

2ch マイクによる CSP 係数の識別に基づく話者の頭部方向の推定*

高島遼一, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

人と人のコミュニケーション,あるいはロボットとのコミュニケーションにおいて,話者の頭部方向は聞き手にとって重要な手がかりの一つであり,我々は話し手の頭部の向きから「誰が話しているのか」だけでなく「誰に向かって話しているのか」という情報まで得ることができる.この「誰が誰に向かって話しているのか」という情報は,特に複数のユーザーが会話をしている状況において有効であり,会議システム,ロボット対話,雑談とシステム要求の判別など,様々なタスクにおいて利用できると期待される.

これまでに,マイクロホンアレーを用いて音源方向や位置を推定する研究が多くなされている.MUSIC (Multiple Signal Classification) や CSP (Cross-power Spectrum Phase) といった手法では,マイクロホンアレーで収録される観測信号間の位相差を用いて音源方向や位置を推定している [1, 2, 3, 4]. また,バイノーラル信号を用いて,両耳間の音圧差や時間差から音源方向を推定する手法についても研究されている [5, 6].

一方,話者の頭部方向の推定へ関心が向けられ出したのは比較的近年のことであり,いくつかの手法が提案されている [7, 8, 9, 10]. これらの手法は複数組のマイクロホンアレーからなるネットワークを用いており,従来の音源位置推定のアルゴリズムを拡張することで話者の頭部方向を推定している.文献 [7] で提案されている手法は,従来の音源位置推定法の一つである CSP 法をベースとした手法であり,従来の CSP の評価関数を,話者の頭部方向に依存する重み係数によって重み付けを行うことにより,話者の位置推定問題から頭部方向推定問題へ拡張している.文献 [8] では話者の頭部の方向ごとに変化する観測信号の音圧のパターンに着目しており,文献 [9] では CSP を用いた手法と音圧パターンを用いた手法の両方を提案し組み合わせることでさらなる精度の向上を示している.また,文献 [10] では各マイクロホンアレーから算出された音源方向推定結果を用いて作成されるヒストグラムから,話者の頭部方向を推定する手法を提案している.

しかしながらこれらの手法は複数のマイクロホンアレーを,ユーザーを囲むようにして部屋の壁などに設置する必要がある,システムが大規模になってしまふという欠点がある.我々はこれまで,観測された音声信号の音響伝達特性が,発話された位置によって異なるという点に着目して,位置毎に発話された音声から音響伝達特性を推定し,それらを識別することにより単一マイクロホンで音源位置を推定する方

法を提案してきた [11]. また,音響伝達特性は音源位置だけでなく,頭部方向にも依存するため,この手法を応用し,音響伝達特性の識別により音源位置と頭部方向を単一マイクロホンで推定する手法についても検討を行った.この手法についての詳細は本発表予稿集の同著者による別稿を参照されたい.この研究により,音響伝達特性が音源位置と頭部方向の識別に有効であることが示された.しかし,この手法では単一マイクで観測された信号から,音響伝達特性のみをブラインドで完全に推定しきれていないため,頭部方向の細かな変化は識別できていなかった.

そこで本稿では,2ch マイクロホンを用いてその CSP 係数を学習・識別することで,複数のマイクロホンアレーを使わずに音源の位置と頭部の方向を推定する手法について検討を行う.従来の音源位置推定や頭部方向推定の先行研究において,CSP 係数はそのピーク値のみが用いられていたが,本手法では,残響によって CSP 係数のピーク周辺の値も相関値が高くなる点に着目し,ピーク値以外の値も情報として用いた.また,CSP 係数は音声信号のパワースペクトルで正規化をしているため,以前に提案した手法に比べて発話内容にロバストであると期待できる.従来の頭部方向の推定法と異なり,本手法はあらかじめ音響伝達特性を学習しておく必要があるが,マイクの位置を任意の場所に設置することができるという利点がある.

2 音源位置と頭部方向の推定

CSP 法は GCC-PHAT (Generalized Cross-Correlation PHase Transform) アルゴリズムとも呼ばれる手法で,音源方向や音源位置推定によく用いられる手法の一つである.CSP 係数は二つのマイクロホンで観測された信号を白色化し,その相互相関によって求めることができる.以下の式では二つの信号のクロススペクトルを,それぞれの信号のパワースペクトルで正規化し,その逆フーリエ変換を取ることで CSP 係数を求めている.

$$CSP(\tau) = DFT^{-1} \left\{ \frac{DFT(o_l(t)) \cdot DFT^*(o_r(t))}{|DFT(o_l(t))| \cdot |DFT(o_r(t))|} \right\} \quad (1)$$

$o_l(t)$ と $o_r(t)$ はそれぞれ左のチャネルと右のチャネルで観測された信号で, τ はこれらの信号間の時間のずれを表す.従来の音源位置推定手法や,前節で述べた頭部方向推定の先行研究においては,この CSP 係数のピーク値を探索することにより,音源位置や頭部方向を推定している.これに対して本手法では,観測信

* Estimation of talker's head orientation based on discrimination of cross-power spectrum phase coefficients.
by Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe univ.)

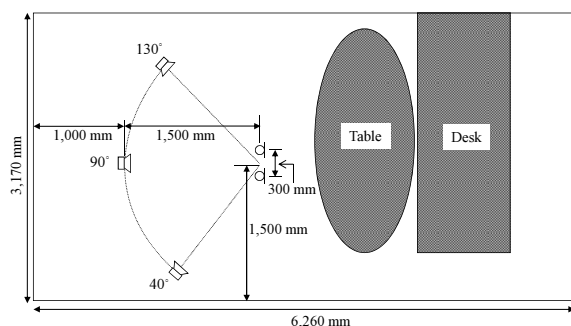


Fig. 1 実験環境 1

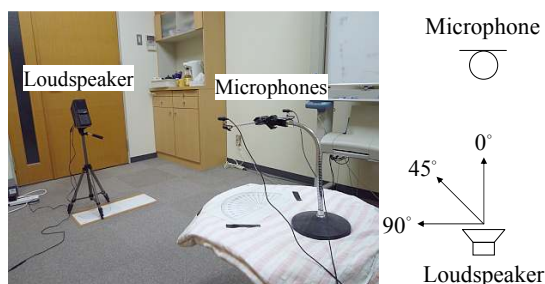


Fig. 2 実験環境 1 における収録の様子とスピーカ-の回転方向

号に含まれる残響により、音源位置に対応する CSP 係数のピーク値以外の点においてもある程度の相関エネルギーが存在すると考え、CSP 係数の全次元を残響を表す特徴量として用いた。

本手法の流れは、まずあらかじめ学習用の音声データを用意し、各音源位置・頭部方向についてその CSP 係数を算出し、これらを SVM (Support Vector Machine) により学習する。そして評価したい音声データについても同様に CSP 係数を算出し、これを識別することで音源位置・頭部方向を推定する。

3 評価実験

2 種類の収録環境にて実験を行い、本手法の有効性を評価した。実験環境 1 では別稿にて提案している単一マイクロホンによる音源位置・頭部方向推定法と比較を行い、実験環境 2 では本稿における提案法についてさらに詳細に分析を行った。

3.1 実験環境 1 における実験結果

実験環境 1 を Fig. 1 に、収録環境とスピーカ-の回転の様子を Fig. 2 に示す。約 6.3 m × 3.2 m × 2.8 m (W × D × H) の部屋において、ある位置にスピーカ-を設置し、スピーカ-の向きを変えながら特定話者の音声を再生し、これを 2ch マイクロホンで収録した。単一マイクによる手法ではこれらの内、一方のマイクロホンで収録された音声のみを用いた。部屋の残響時間は約 350 msec、マイクロホンとスピーカ-の距離は約 1.5 m である。スピーカ-は BOSE Mediamate II を、マイクロホンには指向性マイク (SONY ECM-66B) を使用した。音源位置の候

Table 1 実験環境 1 における、頭部方向毎の音源位置・頭部方向推定精度 [%]

head orientation	0°	45°	90°	mean
single-channel	68.3	32.1	80.2	60.2
2-channel	100.0	100.0	100.0	100.0

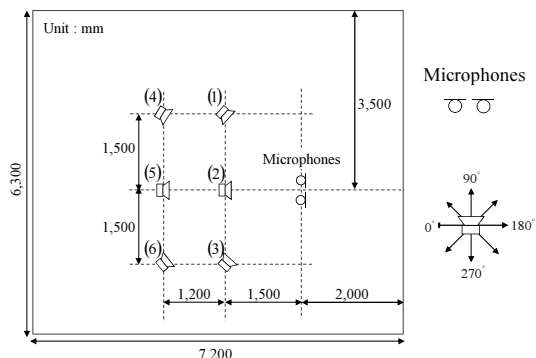


Fig. 3 実験環境 2. 括弧内の数字は音源位置のインデックスを表す。

補は 40°, 90°, 130° の 3 種類、スピーカ-の回転方向 (頭部方向) は 0°, 45°, 90° の 3 種類で、これらの組み合わせは計 9 通り存在する。

音声データは ATR 研究用日本語音声データベース セット A より男性話者 1 名の単語音声を用い、サンプリング周波数 12 kHz, 窓幅 32 msec, フレームシフト 8 msec の条件下でそれぞれの手法における特徴量を算出した。本稿における提案法では、各フレームに対して 512 次元の CSP 係数を求め、それらの平均をその単語の特徴量とした。各位置・頭部方向の学習には 50 単語を、評価には 166 単語を、組み合わせを変えて 4-fold のクロスバリデーションにより推定精度を算出した。位置及び頭部方向の学習と、評価に用いたデータはそれぞれ異なる発話内容の単語を使用している。SVM には SVM^{light} を、カーネル関数に RBF (Gaussian) カーネルを使用し、one-vs-rest 法によりマルチクラス識別を行った。

各手法における、頭部方向毎の音源位置・頭部方向推定結果 (9 クラス識別) を Table 1 に示す。単一マイクによる手法では頭部方向毎に認識率にばらつきがあり、特に頭部方向 45° の認識率が低い。一方本稿における提案法ではいずれの頭部方向においても 100 % の認識率で音源位置・頭部方向の推定が行えている。

3.2 実験環境 2 における実験結果

実験環境 2 とスピーカ-の回転方向を Fig. 3 に示す。部屋の広さは約 7.2 m × 6.3 m × 2.8 m (W × D × H) の部屋で、マイク間隔は環境 1 と同様 30 cm, 残響時間は約 1,220 msec であった。図中における括弧内の数字は音源位置の番号を表しており、位置の候補は 6 種類存在する。以降、位置についてはこの番号を用いて呼ぶことにする。回転方向は 45° 刻みの

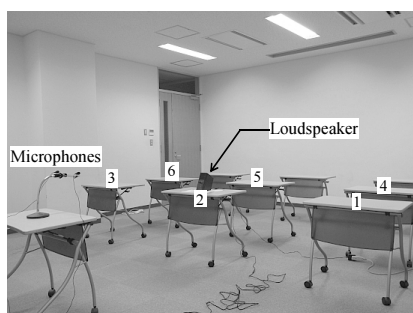


Fig. 4 実験環境 2 における収録の様子．数字は音源位置を表す．

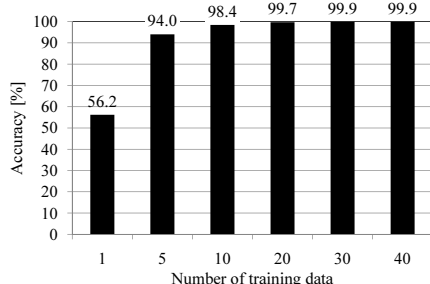


Fig. 5 学習データ数毎の音源位置と頭部方向の認識率

8 種類とし、スピーカーから見てマイクロホンの方向を 90° とする．位置と頭部方向の組み合わせとしては計 48 種類存在することになる．収録環境の写真を Fig. 4 に示す．

各位置、各頭部方向において 50 単語を収録し、内 10 単語を評価用のデータに用いる．残りの 40 単語は学習用のデータとし、これらから 1, 5, 10, 20, 30, 40 単語と学習データ数を変えて用いた．これらのテストを評価と学習のセットの組み合わせを変えて、5-fold のクロスバリデーションにより推定精度を算出した．合計のテストデータの数 は 2,400 (50×48) となる．その他の分析条件は実験環境 1 のものと同様である．

音源位置と頭部方向の組み合わせ全 48 クラスに対する学習データ数毎の識別結果を Fig. 5 に示す．図より、学習データが位置と頭部方向の組み合わせ毎に 5 単語以上あれば 9 割以上の認識精度が得られており、30 単語以上であればほぼ 100 % で認識が行っていた．

従来の音源位置推定の研究や頭部方向推定の先行研究において、CSP 法を用いる手法では CSP 係数のピーク値のみに着目していたのに対し、本手法では CSP 係数全次元を用いて識別を行っている．そこで、CSP 係数全次元の内、どの辺りの次元が識別に有効に働いているのかを調べるために、以下の実験を行った．Fig. 6 のように、CSP 係数のピークのみを用いた場合、ピーク値とその周囲 50, 100, 200, 300, 400 次元を用いた場合、512 次元全てを用いた場合で特徴ベクトルの次元数を変え、音源位置を 2 に限定してスピーカーの頭部方向の認識の評価を行った．なお、

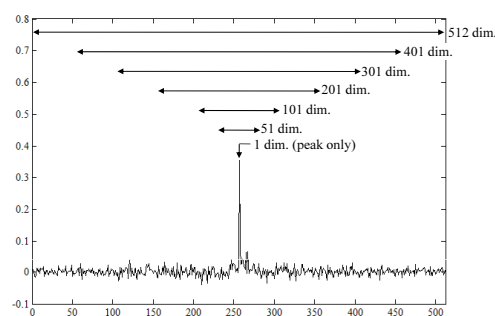


Fig. 6 特徴ベクトルの次元数と CSP 係数の使用範囲

Table 2 CSP 係数の使用する次元数 (dim.) とそれぞれにおける学習データ数 (num.) 毎の頭部方向の認識精度 [%]

dim. \ num.	1	5	10	20	30	40
1 (peak only)	22.0	24.3	24.5	26.3	26.8	22.5
51	72.3	70.8	78.8	95.8	93.0	93.0
101	82.5	94.5	97.5	87.8	97.5	95.5
201	91.3	92.8	96.5	99.5	99.5	99.0
301	92.5	94.0	95.8	99.3	99.3	99.5
401	92.5	98.3	95.3	99.3	99.5	99.5
512	91.8	97.8	95.5	99.5	99.5	99.5

この実験では位置を限定しているので頭部方向のみの 8 クラスの識別問題となる．

結果を Table 2 に示す．表より、ピークのみを用いた場合に対して、周囲 50 次元を加えることにより精度が大幅に向上しているのが分かる．さらに次元を増やすことで精度は向上するが、201 次元以上は大幅な精度の変化は見られなかった．また、ピークのみを用いた場合では、ほとんど頭部方向の識別が行えていないことが分かる．これは、ピーク値の大小のみで頭部方向を識別しようとした場合、 90° のようなピーク値の変化が顕著な頭部方向しか認識できていなかったためである．

以上のことから、先行研究では複数のマイクロホンアレーを設置し、各マイクロホンアレーにおける CSP 係数のピーク値を用いることで頭部方向を推定することができるが、2ch マイクロホンのみでは、ピーク値のみで頭部方向を識別することはできず、ピーク値以外の CSP 係数を用いることで頭部方向の識別が可能になると言える．

次に、収録したデータにピンクノイズを足し合わせて、雑音環境下での精度評価を行った．左右のマイクロホンで収録した音声データの内、テストデータのみと同じピンクノイズを SNR が 5 dB, 10 dB, 20 dB になるように足し合わせ、学習データ数を 40 単語として音源位置と頭部方向の 48 クラス識別を行った．結果を Fig. 7 に示す．図より、雑音を足すことによって精度が大幅に減少しており、SNR が 5 dB や 10 dB の場合においては 48 クラスをランダムに識別したときの期待値である 2.1 % とほとんど変わらない精度となっている．SNR が 20 dB の場合について

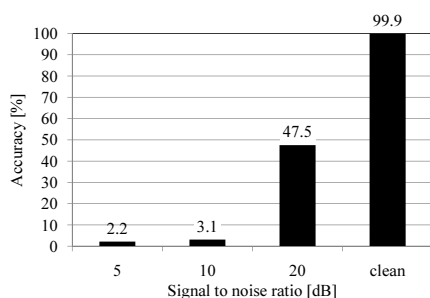


Fig. 7 雑音環境下での音源位置と頭部方向の認識率

音源位置を2に限定し，頭部方向のみの8クラス識別を行ったところ，認識率は14.8 %で，ほぼすべてのテストデータに対して頭部方向を90°と認識していた．

これは，両チャンネルで収録したデータに対して同じピンクノイズを足し合わせたことによって，CSP係数が位相差0の位置において高いピークを取ってしまったためである．雑音を足していないデータの頭部方向90°におけるCSP係数，0°におけるCSP係数，及び雑音を足したデータの0°におけるCSP係数をFig. 8に示す．図のように，0°におけるCSP係数が，雑音を足したことによって中央のピークが加算され，90°のCSP係数と似た形状になっていることが分かる．SNRが5 dBと10 dBにおいては，さらに中央のピークが高くなっており，位置2と5以外の本来中央にピークを持たない位置の音声データでも，位置2，頭部方向90°と識別され，認識精度が著しく低下する原因となった．

4 おわりに

本稿では，音声信号に畳みこまれる残響が話者の位置や頭部の回転方向によって異なる特性を持つ点に着目し，各音源位置・頭部方向における残響の性質をCSP係数で表現し，これをSVMにより学習・識別することで，音源の位置と頭部方向を2chマイクロホンで推定する手法について検討を行った．実環境における実験では，音源位置の候補が6種類，頭部方向の候補が8種類の，全48通りの組み合わせにおいて，最大で約99 %の精度で識別が行えた．また，CSP係数のピーク値のみを用いた場合との比較を行い，先行研究の枠組みでは複数のマイクロホンアレーが必要であるのに対して，ピーク値以外の情報も用いれば2chマイクロホンのみでも音源位置と頭部方向を表現することが可能であることを示した．しかしながら，雑音に加わると精度が大幅に低下するという問題があるため，これに対しては雑音と音声の区別などの処理を行う必要がある．

今後は，学習した位置や頭部方向に対してテストした位置や頭部方向がずれた場合の評価や，未知の位置や頭部方向の推定手法について検討を行う．

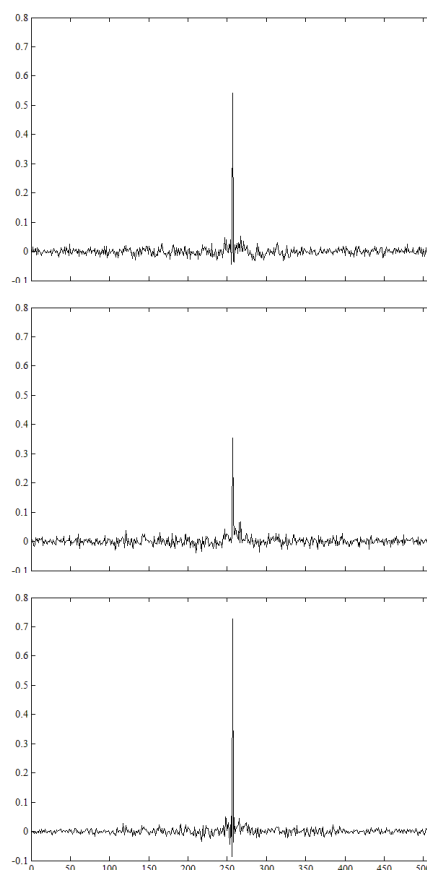


Fig. 8 上段:位置2，頭部方向90°におけるCSP係数．中段:位置2，頭部方向0°におけるCSP係数．下段:位置2，頭部方向0°の収録データに雑音を足したデータにおけるCSP係数．

5 謝辞

本研究は日本学術振興会特別研究員奨励費(23-2495)の助成を受けたものである．

参考文献

- [1] D. Johnson and D. Dudgeon, "Array Signal Processing," Prentice Hall, 1996.
- [2] M. Omologo and P. Svaizer, Proc. ICASSP96, pp. 921-924, 1996.
- [3] F. Asano, et al., IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, E83-A, pp. 2286-2294, 2006.
- [4] Y. Denda, et al., IEICE Trans. on Information and Systems, E89-D, pp. 1050-1057, 2000.
- [5] F. Keyrouz, et al., Proc. ICASSP06, pp. V-341-V-344, 2006.
- [6] M. Takimoto, et al., The Fourth Joint Meeting ASA and ASJ, pp. 3216, 2006.
- [7] A. Brutti, et al., Proc. Interspeech05, pp. 2337-2340, 2005.
- [8] J. M. Sachar and H. F. Silverman, Proc. ICASSP04, vol. 4, pp. 65-68, 2004.
- [9] C. Segura, et al., Proc. Interspeech08, pp. 1325-1328, 2008.
- [10] M. Togami and Y. Kawaguchi, Proc. ICASSP10, pp. 133-136, 2010.
- [11] R. Takashima, et al., ICASSP2010, pp. 2830-2833, 2010.