家庭環境における補正音源定位と統計的音声区間検出の統合 –DIRHA コーパスの利用– *

○太刀岡勇気, 成田知宏 (三菱電機・情報総研), 渡部晋治, ルルージョナトン (MERL)

1 はじめに

遠隔音声認識が、種々の装置に搭載されるようになってきている。例えば、自動音声認識による家電の操作がある。そのような状況では、目的音声だけを強調する必要があるが、その前段階として、音源定位と音声区間検出の技術が重要かつ有効である。Distant-speech Interaction for Robust Home Applications (DIRHA) プロジェクト [1] では、この課題に取り組んでおり、DIRHA コーパスでは、音源定位と音声区間検出の二つの課題が設定されている。

音源定位では、2、3 次元を対象としている。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、方向のみの推定に比べ難易度が高いものの、応用上はより重要である。近年、いくつか方式が提案され、2D Cross Spectrum Phase (2D-CSP) 法 [2] は単純ながら効果的である。ただし、残響がある環境で誤差の影響で性能が大きく低下することが知られており、本報では、この残響による誤差を補正するため、テンプレートに基づく方法を導入する [3]。(3 節)

音声区間検出では、統計モデルに基づく手法 [4, 5] が成果を挙げている。(4 節) これらは騒音に頑健であるが、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。(5 節)

2 システムの概略

図 1 には、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 個のマイク入力より選択した M 組に対して、CSP 法により対応する 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 と比較し、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。テンプレートに基づく方法では、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 [4] とスイッチングカルマンフィルタ (SKF) に基づく手法 [5] を使った。検出はマイクごとで、結果は多数決により統合される。システムの応答は上記の検出結果から除外した。最後に、検出結果は、最小コスト基準あるいは各部屋のコスト P と平均パワーを入力とする分類器により、修正される。

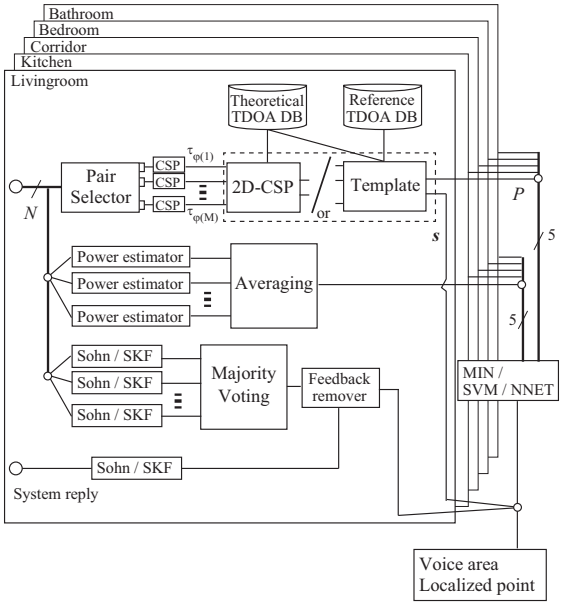

Fig. 1 Schematic diagram of the proposed system for the “Livingroom” localization and detection. (CSP: cross spectrum phase analysis, TDOA: time difference of arrival, Sohn: Sohn’s speech detection, SKF: switching Kalman filter based speech detection, MIN: minimum cost criterion, SVM: support vector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

3 音源定位手法

3.1 2D-CSP 法

通常 CSP 法 [6] は、平面波仮定に基づき音声の到来方向を推定する。話者が存在する範囲にマイクが配置されていれば、三角測量の原理で話者位置を特定できる。一方、2D-CSP 法 [2] は、球面波仮定に基づき話者位置 $\mathbf{s}$ を推定する。N マイク中で i 番目のマイク位置を $\mathbf{r}_i$ としたとき、c を音速として、マイク i, j $(1 \le i, j \le N)$ の間の理論的な TDOA τ_{ij}^{theo} は

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

* Ensemble integration of calibrated speaker localization and statistical speech detection in domestic environments: DIRHA corpus, by TACHIOKA, Yuuki, NARITA, Tomohiro (Mitsubishi Electric Corporation), WATANABE, Shinji, Le Roux, Jonathan (MERL).

のようにあらわされる。ここで TDOAτ_{ij}^{csp} は、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以下の (2) の最適解として求まる [6]。

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

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ベクトル間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 ($2 \leq M \leq {}_NC_2$) のマイクペアで観測した TDOAτ^{csp} と、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、コスト関数 P は小さい値を取るので、(3) 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。

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

ここで $\varphi(m)$ は、m 組目のマイクペアである。一般に 2 次元の音源定位には、2 組以上の異なるマイクペア (つまり 3 つ以上のマイク) が必要である。

3.2 テンプレートに基づく手法

実環境では、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。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。

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

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、式 (5) のように観測は不可避的に誤差 ϵ を含む。

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

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導入する [3]。提案法では、参照値となる TDOA$\tau_{\varphi(m)}^{ref}$ を、Eq. (6) から求められる TDOA に替え、誤差 ϵ は開発セット中の全ての点 $\mathbf{s} \in \mathbf{S}$ に対して計算される。

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

参照値を修正することで、誤差の影響を低減できる。

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 [4] 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、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、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のいずれであるかは、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から構築し、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、音声と騒音のパワーが加法的であると仮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 [5] が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の特徴量の対数和として表現されるためである。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。GMM の平均、分散と混合重みは、SKF により更新される。尤度比の計算は式 (8) と (9) と同様に行われる。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。

5 音源定位と音声区間検出の統合

DIRHA 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 2 手法を提案する。

5.1 コスト最小基準

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。話者が複数の部屋に定位された場合は、コストの最小の位置を選ぶのが理に適っている。しかしながら、コストは室の形状やマイク設定に依存するため、単純な比較では誤棄却の可能性がある。よって、許容パラメータ η' を導入し、各フレームにおいて、全部屋の中で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。これにより、誤棄却を減らすことができる。

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

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

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。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、毎フレームフラグを推定する。

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

これらのフラグは、5.1 と同様にして統合される。

6 実験条件

6.1 DIRHA コーパスについて

40 マイクにより、同期録音された 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ている。5 部屋 [1] あるが、音源定位と音声区間検出の対象は 2 室 [2] に限定されている。この 2 室には、室中心に 6 個の円形マイクが据え付けられており、加えて、すべての部屋に、2,3 個のマイクから成るマイクアレイが複数個、部屋を取り囲むように壁に取り付けられている。マイクペアは各マイクアレイ内で選択した [3]。

開発セット (**dev**) とテストセット (**test**) が提供されている [4]。どちらにも Real と Simulations のサブセットがあるが、本稿ではその平均の結果のみを示している。Real セットでは、1 部屋に 1 話者のみがおり、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、話者位置は変わらない。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

高さの定位は水平面上での定位ほど重要ではないので、2 次元の定位とした [5]。実験には、48 kHz から 16 kHz にダウンサンプリングした音声を用いた。フレーム長は 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を、マルチチャネル CSP 法 [7] および長いフレーム長 (1 秒) の SRP-PHAT 法 [6][8] と比較した。音源にも大きさがあり、位置を誤差なしで推定することはできないので、“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、precision、recall、F 値の観点から評価した。フレームサイズは 960、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 では、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 hangover 手法 [4] を使った。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)[7]、神経回路網 (NNET) に基づく分類には pyBrain (v.0.31)[8] を使った。特徴量の分散が 1 となるように正規化を行った。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、これを学習した。NNET の隠れ層を 2 層とし、隠れ層のノード数は下から 15,10 とした。Real セットでは、1 部屋だけにしか話者はいないため、いずれかの室の音声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

音源定位精度を比較するため、表 1 の 1 段目には、正解の音声区間検出結果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

[1] Kitchen, Livingroom, Corridor, Bathroom, Bedroom
[2] Kitchen と Livingroom
[3] 別のマイクアレイに属しているマイクとは相関が低すぎるため、有用な情報が得られるとは考えられないため。
[4] すべてのパラメータは開発セットにより調整した。
[5] 評価ツールにおいて、 `-2D` オプションを使った。
[6] http://www.lems.brown.edu/array/tools/srplems.m
[7] http://svmlight.joachims.org/
[8] http://pybrain.org/

Table 1 Localization and speech detection results on the development and test set. 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 SAD, utterance-based criteria are used: Precision (P), Recall (Re), and F value. These results are the average of those for REAL and SIMULATIONS subsets.

Methods			AVERAGE(**dev**)						AVERAGE(**test**)					
SAD	LOC	INT	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F
Oracle	2D-CSP	-	306	870	.540	-	-	-	302	965	.497	-	-	-
	Template		200	817	.658	-	-	-	228	972	.592	-	-	-
	M-CSP		348	1409	.202	-	-	-	-	-	-	-	-	-
	SRP-PHAT		257	957	.515	-	-	-	-	-	-	-	-	-
Sohn	2D-CSP	-	305	794	.559	.414	**.919**	.570	302	904	.517	.441	**.949**	.602
	Template	-	197	732	.673	.414	**.919**	.570	225	870	.613	.441	**.949**	.602
		MIN	197	732	.673	.419	**.919**	.575	225	868	.616	.441	.942	.600
		SVM	197	714	.695	.689	.833	.754	204	920	.602	.700	.762	**.730**
		NNET	193	692	.704	**.799**	.729	**.762**	211	889	.588	**.755**	.657	.703
SKF	2D-CSP	-	302	762	.574	.461	.872	.603	301	823	.557	.462	.881	.606
	Template	-	194	686	.689	.461	.872	.603	225	768	.651	.462	.881	.606
		MIN	194	682	.692	.457	.857	.596	225	766	.653	.461	.870	.602
		SVM	196	663	.707	.694	.826	.754	**203**	798	.647	.664	.686	.675
		NNET	**180**	**642**	**.712**	.753	.733	.743	215	**710**	**.666**	.707	.704	.706

レーム長の SRP-PHAT 法の性能よりも高かった。さらに、計算量も少なかったので、ここでは、2D-CSP 法をベースラインとした。提案のテンプレートに基づく手法は 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 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。

音源定位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、F 値の改善が見られた。

8 おわりに

本報では、家庭環境における音源定位と音声区間検出の問題を、DIRHA コーパスに基づき扱った。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を想定したテンプレートに基づく方法を提案し、有効性を確認した。加えて、音声区間検出器のみでは容易に棄却できない他の部屋における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、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た。

参考文献

[1] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," Proc. of LREC, pp.2629–2634, 2014.

[2] 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," Proc. of SPIE, pp.88–99 (1996).

[3] 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.90–93，2012.

[4] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Processing Letters, **6**, 1–3, 1999.

[5] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," IEICE Trans. on Info. and Sys., **E91-D**, 467–477, 2008.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. on Acoustics, Speech, and Signal Processing, **24**, 320–327, 1976.

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

[8] 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," Proc. of ICASSP, pp.121–124, 2007.