空間特徴と音響特徴を併用する音響イベント検出の検討 *

☆陳軼夫，山田武志，牧野昭二 （筑波大）

1 はじめに

我々が日常的に聞いている環境音には様々な情報が含まれている．近年，環境音から抽出した情報を高齢者の見守りや動画のタグ付けなどに利用するシステムが普及しつつある．これらのシステムで利用される技術の一つとして，音響イベント検出は重要な役割を担う．ここで，音響イベント検出は，入力信号から「いつ，何の音がしたか」を検出するものであり，音の種類としては人の声や転倒音，車の走行音などが挙げられる．

従来の音響イベント検出ではモノラル入力を前提としていた．そのため，複数の音響イベントが重畳している場合に正確な検出が難しいという問題があった．それに対して，最近では DCASE（Detection and Classification of Acoustic Scenes and Event）Challenge のように，マルチチャネル入力を前提とすることが増えてきている．これは，空間情報，すなわち各音源の位置の違いに注目することによって，複数の音響イベントが重畳している場合にも対応できるようになるからである．

従来，空間情報を利用する様々な手法が提案されているが，その中でも DCASE 2017 Task 3 [1] において第 1 位を達成した Adavanne らの手法 [2] は高い検出性能を有している．この手法は，音響特徴量として mbe（mel band energy）あるいは bin-mbe（binaural mel band energy）[3] を入力とし，再帰型畳み込みニューラルネットワーク CRNN（convolutional recurrent neural network）用いて音響イベント検出を行う．更に，空間特徴量として GCC-PHAT（generalized cross correlation phase transform）[4] により求めた空間スペクトログラムを追加した手法 [5] が提案された．本稿では，この手法の検出性能をさらに向上させるために，音響特徴量と空間特徴量の組み合わせ方，及び空間特徴量を入力する CNN の構造を最適化する．

2 空間特徴と音響特徴を併用する音響イベント検出

2.1 全体の流れ

空間特徴と音響特徴を併用する音響イベント検出の全体の流れを Fig. 1 に示す．まず，入力ステレオ信号から空間特徴量（GCC-PHAT）と音響特徴量（mbe，あるいは bin-mbe）を抽出する．そして，それぞれの特徴量を CRNN である検出器に入力し，入力ステレオ信号に含まれる音響イベントの名称とその時間区間を出力する．なお，CNN_S は空間特徴量，CNN_A は音響特徴量を処理する．

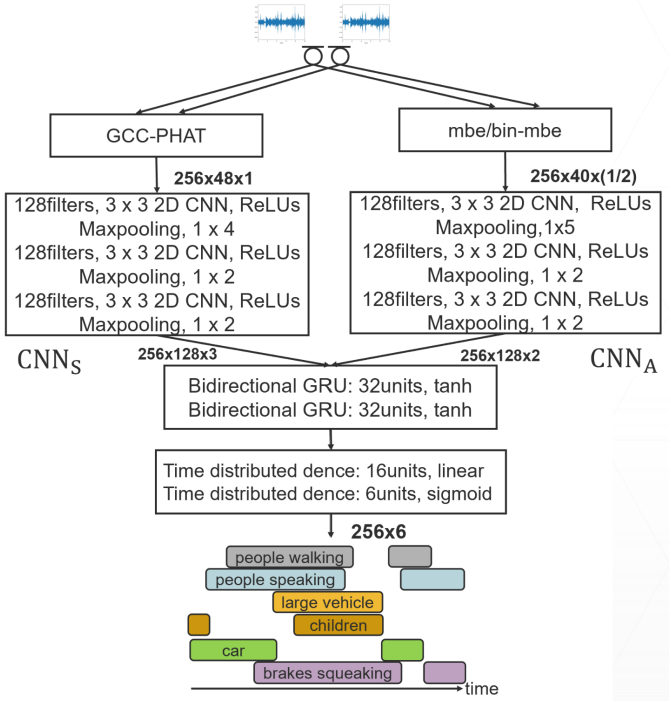

Fig. 1 Process flow

2.2 音響特徴量

音響特徴量には mbe と bin-mbe を用いる．mbe はステレオ信号をモノラル化して抽出した対数メルスペクトログラムであり，音源の方向などの空間情報は含まれていない．一方，bin-mbe はステレオ信号の各チャネルから抽出した 2 チャネル分の mbe である．これは音響情報と空間情報の両方を含むことになる．

2.3 空間特徴量

空間特徴量には GCC-PHAT により求めた空間スペクトログラムを使用する．ある時間区間における GCC-PHAT を式 (1) に示す．

$$GCC(\tau)_{LR} = DFT^{-1}\left(\frac{DFT_L DFT_R^*}{|DFT_L||DFT_R|}\right) \quad (1)$$

* Sound event detection using both spatial and acoustic features. by Yifu CHEN, Takeshi YAMADA, Shoji MAKINO(University of Tsukuba)

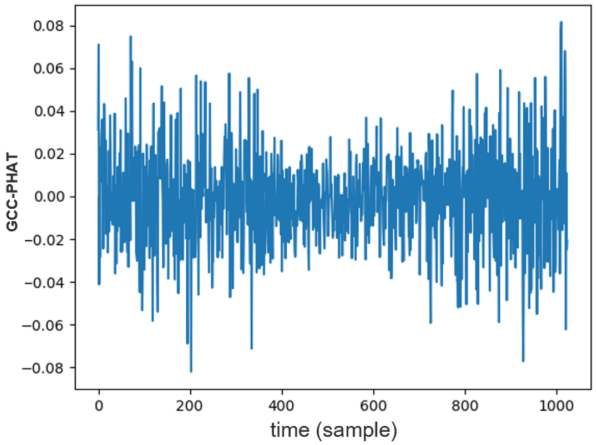

Fig. 2　Example of GCC-PHAT

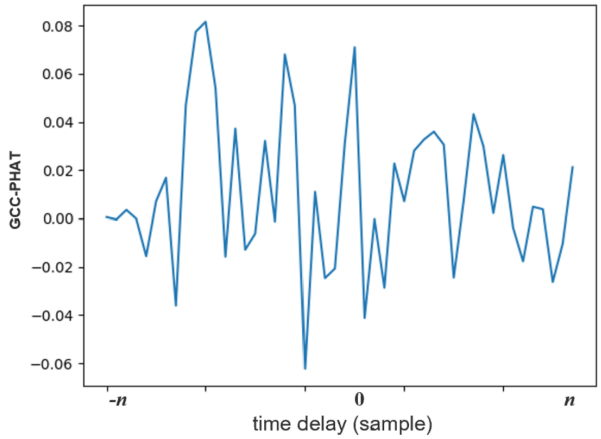

Fig. 3　Rearrangement of GCC-PHAT

ここで，τ は時間サンプル，DFT_L と DFT_R は左右チャネルに対する短時間フーリエ変換，DFT^{-1} は短時間逆フーリエ変換を表す．GCC-PHAT の例を Fig. 2 に示す．ある音源信号が左右のマイクに到達する時間差に対応する時間サンプルにピークが現れる．ただし，実際に起こり得る時間差には上限があり，これは式 (2) に示すように左右のマイク間隔によって決まる．

$$TDOA_{max} = \frac{d}{C} = \frac{n}{Fs} \tag{2}$$

ここで，d はマイク間隔（m），C は音速 (m/s)，n は最大時間差（時間サンプル），Fs はサンプリング周波数（Hz）である．最大時間差 n を超える時間サンプルは意味を持たないため, Fig. 2 の左右から n までの区間をそれぞれ切り取って，Fig. 3 のように再配置したものを GCC-PHAT として用いる．短時間ごとに求めた GCC-PHAT を Fig. 4 のように時間軸に沿って並べて得た空間スペクトログラムを空間特徴量とする．

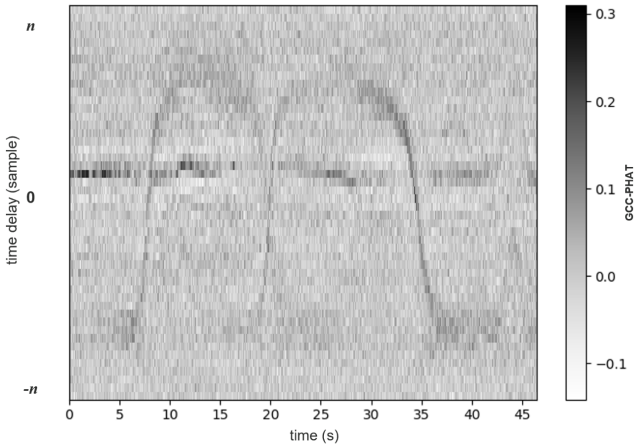

Fig. 4　Spatial spectrogram based on GCC-PHAT

2.4　ネットワーク構造

本手法では検出器に再帰型畳み込みニューラルネットワークである CRNN を用いる．ここで，RNN（recurrent neural network) には BGRU(bidirectional gated recurrent unit）[6] を用いている．

本稿で用いる CRNN の詳細を Fig. 1 を用いて説明する．各 CNN への入力フレーム数は 256 であり，入力信号を 256 フレーム毎に分割して入力する．GCC-PHAT と mbe/bin-mbe の特徴マップのサイズはそれぞれ $256 \times 48 \times 1$ と $256 \times 40 \times (1/2)$ である（フレーム数×特徴量次元×チャネル数）．　CNN の各層では 3×3 の 2 次元フィルタを用いて畳み込みを行う．その後, バッチ正規化を行い, 活性化関数として ReLU 関数を適用する．最終的に 256 フレーム分の特徴量を出力するため, 各畳み込み層の後段ではフレーム次元以外に対して max pooling を行う．BGRU は二層構造であり，各層は前向きと後ろ向き共に 32 ユニットで構成される．なお，活性化関数には tanh を用いる．最終的に二層の dence 層と sigmoid 関数を通して 256×6（フレーム数×音響イベント数）の音響イベントマップを出力する．

Fig. 1 において，$\mathrm{CNN_S}$ からの出力サイズは $256 \times 128 \times 3$ であり，256 はフレーム数，128 は $\mathrm{CNN_S}$ の最後の畳み込み層のフィルタ数，3 は空間分割数である．ここで，空間分割とは，48 次元の GCC-PHAT を 3 回のプーリングを経て 3 次元に集約したこと，すなわちステレオマイクの前方 180 度を 3 つの領域に集約したことを意味する．このパラメータを最適化することにより検出性能を向上できると考えられる．

Table 1　Dataset used for the experiment

	Development dataset	Evaluation dataset
Record environment	Stereo recording(street)	
Microphone	In-ear micphone(L, R)	
Data length	91 mins	25 mins
Sampling rate	44.1 kHz	
Quantization bits	24 bits	
Cross-validation folds	4-folds	None

Table 2　Conditions of features and CRNN

Features	GCC-PHAT	mbe/bin-mbe
Input size	$256 \times 48 \times 1$	$256 \times 40 \times (1/2)$
Frame length	40 ms	
Shift length	20 ms	
Audio block size	256 frames	
Activation function	CNN: ReLU BGRU: tanh	
Optimization method	Adam	
Epoch	500	
Learning rate	0.001	

3　実験

3.1　実験条件

本実験では，DCASE 2017 Task 3 の開発用データセットを用いて，CRNN のハイパーパラメータの調整を行い，評価用データセットを用いて有効性を評価する．Table 1 にデータセットの概要を示す．開発用データセットとして約 91 分, 評価用データセットとして約 25 分のステレオ録音データが提供され, 検出対象となる音響イベントは自動車のブレーキ音, 自動車の走行音, 大型車両の走行音, 子供の声, 会話, 歩行音の計 6 種類である．なお、サンプリング周波数は 44.1kHz, 量子化ビット数は 24 である.

Table 2 は特徴量と CRNN の条件を示している．このデータセットにおいてはマイク間隔 d が不明であるが、インナーイヤー型マイクであるので両耳間隔から d は 15cm〜23cm（サンプル数 n は 20〜30）と見込まれる．よって, 本実験では n を 24 に設定する．その結果，1 フレームあたりの GCC-PHAT は 48 次元となる．音響特徴量については、メル周波数ビン数を 40 とするので、1 フレームあたりの mbe は 40 次元、bin-mbe は 80 次元となる．短時間フレーム分析におけるフレーム長とフレームシフト長はそれぞれ 40 ms と 20 ms である．特徴量時系列を分割する際のブロックサイズは 256 フレームである．検出器の学習における最適化手法には Adam を用いる．学習時のエポック数は 500, 学習率は 0.001 である．これらの条件に関しては文献 [2] で用いられているものと同じである．開発用データで検出器を調整する際は 4-fold クロスバリデーションによる検出精度の平均を用いた．4-fold クロスバリデーションにおけるデータの分割方法は DCASE 2017 Task 3 で指定されたものと同様である．なお、本手法の実装には Keras を用いた．

本稿では, 空間特徴量を処理する $\mathrm{CNN_S}$ の構造が音響イベント検出にどの程度の影響を与えるかを検証するため，$\mathrm{CNN_S}$ の出力が以下になるように各層のフィルタ数とプーリングサイズを調整した．

- $\mathrm{CNN_S}$ A : 128×2（128 フィルタ，空間分割数 2）
- $\mathrm{CNN_S}$ B : 128×3（128 フィルタ，空間分割数 3）
- $\mathrm{CNN_S}$ C : 128×4（128 フィルタ，空間分割数 4）
- $\mathrm{CNN_S}$ D : 32×6 （32 フィルタ，空間分割数 6）
- $\mathrm{CNN_S}$ E : 32×8 （32 フィルタ，空間分割数 8）

ここで，下にあるものほど空間分割数が大きくなっている．なお，空間分割数を大きくする際には，全体の次元が激増するのを避けるためにフィルタ数を減らしている．

3.2　評価基準

本実験では，DCASE 2017 Task 3 における評価尺度である ER（Error Rate）と F-score により評価する．それぞれの範囲は 0〜1 であり，完全に検出できた場合には ER は 0，F-score は 1 となる．ER と F-score はそれぞれ式 (3)(4) のように表される．

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \quad (3)$$

$$F = \frac{\sum_{k=1}^{K} TP(k)}{2\sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)} \quad (4)$$

ここで，k はフレーム番号であり，$S(k)$ は置換エラー数（イベント名を誤った数），$D(k)$ は削除エラー数（存在するイベントを検出できなかった数），$I(k)$ は挿入エラー数（存在しないイベントを誤検出した数），$N(k)$ は存在するイベント総数である．また，$TP(k)$ はイベントが存在すると予測して実際に実在する数，

Table 3　Experimental results

手法	特徴量	Development dataset		Evaluation dataset	
		ER	F-score	ER	F-score
CNN_S A	GCC+mbe	0.481	70.0%	0.790	42.3%
	GCC+bin-mbe	0.474	70.3%	0.785	42.7%
CNN_S B	GCC+mbe	0.470	71.1%	0.775	42.7%
	GCC+bin-mbe	0.477	70.3%	0.783	42.4%
CNN_S C	GCC+mbe	0.481	70.3%	0.786	42.2%
	GCC+bin-mbe	0.474	70.4%	0.797	42.2%
CNN_S D	GCC+mbe	0.471	70.5%	0.788	41.2%
	GCC+bin-mbe	0.490	69.5%	0.801	40.5%
CNN_S E	GCC+mbe	0.466	71.6%	0.768	43.2%
	GCC+bin-mbe	0.496	69.0%	0.808	40.9%
Adavanne らの手法 (DCASE 2017 Challenge Ranking 1st)[2]	mbe	0.55	69.3%	0.791	41.7%
	bin-mbe	0.52	69.1%	0.806	42.9%

$FP(k)$ はイベントが存在すると予測したが実際には存在しない数，$FN(k)$ はイベントが実在しないと予測したが実際には存在する数である．

3.3　実験結果と考察

Table 3 は手法と特徴量の各組み合わせにおける検出精度を示している．表の 1 列目は CNN_S の構造，2 列目は特徴量を示す．また，3，4 列目は開発用データセットにおける ER と F-score，5，6 列目は評価用データセットにおける ER と F-score である．

最も良い結果（赤字）を DCASE 2017 Task 3 の第 1 位の手法（bin-mbe）と比較すると，開発用データセットでは ER は約 0.054 低減し，F-score は約 2.5% 向上した．同様に評価用データセットでも，ER は約 0.038 低減し， F-score は約 0.3%向上した．

最も良い結果は CNN_S E であり，GCC+mbe を用いたときであった．これは，空間分割数をある程度大きくすること，及び音響特徴量には空間情報を含めない方が良いことを示唆している．特に後者の効果は CNN_S D と CNN_S E を比較すれば分かるように空間分割数が大きいときほど顕著である．

4　おわりに

本稿では，空間特徴と音響特徴を併用する音響イベント検出手法における音響特徴量と空間特徴量の組み合わせ方，及び空間特徴量を入力する CNN の構造を最適化し，実験によりその有効性を確認した．今後は GCC-PHAT 以外の空間特徴量の使用について検討する．

謝辞　本研究は JSPS 科研費 19H04131 の助成を受けた．

参考文献

[1] http://dcase.community/challenge2017/task-sound-event-detection-in-real-life-audio.

[2] S. Adavanne, T. Virtanen, “A report on sound event detection with different binaural features,” DCASE 2017 Challenge Technical Report, 2016.

[3] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, “Sound event detection in multichannel audio using spatial and harmonic features,” DCASE 2016 Challenge Technical Report, 2016.

[4] M. Azaria, D. Hertz, “Time delay estimation by generalized cross correlation methods,” IEEE Trans. ASSP, Vol. 32, Issue 2, pp. 280–285, Apr. 1984.

[5] S. Adavanne, P. Pertila, and T. Virtanen, “Sound event detection using spatial features and convolutional recurrent neural network,” Proc. ICASSP 2017, pp. 771–775, Jun. 2017.

[6] K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougres, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN encoder-decoder for statistical machine translation,” Proc. EMNLP 2014, pp. 1724–1734, Oct. 2014.