社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音響モデルを利用したシングルチャネルによる音源方向推定

住田　雄司†　　滝口　哲也†　　有木　康雄†

† 神戸大学工学部　〒 657-8501 神戸市灘区六甲台町 1-1
E-mail: †yuji@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　本稿では，音響モデルを利用することにより，単一マイクロホンで音源方向を推定する方法を検討する．あらかじめクリーン音声の音響モデルを作成しておき，各方向から到来する数単語の音声を用いて，EM アルゴリズムに基づきクリーン音声モデルと音響伝達特性の分離を行う．また本稿では，音響伝達特性のモデルとして GMM（Gaussian Mixture Model）を用いる事により，短時間分析における音響伝達特性のばらつきの影響に対処する方法も検討する．

キーワード　音源方向推定，シングルチャネル，音響モデル，最尤推定

Single-channel voice localization using acoustic model

Yuji SUMIDA†, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Faculty of Engineering, Kobe University
1-1 Rokkodai, Nada, Kobe, 657-8501 Japan
E-mail: †yuji@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　This paper presents a voice localization method using only a single microphone, where the GMM (Gaussian Mixture Model) of clean speech is introduced to estimate the acoustic transfer function from any user's position. The sequence of the acoustic transfer function is estimated by maximizing the likelihood of train data (only several words) uttered from an unknown position, where the cepstral parameters are used due to effectively represent useful clean speech information. Using the sequence data of the acoustic transfer function, the GMM of the acoustic transfer function is created to deal with the influence of a long impulse response. Its effectiveness is confirmed by voice (talker) direction experiments in a room environment.

Key words　voice localization, single-channel, acoustic model, ML estimation

1. ま え が き

音源方向を検出するために，これまで様々な手法が提案されてきた．マイクロホンアレーによる方向推定法として，MUSIC 法や CSP 法といった手法が提案されている [1] [2]．また，相互相関関数を基にした各方法について，室内残響下実験により比較した報告もされている [3]．CSP 法における改善手法として，方向推定に用いる信号を音声と仮定して，音声の主要な周波数成分を重視する帯域分割型 CSP 法などがあり，従来の方法より良い結果が得られている [4]．

しかし，これらの方法に総じていえることは，到来信号の時間差・強度差といった情報を用いており，複数のマイクロホンという条件が必要不可欠となっている．

そこで，本稿では，音響モデルを利用することにより，時間差という情報を用いずに単一マイクロホンで音源方向を推定する方法を提案する．単一マイクロホンで方向を推定することができれば，複数の場合と比較して，コストを削減することができ，またマイクロホンの設置も容易であるといった様々な利点があるといえる．

提案手法では，あらかじめクリーン音声の音響モデルを作成しておき，各方向から到来する数単語の音声から EM アルゴリズムを用いることにより，音響伝達特性を推定する．これにより得られた音響伝達特性の時系列データから，各方向における音響伝達特性モデルを作成する．そして実際の入力音声から同様にして音響伝達特性を推定し，これらのモデルとの尤度を求めることで方向の決定を行う．

2. 音響伝達特性の推定

2.1 残響音声のケプストラム表現

図 1 に示すように，ある場所で発声されたクリーンな音声 s は，音響伝達特性 h の影響を受ける．このとき，観測信号 o は以下のように表現される．

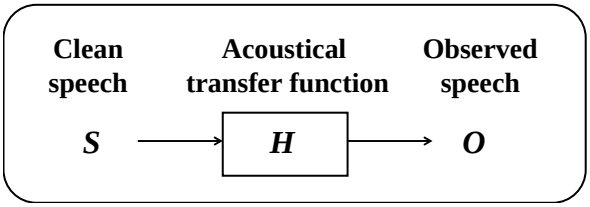

図 1　対象とする環境のモデル

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

ここで，L はインパルス応答長とする．今，観測信号の短時間スペクトルを以下の式で近似する．

$$O(\omega; n) \approx H(\omega)S(\omega; n) \tag{2}$$

ここで，ω は周波数，n はフレーム番号を表す．(2) 式の両辺の対数をとると，次式のように加算の形で表すことができる．

$$\log O(\omega; n) \approx \log H(\omega) + \log S(\omega; n) \tag{3}$$

(3) 式に，逆フーリエ変換を適用する事によりケプストラムが得られる．

$$O_{cep}(i; n) \approx H_{cep}(i) + S(i; n) \tag{4}$$

ここで，i はケプストラムの次元を表す．ケプストラムは，音声情報を効率よく表現できるパラメータの一つであり，音声認識ではよく使われる．本稿では，このケプストラム領域にてクリーン音声モデルを作成する．

(4) 式より，O と S が分かれば H を推定することができる．しかし S を観測することはできないので，本稿では，S の代わりにクリーン音声モデルを用い，ケプストラム領域にて尤度最大基準に基づいて O から H を分離する．

2.2　EM アルゴリズムによる音響伝達特性の推定

(4) 式の音響伝達特性の時系列データを，観測信号に対して，そのモデルの尤度が最大となるようにして求める．

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|\lambda_S, H) \tag{5}$$

ここで，λ はモデルパラメータ（GMM: Gaussian Mixture Model）の集合を表し，添え字の S はケプストラム領域におけるクリーン音声を表す．

EM アルゴリズムは，2 段階の繰り返し処理となる．まず最初に expectation step にて，以下の Q 関数を計算する [5].

$$\begin{aligned} &Q(\hat{H}|H) \\ &= E[\log \Pr(O, b, c|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \sum_b \sum_c \frac{\Pr(O, b, c|H, \lambda_S)}{\Pr(O|H, \lambda_S)} \cdot \log \Pr(O, b, c|\hat{H}, \lambda_S) \end{aligned} \tag{6}$$

ここで，b と c は，各々状態系列と混合要素系列を表す．時系列データ O, b, c の同時確率は，以下の式により計算することが出来る．

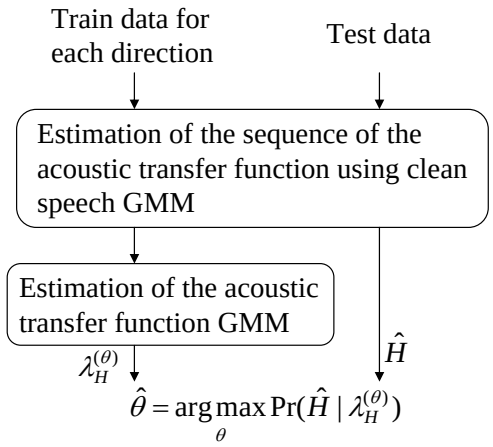

図 2　音源方向推定の概要

$$\Pr(O, b, c|\hat{H}, \lambda_S) = \prod_n a_{b_{n-1}, b_n} w_{b_n, c_n} \Pr(O_n|\hat{H}_n, \lambda_S) \tag{7}$$

ここで，a は状態遷移確率，w は混合重み，O_n はケプストラム領域での n 番目の観測系列を表している．(4) 式で示されているように，音響伝達特性はケプストラム領域における加算性雑音としてみなす事が出来る．従って，観測信号のモデルの平均値は，クリーン音声の平均値に音響伝達特性 H を加算する事により得られる．よって，(7) 式は以下のように書き換える事が出来る．

$$\begin{aligned} &\Pr(O, b, c|\hat{H}, \lambda_S) \\ &= \prod_n a_{b_{n-1}, b_n} w_{b_n, c_n} \cdot N(O_n; \mu_{b_n, c_n} + \hat{H}_n,\ \Sigma_{b_n, c_n}) \end{aligned} \tag{8}$$

ここで，$N(O; \mu, \Sigma)$ はクリーン音声の多次元正規分布を表す．最終的に，Q 関数は以下のように導出される [6].

$$\begin{aligned} &Q(\hat{H}|H) \\ &= \sum_i \sum_j \sum_n \Pr(O_n, b_n = j, b_{n-1} = i|\lambda_S) \log a_{i,j} \\ &\quad + \sum_j \sum_k \sum_n \Pr(O_n, b_n = j, c_n = k|\lambda_S) \log w_{j,k} \\ &\quad + \sum_j \sum_k \sum_n \Pr(O_n, b_n = j, c_n = k|\lambda_S) \\ &\quad \cdot \log N(O_n; \mu_{j,k} + \hat{H}_n, \Sigma_{j,k}) \end{aligned} \tag{9}$$

ここで，H に関する項のみに注目して，Q 関数を以下のように書き換える．

$$\begin{aligned} &Q(\hat{H}|H) \\ &= \sum_j \sum_k \sum_n \Pr(O_n, b_n = j, c_n = k|\lambda_S) \\ &\quad \cdot \log N(O_n; \mu_{j,k} + \hat{H}_n, \Sigma_{j,k}) \\ &= -\sum_j \sum_k \sum_n \gamma_{j,k,n} \sum_{d=1}^{D} \Big[\frac{1}{2} \log(2\pi)^D \sigma_{j,k,d}^2 \\ &\quad + \frac{(O_{n,d} - \mu_{j,k,d} - \hat{H}_{n,d})^2}{2\sigma_{j,k,d}^2} \Big] \end{aligned} \tag{10}$$

表 1　実 験 条 件

サンプリング周波数	12 kHz
窓関数	Hamming
Frame length	32 msec
Frame shift	8 msec
特徴量	MFCC(order 16)

$$\gamma_{j,k,n} = \Pr(O_n, j, k | H, \lambda_S) \tag{11}$$

ここで，D は特徴量の次元数である．次に，maximization step にて，この Q 関数を最大にする $\hat{H}$ を求める．これは $\hat{H}$ について偏微分して解くことにより求める事が出来る．

$$\hat{H}_{n,d} = \frac{\sum_j \sum_k \gamma_{j,k,n} \frac{O_{n,d} - \mu_{j,k,d}}{\sigma^2_{j,k,d}}}{\sum_j \sum_k \frac{\gamma_{j,k,n}}{\sigma^2_{j,k,d}}} \tag{12}$$

2.3　音響伝達特性 GMM の学習

次に，各 θ 方向に対応する音響伝達特性の GMM を求める．まず，(12) 式において，θ 方向から発声した数単語の観測信号 O を用いて，$\hat{H}_n$ を推定する．推定された $\hat{H}_n$ を用いて，EM アルゴリズムに基づき，θ 方向に対応する音響伝達特性を求める．

$$\mu_m^{(H)} = \frac{\sum_v \sum_{n^{(v)}} \gamma_{m,n^{(v)}} \hat{H}_{n^{(v)}}}{\sum_v \sum_{n^{(v)}} \gamma_{m,n^{(v)}}} \tag{13}$$

ここで，v は v 番目の学習データを表し，$n^{(v)}$ は v 番目の学習データにおける n 番目のフレームを表す．また，$\mu_m^{(H)}$ は音響伝達特性 GMM の m 番目の平均値ベクトルを表す．同様にして分散も求める事が出来る．

$$\Sigma_m^{(H)} = \frac{\sum_v \sum_{n^{(v)}} \gamma_{m,n^{(v)}} (\hat{H}_{n^{(v)}} - \mu_m^{(H)})^T (\hat{H}_{n^{(v)}} - \mu_m^{(H)})}{\sum_v \sum_{n^{(v)}} \gamma_{m,n^{(v)}}} \tag{14}$$

2.4　尤度基準に基づいた方向推定

各評価発話毎に，(12) 式により $\hat{H}$ を推定し，尤度最大基準に基づき方向推定を行う．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \Pr(\hat{H} | \lambda_H^{(\theta)}) \tag{15}$$

ここで，$\lambda_H^{(\theta)}$ は，2.3 節で構築した θ 方向に対応する音響伝達特性 GMM である．

図 2 に音源方向推定の概要を示す．各方向に対応する音響伝達特性の時系列データが，(12) 式より計算され，更に (13) 式と (14) 式を用いて音響伝達特性の GMM （$\lambda_H^{(\theta)}$）が構築される．得られた音響伝達特性 GMM を用いて，各評価発話毎に推定された $\hat{H}$ に対して尤度を計算し，最大尤度を与えるモデルを正解方向とする．

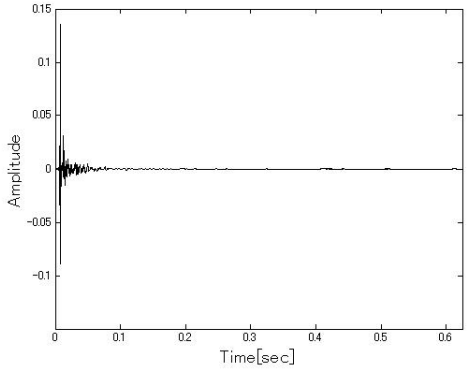

図 3　90 °からのインパルス応答

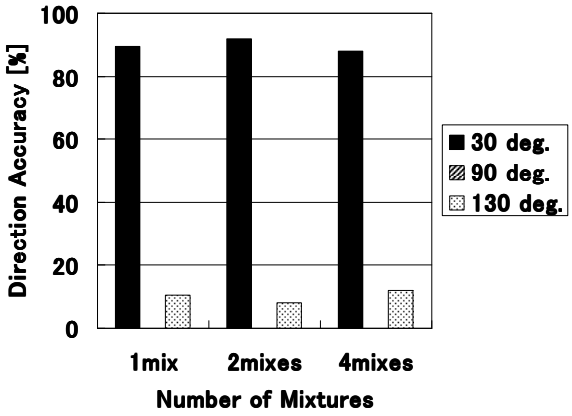

図 4　30 °入力における方向識別率

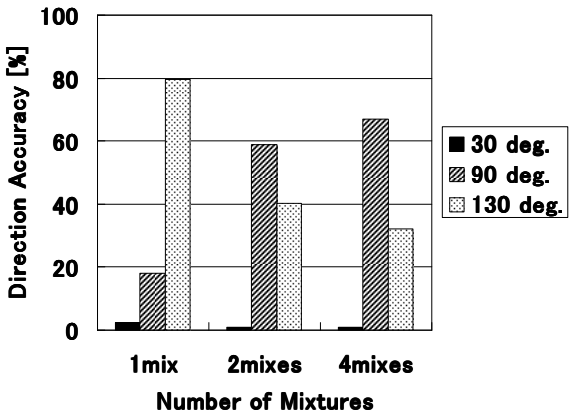

図 5　90 °入力における方向識別率

3.　評 価 実 験

提案手法を評価するために実験を行った．実験条件を表 1 に示す．

音声が到来する方向は，30 °，90 °，130 °のいずれかであるとする．実験環境は，音源とマイクロホンの距離が 2 m，残響時間が 300 msec の環境で収録されたインパルス応答を畳み込むことにより，残響環境をシミュレーションした．例として，90 °におけるインパルス応答を図 3 に示す．このインパルス応答は RWCP 実環境音声・音響データベースを用いた [7].

音声データとして，ATR 音声データベースより男性話者 1

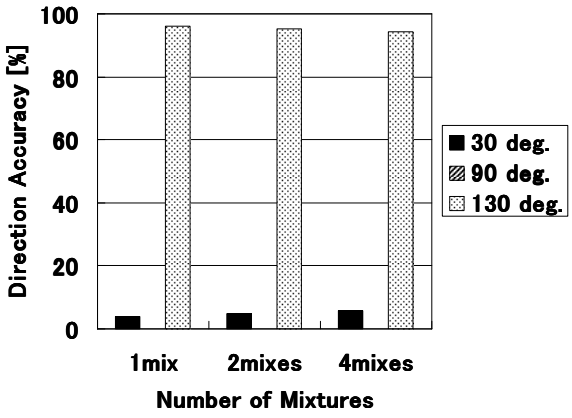

図 6　130 °入力における方向識別率

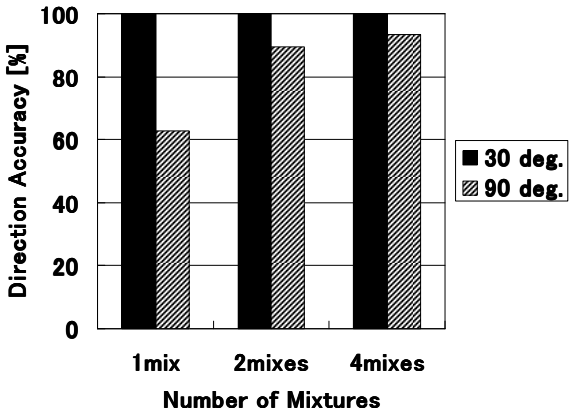

図 7　30 °と 90 °における方向識別率

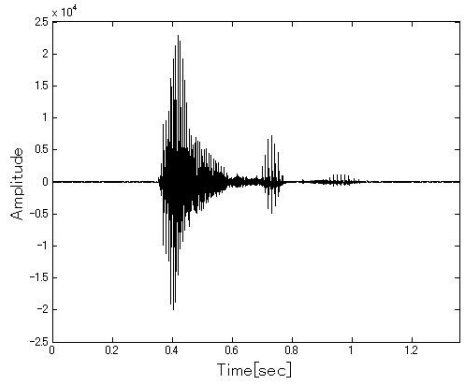

図 8　“a-i-sa-tsu” の波形データ

名，クリーン音声の学習データには 2620 単語，音響伝達特性を推定するためのデータには 10 単語，評価用のデータには 1000 単語を用いた．音響モデルとして，クリーン音声モデルは GMM（64 混合）で作成し，音響伝達特性モデルは GMM の混合数を変化させて比較を行った．

3 方向における実験結果について述べる．30 °，90 °，130 °の方向から音声が到来したときの方向推定結果をそれぞれ図 4〜6 に示す．

発話方向が 30 °，130 °の場合では混合数にかかわらず，それぞれ 90% ，95% 程度の正解率が出ており，正しく方向が推定できているといえる．しかし，90 °の場合では，単一の正規分布において，1000 単語のうちの 8 割が 130 °と誤識別されて

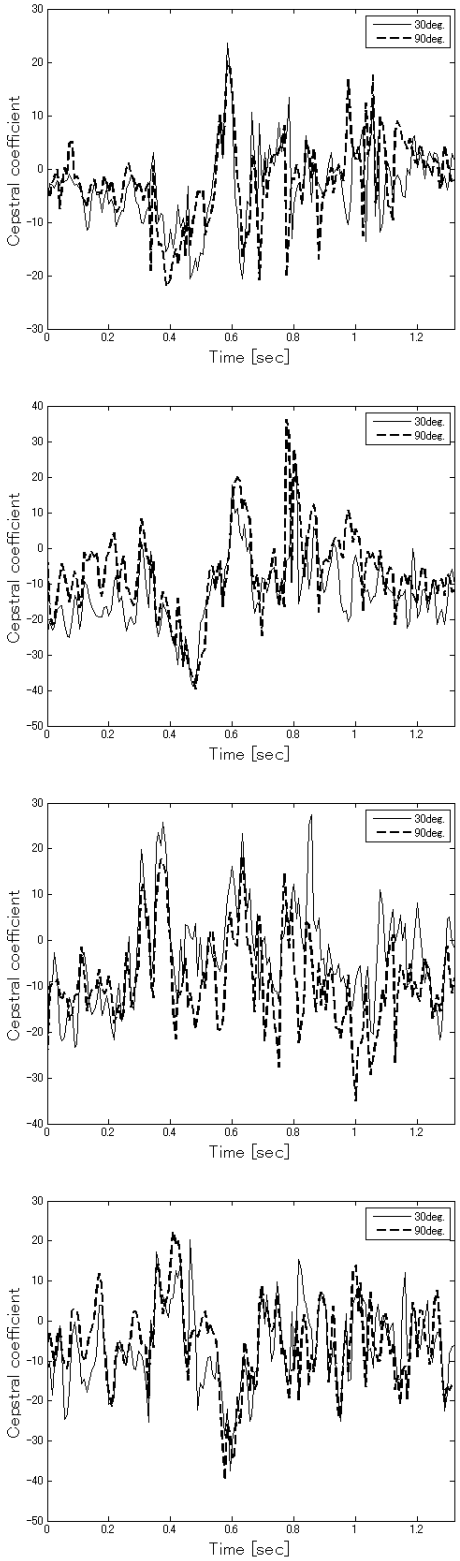

図 9　音響伝達特性の時間変化の様子．上からそれぞれ MFCC 2^{nd}, 4^{th}, 6^{th}, 8^{th} order.

おり，GMM の混合数を増やしても正確率は 60% 程度に留まる結果となった．

また，30 °，90 °の 2 方向に限定した正解率を図 7 に示す．30 °の正解率は混合数によらず 100% ，90 °においても混合数を増やすことにより 90% 程度まで正しく識別できていること

が確認できた．

各々の方向，混合数においても正しく識別された一例として，単語 “a-i-sa-tsu” の波形データを図 8 に示す．この音声が 30 °と 90 °の方向から到来したときに推定した音響伝達特性の時間変化の様子を図 9 に示す．各次元において，時間変化とともにほぼ同じ値をとっていることもあれば，全く違う値をとっていることもあり，方向によって音響伝達特性の違いが出ていることがわかる．

誤識別の原因としては，各方向における音響伝達特性の類似性，EM アルゴリズムにおける近似誤差や残響による過去の音声の重なりの影響があげられる．これらの誤差から値がばらつき，GMM の分散が増大することにより誤識別が起きたものと考えられる．

4. ま と め

音源方向の検出法として，音響モデルを利用した単一マイクロホンによる方向推定法について検討した．評価実験より，提案手法によって単一マイクロホンでも方向推定ができることを示した．

今後の課題としては，方向の間隔を狭くし，かつ数を増やすことにより対応する方位数を増やしていくこと，方向推定に他の手法を取り入れて，今回の手法と比較して精度の向上を目指していくことがあげられる．精度向上のための具体案として，指向性マイクロホンの導入，あるいは LDA を用いることにより特徴成分を強調することなどを検討している．

また，今回の実験は学習・評価を単語毎に行ったため，単語を文章に変えたときの識別率の変化も調べる必要がある．

謝辞

本研究の一部は，公益信託 小野音響学研究助成基金の助成により行われた．

文　　献

[1] 大賀寿郎, 山崎芳男, 金田豊, “音響システムとディジタル処理,” 電子情報通信学会, 1995.

[2] C.H.Knapp and G.C.Carter, “The Generalized Correlation Method for Estimation of Time Delay,” IEEE Trans. On Acoust., Speech and Signal Proc., ASSP-24, 4, pp.320-327, 1976.

[3] 上杉信敏, 金田豊, “音源方向検出法の室内残響下での性能評価について,” 音講論, 3-Q-4, pp.635-636, 2006-3

[4] 傳田遊亀, 西浦敬信, 河原英紀, 入野俊夫, “帯域分割型 CSP 法に基づく話者位置推定法の検討,” 情処研報, SLP-54-29, pp.169-174, 2003.

[5] A.Sankar and C-H.Lee, “A maximum-likelihood approach to stochastic matching for robust speech recognition,” IEEE Trans. Speech and Audio Processing, vol.4, no.3, pp.190-202, 1996.

[6] B.-H. Juang, “Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains,” AT&T Tech. J., Vol. 64, No. 6, pp. 1235-1249, 1985.

[7] RWCP 実環境音声・音響データベース, “http://tosa.mri.co.jp/sounddb/”