事前分布を用いたCSP法による到来音方向推定*

◎太刀岡勇気, 成田知宏, 岩崎知弘 (三菱電機・情報総研)

1 はじめに

実環境・高騒音下で遠隔マイクによる音声認識を行うためには、音声から話者位置を推定し目的音を強調する必要がある。方向音推定に最低限必要な 2 ch システムで演算量の少ない方向推定手法としては、CSP (Cross-power Spectrum Phase Analysis) 法 (PHAT 法とも呼ぶ)[1] が有力である。CSP 法は CSP 係数列のピークから音の到来方向を推定するため、エネルギーの大きな騒音や他の方向性雑音、周期性を持つ騒音の影響を受けやすい。音声のあらわれやすさでスペクトルを重みづけしたり、SS のように CSP 係数を騒音区間のそれで引き去る CSP Coefficient Subtraction[2] が提案されているが、音声スペクトルと騒音のスペクトルが重なった場合や、非定常な騒音に対して推定性能が低下する。また 3 ch 以上のマイクの 2 ch ずつのペアから算出した CSP 係数を重ね合わせる CSP 係数加算法 [3, 4] も提案されているが、装置規模および演算量が増加する。

そこで本報では CSP 係数の履歴から音源位置の事前分布を推定し、これにより CSP 係数から外乱を取り除く方法を提案する。この方法は音声の特徴を用いないため任意の音源に対応でき、マイクの本数を増やす必要がないので演算量も小さいという特長がある。

2 CSP 係数の外乱除去法

2.1 CSP 法の概要

CSP 法は、距離 L_m [m] 離れた 2 ch の信号のクロススペクトルから信号間の到来時間差 τ を求め、到来方向を推定する。まず、Eq. (1) により遅れ時間 k $(0 \le k \le k_{max} = \mathrm{INT}(L_m f_s/c) + 1)$ の関数である CSP 係数を算出する。ここで INT は整数に切り捨てる関数、f_s はサンプリング周波数 [Hz]、c は音速 [m/s] である。到来時間差 τ は CSP 係数のピークとしてあらわれるため、Eq. (2) によって計算できる [1]。

$$CSP(i,k) = \mathcal{F}_i^{-1}\left(\frac{\mathcal{F}_i(x_1(t))\mathcal{F}_i(x_2(t))^*}{|\mathcal{F}_i(x_1(t))||\mathcal{F}_i(x_2(t))|}\right) \quad (1)$$

$$\tau = \arg\max_k \big(CSP(i,k)\big) \quad (2)$$

ここで、x_1, x_2 は各 ch の入力、$\mathcal{F}_i$ はフレーム番号 i の短時間フーリエ変換 (STFT)、*は複素共役を表す。求まった到来時間差 τ より音源の方向 θ を $\theta = \sin^{-1}\big(\tau c/(L_m f_s)\big)$ で計算する。

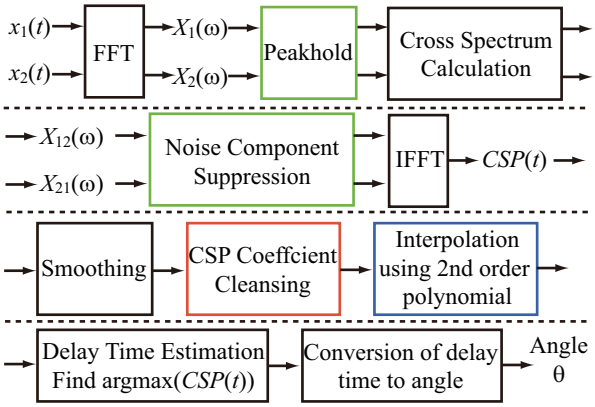

Fig. 1 Schematic diagram of the proposed method.

CSP 法による到来音方向推定の概略図を Fig. 1 に示す。本報ではオリジナルの CSP 法に加えて青枠で示した 2.2 に示した 2 次関数による CSP 係数の内挿 (Interpolation using 2nd order polynomial) を行ったものをベースラインとした (CSP(I) 法と呼ぶ)。既往研究における CSP 法の改良として、緑枠で示したピークホールド処理 (Peakhold)[5] と推定 S/N によるノイズ成分の除去処理 (Noise Component Suppression, 推定 S/N が 0 dB 以下のクロススペクトルを 0 にする) が追加されている (CSP(II) 法と呼ぶ)。これに、2.3, 2.4 に示した事前分布により CSP 係数から外乱を除去する処理 (CSP Coefficient Cleansing) を加えたのが提案法である。

2.2 2 次関数の内挿による CSP 法の精度向上

サンプリング周波数 f_s が十分に高いときは CSP 係数のピークが τ にはっきりとあらわれるため、上記手順で θ を求めることができる。f_s が低いとピークが τ の近傍にあいまいにあらわれるようになってくる。この場合、τ の近傍の $\tau-1$ と $\tau+1$ の 3 点の CSP 係数から 2 次関数による内挿を行い極値から角度を求めることで、推定精度を向上させることができる。

CSP 係数を Eq. (3) の 2 次関数で近似すると、極値 τ' は Eq. (4) のように表される。

$$CSP(i,k) = a_i(k-\tau)^2 + b_i(k-\tau) + c_i \quad (3)$$

$$\tau' = \tau - b_i/(2a_i) \quad (4)$$

ただし τ が 0 か k_{max} のときは片側での 1 次関数補間とする。例えば $\tau = 0$ の場合は、Eq. (5) とする。

$$\tau' = 1 - \frac{CSP(i,0)}{CSP(i,1) + CSP(i,0)} \quad (5)$$

*Direction of arrival estimation by the CSP method using prior distributions.
by TACHIOKA, Yuuki, NARITA, Tomohiro, IWASAKI, Tomohiro (Mitsubishi Electric Corp.)

2.3 事前分布の利用による CSP 係数からの外乱除去

$CSP(i,k)$ を前後 d フレーム (本報では $d=5$) で平均し、Eq. (6) のように $\overline{CSP}(i,k)$ を求める。

$$\overline{CSP}(i,k) = \frac{1}{2d+1}\sum_{j=i-d}^{i+d} CSP(j,k) \qquad (6)$$

$\overline{CSP}(i,k)$ を、遅れ時間 k に対応する方向に音源が存在する尤度と考える。騒音に比べて目的音が長いと仮定して、過去のフレームを加えた尤度 $L(i,k)$ を Eq. (7) により求める。

$$L(i,k) = \sum_{j=0}^{i} \overline{CSP}(j,k) \qquad (7)$$

次に $L(i,k)$ を Eq. (8) のように最大値で除して基準化した $P(i,k)$ $(0 \le P(i,k) \le 1)$ を求める。ここで MAX は引数の最大を返す関数である。

$$P(i,k) = \frac{\mathrm{MAX}\big(L(i,k),0\big)}{\mathrm{MAX}\big(L(i,0),L(i,1),\cdots,L(i,k_{max})\big)} \qquad (8)$$

最後に、重み $P(i,k)$ を乗じた CSP 係数と元の CSP 係数を Eq. (9) のように割合 r で混合し、$\overline{CSP}'(i,k)$ を求める。

$$\overline{CSP}'(i,k) = \big(r+(1-r)P(i,k)\big)\overline{CSP}(i,k) \qquad (9)$$

これを Fig. 2 に模式的に示す。まず図の $\overline{CSP}(i,k)$ が得られたとする。図の中央が音源方向に対応しているのだが、4 フレーム目の左側に騒音による突発的なピークがあらわれ、黒星印が音源位置と誤推定される。ここで発話者が動かないと仮定すれば、突発的に表れたピークと中央のピークの差はわずかで中央のピークが音源である確率が高い。そこで CSP 係数の履歴から Eq. (7) に従い、図の $P(i,k)$ のように尤度関数を計算する。尤度関数の値は、過去にピークのあった図の水色部分が他の部分よりも大きくなる。これを事前分布として $\overline{CSP}(i,k)$ に掛け合わせ、Eq. (9) に従い割合 r で元の CSP 係数と混合することで、図の赤線 $\overline{CSP}'(t,k)$ のように外乱を取り除き、中央の赤丸で示す音源位置に定位できる。

2.4 音声区間検出情報の利用による推定精度向上

音声区間検出の情報が利用できる場合は、非音声区間に出ているピークは外乱によるものであるとわかる。音声区間検出 (VAD) 情報に応じて、音声区間では 1 を返し、非音声区間では 0 を返す関数 $\delta(i)$ を考え、修正された尤度 $L'(i,k)$ を Eq. (10) により求める。これは非音声区間の CSP 係数の符号を反転してペナルティ α を乗じていることに相当し、非音声区

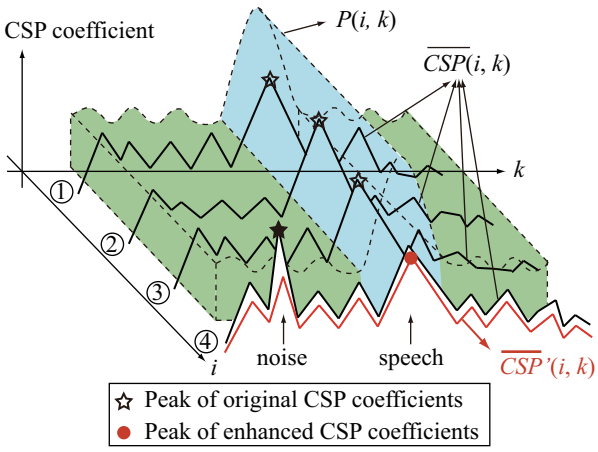

Fig. 2 Procedure of calculating a likelihood function of CSP coefficients and eliminating the disturbance of noises.

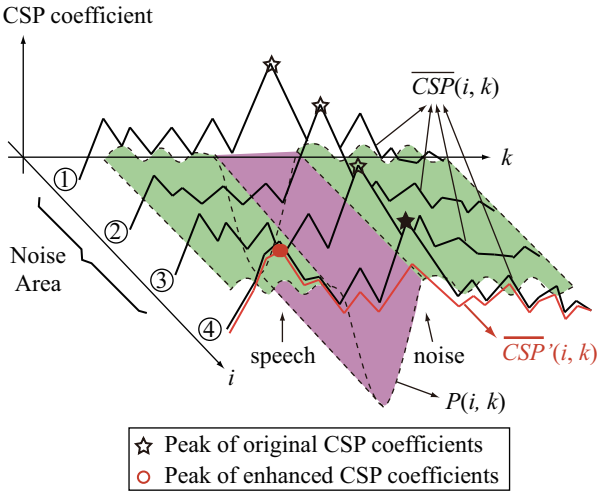

Fig. 3 Procedure of calculating a likelihood function of CSP coefficients with VAD information and eliminating the disturbance of noises.

間においてピークを持っていた外乱を抑制しつつ音声のピークを強調できる。

$$L'(i,k) = \sum_{j=0}^{i}\big((1+\alpha)\delta(j)-\alpha\big)\overline{CSP}(j,k) \qquad (10)$$

Fig. 3 に模式的に説明する。図のようにはじめの 3 フレームはノイズであることが別手段で分かっている。4 フレーム目の左側に音声によるピークがあらわれているものの、中央付近の黒星印のノイズによるピークにマスクされて誤推定されている。このような場合にはノイズ区間の CSP 係数を Eq. (10) に従い符号を反転させ、ノイズ区間において支配的であった成分の尤度が低い図の紫で示す尤度関数 $P(t,k)$ を設定する。その後 2.3 と同様にして、図の赤線 $\overline{CSP}'(t,k)$ のように外乱に埋もれている音声のピークを強調し、図の赤丸で示す音源位置に定位できる。

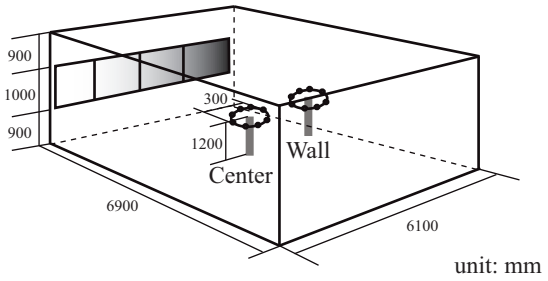

Fig. 4　Geometry of a meeting room.

Table 1　Recorded noises added to evaluation data. ($S/N = 6, 24$ [dB])

Name	Explanation
AirCon	Air Conditioner noise.
180F (270F)	Foot noise generated by a person who stamped at 180 (270) degree.
OpenWin	Environmental noise when the windows were open.
RotateF	Foot noise generated by a person who rotated around the microphone array.

3　実験による検証

3.1　実験条件

Fig. 4 に示す会議室の室中央と室隅で、音源を 30 度毎に移動し、音源とアレー (8 ch の円形アレー (直径 30 cm)) 中心の距離 L_{sr} が {1, 2} m の場合のインパルス応答を測定した。オールパスの残響時間 (T_{30}) は 0.68 秒で、残響減衰曲線に折れ曲がりは見られなかった。これを機器操作語の音声に畳み込み評価データを作成し、対角の 2 ch の音声から音源方向を推定した。同時に収録した下記 Table 1 に示す騒音を S/N が {6, 24} dB になるよう重畳した。f_s は 16 kHz、STFT の窓長は 60 ms、フレームシフトは 30 ms とした。予備検討により、Eq. (9) の r は 0.3、Eq. (10) の α は 1 としたが、パラメータによる性能差は小さかった。

3.2　推定精度の検証

3.2.1　従来法との比較

音源・受音点が室中央の場合のフレーム単位 (音声区間) の推定精度 (許容誤差 ±15°) を算出した。Fig. 5 は最も条件の緩い場合 (S/N 24 dB、L_{sr} 1 m)、Fig. 6 は最も条件の厳しい場合 (S/N 6 dB、L_{sr} 2 m) の全方位の平均である。CSP(I) 法に提案法 (CSP C) を用いると (CSP(I)+CSP C)、Fig. 5 では推定精度が向上したが、Fig. 6 では向上していない。CSP 係数のノイズフロアが高いため、提案法の効果が表れにくいか逆効果になったと考えられる。CSP(II) 法に提案法を用いることで性能が向上した。提案法の処理前に、CSP 係数から騒音の影響をある程度除くことが必要である。音声区間検出情報を用いると (CSP C with VAD)、騒音の学習により騒音の影響を低減でき、推定精度が大きく向上した。検出器は文献 [6] のものを用いた。正解音声区間を与えたもの (CSP C with Ideal VAD) との性能差はあまりない。

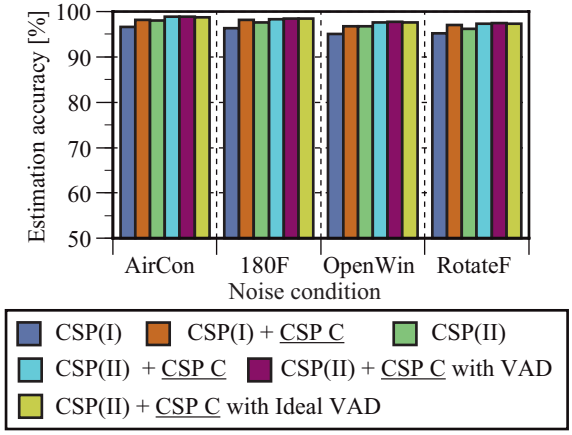

Fig. 5　Estimation accuracy of direction of arrival under condition that S/N is 24 dB and the distance between the source point and the center of the microphone array L_{sr} is 1 m.

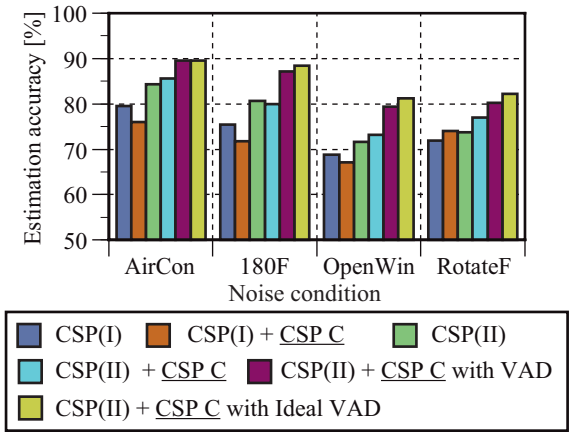

Fig. 6　Estimation accuracy (S/N 6 dB, L_{sr} 2 m).

提案法の効果を見るために、方向性雑音 180F で 60 度方向に話者がいる場合 (S/N 6 dB, L_{sr} 2 m) の時間-角度平面での CSP 係数を Fig. 7 に示す。右側の外乱除去後の CSP 係数は、左側の除去前に比べて話者 (60 度) 方向の CSP 係数に悪影響を与えることなく、足音 (180 度) 方向の CSP 係数を抑圧できている。次に 1.02 秒時点での CSP 係数 (Fig. 7 の A-B 面での断面図) を Fig. 8 に示す。従来法では足音の影響で 160 度方向にピークがあるものが、提案法では 60 度方向にピークがあり音源位置が正しく推定できている。

3.2.2　4ch, 8ch の CSP 係数加算法との比較

3 ch 以上のマイクを用いた場合、2 ch ずつの CSP 係数を加算することで騒音の影響を軽減できる [3, 4]。4 ch の全 6 ペアと 8 ch の対角 4 ペアを提案法と比較した結果を Fig. 9 に示す。提案法は 4 ch のものより性能が良く、8 ch のものと同程度の性能である。

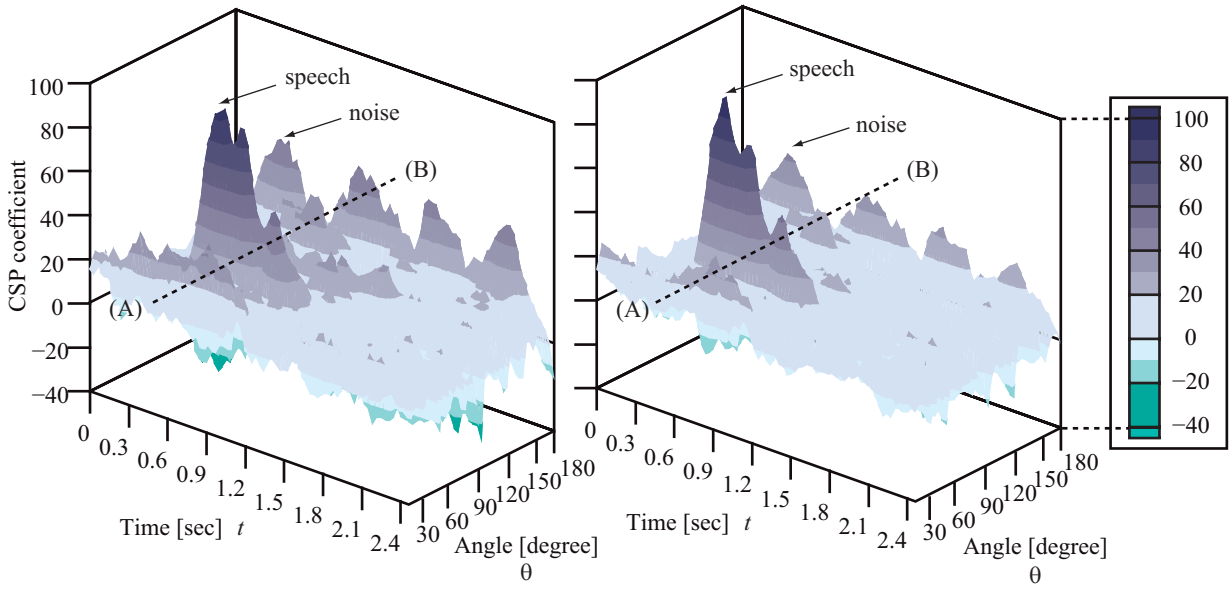

Fig. 7　Original and enhanced CSP coefficients on the time-angle plane. (left: original, right: enhanced)

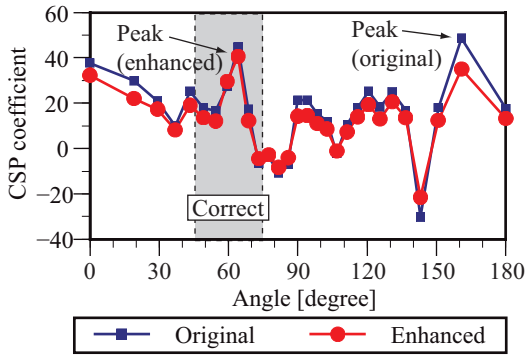

Fig. 8　Enhanced CSP coefficient ($t = 1.02$ [sec]).

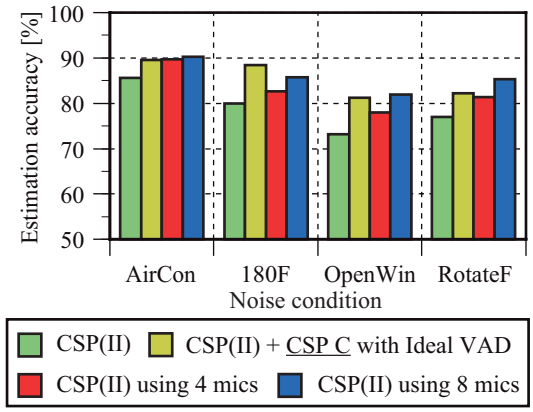

Fig. 9　Estimation accuracy (S/N 6 dB, L_{sr} 2 m) compared with that using 4 and 8 ch microphones.

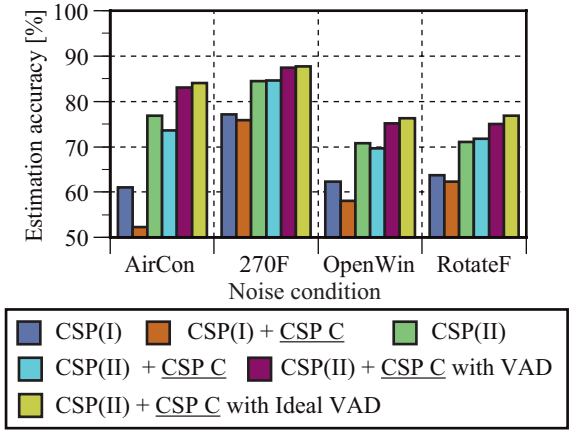

Fig. 10　Estimation accuracy (S/N 6 dB, L_{sr} 2 m). Source and receivers are located near the wall.

3.2.3　受音点位置による影響

Fig. 10 に壁面付近に音源と受音点を置いた場合 (S/N 6 dB、L_{sr} 2 m) の推定精度を示す。1 次反射音の影響が顕著に表れるため推定はより難しいが、騒音の影響を除去後 (CSP(II) 法) に提案法を用いると推定精度が向上した。

4　まとめと今後の課題

CSP 法で音源方向を推定する際に、CSP 係数の履歴から事前分布を推定し騒音の影響を低減させる方法を提案し、拡散性・方向性雑音いずれにも効果がみられた。また音声区間情報を用いて騒音の学習を行うと、推定精度が大幅に向上した。今後、提案の方向推定法を用い強調した音声の音声認識性能を検討する。

参考文献

[1] Knapp *et al.*, IEEE Trans. ASSP24(4) pp. 320–327, 1976. 8.
[2] Denda *et al.*, IEICE Trans. E89-D(3) pp. 1050–1057, 2006. 3
[3] 西浦他, 信学論 J83-DII(8), pp. 1713–1721, 2000.
[4] 中村他, 信学技報 EA2001(4), pp. 25–32, 2001. 4.
[5] 鈴木他, 音響学会誌 65(10), pp. 513–522, 2009.
[6] Sohn *et al.*, IEEE Signal Processing Letters 6(1), pp. 1–3, 1999.