Vol.2019-MUS-123 No.33
Vol.2019-SLP-127 No.33
2019/6/22

i-vectorに基づく賑わい音の推定方式の検討

呉セン陽[1,a)]　朝田興平[2]　原直[1,b)]　阿部匡伸[1,c)]

概要：本報告では，i-vector に基づく賑わい音の推定方式を提案する．本研究で行う賑わい音の推定は，賑わっている状況という音場面の推定である．賑わっている状況とは，商業施設や観光地の周辺や催し物などに多くの人が集まり，活気のある状況であり，このような状況で発生する音を本研究では「賑わい音」と定義する．環境音から賑わい音を検出することにより，街の賑わっている場所を検出することができると考えられる．本報告では i-vector に基づく話者識別手法を応用した賑わい音の推定方式を提案する．提案方式と GMM-UBM に基づく賑わい音の推定方式とを比較するとともにパラメータを変化させて提案方式の性能評価を行った．i-vector に基づく賑わい音の推定方式は GMM-UBM に基づく賑わい音の推定方式よりも最高性能が高い結果を得られた. パラメータを変化させた評価実験の結果では，MFCC の次元数 59，混合数 128，TvRank 数 600 の場合が最も良く，F 値の最大値 0.7899 となったことから，賑わい音の識別性能が優れていることが明らかとなった．

A study of estimation method for hubbub sound based on i-vector

1. はじめに

人間の生活を取り巻く環境音は，多種多様な音源やその場の状況や雰囲気など，多くの情報を含んでいる．環境音は音源の種類そのものの情報以外にも，同じ種類の音源の数や音が発生した状況の情報を含んでいる．例えば，人が歩く足音を聞くと，歩いている人が一人か複数かという数の情報や，歩いている路面が濡れているのか，砂利なのか，という音の発生状況の情報が得られる．このように環境音を解析することで，環境音が収録された付近の様子を推測することができる．そのため，環境音を分析してその中に含まれている情報を取り出し，活用する研究が行われている．

本研究で行う賑わい音の推定は，賑わっている状況という音場面の推定である．賑わっている状況とは，商業施設や観光地の周辺や催し物などに多くの人が集まり，活気のある状況であり，このような状況で発生する音を本研究では「賑わい音」と定義する．環境音から賑わい音を検出することにより，街の賑わっている場所を検出することができると考えられる．

本研究では，環境音から周囲の賑わい度を推定し，図 1 のように地図上に可視化するようなシステムの実用化を目指す．この手法は i-vector に基づく話者識別手法を用いて，スマートデバイスで収録した環境音を賑わい音識別器に入力し [1]，賑わい音の識別を行う．

我々の研究の 1 つの特徴としては，スマートフォンで収集した音を利用していることである．一般に環境音の収録では，収音機器をある場所に設置して音が録音されていた．スマートフォン普及にしたがって，スマートフォンは多くの人が持っている．したがって，いろいろなところで音を収音できる．「スマートフォンの出現で，環境音が収録しやすくなったので，これを利用して賑いを推定する」と言う点が，新しい着眼点である．

一般的な音声認識システムとして，まだ多くの欠陥がある．たとえば，話者のロバストネス（robustness）をうまく解決することはできない．現在のところ、ｉ-vector に基づくシステムは最も広く使用されている話者認識システムである [2]．i-vector は話者認識において高い性能を達成している発話表現の一つで，発話ごとに算出した混合ガウス分布（GMM）に因子分析を適用することで得られる．このとき得られるベクトルは，LDA (Linear discriminant analysis)[4] や WCCN (Wihin class covariance normaliza-

[1] 岡山大学　大学院ヘルスシステム統合科学研究科
[2] 岡山大学　大学院自然科学研究科
[a)] pqzq2tp8@s.okayama-u.ac.jp
[b)] hara@okayama-u.ac.jp
[c)] abe-m@okayama-u.ac.jp

Vol.2019-MUS-123 No.33
Vol.2019-SLP-127 No.33
2019/6/22

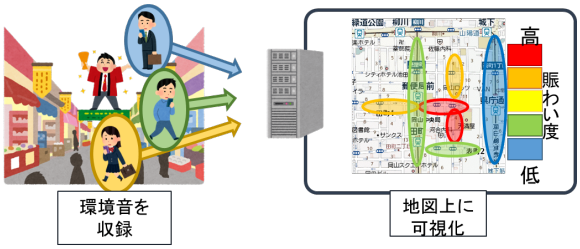

図 1 賑わい度可視化

tion)[6]，PLDA (probabilistic LDA)[7] といった統計処理を適用することで，セッションやチャネルの違い等，話者に依存しない変動の影響を低減できることが知られている．また，雑音環境においても，LDA[4] や PLDA[7] で用いる射影行列を雑音を含んだ音声で学習することで，高精度な話者照合が実現できることが示されている [8][9]．

i-vector を用いた識別方式は，話者識別の分野においても高い性能を示しているため，本研究の目的である環境音からの賑わい音の識別でも高い識別性能が得られる事が期待できる．そこで，i-vector に基づく話者識別手法を応用した賑わい音の推定方式を提案する．

本報告では，提案方式の有効性を示すために，観光地で収録された環境音を用いて賑わい音の推定実験を行った．提案方式の使用により，i-vector の賑わい音識別器システムを構築することは可能であると考えられる．従来方式の GMM-UBM に基づく賑わい音の推定方式と提案方式を比較することにより，提案方式の有効性を示す．

本報告は以下の通りの構成である．2 章は GMM-UBM に基づく賑い音推定について述べる．3 章では提案方式について述べる．4 章では実験データについて述べる．5 章では実験と考察について述べる．6 章では結論と今後の課題について述べる．

2. GMM-UBM に基づく賑わい音推定

2.1 GMM に基づく音響モデル

GMM は複数の正規分布の重ね合わせによって特徴ベクトルの分布を表現する統計的なモデルのひとつであり，環境音の特徴を表現できると想定する．GMM M によってモデル化される，ある音データの尤度 L は次の式で計算される．

$$L(\boldsymbol{o}; M) = \sum_{m} \lambda_m f(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\sigma}) \qquad (1)$$

ここで，$\boldsymbol{o}$ は，ある音データの特徴ベクトル，M は学習データにより学習された GMM，λ_m は分布の構成要素の混合重みであり，GMM の共分散行列は対角行列とした．

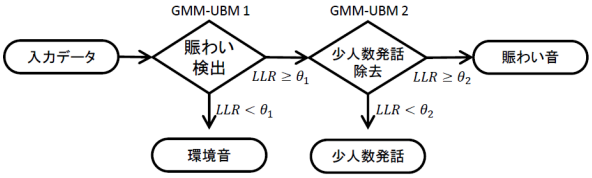

図 2 GMM-UBM に基づく賑い音識別のフローチャート

2.2 GMM-UBM に基づく賑わい音識別

UBM は識別したい空間全体を表すモデルである．例えば，話者照合システムであれば，話者の平均的な特徴を表すためにはあらゆる話者の声を含むことが理想であるため，不特定多数の話者の声で学習する．話者照合システムでは人の声の分布がガウス分布に従うと仮定し，GMM により UBM を作成した GMM-UBM が用いられる [3]．

GMM-UBM に基づく賑い音識別 [1] は，GMM-UBM を環境音に適用し，様々な環境音の中から賑わい音を検出する手法である．賑わい音を大勢の人が集まっていて，さらに多くの人が話していることにより，多くの人の声が重複している状態の音と定義している．このことから賑わい音は人の声としての性質を持っていると考えられる．そのため，賑わっていない状況で収録された音に少人数の話し声が含まれている場合，その音を賑わい音であると誤識別する可能性がある．そこで，1 段目では賑わい音と環境音の識別を行い，2 段目では賑わい音と少人数発話の識別を行う．フローチャートを図 2 に示す．

GMM-UBM に基づく賑い音識別の学習と識別の手順を述べる．

(1) 全学習データから M_{ubm} を学習する．
(2) 特定話者のモデル M_{spk} を学習する．
(3) M_{spk} の学習には M_{ubm} の分布を利用し，話者ごとの学習データを再学習させることによってモデルを作成する．
(4) 入力データ o と各モデル M の対数尤度比 (Log-likelihood ratio; LLR) を用いて，M_{ubm} と尤度を比較する．
(5) 入力データ o が与えられたとき，o に対するモデル M との対数尤度比 $LLR(o; M)$ を計算する．

o に対するモデル M との対数尤度比 $LLR(o; M)$ は以下の式で計算する．

$$LLR(\boldsymbol{o}; M_{spk}) = \log \frac{L(M_{spk}; \boldsymbol{o})}{L(M_{ubm}; \boldsymbol{o})} \qquad (2)$$

o が各話者のクラス spk に属する (+1) か否 (-1) かを判定する識別器 $C_{spk}(\cdot)$ は，次の式により計算される．

$$C_{spk}(\cdot) = \begin{cases} +1 & if\ LLR(\mathrm{o}; M_{spk}) \geq 0 \\ -1 & if\ LLR(\mathrm{o}; M_{spk}) < 0 \end{cases} \qquad (3)$$

本研究では，式に，検出器の感度を調整するパラメータ θ を導入する．θ によって識別を行う際の式を以下に示す．

$$C_{spk,\theta}(o) = \begin{cases} +1 & if\ LLR(\mathrm{o}; M_{spk}) \geq \theta \\ -1 & if\ LLR(\mathrm{o}; M_{spk}) < \theta \end{cases} \quad (4)$$

2.2.1 賑わい音と環境音の識別

1 段目の賑わい検出では，様々な環境音の中から賑わい音をおおまかに検出する．全データを表現する M_{UBM} は，賑わい音と環境音を用いて GMM 学習によって作成する．賑わいモデル M_{bubble} は，M_{ubm} と賑わい音を用いて GMM 再学習によって作成する．識別は，式により，評価データ o の対数尤度比 LLR($\boldsymbol{o}$; M_{bubble}) を計算して行う．LLR($\boldsymbol{o}$; M_{bubble}) が閾値 θ_1 より大きければ賑わい音，θ_1 より小さければ環境音として分類する．

2.2.2 賑わい音と少人数発話の識別

2 段目の少人数発話除去では，1 段目によって賑わい音として検出されたデータから少人数発話を除去する．少人数発話とは，「賑わいではないが少人数が会話している環境音」である．識別は 1 段目と同様に LLR($\boldsymbol{o}$; M_{bubble}) を計算して行う．LLR($\boldsymbol{o}$; M_{bubble}) が閾値 θ_2 より大きければ賑わい音，θ_2 より小さければ少人数発話として分類する．

3. 提案方式

3.1 i-vector に基づく賑わい音識別

i-vector に基づく話者識別手法では，因子分析を用いて GMM スーパーベクトルから話者ごとに固有の特徴を抽出し，得られた特徴を登録話者の特徴と比較することで話者を識別する．GMM スーパーベクトルは，時系列データである発話をベクトル空間上の一点として表現するものである．i-vector も，この GMM スーパーベクトルを基礎としている．

提案方式でも，i-vector を用いることで，GMM-UBM による賑わい音識別 [1] よりも高い性能が得られることが期待される．図 3 は提案方式である．提案方式によって処理の流れを示す．

(1) 賑わい音と環境音の学習データから GMM 学習により M_{UBM} を作成する．
(2) 作成した M_{UBM} と学習データを用いて全変動行列 T を計算する．
(3) M_{UBM} と T を用いて，賑わい音と環境音の学習データから賑わい音の i-vector W_{env}，環境音の i-vector $W_{bubbble}$ を抽出する．
(4) 識別の際には，入力データに対して M_{UBM} と T を用いて入力データの i-vector W_o を抽出し，W_o と W_{env}，W_{bubble} の対数尤度比 LLR_{env}，LLR_{bubble} をそれぞれ計算する．
(5) 賑わい音を判定する．

識別の際には，入力データ o に対して M_{UBM} と T を用いて入力データの i-vectorw_o を抽出し，w_o と w_{env}，w_{bubble} の対数尤度比 LLR_{env}，LLR_{bubble} をそれぞれ計算する．賑わい音 (+1) か否 (−1) かは以下の式により判定する．

$$C_{bubble,\theta}(o) = \begin{cases} +1 & if\ LLR_{bubble} - LLR_{env} \geq \theta \\ -1 & if\ LLR_{bubble} - LLR_{env} < \theta \end{cases} \quad (5)$$

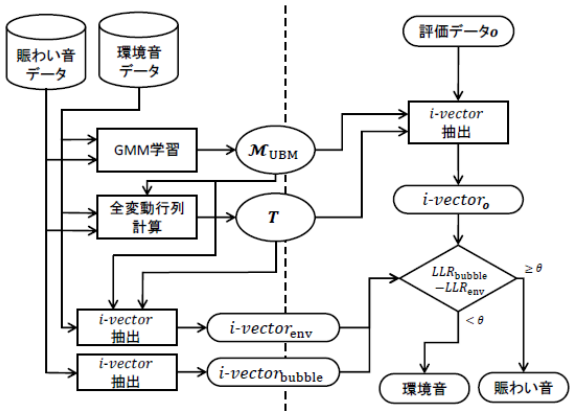

図 3 提案方式

表 1 収録条件

	環境音収録	賑わい音収録
日程	2014/6/25 ～ 2015/1/31	2016/1/10 ～ 7/16
場所・行事	住宅街	成人式
	駅の近辺	センター試験
	商店街	入学式
		花火大会
データ数	7,499	263

3.2 音響特徴量の抽出

本研究で使用されている特徴量は MFCC(Mel-Frequency Cepstral Coefficients) である．MFCC の分析次数は大きくすることで，環境音からより詳細な情報が得られると考えられる．

4. 実験データ

4.1 学習に用いるデータ

学習に用いるデータは文献 [1][11] で環境音と賑わい音を収録したデータである．環境音収録では，車の音や鳥の声などの賑わいを感じない音を主に収録した．賑わい音収録では，祭りや成人式などの行事で収録を行い，賑わいを感じる音を収録した．収録方法では収録者端末を手に持った状態で収録を行った．環境音収録，賑わい音収録の収録音はそれぞれ 10 秒，15 秒であり，合計で約 20 時間分のデータが収録された．収録条件を表 1 を示す．

環境音は全部で 7,499 ファイル，賑わい音は全部で 263 ファイルである．収録端末には Google Nexus 7 を用いて，収録アプリにはオトログマッパーを用いた [11].

4.2 評価データ

評価データでは 2017 年 9 月 23 日に，催し物のない休日の収録されたデータ．倉敷美観地区は有名な観光地であるため，催し物がなくても休日になると多くの人で賑わう．午前 7 時から午後 6 時 15 分までの収録時間の間に，収録により集まったデータは 1,139 データである [11].

5. 実験

5.1 評価実験条件

音響モデルの特徴量には，3.2 節で述べた MFCC を用いた．分析帯域幅は 0 Hz から 16 kHz とした．抽出時のフレームサイズは 25 msec，フレームシフトは 10 msec，分析窓の窓関数はハミング窓とした．学習には HTK3.4.1 を用いる [10].

5.1.1 提案方式と GMM-UBM に基づく賑わい音の推定方式の比較実験条件

MFCC とその一次差分を 12 次元，さらにエネルギーの一次差分の計 25 次元を用いた．i-vector に基づく賑わい音識別器は，特徴量次元数と GMM の混合数を GMM-UBM に基づく賑わい音識別器と揃えるため，特徴量次元数は 25，混合数は 256 として構築した．このとき，TvRank は 100 とした．

5.1.2 提案方式による実験条件

i-vector の抽出に用いる M_{UBM} と T の作成条件について述べる．M_{UBM} の作成のための音響特徴量には MFCC 次元を用いた．本報告では，38，59 次元の MFCC を用いる．

(1) MFCC とその一次 · 二次差分を各 12 次元，さらにエネルギーの一次 · 二次差分の計 38 次元.
(2) MFCC とその一次 · 二次差分を各 19 次元，さらにエネルギーの一次 · 二次差分の計 59 次元.

特徴量次元数は 38，59，混合数は 128，256，512，1024 として構築する．T のランクは 100，400，600 とした．

5.2 性能評価尺度

本研究の評価指標として F 尺度を用いる．また，ROC 曲線を描いてその特性を用いる．F 尺度は予測結果の評価尺度の一つである．再現率と適合率の調和平均によって求められる．最大値 1 に近いほど高評価になる．

$$F - measure = \frac{2Precision \cdot Recall}{Precision + Recall} \qquad (6)$$

$F - measure$，適合率 ($Precision$)，再現率 ($Recall$) を用いて検出性能の評価をおこなう．適合率は，正例と予測したデータのうち，実際に正例であるものの割合を示し，再現率は，実際に正例であるもののうち，正例であると予測されたものの割合を示す．

ROC（Receiver Operating Characteristic）は受信者操作特性とも呼ばれる．曲線より下側の面積が広いほど性能が高くなる．縦軸は TPR(=Recall)，横軸は FPR としてプロットする．

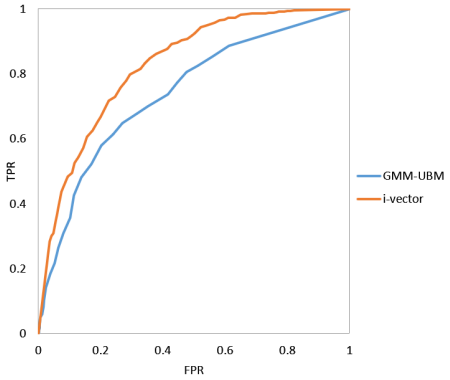

図 4 i-vector と GMM-UBM 二つ方式の比較結果

表 2 i-vector と GMM-UBM 二つ方式の比較結果

	F 尺度
i-vector	0．74
GMM-UBM	0．67

5.3 実験結果

5.3.1 提案方式と GMM-UBM に基づく賑わい音の推定方式の比較実験結果

表 2 より，評価データにおいて，i-vector に基づく賑わい音識別器は GMM-UBM に基づく賑わい音識別器よりも F 尺度の最大値が高い．したがって，i-vector に基づく賑わい音識別器は GMM-UBM に基づく賑わい音識別器よりも最高性能が高い結果が得られた．

図 4 は評価データを用いて i-vector に基づく賑わい音識別器と GMM-UBM に基づく賑わい音識別器の ROC 曲線である．ROC 曲線による可視化により，全ての FPR の範囲において，i-vector に基づく賑わい音識別器の曲線が GMM-UBM に基づく賑わい音識別器の曲線よりも上側を通っている．そのため，全ての FPR の範囲において i-vector に基づく賑わい音識別器の TPR が高いため，優れていると言える．

したがって，i-vector に基づく賑わい音識別器は GMM-UBM による賑わい音識別器よりも優れた性能を発揮することが明らかとなった．

5.3.2 提案方式による実験結果

表 3，4 より，評価データにおいて，提案方式によって F 尺度が大きな違いは見られない，特徴量次元数 MFCC や全変動行列 TvRank を増やして性能は上がる．混合数 Mix を増やして性能は上がらない．全変動行列 TvRank 不変な条件下では，特徴量次元数の増加とともに F 尺度の最大値はあまり変化しないが，ただし，計算量は増える．

特に全変動行列 TvRank を増加させると学習や識別に時間が掛かる．特徴量次元数と混合数 Mix が一定の場合，全変動行列 TvRank の増加とともに F 尺度の最大値はわ

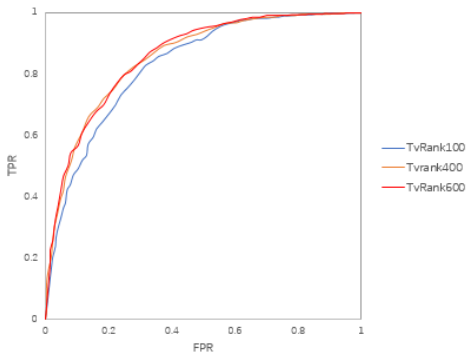

図 5 MFCC38 Mix128

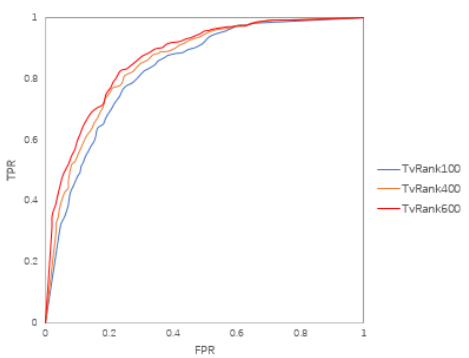

図 6 MFCC38 Mix256

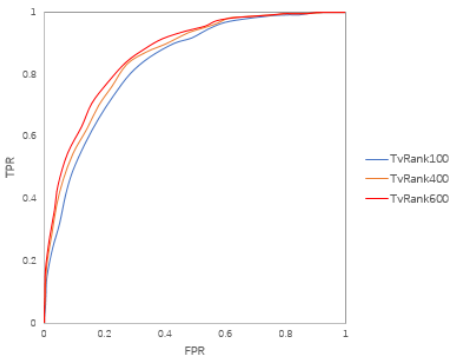

図 7 MFCC38 Mix512

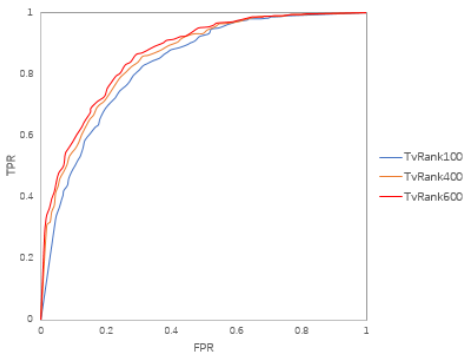

図 8 MFCC38 Mix1024

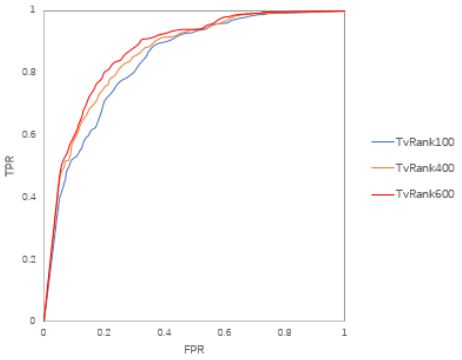

図 9 MFCC59 Mix128

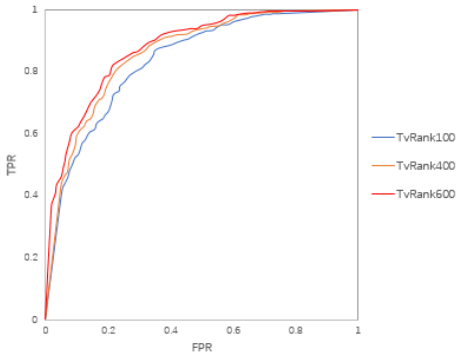

図 10 MFCC59 Mix256

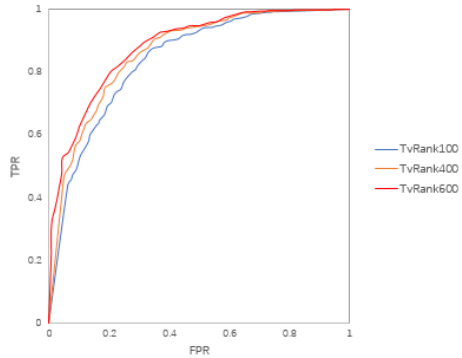

図 11 MFCC59 Mix512

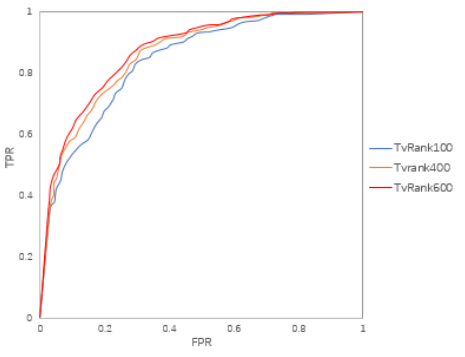

図 12 MFCC59 Mix1024

表 3 Fmeasure の最大値

MFCC38	TvRank100	TvRank400	TvRank600
128mix	0.7496	0.7632	0.7686
256mix	0.7515	0.7695	0.7833
512mix	0.7524	0.7711	0.7759
1024mix	0.7509	0.7684	0.7801

表 4 Fmeasure の最大値

MFCC59	TvRank100	TvRank400	TvRank600
128mix	0.7631	0.7720	0.7899
256mix	0.7589	0.7772	0.7893
512mix	0.7640	0.7776	0.7863
1024mix	0.7601	0.7772	0.7851

ずかに改善され，性能はわずかに改善された．同じ条件下で，特徴量次元数が増加すると，F 尺度の最大値も増加するが，増加はわずかである．

図 5-12 は評価データを用いて提案方式よりの ROC 曲線である．図 5-12 より，全ての FPR の範囲において，全変動行列 TvRank の増加とともに F 尺度の最大値はわずかに改善される．

また，用いた最大全変動行列 TvRank の曲線が他の曲線よりも上側を通っている．そのため，全ての FPR の範囲において全変動行列 TvRank が大きいほど TPR が高いため，優れていると言える．i-vector に基づく賑わい音識別器は GMM-UBM に基づく賑わい音識別器よりも最高性能が高い結果が得られた．

6. まとめと今後の課題

本報告では i-vector に基づく話者識別手法を応用した賑わい音の推定方式を提案した．

GMM-UBM に基づく賑わい音推定方式は F 尺度の最大値が 0.67 であったのに対し，i-vector に基づく賑わい音識別方式は 0.74 であった．ROC 曲線による比較でも i-vector に基づく賑わい音識別方式は GMM-UBM に基づく賑わい音識別方式よりも優れていた．

i-vector に基づく賑わい音の推定方式について，より詳細に分析パラメータを変更したところ，F 尺度の最大値 0.7899 であった．また，わずかではあるが，特徴量次元数 MFCC や全変動行列 TvRank を増やすことで識別性能は上がる傾向にあることが示唆された．特徴量次元数の増加とともに計算量は増える．特に全変動行列 TvRank を増加させると学習や識別に時間が掛かる．したがって，本研究の実験条件においては，計算量の少ない特徴量次元数 MFCC59，混合数 256mix, Tvrank 600 が最良であると考えられる．今後は，i-vector に基づく他の賑わい音を推定できる手法の実現について検討している．

謝辞 本研究は JSPS 科研費 18K02862 の助成を受けて実施したものである．

参考文献

[1] T.Tanaka, S.Hara and M.Abe, A Classification Method For Crowded Situation Using Environmental Sounds Based On GMM-UBM, 5th Joint Meeting, Acoustical Society of America and Acoustical Society of Japan, pp. 3110, Nov. 2016
[2] N.Dehak, P.J.Kenny, R.Dehak, P.Dumouchel and P.Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, No. 4, pp. 788–798, 2011.
[3] D.A.Reynolds, T.F.Quatieri, R.B.Dunn, Speaker Verfication Using Adapted Gaussian Mixture Models, Digital Signal Processing, vol. 10, pp. 19–41, 2000.
[4] A.Kanagasundaram, D.Dean, R.Vogt, et al. Weighted LDA techniques for i-vector based speaker verification, IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, Japan, pp. 4781–4784, 2012.
[5] W.M.Campbell, D.E.Sturim, D.A.Reynolds, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation, IEEE International Conference on Acoustics Speech, and Signal Processing, Speech, and Signal processing. Philadelphia, USA:IEEE, pp. 97–100, 2005.
[6] AO Hatch, S Kajarekar, A Stolcke, Within-class covariance normalization for SVM-based speaker recognition, International conference on interspeech, pp. 1471–1474, 2006.
[7] Machlica L,Zajic Z, An efficient implementation of probabilistic linear discriminant analysis, IEEE international conference on acoustics, Speech, and Signal processing, Vancouver, Canada:IEEE, pp. 7678–7682, 2013.
[8] D.Garcia-Romero, Xinhui Zhou and Carol Y.Espy-Wilso, Multicondition training of Gaussian PLDA models in i- vector space for noise and reverberation robust speaker recognition, ICASSP, pp. 4257–4260, 2012.
[9] Yun Lei, Lukas Burget, Luciana Ferrer, Martin Graciarena and Nicolas Scheffer, Towards noise-robust speaker recognition using probabilistic linear discriminant analysis, ICASSP, 2012.
[10] The HTK Book, http://htk.eng.cam.ac.uk/
[11] 朝田興平, 原直, 阿部匡伸, クラウドソーシングによる賑わい音識別方式のフィールド実験評価, 2018 年日本音響学会春季研究発表会, pp. 79–82, March, 2018.