Vol.2018-MUS-119 No.38
Vol.2018-SLP-122 No.38
2018/6/16

屋外拡声システムの主観的聴き取りにくさの客観計測器の提案

野口 啓太[1,a)]　小林 洋介[1,b)]　岸上 順一[1]　栗栖 清浩[2]

概要：屋外拡声システムの主観品質値の予測をする, 客観計測器を提案する. 提案では, 2 台のシングルボードコンピュータに屋外拡声器から放送された音声の主観品質値を教師として学習したモデルを組み込み, リアルタイム動作で主観品質値の予測値を表示する. 主観品質の指標には,「聴き取りにくさ」指標を用いた. 予測モデルには, 音響特徴量として MFCC を用い, 機械学習アルゴリズムの 1 つであるランダムフォレストを用いた. 作成したモデルの精度を評価したところ, 主観値との相関係数で 0.88 の性能であった.

1. はじめに

2011 年 3 月に発生した東日本大震災では, 20%の市民が防災行政無線の屋外拡声音をよく聴き取れず [1], 屋外拡声システムにおける基準の提案に繋がった [2]. この基準では, 屋外拡声システムの性能確認は, 拡声音を聴取することが求められている. しかし, 聴取実験には多数の被験者が必要であり, コストが掛かってしまう.

拡声システムの品質は, Steenken が発表した音声伝達指標 (STI: Speech Transmission Index)[3] を用いる事が多い. 一般に STI は, 空間のインパルス応答から求められるため, 実際の屋外拡声器から試験信号を放送する必要がある. 小林らは, 機械学習アルゴリズムである SVM(Support vector machine)[4] で主観評価実験の被験者を模擬する単語了解度の予測を提案をした [5]. その結果, 予測値と主観評価値との RMSE が 0.035 と誤差が非常に小さい. しかし, 了解度は, 単語の意味理解に重点を置いているが, 聴取品質を十分考慮しているとは言えない. 栗栖は, 実際の屋外拡声器から試験放送を行い, 主観品質である「聴き取りにくさ」の指標 LDR(LDR: Listening Difficulty Rating)[6] を, DLF_{BLS} (Band Limited Sum of Depth of Loudness Fluctuation) を用いた推定法を提案した [7].

[1] 室蘭工業大学
Muroran Institute of Technology, Mizumoto-cho 27-1, Muroran, Hokkaido, Japan
[2] TOA 株式会社
TOA Corporation, Takamatsu-cho 2-1, Takarazuka, Hyogo, Japan
[a)] 18043037@mmm.muroran-it.ac.jp
[b)] ykobayashi@csse.muroran-it.ac.jp

これらの先行研究は, 録音された音源の解析のみであり, リアルタイムで実運用できるシステムとはなっていない. そこで, 本研究では, 屋外拡声音声の品質をハンドヘルドで計測可能な計測器のプロトタイプを提案する.

2. 提案システムの概要

図 1 に提案する計測器を示す. 計算リソースを確保するため, 2 台のシングルボードコンピュータとモバイルバッテリー, オーディオインタフェース, 測定用マイクロホンで構成した.

図 2 に提案する計測器のフローを示す. まず, マイクロホン (BEHRINGER, ECM8000) に入力された音源をオーディオインタフェース (BLUE, ICICLE) を通して, Board 1(Rapberry Pi Foundation, Rapberry Pi 3 model B) で 1 sec. ごとに録音し, フレーム長を 100 msec. として 12 次元の MFCC(Mel Frequency Cepstrum Coefficients) とパワーおよび, Δ, Δ^2 パラメータの合計 39 次元を算出する. 次に, Board 2(Asus, Tinker Board) で RF(Random Forest)[8] で作成した予測モデルを使用して, VAD と LDR 予測を行う. 最後に, Board 1 の LCD ディスプレイに VAD の結果と予測 LDR を表示する.

3. 屋外での実音源の録音

機械学習モデルを作成する際, 多数の音源が必要である. また, 提案する計測器で用いるマイクロホンの特性を考慮した録音である必要もある.

そこで, LDR 学習音源と VAD 学習音源を作成するために, インパルス応答を図 3 に示す室蘭工業大学 V/R 棟前広

Vol.2018-MUS-119 No.38
Vol.2018-SLP-122 No.38
2018/6/16

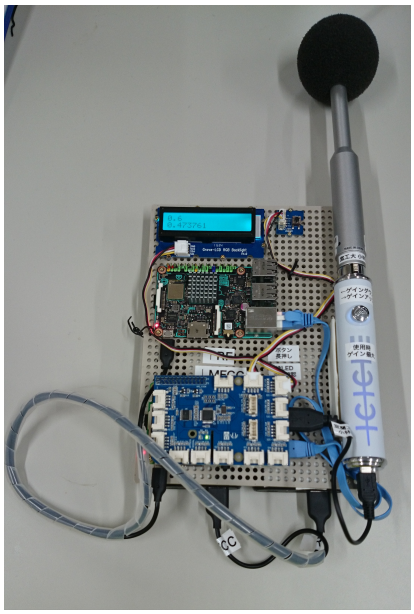

図 1　提案する計測器

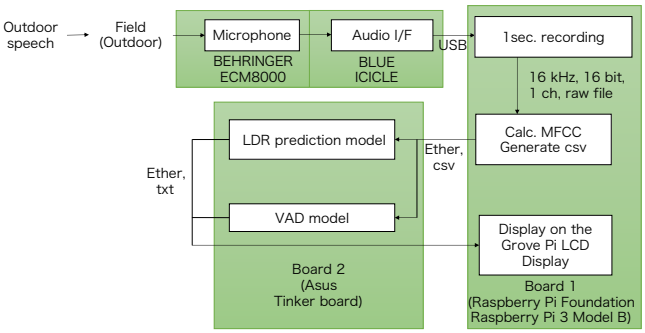

図 2　提案する計測器のフロー図

場で取得する実験を, 2017 年 11 月 13, 14 日に行った. 拡声のためのメガホン (TOA, ER-2830W) は, 地上から 1.50 m の地点に図 3 の矢印の向きに設置し, M 系列ノイズを放送し, 本計測器で使用するマイクロホン (BEHRINGER, ECM8000) で録音した. この出力レベルは, メガホンから 2.00 m の地点の騒音レベルを, M 系列ノイズで 92 dB となるように固定した. フィールドの対角線は約 40 m あり, 2.82 m (約 $2\sqrt{2}$ m) 間隔でメッシュを作成し, 格子点でインパルス応答を収録し, 騒音レベルを記録した.

B.G.N.(背景騒音) はフィールドの中心地点にあたる, 拡声器から直線上 20 m 地点で, マイクロホン (BEHRINGER, ECM8000) の先端を地上から 1.50 m の地点に垂直上向きに設置し, PCM レコーダー (TASCAM, DR-60DMKII) で録音した. 収録は 2017 年 11 月 15 日 11 時 38 分から 5 分 20 秒間行い, サンプリング周波数 48 kHz, 量子化ビット数 16 bit で録音した. 平均騒音レベルは 42 dB であった. このインパルス応答と B.G.N. および, 先行研究 [7] の音源を使い, VAD と LDR 予測のモデルを学習した.

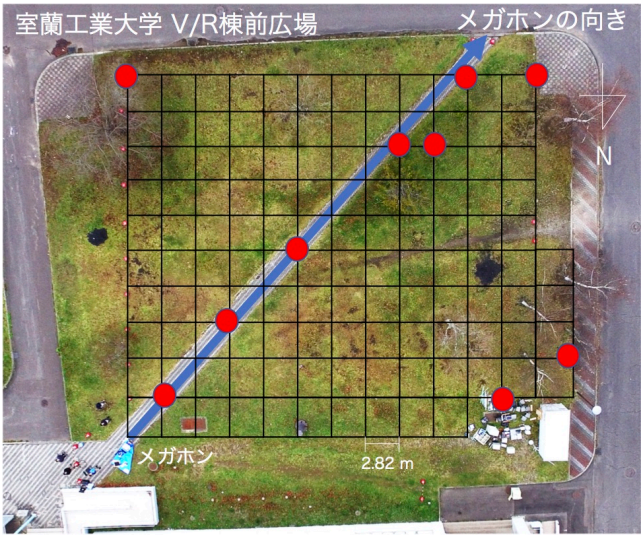

図 3　屋外での実音源の録音: 格子点でインパルス応答, 騒音レベル, STI を取得した. 丸印は主観評価実験で用いた地点

4. RF による VAD

VAD の学習音源は, 表 1 に示す条件で作成した. 音声区間はインパルス応答と文章音声を畳み込み, ランダムな区間の騒音を加算して作成した. VAD 学習音源の特徴量として, MFCC をフレーム長 100 msec., MFCC12 次元, パワー, およびそれらの Δ, Δ^2 パラメータの合計 39 次元を求めた. 音声区間および非音声区間として, 表 1 に示す VAD 用音源 3,593,038 音のうち約 70 %をランダムに選びトレーニングデータとし, モデルの学習に用いた.

Δ パラメータの重要度が 0.015 以下と低く, Δ を除いた 13 特徴量の重要度の積算が 0.927 であったため, Δ パラメータを使用せず, モデルを作成した. シングルボードコンピュータに実装することを考慮し RF の決定木の数を 50 とした.

表 2 にオープンテストの混合行列を示す. 誤分類率は 0.004, 適合率は 0.996, 再現率は 0.991 であった. 誤分類率が非常に低く, 適合率と再現率が高いため, 3 節でインパルス応答を畳み込んだ音源に対しては VAD が可能であることを示している.

5. LD の主観評価

5.1 主観評価指標

提案システムの評価指標は先行研究 [7] でも用いた LDR を使用する. LDR は, 被験者に評価音源を聴取した際に感じた「聴き取りにくさ」を表 3 に示す 4 段階から評価させ, 式 (1) に示す聴き取りにくいと回答した割合を算出する.

$$\mathrm{LDR} = \frac{\mathrm{L2} \sim \mathrm{L4}\ \text{の回答数}}{\text{回答総数}} \tag{1}$$

表 1 音源の作成条件

用途	VAD		LDR	
	非音声区間	音声区間	主観評価	予測モデル学習
インパルス応答	なし	室蘭工大で計測した 145 地点	STI が 0.596~0.732 の 5 地点, 拡声器の直線上 5 地点	
文章	なし	ATR 音素バランス文 A, F, G セット 150 文		
音声音源	なし	ASJ-JIPDEC ECL0001 ~ 1004 男性 2 名, 女性 2 名		
騒音	室蘭工大で録音した音源, JEIDA-NOISE から 6 音源 (駅 (通路), 幹線道路, 交差点, 人混み, 列車 (在来線), 空調機 (大型))		室蘭工大で録音した音源	
音声レベル	-	50 dB ~ 80 dB まで 10 dB ごと		
騒音レベル	40 dB ~ 80 dB まで 1 dB ごと	40 dB, 45 dB, 50 dB	40 dB	
サンプリングレート	16 kHz		48 kHz	16 kHz
音源長	100 msec.		文章長 + 遷移区間	1.0 sec.
総音源数	510,860 (騒音数 × 騒音レベル × 1,780 (ノイズを切り出す数))	3,082,178 (騒音数 × 騒音レベル × インパルス応答 × 音声レベル × 話者数 × 音声音源長 (msec.) ÷ 100(msec.))	160	744

表 2 VAD の結果

真のクラス \ 予測クラス	音声	非音声
音声	924,059	811
非音声	3,840	149,202

表 3 LD の 4 段階カテゴリー

L1	聴き取りにくくはない
L2	やや聴き取りにくい
L3	かなり聴き取りにくい
L4	非常に聴き取りにくい

5.2 LD の主観評価の設定

LD の主観評価音源は, 表 1 に示す条件で作成した 160 音を使用する. 用いたインパルス応答は図 3 に示す地点であり, 計測した全地点の中から STI が 0.596~0.732 と低い 5 地点と拡声器の直線上 5 地点を用いた. 音源長のうち, 遷移区間は, 急激な音量の変化から聴覚を保護するための騒音の立ち上がり・立ち下がり区間であり, 評価音源の前後に設定した. 主観評価は防音ブース内でラップトップマシンに接続したオーディオインタフェース (Roland, UA-25 EX) からヘッドホン (SENNHEISER, HDA300) を用いてダイオティックに被験者へ提示した.

被験者は日本語話者 20 代 25 人 (男性 24 人, 女性 1 人) である. 音量は 1 kHz 94 dB のキャリブレーション信号を 94 dB で提示できるようにダミーヘッド (サザン音響, SAMURA HATS Type3700E) に組み込んだイヤーシミュレータ (アコー, Type2128E) を用いて音量を調整した. 被験者には, 評価中に音量を変更しないように指示した. 被験者は, 各音源に対して表 3 に示す 4 段階をラップトップ画面上の該当箇所をクリックする専用 GUI で回答させた. 実験の前に, 聴取と回答の練習を兼ねて, 試験音源に用いていない 2 地点分のインパルス応答と未使用の文章音声を畳み込んで作成した 8 音源を提示し, 操作練習を行った. 被験者の疲労を考慮し, 評価はブース内で着席して行い, 適時休憩を取れるように配慮した. なお, 本実験は室蘭工業大学研究倫理審査委員会の承認のもと行なわれた.

6. RF による LDR 予測モデル

6.1 RF モデル作成条件

5.2 節の主観評価音源を 1 sec. ごとに切り出し, Δ, Δ^2 を含む 39 次元の特徴量を 100 msec. ごとに求めて, LDR を予測する. よって, 1 sec. のごとに 390 次元の特徴量となる. 学習に用いる音源数を増やすため, 実際の屋外拡声音声を録音した先行研究 [7] の音源も同様に特徴量を求め, 学習音源とした. 学習音源を 1 sec. に切り出したため, 5.2 節の主観評価音源数は 744 音であり, 先行研究 [7] の音源数は 392 音である. LDR 予測モデルは, これらの音源のうちそれぞれ約 70%をトレーニングデータとしてモデルの学習に使う. 5.2 節の主観評価音源数の残りの約 30%をバリデーションデータ, 先行研究 [7] の音源をテストデータとして性能評価に使う. モデルの精度評価は, 主観評価値と予測値との RMSE, 相関係数, 決定係数を用いた.

6.2 バリデーションデータによるモデル最適化

RF のハイパーパラメータである決定木の数と深さは, 表 4 の条件でグリッドサーチで決定する. 図 4 に各決定木の深さごとの木の数と RMSE を示す. 木の深さが 10, 木の数

Vol.2018-MUS-119 No.38
Vol.2018-SLP-122 No.38
2018/6/16

表 4　RF のパラメータ調整

パラメータ	設定
決定木の数	10, 25, 50, **75**, 100, 250, 500, 750, 1,000
決定木の深さ	**10**, 20, 30

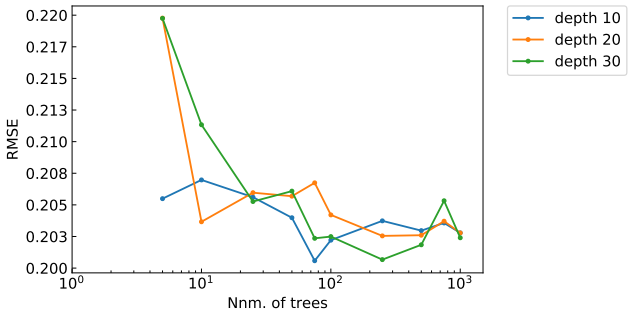

図 4　決定木の深さごとの RMSE と決定木の数

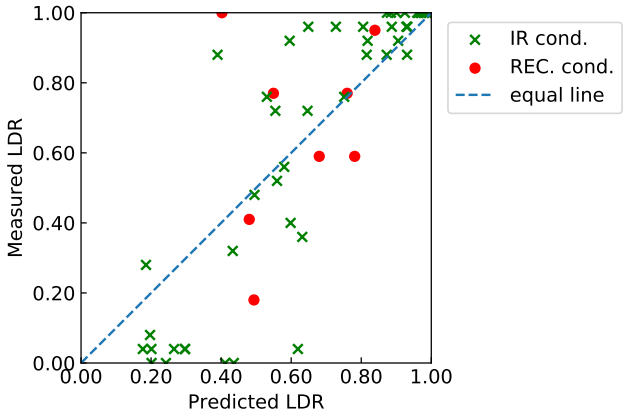

図 5　評価実験による LDR と, 予測モデルによる LDR 予測値: IR cond. は 5.2 節で作成した音源, Rec. cond. は先行研究 [7] の音源

が 75 の時, RMSE が 0.20 と誤差が最も小さくなった. この値で学習してモデルを組み込むこととする.

6.3　テストデータによる性能評価

図 5 に主観評価による LDR(Measured LDR) と予測した LDR(Predicted LDR) を示す. 5.2 節の主観評価音源でのバリデーションデータに対して, equal line に漸近するが, 先行研究 [7] のテストデータに対して, RMSE が 0.60 と大きく外れるデータが存在する事が示された. 表 5 に先行研究 [7] と本稿における RMSE, 相関係数, 決定係数を示す. 5.2 節の音源に対して, 相関係数が 0.88 と強い相関がある. しかし, 先行研究 [7] の音源に対して, 相関係数が 0.25 と低い. また, 決定係数が −0.08 とマイナスとなり精度が悪い. 5.2 節の音源と先行研究 [7] の音源数に偏りがあるため, 精度に差が生じたと考えられる.

表 5　LDR 予測モデルの精度

	音源	RMSE	相関係数	決定係数
先行研究 [7] の精度	男声	-	0.8971	0.8048
	女声	-	0.7461	0.5566
提案予測モデルの精度	先行研究 [7] の音源	0.27	0.25	−0.08
	5.2 節の主観評価音源	0.20	0.88	0.74

7.　まとめ

屋外拡声システムの聴き取りにくさを予測する計測器のプロトタイプとして, MFCC を特徴量とし, RF による VAD と LDR 予測モデルを用いることを提案した. その結果, インパルス応答を畳み込み, 暗騒音を加算した音源に対して, 相関係数が 0.88 と強い相関が示された. しかし, 実際の屋外拡声システムから放送された録音音源に対して, 相関係数が 0.25 と低い相関が示された. 今後は, 学習音源を増やし, さらなる改良を行う.

謝辞　本研究の一部は JSPS 科研費 (16K21584), (公財) 人工知能研究振興財団, (公財) 電気通信普及財団, (公財) 国際科学技術財団, (公財) 立石科学技術振興財団, (公財) 矢崎科学技術振興記念財団, 東北大学電気通信研究所共同研究プロジェクト (H29/A18) の助成を受けた. 関係各位と被験者各位に感謝する.

参考文献

[1] 東北地方太平洋沖地震を教訓とした地震・津波対策に関する専門調査会 (第 7 回). “平成 23 年東日本大震災における避難行動等に関する面接調査（住民）単純集計結果,” 2011.
[2] “災害等非常時屋外拡声システムのあり方に関する技術調査研究委員会,” ASJ 屋外拡声規準 第 1 版, 2017.
[3] H.J.M Steeneken and T.Houtgast. “A physical method for measureing speech transmission quality,” *J. Acoust. Soc. Am.*, pp. 318–326, 1980.
[4] Vapnik, Vladimir N. *“The Nature of Statistical Learning Theory”*, Springer-Verlag, 1995.
[5] 小林洋介, 太田健吾, 近藤和弘. “機械学習と音声認識を用いた屋外拡声器の音声品質予測,” 研究報告音楽情報科学 (MUS), pp. 1–4, 2016.
[6] Masayuki Morimoto, Hiroshi Sato, and Masaki Kobayashi. “Listening difficulty as a subjective measure for evaluation of speech transmission performance in public space,” *J. Acoust. Soc. Am.*, pp. 1607–1613, 2004.
[7] 栗栖清浩. “変動量解析による屋外拡声音声の明瞭性評価,” 日本音響学会聴覚研究会資料, vol45, No5, pp. 405–411, 2015.
[8] Leo Breiman. “Random Forest,” *Machine Learning*, pp. 5–411, 2001.