

CSP 係数の識別に基づく話者の頭部方向推定の検討

高島 遼一[†] 滝口 哲也^{††} 有木 康雄^{††}

[†] 神戸大学大学院システム情報学研究科 〒657-8501 兵庫県神戸市灘区六甲台町 1-1

^{††} 神戸大学自然科学系先端融合研究環 〒657-8501 兵庫県神戸市灘区六甲台町 1-1

E-mail: [†]takashima@me.cs.scitec.kobe-u.ac.jp, ^{††}{takigu,ariki}@kobe-u.ac.jp

あらまし 本稿では 2ch マイクロホンによる話者の音源位置と頭部方向の識別手法について検討する。近年提案されつつある話者の頭部方向の推定手法では、複数のマイクロホンアレーからなるネットワークを用いて、各マイクロホンアレーから得られる CSP (Cross-power Spectrum Phase) 係数のピーク値や音圧情報などから音源位置と頭部方向を推定しており、多数のマイクロホンを配置する必要があった。本手法では、話者の位置や頭部方向によって変化する残響が CSP 係数の形状に影響を及ぼすという点に着目し、CSP 係数のピーク値以外の点も特徴量として扱い、それを識別することにより 2ch マイクロホンで話者の音源位置と頭部方向の同時推定を行う。

キーワード 頭部方向推定, 音源位置推定, CSP 係数, 残響, マイクロホンアレー

Estimation of Head Orientation Based on Discrimination of Cross-power Spectrum Phase Coefficients

Ryoichi TAKASHIMA[†], Tetsuya TAKIGUCHI^{††}, and Yasuo ARIKI^{††}

[†] Graduate School of System Informatics, Kobe University
Rokkodaicho 1-1, Nada-ku, Kobe, Hyogo, 657-8501 Japan

^{††} Organization of Advanced Science and Technology, Kobe University
Rokkodaicho 1-1, Nada-ku, Kobe, Hyogo, 657-8501 Japan

E-mail: [†]takashima@me.cs.scitec.kobe-u.ac.jp, ^{††}{takigu,ariki}@kobe-u.ac.jp

Abstract This paper presents a talker's head orientation estimation method using 2ch microphones. In recent researches, some approaches based on a network of microphone arrays have been proposed in order to estimate the talker's head orientation. In those methods, the talker's head orientation is estimated using the sound amplitude or peak value of CSP coefficients obtained from each microphone arrays. However, microphone array network systems need a lot of microphones to be set along the walls of a given room so that sub-microphone arrays surround the user. In this paper, we focus on the shape of the CSP coefficients affected by the reverberation which depends on the talker's position and the head orientation. In our proposed method, we use not only the peak value but also the other values of the CSP coefficients as feature vectors, and the talker's position and the head orientation are estimated by discriminating the CSP vector. The effectiveness of this method has been confirmed by talker localization and head orientation estimation experiments performed in a real environment.

Key words head orientation, talker localization, CSP coefficients, reverberant speech, microphone arrays

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、話者の頭部の回転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することができる。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]~[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]~[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す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り、システムが大規模になってしまうという欠点がある。そこで我々は以前、発話された音声信号に畳みこまれる残響 (音響伝達特性) が、話者の位置や頭部方向に依存するという点に着目し、観測された音声信号の音響伝達特性を推定して、これを学習・識別することにより、音源の位置と頭部方向を単一マイクのみで推定する手法を提案し、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を示した [11]。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できていないため、頭部方向の細かな変化は識別できていなかった。

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

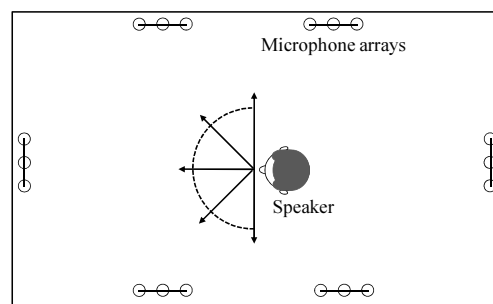


図 1 マイクロホンアレー・ネットワークを用いた頭部方向推定

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

2.1 先行研究

本節では、近年提案されたマイクロホンアレーのネットワークによる話者の頭部方向の推定手法と、以前に我々が提案した単一チャンネルによる頭部方向推定の手法について述べる。マイクロホンアレー・ネットワークを用いる手法では、図 1 のように複数のマイクロホンアレーを話者を中心に取り囲むように設置している。この図において、仮に話者が左方向を向いていた場合、左側に設置したマイクロホンアレーにおいては、その CSP 係数のピーク値や音圧は高くなり、逆に右側のアレーでは低くなる。このような、各マイクロホンアレーから得られる CSP 係数や音圧情報を用いることにより、話者の頭部方向を推定している。文献 [7] や [9] で提案されている手法は、従来の音源位置推定における CSP 係数と同じ枠組みになっており、音源位置の関数である CSP 係数に、頭部方向とアレーの方向が近くなるほど大きくなるような重みを掛け、それらが最大となる位置と頭部方向を求めている。しかし、これらの手法は複数のマイクを部屋の周囲に設置する必要があり、システムが大規模になるという問題がある。

我々はこれまで、観測信号に含まれる音響伝達特性が音源の位置に依存するという点に着目し、音響伝達特性を識別することで音源の位置を単一マイクロホンで推定する手法を提案しており、その有効性を示してきた [12]。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という点にも着目し、これまでの枠組みを用

いて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提案を行った[11]。前述の先行研究とは異なり、この手法はあらかじめ候補となる音源位置と頭部方向毎の音響伝達特性を学習しておく必要があるが、複数のマイクを必要とせず、またマイクを任意の場所に設置することができるという利点があった。この手法では、観測信号から音韻成分を除去し、音響伝達特性のみを分離する必要がある、そのために特定話者の音素 HMM (Hidden Markov Model) を学習しておき、最尤推定法により音響伝達特性を観測信号から推定していた。しかし完全に音韻成分を取り除くことができていないため、残った音韻成分がノイズとなり、頭部方向の大きな変化以外では識別が困難であった。

そこで、より音韻成分の影響を抑えるために、本稿では 2ch マイクロホンを用いて CSP 係数を音響伝達特性の情報として扱い、これを識別することにより音源位置と頭部方向を推定する手法を提案する。

2.2 提案手法

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ンネルと右のチャン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

3.1 実験環境

提案手法を評価するために、特定話者による実験環境

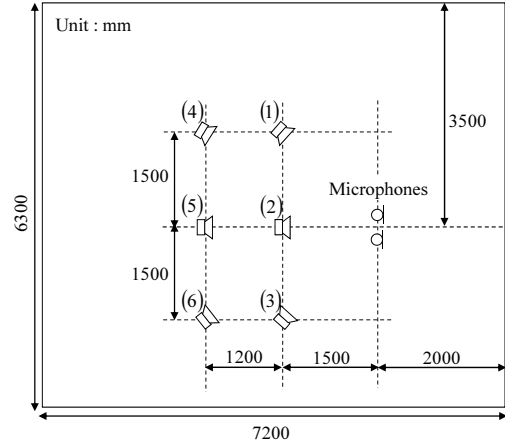


図 2 実験環境。括弧内の数字は音源位置のインデックスを表す。

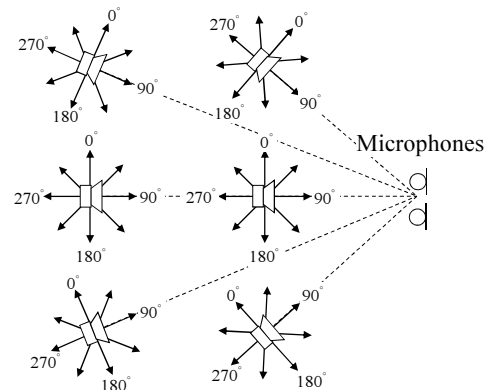


図 3 各位置におけるスピーカークの回転方向

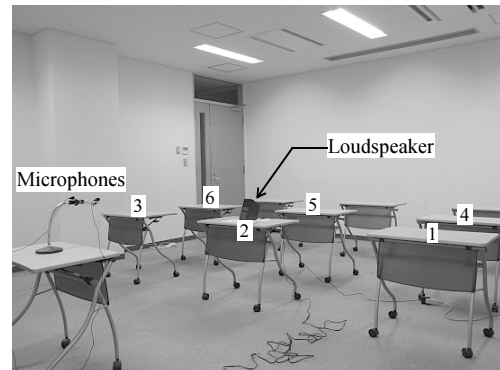


図 4 収録環境。数字は音源位置を表す。

験を行った。約 7.2 m × 6.3 m × 2.8 m (W × D × H) の部屋において、ある位置にスピーカークを設置し、スピーカークの向きを変えながら特定話者の音声を再生し、これをマイク間隔 30 cm の 2ch マイクロホンで収録した。実験環境を図 2 に示す。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。スピーカークの回転方向について図 3 に示す。回転方向は 45° 刻みの 8 種類とし、スピーカーク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

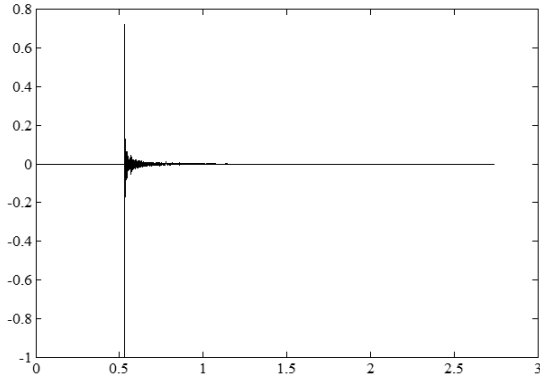


図 5 音源位置 2, スピーカの回転方向 90° におけるインパルス応答

図 4 に示す。

TSP (Time-Stretched Pulse) 法 [13] を用いて求めた位置 2, スピーカー方向 90° におけるインパルス応答を図 5 に示す。残響時間は約 1220 msec であった。スピーカーは 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た。サンプリング周波数 12 kHz, 窓幅 32 msec, フレームシフト 8 msec の分析条件の下で、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、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 2400 (50 × 48) となる。SVM には SVM^{light} [14] 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。

3.2 実験結果

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図 6 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 % で認識が行えていた。

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。図 7 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

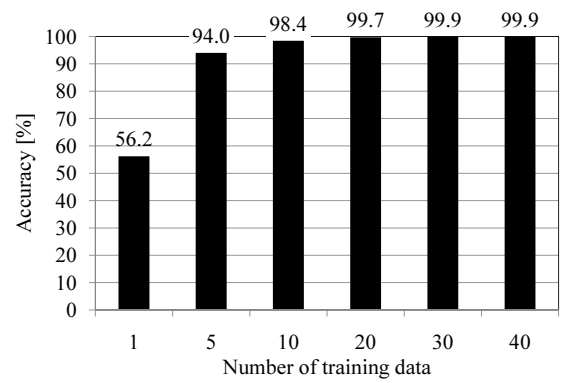


図 6 学習データ数毎の音源位置と頭部方向の認識率

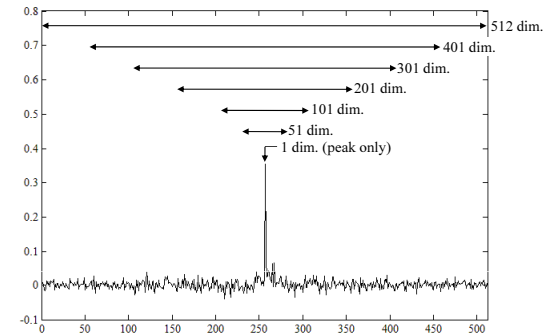


図 7 特徴ベクトルの次元数と CSP 係数の使用範囲

表 1 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

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、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。

結果を表 1 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。学習データ数 40, ピークのみを用いた場合における Confusion matrix を表 2 に示す。表より、頭部方向 90° においては 9 割以上の精度で認識できているが、90° に近い 45° と 135° は多数が 90° と誤認識されており、また他の頭部方向はほとんどが 225° と認識されていることが分かった。

このように、ピークのみを用いた場合では、そのピーク

表 2 学習データ数 40, ピークのみを用いた場合における Confusion matrix [%]

	degree	Predicted							
		0	45	90	135	180	225	270	315
Actual	0	0	0	20	0	0	60	0	20
	45	0	8	74	0	0	10	0	8
	90	0	0	92	0	0	2	0	6
	135	0	8	60	0	0	22	0	10
	180	0	0	20	0	0	60	0	20
	225	0	0	20	0	0	60	0	20
	270	0	0	20	0	0	60	0	20
	315	0	0	20	0	0	60	0	20

値が高いか低いかの判別しか出来ていないと言える。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、テストデータのように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。結果を図 8 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 の結果について、さらに位置毎に精度を分析した結果を図 9 に示す。図より、音源位置が 2 と 5 のとき、つまりマイクに対して正面位置における認識精度が特に低いことが分かる。そこで、SNR が 20 dB の場合について、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 % であった。このときの Confusion matrix を表 3 に示す。表より、ほぼすべてのテストデータに対して、頭部方向を 90° と認識していることが分かる。

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 係数を図 10 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

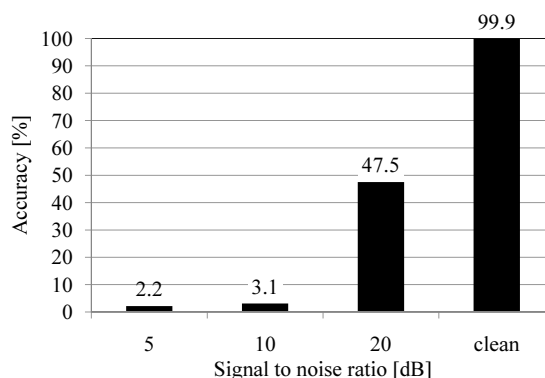


図 8 雑音環境下での音源位置と頭部方向の認識率

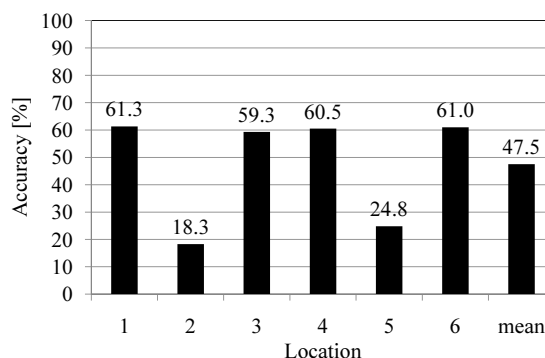


図 9 SNR が 20 dB の場合における、位置ごとの音源位置と頭部方向の認識率

表 3 SNR が 20 dB, 音源位置 2 においての、頭部方向の認識における Confusion matrix [%]

	degree	Predicted							
		0	45	90	135	180	225	270	315
Actual	0	4	0	96	0	0	0	0	0
	45	0	2	98	0	0	0	0	0
	90	0	0	100	0	0	0	0	0
	135	0	0	100	0	0	0	0	0
	180	0	0	98	0	2	0	0	0
	225	0	0	96	0	0	4	0	0
	270	0	0	98	0	0	0	2	0
	315	0	0	96	0	0	0	0	4

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

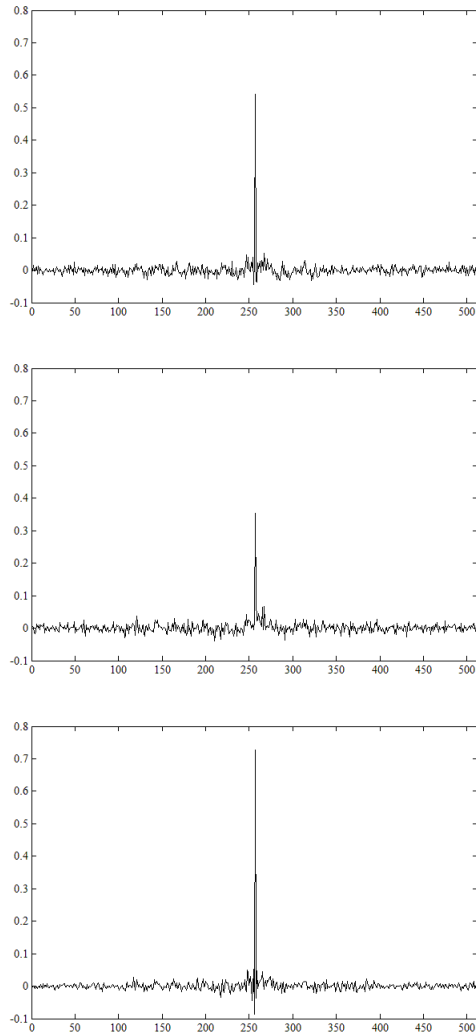


図 10 上段:位置 2, 頭部方向 90° における CSP 係数. 中段:位置 2, 頭部方向 0° における CSP 係数. 下段:位置 2, 頭部方向 0° の収録データに雑音を足したデータにおける CSP 係数.

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, 雑音を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.

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.

文 献

- [1] D. Johnson and D. Dudgeon, "Array Signal Processing," Prentice Hall, 1996.
- [2] M. Omologo and P. Svaizer, "Acoustic event localization in noisy and reverberant environment using CSP analysis," Proc. ICASSP96, pp. 921-924, 1996.
- [3] F. Asano and H. Asoh and T. Matsui, "Sound source localization and separation in near field," IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, E83-A, pp. 2286-2294, 2006.
- [4] Y. Denda and T. Nishiura and Y. Yamashita, "Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation," IEICE Trans. on Information and Systems, E89-D, pp. 1050-1057, 2000.
- [5] F. Keyrouz and Y. Naous and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," Proc. ICASSP06, pp. V-341-V-344, 2006.
- [6] M. Takimoto and T. Nishino and K. Takeda, "Estimation of a talker and listener's positions in a car using binaural signals," The Fourth Joint Meeting ASA and ASJ, pp. 3216, 2006.
- [7] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," Proc. Interspeech05, pp. 2337-2340, 2005.
- [8] J. M. Sachar and H. F. Silverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," Proc. ICASSP04, vol. 4, pp. 65-68, 2004.
- [9] C. Segura, A. Abad, J. Hernando and C. Nadeu, "Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR," Proc. Interspeech08, pp. 1325-1328, 2008.
- [10] M. Togami and Y. Kawaguchi, "Head orientation estimation of a speaker by utilizing kurtosis of a DOA histogram with restoration of distance effect," Proc. ICASSP10, pp. 133-136, 2010.
- [11] 高島遼一, 滝口哲也, 有木康雄, "音響伝達特性の識別に基づく話者の頭部回転方向の推定," 電子情報通信学会技術研究報告, vol. 111, no. 28, SP2011-29, pp. 167-172, 2011.
- [12] R. Takashima, T. Takiguchi and Y. Ariki, "HMM-based Separation of Acoustic Transfer Function for Single-channel Sound Source Localization," ICASSP2010, pp. 2830-2833, 2010.
- [13] Y. Suzuki, F. Asano, H.-Y. Kim and Toshio Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," J. Acoust. Soc. Am. Vol. 97(2), pp. 1119-1123, 1995.
- [14] T. Joachims, "Making large-scale SVM learning practical," in Advances in Kernel Methods - Support Vector Learning, B. Scholkopf, C. Burges and A. Smola, Ed. MIT Press, 1999.