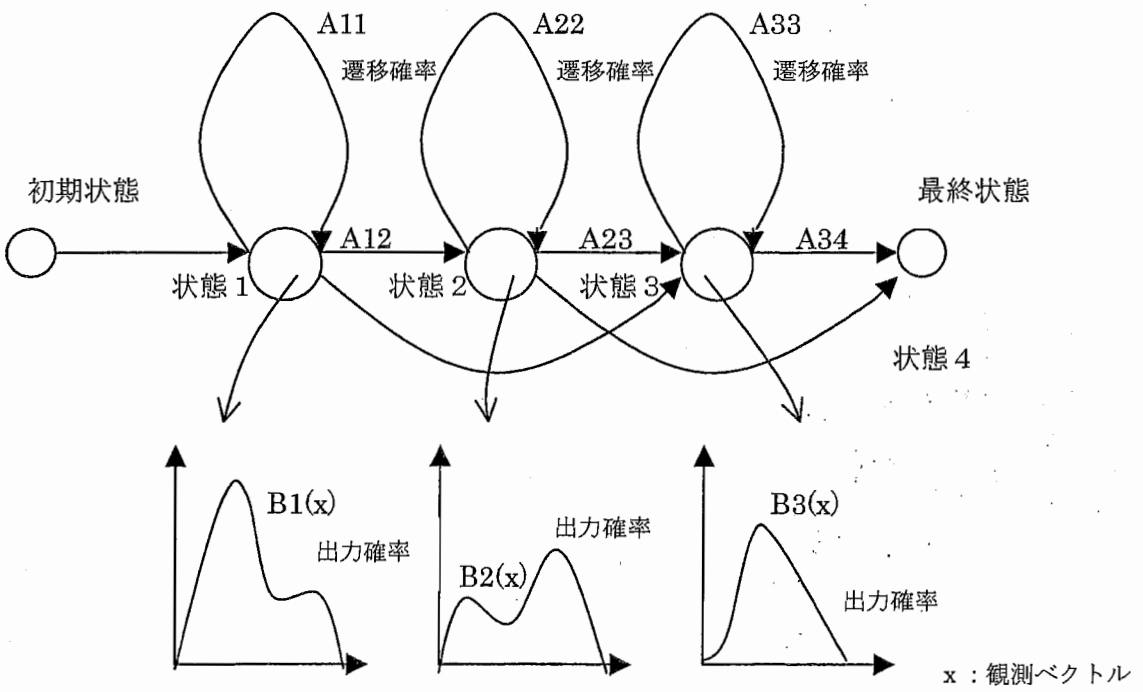

図1　音韻HMMの構造

２．２　　HMM 合成法による雑音の適応

HMM 合成とは、クリーン音声を用いて学習された HMM と雑音を用いて学習された HMM を合成し、雑音重畳音声に対する HMM を作成する方法である。それぞれの HMM の状態は確率分布を有するので、その構造はそれぞれの HMM の直積で定義される。遷移確率は、対応する遷移確率の積で求められる。状態の雑音モデルを用いた場合を図２に示す。

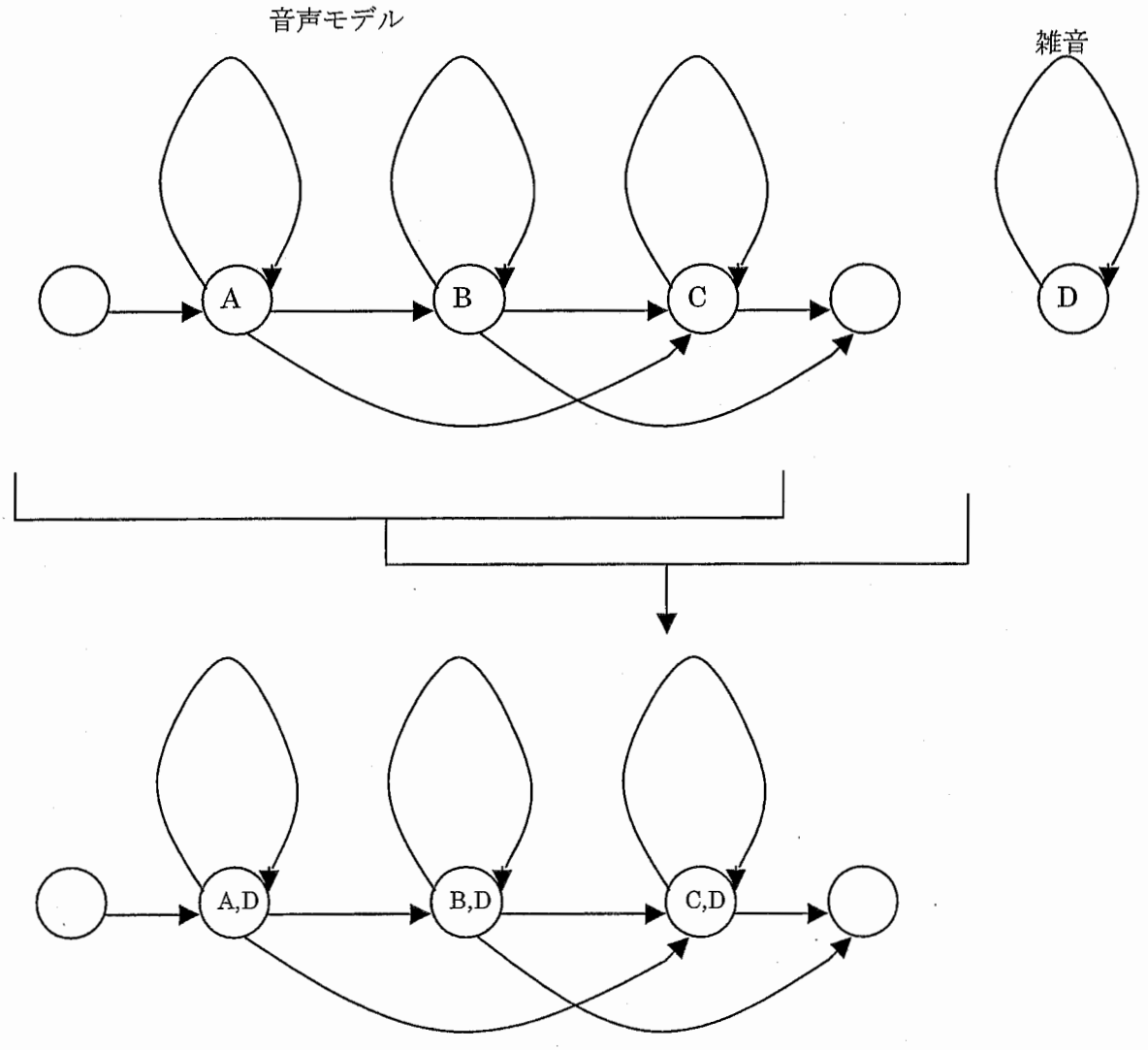

図２　　合成 HMM の構造

合成HMMの出力確率分布O_{cep}は、音声の出力分布Sと雑音の出力確率分布Nを結合することで求められる。各出力分布を単一正規分布とすると、正規分布は再生性が保証されるため確率変数の和は分布の畳み込みとなり、合成された確率密度関数は各密度関数の平均、分散の和によって$O_{cep}:\{\mu_s+\mu_n,\Sigma_s+\Sigma_n\}$のように与えられる。一般に、音声と周囲の雑音は線形スペクトル領域において加法性が成り立つが、音声認識では特徴量がケプストラムで表現されているため、これらの特徴量にコサイン変換および指数変換を行って、線形スペクトル領域に変換し、この結合（出力確率の計算）を行う必要がある。

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

但し、Fは離散フーリエ変換を示す。

ここでkはSNRに応じた雑音重畳音声とのレベル調整係数である。一般に、信号レベルは学習データとテストデータで異なるので、SN比を調節する必要があり、この係数としてkを導入する。このkは、無音区間において雑音のパワーを推定する方法や直接最尤推定することで求められる。

以下、処理の手順を示す。

① ケプストラムを用いて音声と雑音のHMMを推定する。
② 各HMMの平均と分散を以下の式を用いてコサイン変換し、対数パワースペクトル領域に変換する。

$$\mu_{\log}=\Gamma\mu \qquad \text{（2）}$$

$$\Sigma_{\log}=\Gamma\Sigma\Gamma^{T} \qquad \text{（3）}$$

但し、Γ：コサイン変換行列、　μ：ケプストラム上での平均値
Σ：ケプストラム上の共分散行列

$_{\log}$：対数パワースペクトル上であることを示す

③ 指数変換を行うことにより線形スペクトル領域に変換する。正規分布に従う確率変数を指数変換した確率変数 $Z = \exp^{y}$ は、対数正規分布に従う。

１，２次モーメントを合わせることにより対数正規分布の平均、分散は次のように与えられる。

$$\mu_{(linear),i} = \exp\{\mu_{(\log),i} + \frac{\sigma^2{}_{(\log),ii}}{2}\} \qquad \text{（4）}$$

$$\sigma^2{}_{(linear),i,j} = \mu_{(\log),i}\mu_{(\log),j}\{\exp(\sigma^2{}_{(\log),i,j} - 1)\} \qquad \text{（5）}$$

但し、$_{(linear)}$ は、線形パワースペクトル上であることを示す

④ 確率変数が独立と仮定し、式（１）にしたがって２つの分布を合成する。

$$\mu_{(lin\_comp)} = \mu_{(lin\_speech)} + k \cdot \mu_{(lin\_noise)} \qquad \text{（6）}$$

$$\Sigma_{(lin\_comp)} = \Sigma_{(lin\_speech)} + k \cdot \Sigma_{(lin\_noise)} \qquad \text{（7）}$$

但し、$_{(lin\_speech)}$：線形スペクトル上の音声の分布

$_{(lin\_noise)}$：線形スペクトル上の雑音の分布

$_{(lin\_comp)}$：合成した分布　　であることを示している

⑤ 分布を対数変換する

$$\mu_{(\log\_comp),i} = \log \mu_{(lin\_comp),i} - \frac{1}{2}\{\frac{(\sigma_{(lin\_comp),ij})^2}{(\mu_{(lin\_comp),i})^2} + 1\}$$

$$\sigma^2{}_{(\log\_comp),ij} = \log\{\frac{\sigma^2{}_{(lin\_comp),i,j}}{\mu_{(lin\_comp),i}\mu_{(lin\_comp),j}} + 1\}$$

⑥ 逆フーリエ変換を行って、ケプストラム領域に変換する。

$$\mu = \Gamma^{-1}\mu_{(\log\_comp)}$$

$$\Sigma = \Gamma^{-1}\Sigma_{(\log\_comp)}(\Gamma^{-1})^{T}$$

この分布による HMM を用いればケプストラム上で表現された雑音混じりの音声信号を高精度に認識できる。

これらの処理を図示すると、図３のようになる。

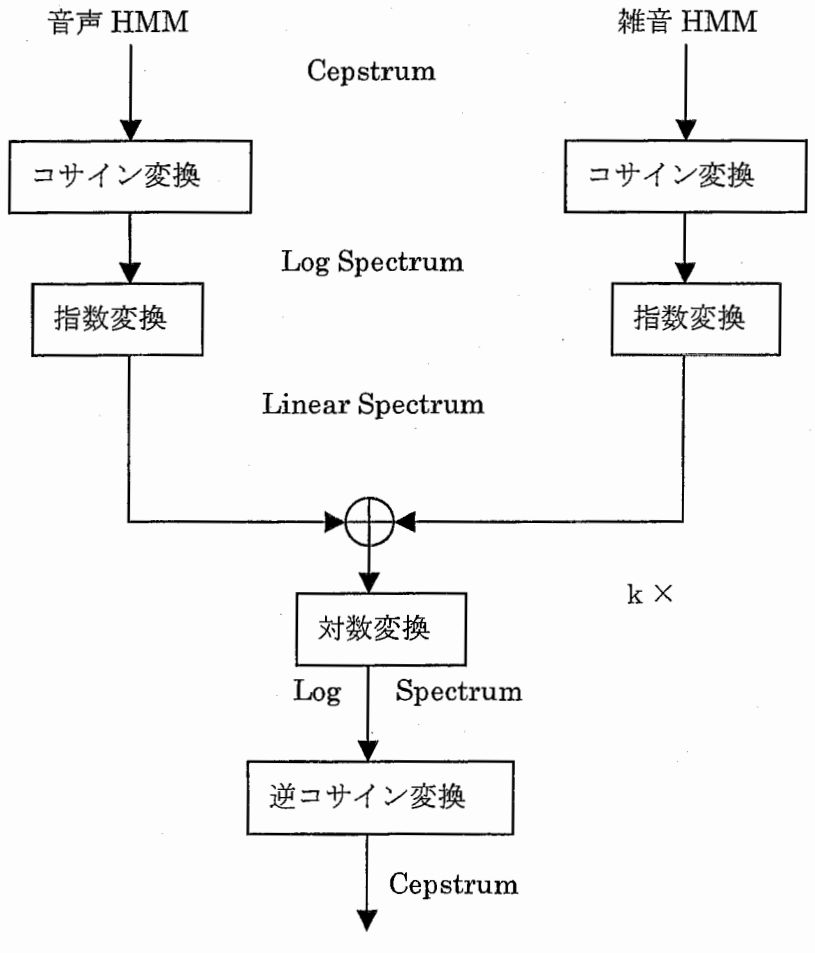

図３　　出力確率の合成アルゴリズム

3．評価用音声を用いた音響モデルの評価

HMM 合成により雑音対策を施した音響モデルの性能評価を行う。

3．1　実験条件

認識タスクは、旅行対話タスクを用いた。雑音を含まない評価セット（ clean speech ）と実際の展示会場の雑音の混入した評価セット（ noisy speech ）を用い、比較する。

評価用音声の発声内容は、ATR 旅行対話音声評価セット．４２話者である。clean speech セットは、防音室において収録された。noisy speech セットは、実際の展示会場において、clean speech を再生、再収録して雑音の混入した評価セットを作成した。再生・再収録には、ユーザが使用している環境を忠実に再現するため、ダミーヘッドを用いた。実際の展示会場として、ATR が出展した「２１世紀夢の技術展」（以下、夢テク、２０００年7月２１日～8月６日東京ビッグサイトにて開催）において、雑音の混入した音声データ、および雑音データの収録を行った。

音響分析条件を表１に示す。

表１　音声データ分析条件

サンプリング周波数	16 kHz
特徴分析	25 次元 (12MFCC+12ΔMFCC+Δpower)
CMS	なし
量子化	16 bit
分析フレーム長	20 msec
フレームシフト	10 msec

認識実験に用いた音響モデルの構成を表２に示す。

この構成条件で、いくつかの音響モデルを作成し、比較評価する。この作成法について、表３に示す。

表２　認識性能評価に用いた音響モデルの構成

26 phoneme context dependent (tri-phone)	
gender dependent	
総状態数	１４０３状態
総分布数	７０３０分布
各状態の混合分布数	１０または５
各音韻モデルの最大状態数	４

表３　各音響モデルの作成法

clean モデル	防音室で収録された ATR 対話音声DB４０７話者の学習データを用いて学習
重畳モデル	ATR 対話音声 DB４０７話者に、SN 比１５ｄB となるよう振幅調整した電子協騒音DB の展示会雑音を重畳したものを用いて学習
電子協騒音 HMM 合成モデル	SN 比１５ｄB となるよう、振幅調整した電子協騒音 DB の展示会場雑音１０秒間を用いて、１状態１混合の雑音モデルで学習し、これを clean モデルとを HMM 合成したもの
夢テク雑音 HMM 合成モデル	夢テク会場において収録した実際の騒音データ１０秒間を用いて、１状態１混合の雑音モデルを学習し、これを clean モデルとを HMM 合成したもの。

３．２　　実験結果

３．２節の音響モデルと、３．３節の評価用音声データを用いて、音声認識実験を行った。評価には Word Accuracy（単語認識率）を用いた。
結果を表５に示す。

表５　　音声認識率

音響モデル	clean speech	noisy speech
clean モデル	85.3	54.9
重畳モデル		72.9
電子協 HMM 合成モデル		56.2
夢テク HMM 合成モデル		58.6

３．３　　考察

・ noisy 音声データ × 重畳モデル

clean 音声データ × clean モデルの認識率より１０％以上低い認識率であった。この原因は、重畳モデルを作成した環境と入力音声環境に不適合が存在したためと考えられる。

重畳モデルは、事前に準備しておく必要があるので、入力音声環境の予測が難しい場合には認識率の低下が避けられない。

・ noisy 音声データ × 電子協 HMM 合成モデル

重畳も出る作成に用いた雑音と同じ雑音で HMM 合成法を用いた場合、認識率が約１５％低下した。

これは、重畳モデルに比べて、雑音データのデータ量や、モデルの表現力において劣るためと考えられる。

・ noisy 音声データ × 夢テク HMM 合成モデル

HMM 合成法は、重畳モデルに比べて、比較的高速にモデルの適応化ができる。したがって、現地の騒音データを使用することも可能である。現地で収録した騒音データを用いて、HMM 合成を行うことで、わずかな性能向上が見られた。

性能向上が２％程度にとどまった原因として、音声収録時に混入した雑音と雑音モデル学習に用いた雑音が同じ夢テク会場で異なる日時に収録された雑音、すなわち多少異なる雑音である点が考えられる。この問題を解消する一つの方法として、入力される音声に対して随時音響モデルを適応化していく方法が考えられる。

また、音響モデルを作成する際に使用した雑音データが１０秒と少なかったことと、雑音データを学習する際の状態数・混合数などの表現力が弱かったことが挙げられる。これらの原因により、混入する雑音の時間的な変動に対する耐性が弱くなったと考えられる。

以上より、今日採用していた１状態１混合の雑音モデルを用いる HMM 合成では、実際の展示会場における音声認識実験においては効果が小さかった。今後、先に挙げた対策により、HMM 合成による性能向上を目指す。

4 まとめ

実際の展示会場で収録された音声を用いて、HMM 合成を用いた適応化音響モデルの評価を行った。

結果、１状態１混合の簡単な雑音モデルを用いる方式では、効果はわずかにとどまった。