# 残響下音声認識における HMM 分離法の改良

**滝口 哲也　　西村 雅史**

日本アイ・ビー・エム　東京基礎研究所
〒 242-8502 神奈川県大和市下鶴間 1623 番 14
E-mail: {takigu, nisimura}@jp.ibm.com

**あらまし**

ユーザがマイクロフォンから離れて音声を入力するハンズフリー音声認識では、残響などの影響を受けて認識精度が劣化する。これまでに室内を移動するユーザの音声認識を実現するために、HMM 分離・合成法を提案し、その有効性を示してきた。HMM 分離法では、観測音声信号の尤度最大化基準により音響伝達特性の推定を行うが、インパルス応答が長くなると、その効果が低下していた [1]。そこで、従来単一ガウス分布で表現していた音響伝達特性モデルを、混合ガウス分布へ拡張することを検討してみた。認識実験の結果、本手法の有効性が得られたので、その推定方法及び実験結果について本稿にて報告する。

**キーワード** ハンズフリー音声認識、残響、モデル適応、HMM 分離

# An Improvement of HMM Separation for Reverberant Speech Recognition

Tetsuya Takiguchi and Masafumi Nishimura

IBM Research Tokyo Research Laboratory
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502
E-mail: {takigu, nisimura}@jp.ibm.com

**Abstract**

In distant-talking speech recognition, the recognition accuracy is seriously degraded by the influence of reverberation and environmental noise. We have proposed a robust speech recognition technique for acoustic model adaptation based on HMM separation and composition methods, which realizes recognition of the distant moving speaker. In HMM separation, the model parameters of the acoustic transfer function are estimated by maximizing the likelihood of adaptation data uttered from an unknown position. However, the improvement was smaller than expected for the impulse response with long reverberations. This paper investigates modeling of the acoustic transfer function based on the Gaussian mixture components. The results of experiments clarify the effectiveness of the proposed method.

**key words** hands-free speech recognition, reverberation, model adaptation, HMM separation

# 1 はじめに

現在、会議などの書き起こし、ロボットとの対話などハンズフリーでの音声認識機能を使用するタスクに関する要求が存在する。しかしながら、現状のシステムではユーザがマイクロフォンから離れて発話すると、入力音声は周囲雑音及び残響の影響を受けて認識性能が劣化してしまう。またデスクトップマイクロフォンやピンマイクロフォンを用いた場合でも、ユーザが横を向くと音響伝達特性の影響により音声がひずみ、認識性能が劣化する場合がある。

従来、音声の伝達経路による影響に対処する方法として、ケプストラム平均減算法 (Cepstrum Mean Subtraction: CMS) などが使われている。この手法は、例えば電話回線の影響などのように、伝達特性のインパルス応答が比較的短い場合には有効であるが、室内にてマイクロフォンから離れて発話した際には、残響の影響を受けて十分な性能が得られない。これは、一般に室内の残響の伝達特性の長さが、音声認識に用いられる短区間分析の窓幅よりも長くなるためである。このような残響成分を除去する方法として、長い分析窓と短い分析窓を組み合わせる方法 [2] が提案されているが、残響成分の除去と同時に音声がひずむ可能性がある。複数のマイクロフォンを利用し、逆フィルタを設計して観測信号から残響成分を除去する方法 [3] も提案されているが、音響伝達特性のインパルス応答が、最小位相とならない場合があり逆フィルタの設計は難しい。また使用環境下においてコストや物理的な配置状況により、複数のマイクロフォンを設置できない場合がある。その他、隣接する分析フレーム間の関係を考慮した残響成分補正についての検討も行われている [4]。

これまでに、室内を移動するユーザの音声認識を実現するために、HMM 分離・合成法を提案し、その有効性を示してきた [1]。HMM 分離法では、あらかじめインパルス応答を測定しておく必要はなく、観測音声信号の尤度最大化基準により音響伝達特性の推定を行う。インパルス応答が長い場合、分析フレーム毎での影響にばらつきが生じるが、これまで音響伝達特性モデルの分散にて対処することを検討してきた。しかしながら、まだ十分な精度が得られてはなく、より複雑なモデル化を検討する必要があった。そこで、従来単一ガウス分布で表現してきた音響伝達特性モデルを、混合ガウス分布へ拡張することを検討してみた。本稿では、その推定方法及び混合ガウス分布の有効性について報告する。

# 2 残響環境下での音声認識

残響環境下での観測信号 $o(t)$ は以下のように表現される。

$$o(t) = \sum_{l=0}^{L} s(t-l) \cdot h(l) \qquad (1)$$

ここで $s(t)$ はクリーン音声、$h(l)$ はインパルス応答（残響特性）、$L$ はインパルス応答長とする。今、観測信号の短区間スペクトルを以下の式で近似する。

$$O(\omega; n) \approx S(\omega; n) \cdot H(\omega) \qquad (2)$$

ここで、$\omega$ は周波数、$n$ はフレーム番号を表す。HMM 合成法は、加算条件の成立する領域において適用されるので、式 (2) を次のように書き換える。

$$O_{cep}(c; n) \approx S_{cep}(c; n) + H_{cep}(c) \qquad (3)$$

$O_{cep}(c; n)$、$S_{cep}(c; n)$、$H_{cep}(c)$ はそれぞれ観測信号、クリーン音声、音響伝達特性のケフレンシー $c$ におけるケプストラムを表している。従って、合成 HMM の出力確率分布は以下の式により求めることができる。ただし、本稿では音響伝達特性のモデルも混合ガウス分布で表現する (図 1)。

$$\mu^{(O)}_{p,j,k} = \mu^{(S)}_{p,j,m} + \mu^{(H)}_{q} \qquad (4)$$

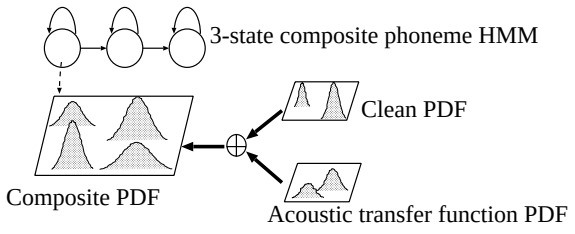


図 1: 音響伝達特性との合成例。音響伝達特性も混合ガウス分布にて表現される。

$$\Sigma^{(O)}_{p,j,k} = \Sigma^{(S)}_{p,j,m} + \Sigma^{(H)}_{q} \tag{5}$$

ここで、$(\mu^{(S)}_{p,j,m}, \Sigma^{(S)}_{p,j,m})$、$(\mu^{(H)}_{q}, \Sigma^{(H)}_{q})$、$(\mu^{(O)}_{p,j,k}, \Sigma^{(O)}_{p,j,k})$ はそれぞれクリーン音声、音響伝達特性、合成 HMM の平均ベクトル、共分散行列である。また音素数 $P$、状態数 $J$、混合数 $K$(合成 HMM)、$M$(クリーン音声)、$Q$(音響伝達特性)である。混合数 $K$ は、クリーン音声の混合数 $M$ と音響伝達特性の混合数 $Q$ の積で与えられる。

$$K = M \cdot Q \tag{6}$$

従って、変数 $k$ は、$m$ と $q$ の組み合わせによって得られる。

$$k = (m, q) \tag{7}$$

$$m = 1, \ldots, M, \quad q = 1, \ldots, Q \tag{8}$$

次に、音響伝達特性モデルの推定方法について述べる。

## 3 HMM 分離による音響伝達特性の推定

音響伝達特性を EM アルゴリズムを使い最尤推定により求める。

$$\begin{aligned} \hat{\lambda}_H &= \underset{\lambda_H}{\operatorname{argmax}} \Pr(O|\lambda_H, \lambda_S) \\ &\triangleq \lambda_O \ominus \lambda_S \end{aligned} \tag{9}$$

ここで、$\lambda$ はモデルパラメータの集合を表し、HMM 分離を $\ominus$ で定義する。式 (9) は、$\lambda_O$ から $\lambda_H$ の分離を行うことを意味する。$\lambda_O$ は観測信号のモデルパラメータの集合であり、このように HMM 分離法では、観測信号の統計量をいったん計算してから、EM アルゴリズムにより音響伝達特性の推定を行う。以下に処理の流れを示す（図 2）。

1) クリーン音声 HMM $\lambda_S$ と音響伝達特性モデル $\lambda_H$ との合成 HMM の作成（$\lambda_H$ の初期値として、平均と分散を 0 とした）。

$$\lambda_O = \lambda_S \oplus \lambda_H \tag{10}$$

2) $\hat{\lambda}_O$ の推定。

$$\hat{\lambda}_O = \underset{\lambda_O}{\operatorname{argmax}} \Pr(O|\lambda_O) \tag{11}$$

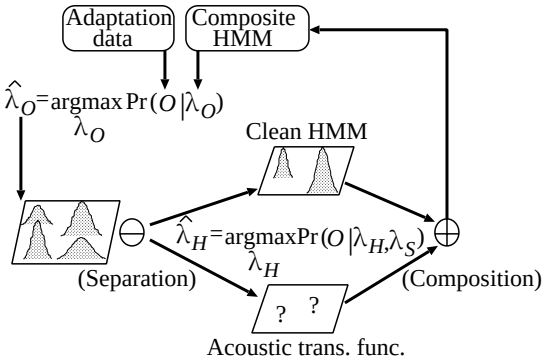


図 2: 音響伝達特性の推定

3) $\hat{\lambda}_O$ から音響伝達特性 $\hat{\lambda}_H$ の分離。

$$\hat{\lambda}_H = \hat{\lambda}_O \ominus \lambda_S \tag{12}$$

4) 処理 1) に戻り、適応データに対する合成モデルの尤度が収束するまで処理を行う。

ここで式 (10) は、式 (4) と (5) により求める。また $\hat{\lambda}_O$ の推定は、

$$\hat{\mu}^{(O)}_{p,j,k} = \sum_{v}^{W^{(p)}} \sum_{n}^{N^{(p,v)}} \gamma_{p,v,j,k,n} O_{p,v,n} \Big/ \gamma_{p,j,k} \tag{13}$$

$$\hat{\Sigma}^{(O)}_{p,j,k} = \textstyle\sum_v \sum_n \gamma_{p,v,j,k,n} (O_{p,v,n} - \hat{\mu}^{(O)}_{p,j,k})^2 \Big/ \gamma_{p,j,k} \tag{14}$$

$$\gamma_{p,v,j,k,n} = \Pr(O_{p,v,n}, j, k|\lambda_O) \tag{15}$$

により行う。ここで音素 $p$ は $W^{(p)}$ 個の適応データをもち、音素 $p$ に関連する $v$ 番目の観測系列の長さを $N^{(p,v)}$ とする。

次に、式 (12) の HMM 分離の詳細について述べる。HMM 分離では、EM アルゴリズムによる最尤推定によりパラメータ推定を行う。まず、Expectation step で以下の関数を定義する。

$$\begin{aligned} &Q(\hat{\lambda}_H|\lambda_H) \\ &= E\Big[\log \Pr(O, S, C|\hat{\lambda}_H, \lambda_S) \Big| O, \lambda_H, \lambda_S\Big] \end{aligned} \tag{16}$$

ここで、観測系列を $O$、それに対応する状態系列を $S$、混合要素系列を $C$ とする。今、Q 関数

にて分布重み $\hat{w}$ のみに注目すると、

$$
\begin{aligned}
&Q_{\hat{w}}(\hat{\lambda}_H|\lambda_H) \\
&= \sum_{p}^{P}\sum_{v}^{W^{(p)}}\sum_{j}^{J^{(p)}}\sum_{k}^{K^{(p)}}\sum_{n}^{N^{(p,v)}} \gamma_{p,v,j,k,n}\log w^{(O)}_{p,j,k} \\
&= \sum_{p}\sum_{v}\sum_{j}\sum_{k}\sum_{n}\gamma_{p,v,j,k,n} \\
&\quad \cdot\left\{\log w^{(S)}_{p,j,m} + \log \hat{w}^{(H)}_{q}\right\} \qquad (17)
\end{aligned}
$$

となる。式 (17) を最大 (Maximization step) にする $q$ 番目の分布重み $\hat{w}_q$ は、ラグランジュ未定乗数法を用いて以下のように求められる。

$$
\begin{aligned}
\hat{w}^{(H)}_{q} &= \frac{\sum_p\sum_v\sum_j\sum_{k'_m}\gamma_{p,v,j,k'_m}}{\sum_p\sum_v\sum_j\sum_k\gamma_{p,v,j,k}} \\
&= \frac{\sum_p\sum_v\sum_j\gamma_{p,v,j,q}}{\sum_p\sum_v\sum_j\sum_k\gamma_{p,v,j,k}} \qquad (18)
\end{aligned}
$$

$$
k'_m = (m,q), \quad m = 1,\ldots,M \qquad (19)
$$

ここで、$k'_m$ は、音響伝達特性の $q$ 番目の分布とクリーン音声の $M$ 個の分布の合成によって得られる分布番号とする。

次に、平均と分散の推定式を求める。まず、出力確率分布に関する項に注目した Q 関数は以下のようになる。

$$
\begin{aligned}
&Q_{\hat{\mu},\hat{\Sigma}}(\hat{\lambda}_H|\lambda_H) \\
&= -\sum_p\sum_v\sum_j\sum_k\sum_n\gamma_{p,v,j,k,n} \\
&\quad\cdot\left\{\tfrac{1}{2}\log(2\pi)^D\left(\Sigma^{(S)}_{p,j,m}+\hat{\Sigma}^{(H)}_{q}\right)\right. \\
&\quad\left.+\frac{(O_{p,v,n}-\mu^{(S)}_{p,j,m}-\hat{\mu}^{(H)}_{q})'(O_{p,v,n}-\mu^{(S)}_{p,j,m}-\hat{\mu}^{(H)}_{q})}{2(\Sigma^{(S)}_{p,j,m}+\hat{\Sigma}^{(H)}_{q})}\right\} \qquad (20)
\end{aligned}
$$

式 (20) を最大にする分散を直接求めるのは困難なので、$q$ 番目の確率分布の EM アルゴリズムにおける変化量を $(\Delta\hat{\mu}^{(H)}_q, \Delta\hat{\Sigma}^{(H)}_q)$ とする。

$$
\hat{\mu}^{(H)}_q = \mu^{(H)}_q + \Delta\hat{\mu}^{(H)}_q \qquad (21)
$$

$$
\hat{\Sigma}^{(H)}_q = \Sigma^{(H)}_q + \Delta\hat{\Sigma}^{(H)}_q \qquad (22)
$$

これらの変化量に関して推定式を求める [1]。従って、$\partial Q(\hat{\lambda}_H|\lambda_H)/\partial\Delta\hat{\mu}^{(H)}_q = 0$ より、

$$
\begin{aligned}
&\Delta\hat{\mu}^{(H)}_q \\
&= \frac{\sum_p\sum_v\sum_j\sum_{k'_m}\sum_n\gamma_{p,v,j,k'_m,n}\frac{O_{p,v,n}-\mu^{(S)}_{p,j,m}-\mu^{(H)}_q}{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q}}{\sum_p\sum_j\sum_{k'_m}\frac{\gamma_{p,j,k'_m}}{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q}} \\
&= \frac{\sum_p\sum_j\sum_{k'_m}\gamma_{p,j,k'_m}\frac{\hat{\mu}^{(O)}_{p,j,k'_m}-\mu^{(S)}_{p,j,m}-\mu^{(H)}_q}{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q}}{\sum_p\sum_j\sum_{k'_m}\frac{\gamma_{p,j,k'_m}}{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q}} \qquad (23)
\end{aligned}
$$

また分散に関しては、$\partial Q(\hat{\lambda}_H|\lambda_H)/\partial\Delta\hat{\Sigma}^{(H)}_q = 0$ より、

$$
\begin{aligned}
&\sum_p\sum_j\sum_{k'_m}\gamma_{p,j,k'_m} \\
&\quad\cdot\frac{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q+\Delta\hat{\Sigma}^{(H)}_q-\phi_{p,j,k'_m}}{\left(\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q+\Delta\hat{\Sigma}^{(H)}_q\right)^2} = 0 \qquad (24)
\end{aligned}
$$

$$
\begin{aligned}
\phi_{p,j,k'_m} =& \hat{\Sigma}^{(O)}_{p,j,k'_m} + \hat{\mu}^{2(O)}_{p,j,k'_m} + (\mu^{(S)}_{p,j,m}+\hat{\mu}^{(H)}_q) \\
&\cdot(\mu^{(S)}_{p,j,m}+\hat{\mu}^{(H)}_q - 2\hat{\mu}^{(O)}_{p,j,k'_m}) \qquad (25)
\end{aligned}
$$

ここで、以下のように関数 $F$ を定義する。

$$
F(\Delta\hat{\Sigma}^{(H)}_q) = \frac{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q+\Delta\hat{\Sigma}^{(H)}_q-\phi_{p,j,k'_m}}{\left(\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q+\Delta\hat{\Sigma}^{(H)}_q\right)^2} \qquad (26)
$$

この式を原点におけるテイラー展開を行い、一次の項まで求める。

$$
\begin{aligned}
&F(\Delta\hat{\Sigma}^{(H)}_q) \\
&\approx F(0) + \left.\frac{\partial F(\Delta\hat{\Sigma}^{(H)}_q)}{\partial\Delta\hat{\Sigma}^{(H)}_q}\right|_{\Delta\hat{\Sigma}^{(H)}_q=0}\times\Delta\hat{\Sigma}^{(H)}_q \\
&= \frac{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q-\phi_{p,j,k'_m}}{(\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q)^2} - \frac{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q-2\phi_k}{(\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q)^3}\Delta\hat{\Sigma}^{(H)}_q
\end{aligned}
$$

ここで、EM アルゴリズムにより $\Delta\hat{\Sigma}^{(H)}_q$ は 0 に収束する。従って分散の推定式は以下のようになる。

$$
\begin{aligned}
&\Delta\hat{\Sigma}^{(H)}_q \\
&= \frac{\sum_p\sum_j\sum_{k'_m}\gamma_{p,j,k'_m}\left\{\frac{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q-\phi_{p,j,k'_m}}{(\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q)^2}\right\}}{\sum_p\sum_j\sum_{k'_m}\gamma_{p,j,k'_m}\left\{\frac{\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q-2\phi_{p,j,k'_m}}{(\Sigma^{(S)}_{p,j,m}+\Sigma^{(H)}_q)^3}\right\}} \qquad (27)
\end{aligned}
$$

# 4 認識実験

## 4.1 実験条件

残響下音声を作成するために、RWCP 実環境音声･音響データベースより残響時間300ms のインパルス応答を使用した。収録されているデータは、10 度方向から 170 度まで 20 度きざみで 9 方向である（マイクまでの距離は 2m)。これらのインパルス応答と ATR 音声データベースのクリーン音声と畳み込みを行い、テストデータと適応データを作成した。タスクは語彙 500 単語として、テストデータは男性話者一人が対象語彙を一回発声したものである。特定話者 HMM(54 音素) を使用して認識実験を行う。クリーン音声 HMM は 3 状態 3 ループ、各状態が 4 混合ガウス分布とした。

## 4.2 実験結果

図 3 に HMM 分離・合成による認識結果を示す。目的話者方向は正面 90 度方向、適応データ数は 10 単語である。クリーン音声 HMM での認識率は 54.8%、一方 HMM 分離・合成により平均と分散を適応させた場合 (HMM-sepa.(Mean, Cov))、認識率は 89.8%まで改善した。ただし、この時の音響伝達特性の混合数は 1 である。更に音響伝達特性の混合数を増やすことにより、認識率は 95.2%（混合数 5）まで改善した。この結果より、分析フレーム毎における残響の影響のばらつきが大きい場合、混合数を増やし、より複雑なモデル化を行うことにより、認識率の改善が可能であることが分かる。また、平均値のみを適応した場合 (HMM-sepa.(Mean)) と比べて、分散も適応することにより認識率の改善が得られている。CMS と比較してみたところ、CMS での認識率は 86%となり、十分な改善が得られていない。ここでの CMS は、テストデータ毎（一単語毎）にケプストラム平均値を計算している。図中の “Matched model” は、インパルス応答とクリーン音声と畳み込みを行い作成した学習データ（2620 単語）を用いて、再学習した音響モデルで認識した結果である。この結果と比べると HMM 分離・合成による推定精度は劣っているが、音響伝達特性の混合数を増やすことにより、その差が改善されているのが分かる。

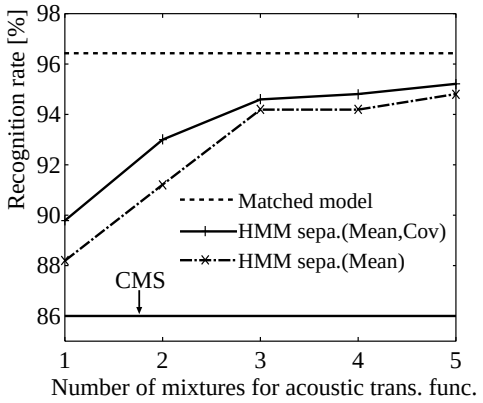


図 3: HMM 分離・合成による認識結果

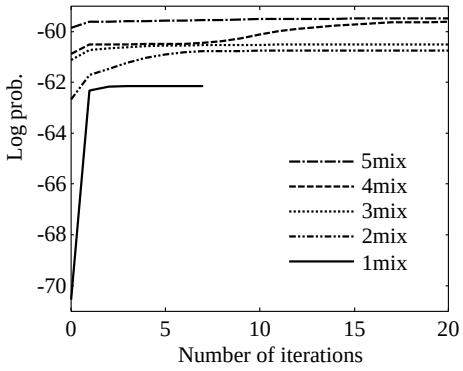


図 4: HMM 分離法の収束性

図 4 に音響伝達特性の混合数が 1 から 5 までの平均対数尤度とアルゴリズムの反復回数を示す（平均のみ適応)。今回、音響伝達特性のガウス数を増やす際、一つ少ない混合数のモデルを初期値として使用している。図より、数回くらいの繰り返しで尤度が収束しているのが分かる。

適応単語数と認識率の関係について図 5 に示す（音響伝達特性の混合数 5)。今回のタスクでは適応データが数単語でも十分な改善が得られた。

次に、話者方向が未知の場合についての結果を示す。話者方向推定は、各々の方向に対応する合成モデルを使い、一単語毎に尤度最大化基準により行う。

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \Pr(O|\lambda_{O_\theta}) \tag{28}$$

ここで、$\theta$ は 10 度から 170 度までの 9 方向とした。図 6 に、合成モデルとして 256 混合の GMM

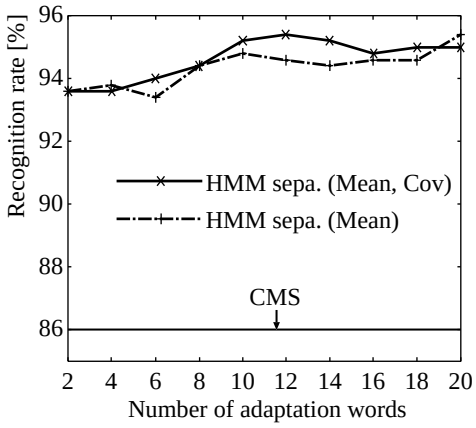


図 5: 適応単語数と認識率の関係

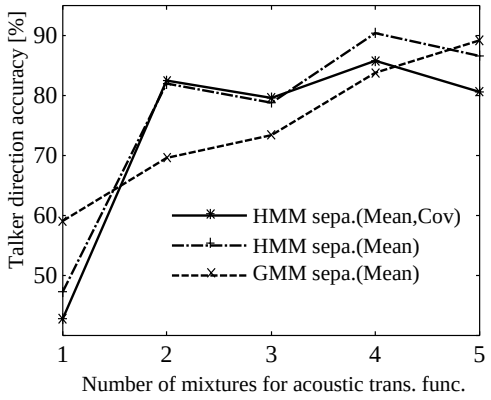


図 6: 話者方向の推定結果

を利用した場合と、認識時に使用する HMM を利用した場合の結果について示す（適応単語数は 10 とした)。また GMM の結果は平均のみ適応、HMM の結果は平均のみと、平均と分散の適応の場合である。いずれの場合にも、音響伝達特性の混合数を増やすことにより、推定精度が改善されている。GMM と HMM の結果を比較すると、音響伝達特性の混合数が 4 以上では、精度の差はあまりなかった。従って、位置推定に関しては GMM を利用することにより、計算量削減が期待できる。この推定方向をもとに、音声認識を行った結果を図 7 に示す。音響伝達特性の混合数が 1 の場合には、方向既知の場合と比べて約 2%近くの認識率の差がある。しかし混合数を増やすことにより、その差がほとんど無くなっているのが分かる。

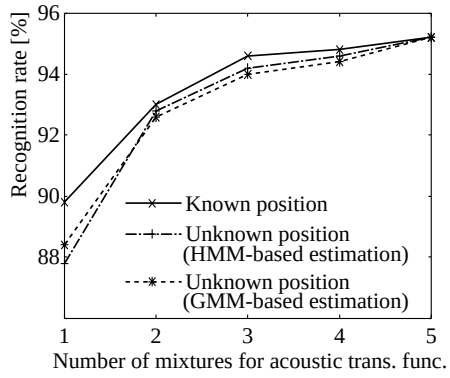


図 7: 音源方向未知の音声認識結果

## 5 まとめ

室内にて話者がマイクロフォンから離れて発話する際には、残響の影響を受けて認識精度が劣化する。発話者からマイクロフォンまでのインパルス応答が長い場合、分析フレーム毎での影響にばらつきが生じ、従来のように単一ガウス分布による音響伝達特性のモデル化では十分な認識精度が得られていなかった。このばらつきに対処する一つの方法として、発話者からマイクロフォンまでの音声伝達経路による影響を、混合ガウス分布により表現することを本稿にて試みた。認識実験の結果、インパルス応答既知とした場合 (matched model) の認識結果に近づけることがわかり、提案手法の有効性が示せた。今後は、雑音環境下における本手法の検討、またフレーム間の関係を考慮した手法の検討などを行っていく。

## 参考文献

[1] T. Takiguchi, S. Nakamura, and K. Shikano, “HMM-Separation-Based Speech Recognition for a Distant Moving Speaker,” IEEE Trans. on SAP, Vol.9, No.2, 2001.

[2] C. Avendano, S. Tivrewala, and H. Hermansky, “Multiresolution channel normalization for ASR in reverberant environments,” Eurospeech, pp. 1107-1110, 1997.

[3] M. Miyoshi and Y. Kaneda, “Inverse Filtering of room acoustics,” IEEE Trans. on ASSP, Vol.36, No.2, 1988.

[4] 杉村、滝口、中村、鹿野、“フレーム間の関係を考慮した残響音声認識の検討”、音講論 3-Q-5、Mar. 1999。