一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2016-132,SIP2016-187,SP2016-127 (2017-03)

［ポスター講演］Rahmonic とメルケプストラムを用いた音響モデルに基づく騒音環境下叫び声検出の性能評価

福森　隆寛†　　中山　雅人†　　西浦　敬信†　　南條　浩輝††

† 立命館大学 情報理工学部 〒 525–8577 滋賀県草津市野路東 1-1-1
†† 京都大学 学術情報メディアセンター 〒 606–8501 京都府京都市左京区吉田二本松町
E-mail: †{fukumori@fc, mnaka@fc, nishiura@is}.ritsumei.ac.jp, ††nanjo@media.kyoto-u.ac.jp

あらまし　本稿では，騒音環境下における Rahmonic とメルケプストラム（Mel-Frequency Cepstrum Coefficients: MFCC）を用いた叫び声検出手法について述べる．MFCC は人間の聴覚特性を考慮したケプストラム係数であり，音韻を特定するための声道特徴量を示している．また Rahmonic は，基本周波数の低調波成分であり，人間の声帯運動に関わる特徴を表現している．これまで，我々は大量の平静音声と叫び声から抽出した MFCC と Rahmonic に基づいて構築した Gaussian Mixture Model（GMM）を用いて叫び声を検出していた．本稿では，この音響モデルを Hidden Markov Model（HMM）や Deep Neural Network（DNN）に拡張して騒音環境下での叫び声検出性能を評価した．評価実験の結果，叫び声の発声機構（声道特性と声帯特性）を MFCC と Rahmonic を用いて効率よく表現できることが確認できた．加えて，ほとんどの騒音環境において音響モデルとして DNN を用いることで GMM や HMM よりも高い叫び声検出性能を達成できた．

キーワード　叫び声検出，騒音環境，Rahmonic，メルケプストラム

Performance evaluation of noisy shouted speech detection based on acoustic model with rahmonic and mel-frequency cepstrum coefficients

Takahiro FUKUMORI†, Masato NAKAYAMA†, Takanobu NISHIURA†, and Hiroaki NANJO††

† College of Information Science and Engineering, Ritsumeikan University. 1-1-1 Nojihigashi, Kusatsu, Shiga, 525–8577, Japan.
†† Academic Center for Computing and Media Studies, Kyoto University. Nihonmatsu-cho, Yoshida, Sakyo-ku, Kyoto, 606–8501, Japan.
E-mail: †{fukumori@fc, mnaka@fc, nishiura@is}.ritsumei.ac.jp, ††nanjo@media.kyoto-u.ac.jp

Abstract　This paper describes a method based on new combined features with mel-frequency cepstrum coefficients (MFCCs) and rahmonic in order to robustly detect a shouted speech in noisy environments. MFCCs collectively make up mel-frequency cepstrum, and rahmonic shows a subharmonic of fundamental frequency in the cepstrum domain. In our previous method, Gaussian mixture models (GMM) is constructed with the proposed features extracted from training data which includes a lot of normal and shouted speech samples. In this paper, evaluation experiments of noisy shouted speech detection were conducted using not only GMM but also hidden Markov models (HMM) and deep neural network (DNN). The results show that MFCCs and rahmonic were effective for representing an utterance mechanism including both vocal tract and vocal cords. In addition, DNN could achieve higher performance in noisy environments than GMM and HMM.

Key words　Shouted speech detection, Noisy environment, Rahmonic, Mel-frequency cepstrum coefficients

1. はじめに

安全安心な暮らしを目指して，カメラなどで撮影した画像情報を用いて異常事態を検知する防犯システムが研究されている [1], [2]．しかしながら，このようなシステムには，カメラの死角で発生した異常事態を検知することが難しいという課題が

残されている．この問題を解決するために，近年はカメラで計測した画像情報以外にマイクロホンで計測した音情報から異常事態を検出するアプローチが注目されている [3]〜[5]．特にカメラの死角の状況を捉えられる音情報を現行の防犯システムに搭載することで，異常事態の検知性能を飛躍的に向上させられると期待できる．

音情報を使って異常事態を検知する手法として，これまでに非日常的な音声である叫び声を検出する手法が数多く提案されてきた [6]〜[9]．しかし，これらの手法には発話内容や評価環境の SNR に大きく依存するという問題があった．このような問題を解決するためのアプローチとして，メルケプストラム（Mel-frequency cepstral coefficients: MFCC）に基づいて構築した GMM（Gaussian Mixture Model）を用いる手法 [10], [11] があり，騒音環境下で頑健に叫び声を検出できることが報告されている．メルケプストラムは音声の発声機構の中でも特に声道情報を重点的に表現しているが，ここで更に声帯情報に関わる音声特徴量を加味しながら叫び声を分析することで叫び声検出性能の向上が期待できる．

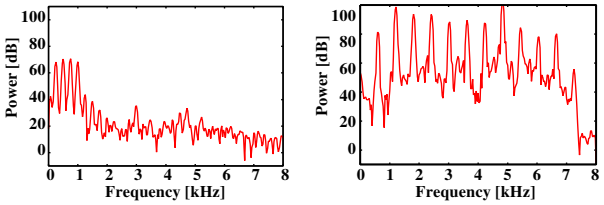
我々は，これまでに Rahmonic と呼ばれる基本周波数の低調波成分が叫び声検出に有効であることを明らかにし，従来のメルケプストラムと併用しながら叫び声を検出する方法を提案した [12]．この手法では，大量の平静音声と叫び声から抽出したメルケプストラムと Rahmonic に基づいて構築した GMM を用いて叫び声を検出していた．本稿では，検出に用いる音響モデルを GMM だけでなく，HMM（Hidden Markov Model）や DNN（Deep Neural Network）に拡張し，それぞれの音響モデルに対して騒音環境下における叫び声の検出性能を評価する．

2. Rahmonic とメルケプストラムを用いた叫び声検出

2.1 音声特徴量（メルケプストラム・Rahmonic）

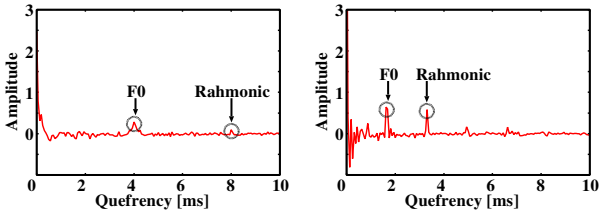
我々は，これまでに Rahmonic とメルケプストラムを用いた叫び声検出法を提案した [12]．メルケプストラムは，人間の聴覚特性を考慮したケプストラム係数であり，音声認識では音韻を特定するための声道特徴量として用いられている [13]．一方，Rahmonic は，基本周波数の低調波成分であり，人間の声帯運動に関わる特徴を表現する [14]．そして，従来研究 [12] において，これらの音声特徴量が平静音声と叫び声で異なることが報告されている．

ここで，図 1 と図 2 に平静音声と叫び声に対する対数パワースペクトルとケプストラムを示す．まず対数パワースペクトルに着目すると，図 1(a) の平静音声よりも図 1(b) の叫び声の調波成分が強調されて表れていることが確認できる．またケプストラムにおいても，図 2(a) の平静音声には顕著に表れなかった Rahmonic を図 2(b) の叫び声では明確に確認することができる．このように周波数領域やケプストラム領域でも平静音声と叫び声の間に差異があることからも，メルケプストラムや Rahmonic を用いることで高精度に叫び声を検出できる可能性があると考えられる．



(a) 平静音声　(b) 叫び声

図 1　平静音声と叫び声の対数パワースペクトル



(a) 平静音声　(b) 叫び声

図 2　平静音声と叫び声のケプストラム

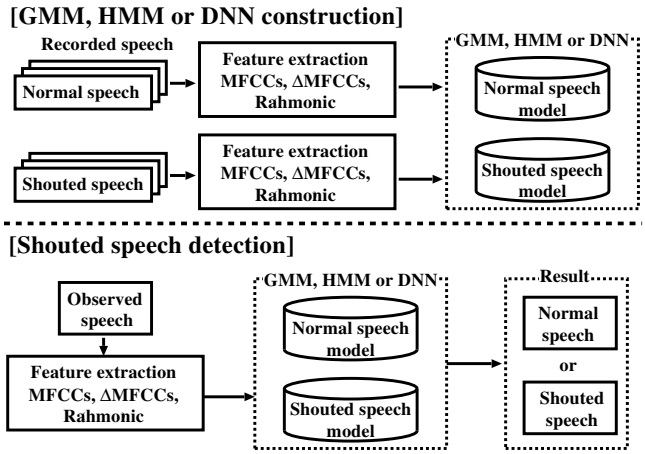


図 3　叫び声検出アルゴリズムの概要

2.2 検出アルゴリズム

図 3 に叫び声の検出手順を示す．叫び声を検出する方法として，はじめに予め収録した平静音声と叫び声から抽出した Rahmonic とメルケプストラムを用いて音響モデルを構築する．次に，実際の評価環境で収録した観測音声から Rahmonic とメルケプストラムを抽出し，これらの音声特徴量と学習した音響モデルを用いて観測音声を平静音声と叫び声のいずれかに分類する．

従来手法は音響モデルとして GMM を利用しているが，本稿では叫び声検出に用いる音響モデルを従来の Gaussian Mixture Model (GMM) から Hidden Markov Model (HMM) や Deep Neural Network（DNN）に拡張して，それぞれの音響モデルが叫び声検出に与える影響を評価する．GMM は観測音声に対する平均的な音声特徴量を用いて叫び声をモデル化している．HMM は音声特徴量の時間的変化を表現できる音響モデルであり，叫び声は平静音声と比べて発話時間やエネルギーの時間変動が異なること [15] から，HMM を用いることで音声特徴量の時間構造も考慮することで叫び声検出の性能改善が期待できる．そして，DNN はニューラルネットワークの 1 つであり，ネット

表 1　実 験 条 件

Training data	Female speaker: 400 samples Male speaker: 400 samples
Testing data	Female speaker: 100 samples Male speaker: 100 samples
Sampling	16 kHz / 16 bit
Acoustic feature	12 orders MFCC 12 orders ΔMFCC 1 order Rahmonic
Acoustic model	1. GMM 2. HMM (3 states) 3. DNN
Noise	White noise, Speech bubble [18]
SNR	0, 10, 20, ∞ dB
Frame length	25 ms (Hamming window)
Frame shift	10 ms

ワーク内で深い層構造を有する．特に入力層を音響特徴量（本稿の場合，メルケプストラムや Rahmonic），出力層を発話様式（平静音声と叫び声）として対応付けることで，DNN を叫び声検出のための音響モデルとして使用することができる．またネットワークに入力された音響特徴量に対して重み付けを行いながら出力層まで伝搬する過程は，評価環境に依存せずに叫び声検出に有効な特徴を重点的に抽出できると考えられる．

3. 評 価 実 験

3.1 実 験 条 件

本実験では，クリーン音声 (男女各 400 発話) に雑音を 4 種類の SNR（∞, 20, 10, 0 dB）で加算した学習音声を用いて性別依存のマルチコンディション音響モデル (GMM，3 状態の HMM，DNN）を構築した．GMM と HMM の混合数は，8 種類（1，2，4，8，16，32，64，128）を用いて評価を行った．また GMM と HMM の構築には HTK [16] を，DNN の構築には Kaldi [17] を用いた．DNN で用いる各音響特徴量の統合フレーム数は，1 フレーム（現在フレームのみ），7 フレーム（前後 3 フレームを含む），11 フレーム（前後 5 フレームを含む）の 3 種類，隠れ層は 3 層（各層の素子数は 20）とした．また発話様式の識別では，話者オープンテストを想定して音響モデルの学習で用いた音声とは異なる話者音声を用いた．音声特徴量として，メルケプストラム単体，Rahmonic 単体，メルケプストラムと Rahmonic 併用の 3 種類とした．雑音は，NOISEX-92 [18] よりホワイトノイズとスピーチバブル雑音を用いた．評価指標として，全ての平静音声と叫び声の内，正しく発話様式が識別された音声サンプル数の割合 (識別率）[%] を用いた．また本実験で用意できた評価音声が少量であることを考慮して，今回は 5 分割交差検定を実施した．

3.2 実 験 結 果

図 4 と図 5 に音響モデルとして GMM と HMM を用いたときの混合数別の平均識別率を示す．まず図 4 の GMM を用いた結果では，メルケプストラムと Rahmonic を用いて混合数 128 の音響モデルを構築した条件，そしてを図 5 の HMM の結果で

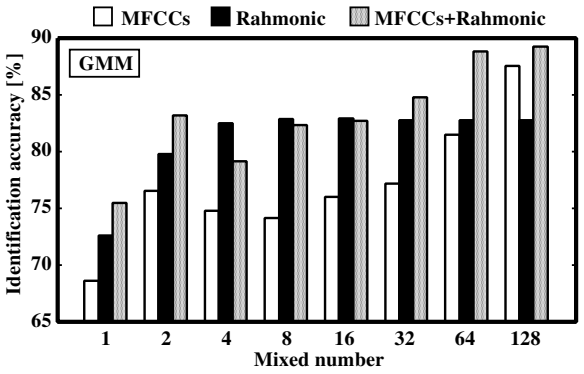


図 4　平均識別率（音響モデル：GMM）

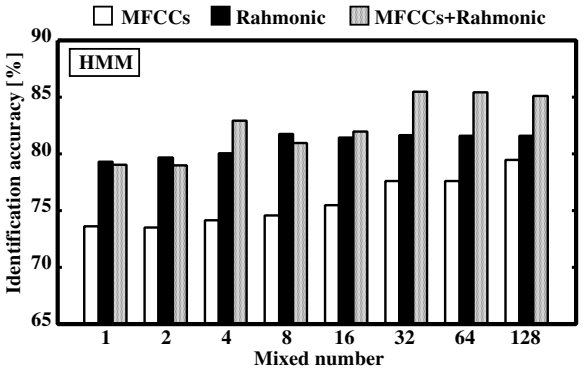


図 5　平均識別率（音響モデル：HMM）

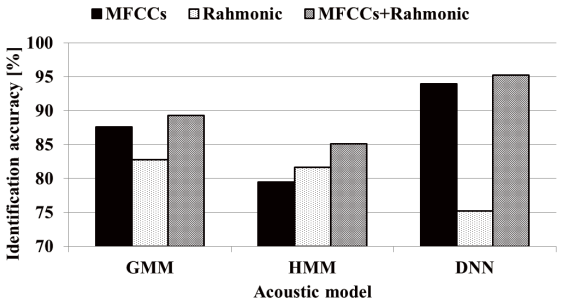


図 6　音響モデルごとの平均識別率

は，メルケプストラムと Rahmonic を用いて混合数 64 の音響モデルを構築した条件において高い識別性能を達成することができた．このことより，メルケプストラムと Rahmonic を併用することで叫び声の特徴を効果的に表現できていることが確認できた．また音響モデルの性能を比較すると，GMM が HMM よりも最高の識別率を上回った．これは HMM では音声特徴量の時間的変化を正確に表現できていなかったことを示しており，今後は平静音声と叫び声の特徴量に対する具体的な時間的変化を分析する必要がある．

次に表 2〜3 に DNN を用いた発話様式の識別率（男女別），図 6 に音響モデルごとの平均識別率を示す．表中の太字は各環境における最高識別率を，そして「M」,「R」,「M+R」は，それぞれメルケプストラム，Rahmonic，両特徴量併用の結果を示す．まず DNN を用いた発話様式の識別率に着目すると，雑音の SNR に関係なく全ての環境において 90 %以上の識別率を達成することができた．特にメルケプストラムと Rahmonic を

表 2　女性話者の識別率 [%] (音響モデル：DNN)

Number of	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
input frames	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
1 frame	92.3	69.5	90.5	94.4	59.4	93.7	92.5	68.4	91.6	90.8	87.4	92.2
7 frames	95.3	68.0	95.6	96.4	65.9	**96.9**	95.3	75.7	95.7	94.2	93.4	95.7
11 frames	95.1	68.6	**95.7**	96.5	66.6	96.8	95.5	75.7	**95.8**	**96.4**	94.9	95.5

*M: MFCCs, R: Rahmonic, M+R: MFCCs and Rahmonic

表 3　男性話者の識別率 [%] (音響モデル：DNN)

Number of	SNR=∞ dB			SNR=20 dB			SNR=10 dB			SNR=0 dB		
input frames	M	R	M+R	M	R	M+R	M	R	M+R	M	R	M+R
1 frame	84.9	73.7	93.4	90.5	70.4	94.8	90.3	73.8	94.3	82.7	61.7	87.1
7 frames	89.4	80.7	**95.1**	93.8	79.1	**96.1**	93.3	80.9	95.7	87.5	54.9	**92.2**
11 frames	91.2	81.3	94.5	94.5	80.1	96.0	94.1	80.7	**96.1**	88.6	53.4	91.5

*M: MFCCs, R: Rahmonic, M+R: MFCCs and Rahmonic

併用して DNN を構築した場合，白色雑音（SNR=0 dB）を除く全ての環境において最も識別率が高かった．この結果からも発声機構（声道特性と声帯特性）をそれぞれメルケプストラムと Rahmonic を用いて効率よく表現できていると考えられる．

そして音響モデルごとの結果に着目すると，GMM や HMM を用いた平均識別率は全て 90 %を下回ったのに対して，メルケプストラムと Rahmonic を併用して DNN を構築した場合の平均識別率が 95 %以上を達成した．これは DNN が GMM や HMM と比較して雑音の影響を受けずに叫び声検出に有効な特徴を重点的に抽出できたためだと考えられる．以上のことより，メルケプストラムと Rahmonic に基づいて構築した DNN が叫び声検出に有効であることを確認できた．

4. おわりに

本稿では，Rahmonic とメルケプストラムを用いた叫び声検出において，発話様式の識別に用いる音響モデルを従来の GMM から HMM や DNN へ拡張して，音響モデルの違いが叫び声の検出性能に与える影響を評価した．実験結果より，どの音響モデルを用いた場合でも Rahmonic とメルケプストラムを併用することで高い叫び声検出性能を実現することができた．さらに叫び声検出に有効な特徴を重点的に抽出できる DNN を用いることで雑音の種類や SNR に関係なく 90 %以上の叫び声検出性能を達成した上に，従来の GMM や HMM よりも叫び声の検出性能が改善した．今後は，実環境を想定して雑音だけでなく残響も混入する環境での叫び声検出評価に取り組む計画である．

謝辞　本研究の一部は，科研費（16K16094）の研究助成を受けた．

文　献

[1] W. Yao-Dong, T. Takeshi, and I. Idaku, “HFR-video-based machinery surveillance for high-speed periodic operations,” *Journal of System Design and Dynamics*, vol. 5, no. 6, pp. 1310-1325, 2011.

[2] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, “Scream detection for home applications,” *5th IEEE Conference on Industrial Electronics and Applications*, pp. 2115-2120, 2010.

[3] M. Cowling, “Comparison of techniques for environmental sound recognition,” *Pattern Recognition Letter*, vol. 24, no. 15, pp. 2895-2907, 2003.

[4] K. M. Kim, J. W. Jung, S. Y. Chun, and K. S. Park, “Acoustic intruder detection system for home security,” *IEEE Transaction on Consumer Electronics*, vol. 51, no. 1, pp. 130-138, 2005.

[5] K. Hayashida, J. Ogawa, M. Nakayama, T. Nishiura, and Y. Yamashita, “Multi-stage identification for abnormal/warning sounds detection based on maximum likelihood classification,” *ICA2013*, PaperID:1pSPb4, 2013.

[6] J. L. Rouas, J. Louradour, and S. Ambellouis, “Audio events detection in public transport vehicle,” *IEEE Intelligent Transportation Systems Conference*, pp. 733-738, 2006.

[7] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, “Audio based event detection for multimedia surveillance,” *ICASSP 2006*, pp. 813-816, 2006.

[8] S. Ntalampiras, I. Potamitis, and N. Fakotakis, “An adaptive framework for acoustic monitoring of potential hazards,” *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[9] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” *IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 21-26, 2007.

[10] J. Pohjalainen, P. Alku, and T. Kinnunen, “Shout detection in noise,” *ICASSP 2011*, pp. 4968-4971, 2011.

[11] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, “Scream detection for home applications,” *Industrial Electronics and Applications 2010*, pp. 2115-2120, 2010.

[12] 柿野 直人, 福森 隆寛, 中山 雅人, 西浦 敬信, 南條 浩輝, “Rahmonic とメルケプストラムを用いた叫び声検出の検討,” 日本音響学会 2013 年秋季研究発表会, pp. 169-170, 2013.

[13] J. Benesty, M. M. Sondhi, and Y. Huang, “Springer handbook of speech processing,” *Springer*, 2008.

[14] A. M. Noll, “Cepstrum Pitch Determination,” *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 203-309, 1967.

[15] C Zhang and J.H.L Hansen “Analysis and classification of speech mode: whispered through shouted,” *INTERSPEECH 2007*, pp. 2289-2292, 2007.

[16] HTK Software Toolkit, http://htk.eng.cam.ac.uk/

[17] Kaldi, http://kaldi-asr.org/doc/index.html

[18] A. Varga and H.J.M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247-251.