様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １９ 日現在

機関番号：３４４１２
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ０４９３５

研究課題名（和文）深層学習を利用した実環境下悲鳴検出システムの開発

研究課題名（英文）Development of a noise-robust scream detection system using deep learning

研究代表者
早坂　昇（Hayasaka, Noboru）

大阪電気通信大学・情報通信工学部・教授

研究者番号：５０５５４５７３

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：本研究の目的は，実環境下での使用に耐えうる悲鳴検出システムを構築し普及させることである．まず，実環境下でも高い検出性能を実現するため，深層学習を利用した悲鳴強調手法を提案した．劣悪な雑音環境を想定したシミュレーションにより，高い悲鳴強調効果および悲鳴検出性能が得られることを確認した．次に，小型PCでもリアルタイムに動作させるべく，演算量の削減に取り組んだ．悲鳴が持つ強い周期性を利用することで，従来と比べ1/19のパラメータ数で同等以上の悲鳴強調効果を得ることに成功した．最後に，類似音と悲鳴の識別を容易にするため，上述の悲鳴強調処理を複数回適用する新たな手法を提案し，その有効性を確認した．

研究成果の学術的意義や社会的意義
本研究により，実環境で使用可能な新たな防犯システムの提供を可能にした．これにより，防犯カメラが設置できないプライバシーに配慮する必要がある場面における安全性が大幅に向上する．他にも，演算コストの削減に成功したことから，小型PCやモバイル端末への実装が可能となり，その結果，悲鳴検出システムの応用先が拡大したといえる．例えば，モバイル端末のアプリケーションとして提供されれば，各個人が所有する端末が通報装置となるため，高い犯罪抑止効果が得られる．その他の利用例として，防犯カメラと併用することで，悲鳴発生源に焦点を当て，より鮮明な証拠映像を捉えることも可能となる．

研究成果の概要（英文）：The purpose of this research is to develop and disseminate a scream detection system that can be used in noisy environments. First, we proposed a scream enhancement method using deep learning to achieve high detection performance even in noisy environments. Through simulations assuming noisy environments, we confirmed that high scream enhancement effects and scream detection performance could be obtained. Next, to enable real-time operation on small PCs, we worked on reducing the computational cost. By utilizing the strong periodicity of screams, we succeeded in achieving the same or better scream enhancement effect with 1/19th of the number of parameters compared to conventional methods. Finally, in order to facilitate the discrimination between scream-like sounds and screams, we proposed a new method that applies the scream enhancement process described above two times, and confirmed its effectiveness.

研究分野：音声信号処理

キーワード：悲鳴検出　悲鳴強調　深層学習

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

防犯・セキュリティ分野において，新たな犯罪防止・抑止技術が望まれている．音声を利用した新しい犯罪防止・抑止システムとして，女性が発する悲鳴を自動で検知する悲鳴検出システムの開発に着手してきた．悲鳴を「身の危険を外部に伝達する音声」と定義した上で，50 名分（20～50 代）の悲鳴約 900 発話を収録し，その諸特性について解析を行った（科学研究費補助金／若手研究 B，課題番号 25750137）．その結果，韻律的情報（音量や音高）は通常の発話と大きく異なること，悲鳴発声時の声道特性には共通した特徴が表れることなどを明らかにし，それらの特徴を考慮した悲鳴検出システムも構築した．しかし，実際の利用シーンでは背景雑音の影響により上記特徴量が大きく歪み悲鳴を検出できない問題や悲鳴と類似する音を誤って検出する問題が存在している．また，当該技術は広く普及することでさらなる犯罪抑止効果が期待できるため，上記悲鳴検出システムを小型 PC やモバイル端末へ実装し利便性を向上させる必要がある．

２．研究の目的

本研究の目的は，上記 1.で述べた問題を解決し実環境下での使用に耐えうる悲鳴検出システムを構築すること，および，当該システムを小型 PC やモバイル端末へ実装し普及させることである．それらの目的を達成するため，以下の 3 点に焦点を当て取り組んだ．

(1) 劣悪な環境下における高い検出性能の達成
(2) 小型 PC やモバイル端末への実装に向けた演算量の削減
(3) 悲鳴と類似音の識別精度向上

３．研究の方法

(1) 劣悪な環境下における高い検出性能の達成

本課題を解決するため，深層ニューラルネットワークを利用した音源分離手法である Wave-U-Net1[1]について検討する．Wave-U-Net は，分離対象信号を周波数領域に変換することなく，時間信号を直接入力する End-to-End の音源分離手法である．文献[1]では，楽曲に対するボーカル音および各楽器音を分離する手法として提案されていたが，本研究では，悲鳴と雑音を分離する悲鳴強調手法として導入する．

Wave-U-Net のアーキテクチャを図 1 に示す．ダウンサンプリングブロック（DB）を繰り返し適用するエンコーダ（図 1 左側）とアップサンプリングブロック（UB）を繰り返し適用するデコーダ（図 1 右側）で構成される．UB では，対応する DB からの信号を結合させてから畳み込みを行うことで高精度な分離結果を得ることができる．本研究では，雑音を重畳した悲鳴を入力，重畳前の悲鳴および重畳した雑音を出力として学習を行う．

(2) 小型 PC やモバイル端末への実装に向けた演算量の削減

Wave-U-Net の演算量は，畳み込み層のフィルタ数およびそのサイズ，ブロック数（図 1 の L）によって決まる．(1)の達成に向けて適用した Wave-U-Net のブロック数は，文献[1]で示されている値を用いているが，この値が大きいほど時間領域において長期に渡る情報を捉えることができる．しかし，図 2 に示すように，悲鳴の時間波形は雑音に比べ非常に強い周期性を持つことから，短期の情報だけで強調が可能であると考えられる．本研究では，ブロック数を削減し，悲鳴強調処理の演算量削減を目指す．

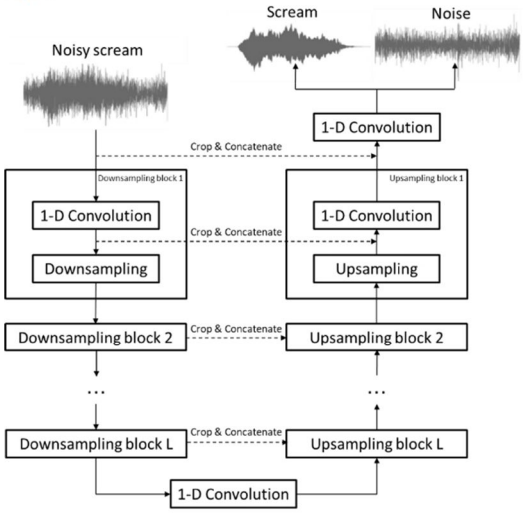


図 1 Wave-U-Net のアーキテクチャ

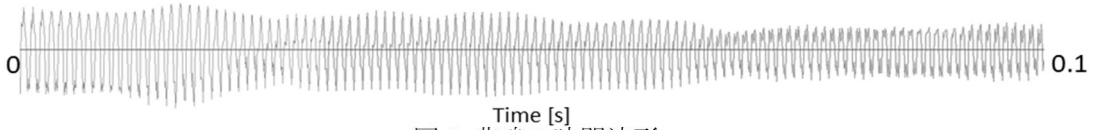


図 2 悲鳴の時間波形

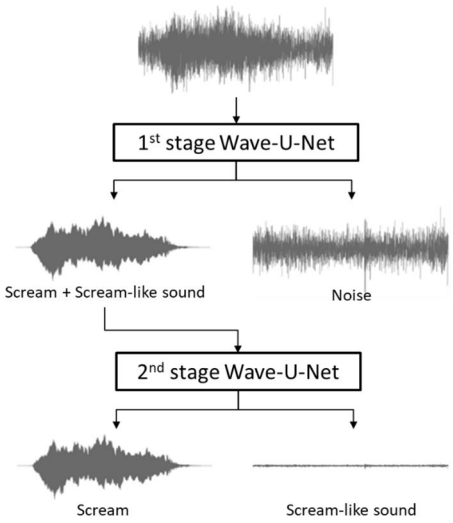


図 3 Two-Stage Wave-U-Net

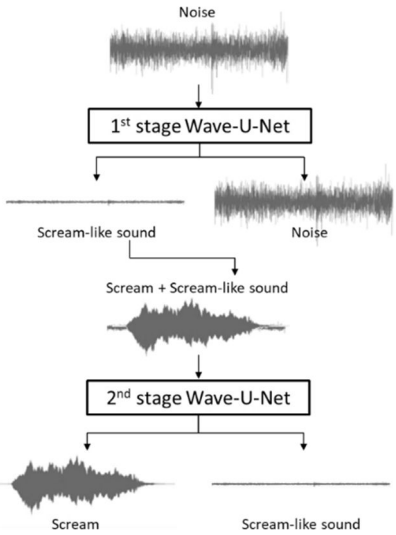


図 4 Two-Stage Wave-U-Net **の学習過程**

(3) **悲鳴と類似音の識別精度向上**

(1)**の悲鳴強調処理では，背景雑音を抑制するように学習されるため，悲鳴と似た周波数成分を持つ信号(類似音)は同時に強調され，誤検出が増大することが懸念される．学習に使用する背景雑音に類似音を加えることも可能であるが，背景雑音と類似音は性質が大きく異なるため，同一のネットワークで学習を行うと十分な分離結果が得られないおそれがある．**

そこで，(1)**で検討した** Wave-U-Net **を** 2 **段で構成する** Two-Stage Wave-U-Net **により解決を図る．**Two-Stage Wave-U-Net **図** 3 **のように構成され，図** 4 **の流れで学習される．**1st Stage Wave-U-Net **では，雑音に含まれる類似音が悲鳴と共に強調されるため，**2nd Stage Wave-U-Net **にて悲鳴と類似音を分離する．**2nd Stage **の** Wave-U-Net **の学習には，悲鳴を含まない雑音のみを入力として** 1st Stage Wave-U-Net **を適用し，悲鳴と抽出された信号を類似音として利用する．**

４．研究成果

(1)〜(3)**に用いるデータセットおよび実験条件について説明する．成人女性** 20 **名から収録した悲鳴** 438 **発話を学習に，学習に用いていない女性** 20 **名から収録した悲鳴** 267 **発話を評価に使用した．また，**1st Stage Wave-U-Net **の学習には** JEITA **騒音データベース**[2]**より「駅」「工場」「幹線道路，交差点」を用い，評価にはそれらに「列車（在来線）」「計算機室（中型）」「空調機（大型）」を加えた．評価用の雑音を重畳した悲鳴は，劣悪な環境を想定し，**Signal-to-Noise Ratio**（以下，**SNR**）が** 0 **または-**5dB **で生成した．**

(1) **劣悪な環境下における高い検出性能の達成の成果**

まず，Wave-U-Net **がどの程度の悲鳴強調性能を有するかを確認する実験を行った．**Wave-U-Net **のパラメータは文献**[1]**で示されている値を用い，比較手法として深層ニューラルネットワークを利用する** SEGAN[3]**，利用しない** MMSE-STSA[4]**を選択した．以下に示す** Segmental SNR **の平均値により評価する．**

$$SSNR_t = 10\log_{10}\frac{\sum|s_t[n]|^2}{\sum|s_t[n]-s_t^e[n]|^2} \quad (1)$$

ここで，$s_t[n]$**は** t **フレーム目の雑音重畳前の悲鳴，**$s_t^e[n]$**は強調後の悲鳴であり，**n **はフレーム内のサンプル番号である．**

結果を表 1 **に示す．表中の「**Noisy**」は悲鳴強調を適用しないときの結果である．表より，学習に利用した既知の雑音環境（実線上部），未知の雑音環境（実線下部）ともに** Wave-U-Net **が最も良好な結果を示し，平均** 14.5 dB **の強調結果が得られた．**

次に，Wave-U-Net **により強調された信号を入力とした悲鳴検出実験を実施する．検出に用いるモデルは混合数が** 32 **の** Gaussian Mixture Model**（**GMM**）であり，特徴量は** 13 **次元の** Mel-Frequency Cepstral Coefficients**（**MFCC**）である**[5]**．悲鳴検出の流れを図** 5 **に示す．雑音を重畳する前の悲鳴および雑音で** GMM **のパラメータを学習し（図** 5(a))**，図** 5(b)**のように学習されたモデルから得られるそれぞれの尤度の差分がしきい値より大きければ悲鳴，そうでなければ雑音と判断する．なお，**SNR = 0 dB **の悲鳴および雑音のみの信号を検出の入力信号とした．**

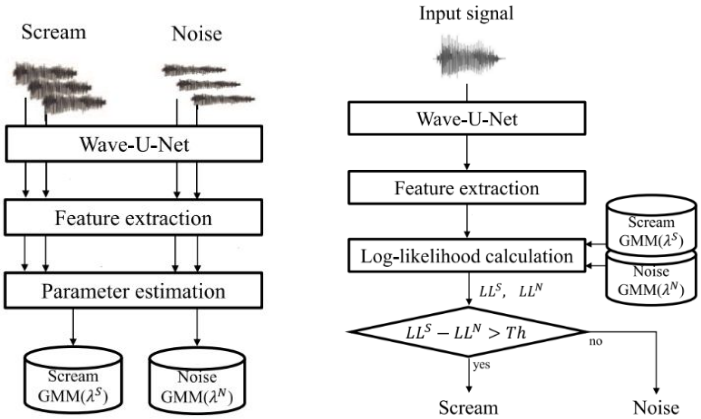
誤検出率を表す FAR **と未検出率を表す** FRR **を以下の式で定義する．**

$$FAR = \frac{\text{誤って悲鳴と検出したフレーム数}}{\text{評価に用いた雑音の総フレーム数}} \quad (2)$$

$$FRR = \frac{\text{未検出の悲鳴数}}{\text{評価に用いた悲鳴の総数}} \quad (3)$$

表 1 悲鳴強調手法の性能比較結果[dB]

Types of noise	Method			
	Noisy	Wave-U-Net	SEGAN	MMSE
Station(0dB)	-1.95	**12.20**	4.63	2.32
Factory(0dB)	-2.06	**10.85**	3.63	1.40
Intersection(0dB)	-1.68	**11.92**	4.89	-0.35
Station(-5dB)	-6.64	**9.49**	2.26	0.26
Factory(-5dB)	-6.80	**8.20**	1.64	0.04
Intersection(-5dB)	-6.36	**9.19**	3.05	-1.91
Train(0dB)	-1.50	**11.16**	3.07	0.67
Computer room(0dB)	-2.12	**10.97**	4.83	1.40
Air conditioner(0dB)	-1.79	**13.01**	4.98	0.28
Train(-5dB)	-6.23	**7.83**	1.38	-2.60
Computer room(-5dB)	-6.87	**8.00**	3.44	0.29
Air conditioner(-5dB)	-6.58	**10.13**	2.53	-2.97



(a) Parameter estimation (b) Detection

図 5 悲鳴検出の流れ

表 2 検出性能評価結果[%]

Types of noise	①	②	③	④
Station	0.447	0.201	0.145	**0.021**
Factory	0.276	0.065	0.016	**0.015**
Intersection	0.271	0.024	0.014	**0.001**
Train	11.583	0.932	0.989	**0.544**
Computer room	0.016	**0.000**	**0.000**	**0.000**
Air conditioner	1.035	0.003	**0.002**	0.007
Average	2.271	0.204	0.194	**0.098**

FAR = FRR となる等誤り率により評価することも可能であるが，悲鳴検出の利用場面を考慮すると，悲鳴を検出できない誤りを表す FRR が 0 であることが望ましいことから，以下に示す *FARmin* により評価を行う．

$$FARmin = \min FAR, \text{ subject to } FRR = 0 \tag{4}$$

以下の 4 通りの比較結果を表 2 に示す．

① Wave-U-Net 未適用
② 検出のみ適用（図 5(b)のみ適用）
③ 検出および悲鳴 GMM の学習に適用（図 5(a)の Scream GMM の学習，図 5(b)に適用）
④ 学習・検出ともに適用

Wave-U-Net 未適用である①の結果は平均 2.27 %と高く，類似音を多く含む「列車（在来線）」では 11 %以上の誤り率であったが, Wave-U-Net を適用する②～④はいずれも大幅に改善しており，悲鳴強調の効果が確認できる．また，④の結果が最良であり，学習・検出どちらに対しても適用することが望ましいことが明らかとなった．

(2) 小型 PC やモバイル端末への実装に向けた演算量の削減の成果

Wave-U-Net のブロックサイズ L を変更し，(1)と同条件で実験を行う．表 3 に強調性能の比較結果，表 4 に各ブロックサイズのパラメータ数を示す．悲鳴は非常に強い周期性を示すことから，ブロックサイズが小さくても十分な強調が可能であるとの予想通り，L = 4，6 で良好な結果となった．また，演算量に強く影響するパラメータ数に関しても，従来と比べ L=4 で約 1/19，L= 6 で約 1/7 程度に抑えることが可能となった．

表 3 ブロックサイズ L の変更による悲鳴強調性能比較結果[dB]

Types of noise	L=12	L=10	L=8	L=6	L=4	L=2
Station(0dB)	12.20	12.52	12.78	**13.03**	12.97	10.84
Factory(0dB)	10.85	11.08	11.36	**11.53**	11.48	9.30
Intersection(0dB)	11.92	12.07	12.33	12.59	**12.69**	10.59
Station(-5dB)	9.49	9.83	10.00	**10.04**	9.70	6.94
Factory(-5dB)	8.20	8.38	8.56	**8.58**	8.15	5.34
Intersection(-5dB)	9.19	9.39	9.55	**9.68**	9.42	6.73
Train(0dB)	11.16	11.82	**12.20**	12.11	12.06	9.69
Computer room(0dB)	10.97	11.07	10.99	11.12	**11.38**	9.39
Air conditioner(0dB	13.01	13.49	**13.72**	13.69	13.66	11.38
Train(-5dB)	7.83	8.38	**8.70**	8.39	8.00	4.99
Computer room(-5dB)	**8.00**	7.98	7.67	7.76	7.81	5.24
Air conditioner(-5dB)	10.13	**10.59**	10.56	10.37	10.25	7.14

表 4 ブロックサイズとパラメータ数

L=12	L=10	L=8	L=6	L=4	L=2
10,263,002	6,180,890	3,354,650	1,553,882	548,186	107,162

表 5 2nd Stage Wave-U-Net の性能評価結果[dB]

Types of noise	L_2 =12	L_2=10	L_2=8	L_2=6	L_2=4	w/o 2nd Stage
Station(0dB)	13.21	13.31	**13.37**	13.33	13.27	12.97
Factory(0dB)	11.77	11.84	**11.90**	11.86	11.81	11.48
Intersection(0dB)	13.04	13.09	**13.13**	**13.13**	13.08	12.69
Station(-5dB)	10.17	10.28	**10.34**	10.24	10.13	9.70
Factory(-5dB)	8.66	8.74	**8.78**	8.64	8.61	8.15
Intersection(-5dB)	9.98	10.06	**10.11**	10.04	9.94	9.42
Train(0dB)	12.06	12.31	12.39	**12.45**	12.35	12.06
Computer room(0dB)	11.38	11.78	11.83	**11.87**	11.79	11.38
Air conditioner(0dB	13.66	14.00	14.12	**14.20**	14.11	13.66
Train(-5dB)	8.30	8.41	**8.45**	8.27	8.18	8.00
Computer room(-5dB)	8.52	**8.61**	**8.61**	8.42	8.28	7.81
Air conditioner(-5dB)	10.78	10.89	**10.98**	10.80	10.66	10.25

(3) 悲鳴と類似音の識別精度向上の成果

図 4 の 2nd Stage Wave-U-Net の学習には，(2)の結果を踏まえ L = 4 のネットワークを 1st Stage として利用し，SNR = 15, 20 dB で類似音を重畳した悲鳴を用いる．(1)の実験と異なる高い SNR に設定した理由は，類似音のスペクトルは雑音と比べ悲鳴と近い性質となるため，SNR = 0，-5dB で重畳すると悲鳴を類似音として除去してしまうためである．

ブロックサイズL_2を変更して得られた結果を表 5 に示す．いずれの雑音環境においても 1st Stage Wave-U-Net と比較し性能が向上し，有効性が確認できる．また，(2)の値よりも大きなブロックサイズ（L_2= 6, 8）で高い性能を示した．悲鳴，類似音ともに強い周期性を持ち，性質が近いそれらの信号を分離するためには，長期に渡る情報も必要になったからだと考えられる．1st Stage が L = 4，2nd Stage がL_2 = 8 の Two Stage Wave-U-Net の総パラメータ数は，従来の Wave-U-Net 1st Stage L =12 の半分以下に抑えられている．

＜引用文献＞

[1] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation,” Proc. of the 19th Int’l Society for Music Information Retrieval Conference (ISMIR), Sep. 2018.
[2] JEIDA Noise Database (ELRA-SD37), http://universal.elra.info/product_info.php?cPath=37_39&products_id=53
[3] S. Pascual, A. Bonafonte, and J. Serrá, “SEGAN: Speech enhancement generative adversarial network,” arXiv:1703.09452, 2017.
[4] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Trans. Acoustics, Speech, and Signal Processing. vol.32, no.6, pp.1109-1121, Dec. 1984.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 HAYASAKA Noboru、KASAI Riku、FUTAGAMI Takuya	４．巻 E107.A
２．論文標題 Noise-Robust Scream Detection Using Wave-U-Net	５．発行年 2024年
３．雑誌名 IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 634～637
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.2023SSL0001	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Riku Kasai, Noboru Hayasaka, Takuya Futagami, Yoshikazu Miyanaga
２．発表標題 Scream Enhancement using Wave-U-Net
３．学会等名 Int'l Workshop on Smart Info-Media Systems in Asia (SISA)（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関