第30回 信号処理シンポジウム
2015年11月4日〜6日（いわき）

音源定位と音声区間検出の有機的統合 –家庭環境を対象に–

Organic integration of speaker localization and speech detection in domestic environments

太刀岡 勇気† 成田 知宏† 渡部 晋治‡ ルルー ジョナトン‡
†三菱電機株式会社 情報技術総合研究所
‡Mitsubishi Electric Research Laboratories

Yuuki TACHIOKA† Tomohiro NARITA† Shinji WATANABE‡ Jonathan LE ROUX‡
†Information Technology R&D Center, Mitsubishi Electric Corporation
‡Mitsubishi Electric Research Laboratories

アブストラクト 本報では家庭環境における音源定位と音声区間検出手法を扱う．実環境では，残響が単純な球面波家庭からの乖離をもたらすため，音源定位は難しい課題である．従来法に含まれている音源定位の誤差を補正するために，テンプレートに基づく手法を提案する．これに加えて，騒音を扱うため，統計的な音声区間検出法を利用する．しかしながら，利用した DIRHA コーパスには，5 つの部屋があり，他の部屋から漏れこんだ発話は棄却しなければならない．この種の棄却は音声区間検出の結果だけを用いたのでは難しい．この問題に対処するため，音源定位と音声区間検出をコスト最小化基準または分類器に基づく方法により，有機的に統合する手法を提案する．提案法は，音源定位において 0.712 の正解率，音声区間検出において 0.743 の F 値を開発セットに対して達成した．ベースラインは，それぞれ 0.559，0.570 であった．テストセットに対してはそれぞれ 0.666，0.706 であり，ベースラインは 0.517，0.602 であった．

1 はじめに

音声を使った遠隔システムを使う際には，音源定位と音声区間検出が重要かつ有効である．一つの応用例としては，遠隔マイクを使った自動音声認識があり，家庭環境に据え付けられた家電などの操作が考えられる．そのような状況では，目的音声を騒音の混ざった音声から強調する必要がある．多くの音声の特徴のみを利用した「ブラインド」音声強調法があるが [1]，話者位置の情報をブラインド手法に加えて利用することで頑健性と有効性の向上を図ることができることが知られている [2], [3]．例えば，音源定位手法により，方向性の雑音を効果的に抑圧することができる．

欧州の公的な支援を受けたプロジェクトである Distant-speech Interaction for Robust Home Applications (DIRHA) プロジェクト [4] では，複数のマイクを使った家庭環境での，遠隔の音声の認識および対話問題に取り組んでいる．DIRHA コーパスはこのプロジェクト由来のコーパスであり，音源定位と音声区間検出の二つの課題からなる．

このコーパスでは，2 次元あるいは 3 次元での音源定位を扱っている．これはかなり難しい課題である．1 次元の推定 (すなわち到来角度だけの推定) は，これに比べてかなり易しい．例えば，相互スペクトル位相 (Cross Spectrum Phase (CSP)) 法 [5] に，事前分布を導入することで [6]，騒音環境下においても実用的な精度で音源方向を知ることができる．一方で，2 次元以上の音源定位は，測定や推定の誤差の影響を受けやすく，方向推定に比べ非常に難易度が高いものの，角度だけの推定よりも応用上は重要である．近年，いくつかの 2 次元での音源定位法が提案されている．それらの中でも，2D-CSP 法 [7] は単純ながら効果的である．この方法は，いくつかの候補点に対して，観測された到来時間差 (time difference of arrival (TDOA)) を理論的な TDOA と比較し，誤差が最小となる点を選択する．この方法は，残響により観測された TDOA が直接波だけから導出した理論的な TDOA と一致しなくなるので，残響がある環境で性能が大きく低下することが知られている．これらの誤差の影響を低減するためには，何らかの受動的な補正が必要である [8]．本報では，この残響による誤差の影響を補正するために，正解点に対する観測 TDOA を参照 TDOA とするテンプレートに基づく方法を提案する [9].

音声区間検出には，統計モデルに基づく手法 [10], [11] が，さまざまなタスクでよい成果を挙げている．ただし，この DIRHA コーパスには，5 つの部屋があり，対象以外の部屋の発話は棄却しなければならないという難しさがある．これらの方法は騒音に対しては頑健であり，騒音

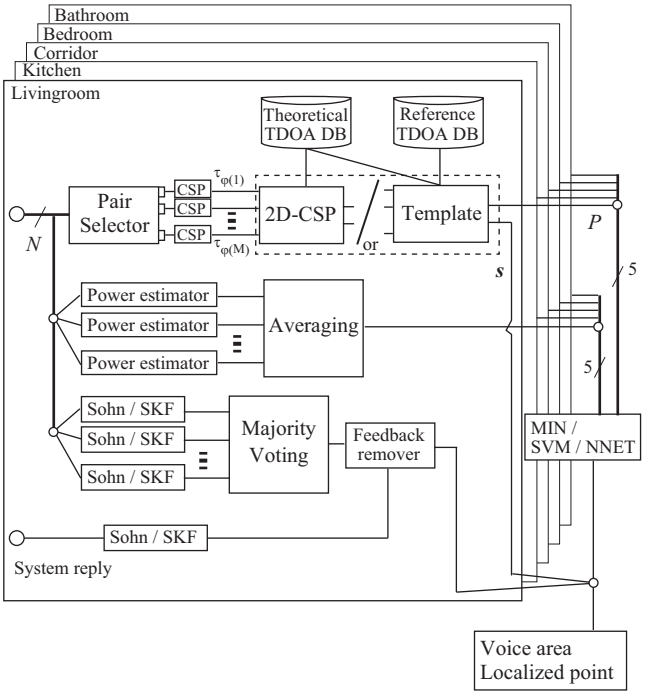

図 1: Schematic diagram of the proposed system for the "Livingroom" localization and detection. (CSP: cross spectrum phase analysis, TDOA: time difference of arrival, Sohn: Sohn's speech detection, SKF: switching Kalman filter based speech detection, MIN: minimum cost criterion, SVM: support vector machine, NNET: neural network)

と音声を区別することはできるが，対象の部屋の音声とそれ以外の部屋の音声を区別することは，同じ音声であるため，それほど単純ではない．この問題に対処するためには，音源定位手法と音声区間検出法の有機的な統合が必要となる．我々は，最小コスト基準もしくは，分類器に基づく方法の 2 つの手法で，音声区間検出のために音源定位の結果を利用する方法を提案する．

本報では，3 節で，従来の 2D-CSP 法による音源定位法 [7] について述べ，ついでテンプレートに基づく手法により誤差を補正する方法を提案する．次に，4 節で，統計的音声区間検出法 [10], [11] について述べ，最後に 5 節で音源定位と音声区間検出を統合する手法を提案する．6 節の実験により，提案のテンプレートに基づく手法が，音源定位の精度を向上させ，分類器に基づく方法が音声区間検出の性能を向上させることを示す．

2 システムの概略

図 1 に，提案法の概観図を示す．音源定位部と音声区間検出部からなる．音源定位部には，N 個のマイクからの入力より選択した M 組に対して，CSP 法を用いて対応する M 個の TDOA τ を計算する．これらの TDOA を理論値から計算された TDOA と比較することで，2D-CSP 法により音源の候補点 $\mathbf{s}$ ごとにコスト $P(\mathbf{s})$ を計算する．さらに提案のテンプレートに基づく方法により，参照値となる TODA を使って誤差を修正する．

音声区間検出部においては，尤度比を使う手法を採用した．ここでは，Sohn の手法 [10] とスイッチングカルマンフィルタ (SKF) に基づく手法 [11] を使った．検出はマイクごとに行われ，N 個の検出結果が得られる．これらは多数決により統合される．実収録されたデータには，システムの応答が発話間に存在する．応答音声の記録は別に残っているため，応答発話は別に検出し，該当の発話が上記の検出結果に含まれていれば除外した．最後に，検出結果は，最小コスト基準あるいはそれぞれの部屋のコスト P と平均パワーを入力特徴量とする分類器により，修正される．

3 音源定位手法

3.1 2D-CSP 法

通常，CSP 法 [5] は，平面波仮定に基づき音声の到来方向を推定する．話者が存在する場所を取り囲むようにマイクが配置されていれば，三角測量の原理を用いて話者位置を特定することができる．一方で，2D-CSP 法 [7] は，球面波仮定に基づき到来方向ではなく音源の位置を推定する．話者位置が $\mathbf{s}$ であり，N 個のマイクの中で i 番目のマイク位置を $\mathbf{r}_i$ としたときに，マイク i, j $(1 \leq i, j \leq N)$ の間の理論的な TDOA τ_{ij}^{theo} は

$$\tau_{ij}^{theo}(\mathbf{s}) = \frac{|\mathbf{r}_i - \mathbf{s}| - |\mathbf{r}_j - \mathbf{s}|}{c} \tag{1}$$

のようにあらわされる．ここで c は音速である．CSP 法は TDOAτ_{ij}^{csp} を観測された短時間フーリエ変換 $\boldsymbol{X}_i$ と $\boldsymbol{X}_j$ のクロススペクトルから式 (2) の最適解として計算する [5].

$$\tau_{ij}^{csp} = \arg\max_{\tau} \left[\mathcal{F}^{-1} \left(\frac{\boldsymbol{X}_i \odot \boldsymbol{X}_j^*}{|\boldsymbol{X}_i||\boldsymbol{X}_j|} \right) \right] \tag{2}$$

ここで $\mathcal{F}$ は短時間フーリエ変換，*と $\odot$ はそれぞれ複素共役と 2 つのベクトルの要素ごとの積を表す．

話者位置の候補点 $\mathbf{s}$ に対して，M 組 $(2 \leq M \leq {}_NC_2)$ の任意のマイクペアで観測した TDOAτ^{csp} とそれに対応する理論値 τ^{theo} との差異を加算することでコスト関数 $P(\mathbf{s})$ を計算する．τ^{theo} が τ^{csp} に近いときは，コスト関数 P は小さい値を取る．以下のように，$P(\mathbf{s})$ を最小化するような点を候補点 $\mathbf{S}$ から選択することで，話者位置 $\mathbf{s}$ が決定される．

$$\arg\min_{\mathbf{s}\in\mathbf{S}} P(\mathbf{s}) = \arg\min_{\mathbf{s}\in\mathbf{S}} \sum_{m=1}^{M} \left| \tau_{\varphi(m)}^{theo}(\mathbf{s}) - \tau_{\varphi(m)}^{csp} \right|^2 \tag{3}$$

ここで $\varphi(m)$ は，m 番目のマイクペアである．一般に，2 次元の音源定位では，1 組のマイクペアでは，音源が双曲線上にあることを示すだけなので，2 組以上の異なるマイクペア (つまり 3 つ以上のマイク) が必要とされる．

3.2 テンプレートに基づく手法

実環境では，例えば残響や観測誤差により，理論的な TDOA と観測された TDOA は正解の音源位置に対してすら異なりうる．ここで，式 (3) のコスト関数 P は，以下の最適化問題として一般化される．

$$\arg\min_{\mathbf{s}\in\mathbf{S}} P(\mathbf{s}) = \arg\min_{\mathbf{s}\in\mathbf{S}} \sum_{m=1}^{M} \left| \tau^{ref}_{\varphi(m)}(\mathbf{s}) - \tau^{csp}_{\varphi(m)} \right|^2 \quad (4)$$

ここで $\tau^{ref}(\mathbf{s})$ は，位置 $\mathbf{s}$ に対する参照値となる TDOA である．2D-CSP 法では，理論値から導かれた TDOA が参照値として使われるが，観測は不可避的に誤差 ϵ を含むため，以下のように誤差を含んだ形式となる．

$$\tau^{theo}_{\varphi(m)}(\mathbf{s}) \approx \tau^{csp}_{\varphi(m)} - \epsilon_{\varphi(m)}(\mathbf{s}) \quad (5)$$

誤差の影響を低減するため，我々はテンプレートに基づく手法を提案する．提案法では，参照値となる TDOA$\tau^{ref}_{\varphi(m)}$ を，Eq. (6) から求められるものに替える．これらの誤差 ϵ は開発セット中にあるすべての点 $\mathbf{s}\in\mathbf{S}$ に対して計算される．

$$\tau^{ref}_{\varphi(m)}(\mathbf{s}) \approx \tau^{theo}_{\varphi(m)}(\mathbf{s}) + \epsilon_{\varphi(m)}(\mathbf{s}) \quad (6)$$

このように参照値を修正することで，誤差の影響を打ち消すことが期待される．

4 音声区間検出法

4.1 従来の尤度比検定法 (Sohn の手法)

尤度比検定による音声区間検出法のうち，最も単純でかつ効果的な手法 [10] を，ここに述べる．$\boldsymbol{X} = \{X_k\}_{k=1}^{K_X}$ は，観測された K_X 次元のスペクトルとする．パワースペクトル $|X_k|^2$ は次元ごとに独立であり，騒音のフレーム (H_S) では騒音の混ざった音声モデル λ^S から，音声のない騒音だけのフレーム (H_N) では騒音モデル λ^N から出力されると仮定する．

$$\begin{aligned} p(\boldsymbol{X}|\lambda^S, H_S) &= \prod_{k=1}^{K_X} \frac{1}{\pi[v_k^S + v_k^N]} e^{-\frac{|X_k|^2}{v_k^S + v_k^N}} \\ p(\boldsymbol{X}|\lambda^N, H_N) &= \prod_{k=1}^{K_X} \frac{1}{\pi v_k^N} e^{-\frac{|X_k|^2}{v_k^N}} \end{aligned} \quad (7)$$

ここで v_k^S と v_k^N はそれぞれ，音声と騒音のスペクトルの分散である．k 次元目の音声と騒音の対数尤度比は，式 (8) で表される．

$$\Lambda_k(X_k|\lambda^S, \lambda^N) = \ln \frac{p(X_k|\lambda^S, H_S)}{p(X_k|\lambda^N, H_N)} \quad (8)$$

個々のフレームが音声か騒音いずれであるかは，式 (9) の対数尤度比の幾何平均に基づき決定する．

$$\Lambda(\boldsymbol{X}|\lambda^S, \lambda^N) = \frac{1}{K_X} \sum_{k=1}^{K_X} \Lambda_k(X_k|\lambda^S, \lambda^N) \underset{H_N}{\overset{H_S}{\gtrless}} \eta \quad (9)$$

もし $\Lambda(\boldsymbol{X}|\lambda^S, \lambda^N)$ が，事前に定めた閾値 η より大きければ，当該フレームは騒音の混ざった音声状態であり，小さい場合は騒音状態であると推定される．騒音モデルは事前に観測した騒音に基づき構築する．音声モデルは最尤推定により推定する．すなわち $\partial\Lambda_k(X_k)/\partial\lambda_k^S = 0$ であるので，$v_k^S = |X_k|^2 - v_k^N$ の関係に従い，推定することとなる．これは音声のモデル λ_k^S は，音声と騒音のパワーが加法的であると仮定して推定していることになる．

4.2 スイッチングカルマンフィルタに基づく手法

SKF に基づく音声区間検出法 [11] は，定常騒音環境や弱非定常騒音下において有効であることが知られている．この方法では，事前に用意したクリーン音声モデルと，オンラインで推定した騒音モデルから，騒音の混ざった音声モデルをフレームごとに構築する．ここでは特徴量には K_Y 次元の対数メルスペクトル $\boldsymbol{Y} = \{Y_k\}_{k=1}^{K_Y}$ を使った．対数メル領域では，騒音の混ざった観測音声の特徴量はクリーン音声と騒音それぞれの対数和として表現されるためである．騒音の混ざった音声モデルと騒音モデルの尤度はそれぞれ GMM により与えられる．モデルパラメータ λ は，$\lambda = \lambda^S$ もしくは $\lambda = \lambda^N$ である．

$$p(\boldsymbol{Y}|\lambda) = \sum_{m=1}^{M} w_m \prod_{k=1}^{K_Y} \frac{1}{\sqrt{2\pi}\sigma_{m,k}} \exp\left[-\frac{(Y_k - \mu_{m,k})^2}{2\sigma_{m,k}^2} \right] \quad (10)$$

M はガウス混合分布の混合数，w_m，$\mu_{m,k}$ および $\sigma_{m,k}^2$ は，それぞれ m 次元目のガウス分布の混合重み，平均および分散であり，これらは SKF によって更新される．尤度比の計算は式 (8) と (9) と同様にして行われる．但し，X_k についてのガウス分布を Y_k についての GMM で置き換える必要がある．

5 音源定位と音声区間検出の統合

DIRHA コーパスでは，他の部屋での発話は棄却されなければならないので，他の部屋の音源定位結果を用いて棄却する方法を提案する．

5.1 コスト最小基準

1 つめの方法は，対象の部屋における音源定位のコスト P_{in} を他の部屋でのコスト P_{out} と比較するものである．もし話者が複数の部屋に定位された場合には，部屋間でコストを比較して最もコストの小さいものを選ぶのが最

も理にかなっていると考えられる．しかしながら，コスト関数は室の形状やマイクの設定に依存しているため，単純に比較するだけでは誤って棄却してしまう可能性がある．よってここでは許容パラメータ η' を導入する．これにより厳密に最小でなくても最小に近い発話を救うことができる．各フレームに対して，全ての部屋の中でそのフレームのコストが最小に近いかどうかを示すフラグ f を設定する．

$$f = \begin{cases} \text{true} & \forall P_{out}, P_{in} < \eta' P_{out} \\ \text{false} & \text{otherwise} \end{cases}$$

各発話に対して，真値であるフレーム数の発話全体のフレーム数に対する割合が，事前に定めた閾値よりも小さい場合には当該発話は棄却される．

5.2 分類器に基づく方法

2 つめの方法は，対象の部屋の特徴量 $\mathbf{z}_{in}$ とそれ以外の部屋の特徴量 $\mathbf{z}_{out}$ を連結したベクトルを入力とする分類器 $\mathcal{C}$ を使う方法である．開発セットで分類器を学習した後，分類器の出力を閾値 η'' と比較し，フレームごとにフラグを推定する．

$$f = \begin{cases} \text{true} & \mathcal{C}([\mathbf{z}_{in}; \mathbf{z}_{out}]) > \eta'' \\ \text{false} & \text{otherwise} \end{cases}$$

これらのフラグは，5.1 と同様に，当該発話を棄却するか否かを決めるために統合される．

6 実験条件

6.1 DIRHA コーパスの概要

DIRHA コンソーシアムにより，同期して録音された 1-2 分程度の音声ファイルが提供されている．実際の環境を模擬するために，実際の家において収録されたデータベースを使っている．Kitchen, Livingroom, Corridor, Bathroom, Bedroom の 5 つの部屋がある．音源定位と音声区間検出の対象は，Kitchen と Livingroom に限定されている．Kitchen と Livingroom に対しては，室中心に 6 個の円形マイクが据え付けられている．これに加え，すべての部屋に，2,3 個のマイクから成るいくつかのアレイが，部屋を取り囲むように壁に取り付けられている．総計で 40 のマイクが使われている．マイクペアは各マイクアレイ内で選択することとした．別のマイクアレイに属しているマイクは離れているため，ペアを形成してもその相関が低すぎ，音源定位にとって有利な情報が得られるとは考えられないためである．

本コーパスでは学習データはなく，開発セット (**dev**) とテストセット (**test**) が提供されている．規定により，すべてのパラメータは開発セットにより調整することになっている．両セットには Real と Simulations のサブセットがある．Real セットでは，各タスクに対して，1 部屋に 1 話者のみがいるが，部屋の中を自由に動き回っている．話者とシステムの間の対話を模擬するため，システムの応答が時折入るが，それは別個に提供されている．Simulations セットでは，異なる部屋において複数の話者が存在しうるが，話者は動かないことになっている．システムの性能は，提供された評価ツールを使って評価した．

6.2 音源定位

高さの同定は水平面上での同定ほど重要ではないので，今回は 2 次元の同定とした [1]．我々の実験では，音声データは元の 48 kHz から 16 kHz にダウンサンプリングした．フレームサイズは 960，フレームシフトは 800 である．2D-CSP 法と提案のテンプレートに基づく手法の性能を，マルチチャネル CSP 法 [12] および長いフレームサイズ (1 秒) を用いた SRP-PHAT 法 [2][13] と比較した．音源にも幅があるため誤差 0 で推定することは原理的にできないので，”Fine error”(すなわち許容誤差) は 50 cm とした．

6.3 音声区間検出

音声区間検出性能を発話単位で，プレシジョン，リコール，F 値の観点から評価した．16kHz サンプリングとし，フレームサイズは 960 であり，フレームシフトは 160 である．無音の最大継続長は 500 ms，発話の最小継続長は 300 ms とした．SKF においては，ガウス混合分布の数は 32 とし，20 次元のメルスペクトルを使った．Sohn の手法，SKF 両手法に対して，HMM ハングオーバー手法 [10] を使った．各音声ファイルごとに音声区間検出を行ったのち，多数決により，最終的な音声区間検出結果を室ごとに得る．

6.4 音源定位と音声区間検出の統合

音源定位のコスト P とフレーム毎の音声パワーをマイクに対して平均したものを特徴量 $\mathbf{z}_{in}$ と $\mathbf{z}_{out}$ として用いた．分類器に基づく方法には，線形サポートベクトルマシン (SVM) に基づく分類には SVM-light (v.6.02)[3]，神経回路網 (NNET) に基づく分類には pyBrain (v.0.31)[4] を使った．両分類器には，特徴量の分散が 1 となるように正規化を行った特徴量を用いた．SVM と NNET は，話者が対象の室にいるかいないかを示す 2 値を教師信号として，これを出力するように学習した．開発セットによ

[1] 評価ツールにおいて，`-2D` オプションを使った．
[2] http://www.lems.brown.edu/array/tools/srplems.m
[3] http://svmlight.joachims.org/
[4] http://pybrain.org/

表 1: Localization and speech detection results on the development set (**dev**). Methods are indicated for speech activity detection (SAD), source localization (LOC), and their integration (INT). Performance criteria for source localization are Fine Error (FE), Gross Error (GE), and Percentage of Correct localization (PCor). For speech detection, utterance-based criteria are used: Precision (P), Recall (Re), and F value.

Methods			Real						Simulations						Average					
SAD	LOC	INT	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F
Oracle	2D-CSP	-	298	602	.685	-	-	-	309	925	.504	-	-	-	306	870	.540	-	-	-
	Template		303	592	.719	-	-	-	160	864	.643	-	-	-	200	817	.658	-	-	-
	M-CSP		347	1307	.177	-	-	-	348	1433	.208	-	-	-	348	1409	.202	-	-	-
	SRP-PHAT		*289*	826	.537	-	-	-	248	987	.509	-	-	-	257	957	.515	-	-	-
Sohn	2D-CSP	-	295	565	.709	.693	**.957**	.804	308	836	.525	.354	**.905**	.509	305	794	.559	.414	**.919**	.570
	Template	-	301	537	.746	.693	**.957**	.804	161	769	.657	.354	**.905**	.509	197	732	.673	.414	**.919**	.570
		MIN	301	537	.748	.744	**.957**	**.837**	161	769	.657	.354	**.905**	.509	197	732	.673	.419	**.919**	.575
		SVM	304	528	.757	.740	.826	.781	159	749	.681	.670	.836	.744	197	714	.695	.689	.833	.754
		NNET	299	**498**	**.779**	.797	.826	.811	151	732	.685	**.800**	.693	.743	193	692	.704	**.799**	.729	**.762**
SKF	2D-CSP	-	300	559	.699	.697	.812	.750	303	798	.548	.416	.894	.568	302	762	.574	.461	.872	.603
	Template	-	306	532	.744	.697	.812	.750	158	714	.678	.416	.894	.568	194	686	.689	.461	.872	.603
		MIN	306	528	.752	.699	.768	.732	158	709	.679	.414	.889	.565	194	682	.692	.457	.857	.596
		SVM	310	535	.741	.823	.783	.802	157	688	.699	.661	.841	.740	196	663	.707	.694	.826	.754
		NNET	**292**	503	.756	**.837**	.609	.705	**149**	**663**	**.704**	.733	.778	**.755**	**180**	**642**	**.712**	.753	.733	.743

表 2: Localization and speech detection results on the test set (**test**).

Methods			Real						Simulations						Average					
SAD	LOC	INT	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F
Oracle	2D-CSP	-	301	622	.582	-	-	-	302	1076	.461	-	-	-	302	965	.497	-	-	-
	Template		297	584	.658	-	-	-	186	1094	.564	-	-	-	228	972	.592	-	-	-
Sohn	2D-CSP	-	298	585	.610	.868	.962	.913	303	1004	.479	.368	**.944**	.530	302	904	.517	.441	**.949**	.602
	Template	-	293	550	.673	.868	**.962**	.913	185	969	.590	.368	**.944**	.530	225	870	.613	.441	**.949**	.602
		MIN	293	545	.677	.882	**.962**	**.920**	186	970	.591	.365	.934	.525	225	868	.616	.441	.942	.600
		SVM	299	**505**	.678	**.917**	.316	.470	185	961	.592	.678	.939	**.788**	204	920	.602	.700	.762	**.730**
		NNET	287	542	.657	.900	.532	.668	**178**	969	.567	**.720**	.707	.714	211	889	.588	**.755**	.657	.703
SKF	2D-CSP	-	296	846	.624	.657	.937	.772	304	922	.526	.411	.859	.556	301	823	.557	.462	.881	.606
	Template	-	292	513	.683	.657	.937	.772	184	859	.637	.411	.859	.556	225	768	.651	.462	.881	.606
		MIN	292	512	**.684**	.651	.937	.768	184	857	.639	.411	.843	.553	225	766	.653	.461	.870	.602
		SVM	299	518	.668	.571	.367	.447	180	838	.644	.684	.813	.734	**203**	798	.647	.664	.686	.675
		NNET	**284**	507	.662	.692	.608	.647	187	**768**	**.667**	.712	.742	.727	215	**710**	**.666**	.707	.704	.706

り，SVM と NNET のパラメータおよび閾値を調整した．NNET は，隠れ層を 2 層とし，隠れ層のノード数は下から 15,10 とした．最後に Real セットに対しては，一つの部屋だけにしか話者はいないことがわかっているため，Livingroom と Kitchen における検出された発話の音声パワーを比較して，よりパワーの大きい方を採用した．

7 実験結果

7.1 正解の音声区間を与えた場合の音源定位精度

上述の手法の音源定位精度を比較するため，表 1 と 2 の 1 段目には，音声区間検出結果に正解のものを与えた場合を示した．2D-CSP 法の性能は，マルチチャンネル CSP 法や長いフレームサイズの SRP-PHAT 法よりも高かった．さらに，必要な計算量もマルチチャンネル CSP 法や SRP-PHAT 法よりも少かったので，ここでは，2D-CSP 法をベースラインとした．テンプレートに基づく手法の性能は 2D-CSP 法よりも優れ，家庭内環境における音源定位タスクにおける有効性が示された．

7.2 音声区間検出精度

表 1 と 2 の 2 段目，3 段目は音声区間検出の結果を示す．SKF の性能は Sohn の手法よりも若干高かった．どちらの手法も単独では，他の部屋から漏れこんだ発話や騒音を棄却するのにそれほど有効ではなかった．

話者位置の同定結果と統合する方法については，最小コスト基準に基づく方法はベースラインと有意な差が見られなかったものの，SVM や NNET を使った分類器に基づく方法は有効であった．分類器は開発セットで学習したので，テストセットの結果でも比較してみると，平均的に見て SVM，NNET ともに，Sohn の手法と SKF のいずれおいても，F 値が改善した．

8 おわりに

本報では，DIRHA コーパスを利用して，家庭環境における音源定位と音声区間検出の問題を扱った．音源定位の問題に対しては，残響の影響で理論的な球面波仮定が成り立たなくなる場合に有効なテンプレートに基づく方法を提案し，実環境において有効に働くことを示した．これに加え，他の部屋における発話といった，従来の音声区間検出器から得られる結果のみからは容易に棄却することのできない発話を棄却するために，音源定位と音声区間検出の結果を統合する手法を提案した．サポートベクトルマシンや神経回路網といった分類器を使うことで，音声区間検出の性能を向上させることができた．

参考文献

[1] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” Multimodal Technologies for Perception of Humans, pp.509–519, Springer, 2008.

[2] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, “Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds,” Computer Speech and Language, vol.27, pp.851–873, 2013.

[3] Y. Tachioka, S. Watanabe, J. Le Roux, and J. Hershey, “Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark,” Proceedings of the 2nd CHiME Workshop on Machine Listening in Multisource Environments, pp.19–24, June 2013.

[4] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, “The DIRHA simulated corpus,” Proceedings of LREC, pp.2629–2634, May 2014.

[5] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.24, pp.320–327, Aug. 1976.

[6] Y. Tachioka, T. Narita, and T. Iwasaki, “Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information,” Acoustical Science & Technology, vol.33, pp.68–71, Jan. 2012.

[7] D.V. Rabinkin, R.J. Renomeron, A. Dahl, J.C. French, J.L. Flanagan, and M.H. Bianchi, “A DSP implementation of source location using microphone arrays,” Proceedings of SPIE, pp.88–99, 1996.

[8] K. Ho and L. Yang, “On the use of a calibration emitter for source localization in the presence of sensor position uncertainty,” IEEE Transactions on Signal Processing, vol.56, pp.5758–5772, 2008.

[9] 太刀岡勇気，成田知宏，石井　純，“音源距離推定方式の比較検討とコスト関数の一般化，” 日本音響学会研究発表会講演論文集 (秋季)，pp.90–93，Sept. 2012.

[10] J. Sohn, N.S. Kim, and W. Sung, “A statistical model-based voice activity detection,” IEEE Signal Processing Letters, vol.6, pp.1–3, Jan. 1999.

[11] M. Fujimoto and K. Ishizuka, “Noise robust voice activity detection based on switching Kalman filter,” IEICE Transactions on Information and Systems, vol.E91-D, pp.467–477, March 2008.

[12] K. Hayashida, M. Morise, and T. Nishiura, “Near field sound source localization based on cross-power spectrum phase analysis with multiple channel microphones,” Proceedings of INTERSPEECH, pp.2758–2761, Sept. 2010.

[13] H. Do, H. Silverman, and Y. Yu, “A real-time SRP-PHAT source location implementation using stochastic region contraction(src) on a large-aperture microphone array,” Proceedings of ICASSP, vol.1, pp.121–124, April 2007.