Vol.2019-MUS-123 No.53
Vol.2019-SLP-127 No.53
2019/6/23

屋外拡声品質予測モデルの中間特徴量の検討

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

概要：東日本大震災では，20%の市民が屋外拡声音をよく聴き取れなかった事が報告されている。このため，屋外拡声器の品質向上が議論されているが，聴取実験のコストがかかるなど解決すべき点は多い。我々はこれまでに，MFCC を音響特徴量として，主観評価指標である LDR (Listening difficulty rating) を予測する機械学習モデルを提案したが，教師となる主観評価数が少なく，RMSE (Root means squared error) が 0.20 と満足な性能が得られなかった。そこで本稿では，MFCC から中間特徴量となる客観評価値を予測するモデルと，中間特徴量から主観評価値を予測する 2 モデルの組み合わせを提案する。中間特徴量に用いる指標に，Short time objective intelligibility，Speech intelligibility prediction based on mutual information，Extended short time objective intelligibility の 3 指標を比較した。その結果，最適な中間特徴量は SNR により異なったものの，SNR が 0 dB から 30 dB の範囲では以前の検討よりも良い，RMSE が 0.14 以下を達成した。

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で，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，コストがかかる。

音声了解度の予測指標として，原音と雑音抑圧音声との時間-周波数分析に基づく知覚的歪みをモデル化した STOI (Short time objective intelligibility) が提案された [3]。また，STOI を改善した指標として，原音と雑音抑圧音声との相互情報量を用いた SIMI (Speech intelligibility prediction based on mutual information) [4]，STOI の相関係数計算部を正規化指標とした ESTOI (Extended STOI) [5] がそれぞれ提案されている。これら STOI-type の指標は音声了解度と相関が高い事が示されている [6] が，3 指標ともその計算過程で原音が必要であり，実運用を考慮した際に原音の入手に関して課題が残る。

この問題を解決するため，我々は特徴量として拡声音の MFCC (Mel frequency cepstrum coefficients) を求め，機械学習アルゴリズムの RF (Randam forest) [7] で主観評価指標 LDR (Listening difficulty rating) [8] の予測を提案した [9]。この提案では，学習の教師データとして，主観評価値を直接用いたが，教師データが 160 音源と少なく，RMSE (Root mean squared error) が 0.20 と満足な精度ではなかった。一方で，Deokgyu *et al.* は音響特徴量に MFCC を利用し，深層学習で STOI を予測するモデルを作成している [10]。その結果，騒音と残響を考慮した音源に対し RMSE が 0.147 と我々の提案よりも誤差が小さく，より実用的と考えられる結果が得られている。

そこで，MFCC から主観評価値と相関の高い客観評価指標を中間特徴量と定義し，中間特徴量を予測するモデルを大規模なデータより学習する事で作成する。次に，中間特徴量から主観評価値を予測するモデルを学習し，2 つのモデルを組み合わせた LDR 予測システムを提案する。本稿では，中間特徴量に，STOI，SIMI，ESTOI を比較したので報告する。

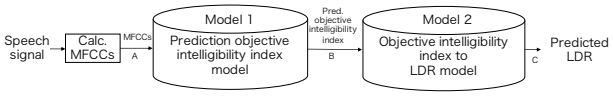

図 1　提案する LDR 予測システムフロー

2. 提案システム

2.1 概要

図 1 に提案する LDR 予測システムのフローを示す。まず，マイクロホンに入力された拡声音をオーディオインタフェースを通じて，1 秒ごとに録音し，フレーム長を 0.1 秒として 12 次元の MFCC，パワー，それらの delta パラメータの合計 26 次元を入力特徴量として計算する。次に入力特徴量を中間特徴量に変換するモデル (Model 1) と中間特徴量から LDR を予測するモデル (Model 2) を用いる。

2.2 Listening difficulty rating

提案システムの主観評価指標は既報 [9] で用いた LDR の評価値を使用する。LDR は，式 (1) に示す，表 1 の評価値 $L1$ から $L4$ の総集計数 T に対して，評価値 $L1$ 以外の割合

[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

© 2019 Information Processing Society of Japan

表 1　聴き取りにくさの指標

$L1$	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

である。

$$\mathrm{LDR} = \frac{T - count(L1)}{T} \tag{1}$$

3. 主観評価と音源の作成

3.1 主観評価の設定

Model 2 の学習に用いる音源は，表 2 (主観評価音源) に示す条件で作成した 720 音である。遷移区間は，急激な音圧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 10 名 (男性 9 名，女性 1 名，平均年齢 22.5 歳) であり，提示音圧を変更しないように指示した。聴取者は各音源に対して表 1 に示す 4 段階評価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 で回答した。

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考慮した。本評価実験は室蘭工業大学ヒトを対象とした研究倫理審査委員会の承認のもと行われた。

3.2 中間特徴量予測モデルの評価音源

Model 1 の学習に用いる音源と評価音源は，表 2 に示す条件で作成した。入力信号を 1 秒ごとに切り出し，26 次元の音響特徴量を 0.1 秒ごとに求めて 1 セットとした。中間特徴量予測モデルの入力次元数は，10 セットの合計 260 次元とした。

4. モデルの学習

4.1 RF による中間特徴量予測 (Model 1)

RF による予測モデルは，決定木の数を最適化する必要がある。そこで，決定木の数をハイパーパラメータとして調整し，実測の客観評価値と予測した客観評価値の RMSE で予測モデルの精度を評価した。決定木の数を，10 から 100 まで 10 刻み，150，200 から 1000 まで 100 刻みとし，モデルの精度を比較した。決定木の数が 100 以上の時，RMSE の減少率が飽和したため決定木の数を 100 とした。

図 2 ～ 4 に STOI，SIMI，ESTOI の予測モデルの性能をそれぞれ示す。3 指標とも SNR が 5 dB の場合，実測の中間特徴量より大きい値を予測する傾向が示された。表 3 に実測の中間特徴量と予測した中間特徴量の相関係数を SNR 別に示す。25 dB を除くと ESTOI 予測モデルの性能が他指標と比べて最も高い。また，ESTOI は，45 dB の相関係数が 0.57 と他指標よりは良いが，ほかの SNR と比べると相関が低い。よって，中間特徴量の予測モデルは，

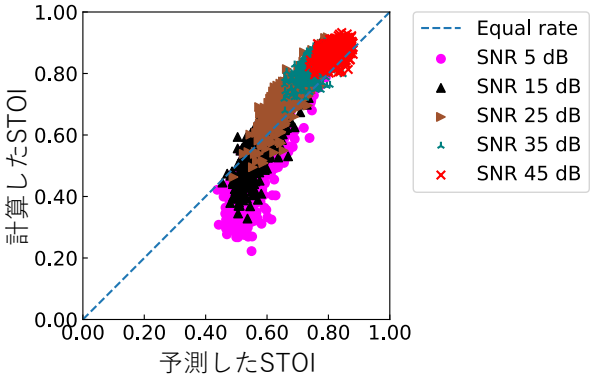

図 2　表 2 のバリデーションセットを用いた STOI 予測結果

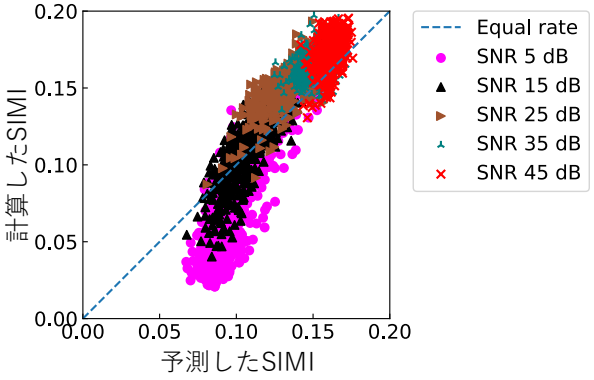

図 3　表 2 のバリデーションセットを用いた SIMI 予測結果

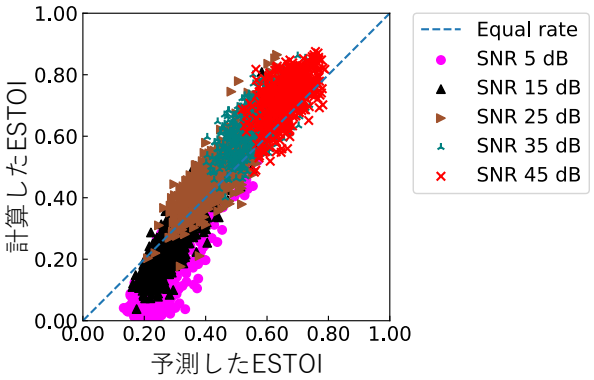

図 4　表 2 のバリデーションセットを用いた ESTOI 予測結果

ESTOI を用いることが最適だが，SNR が大きい時に相関が低くなるため，SNR ごとに中間特徴量を最適に選択する必要がある。

4.2 中間特徴量から LDR 予測 (Model 2)

客観評価指標から主観評価指標である LDR への予測は，3.1 節の主観評価結果を中間特徴量から予測する。本稿では中間特徴量として用いた了解度指標を主観評価値に回帰するため GLM (General linear model) を用いた。

式 (2) にモデル式を示す。ここで，回帰係数 a と b は最尤推定で求める。

© 2019 Information Processing Society of Japan

表 2　音源作成条件

データセット	学習音源	バリデーション音源	主観評価音源
文章	ATR 音素バランス 503 文 A, F, G セット (150 文)		
インパルス応答数	101	44	5
背景雑音	室蘭工大で録音した 1 音源と JEIDA-NOISE から 6 音源 (駅コンコース，幹線道路，交差点，人混み，在来線，空調機)		
ASJ-JIPDEC 話者セット	ECL0001, ECL1003	ECL0002, ECL1004	
発話レベル (dB)	40, 50, 60, 70, 80, 90	45, 55, 65, 75, 85	40, 50, 60, 70, 80
雑音レベル (dB)	40, 50	40	40, 50, 60
SNR (dB)	0, 10, 20, 30, 40, 50	5, 15, 25, 35, 45	0, 10, 20, 30, 40
音源数	16800	3150	720

表 3　図 1 A から B への相関係数

SNR (dB)	5	15	25	35	45
STOI	0.85	0.89	**0.86**	0.68	0.41
SIMI	0.87	0.88	0.78	0.46	0.42
ESTOI	**0.90**	**0.92**	0.85	**0.71**	**0.57**

表 4　モデル式のパラメータ

	a	b
STOI	-14.19	-22.32
SIMI	-11.81	-87.19
ESTOI	-6.99	-19.16

$$f(d) = \frac{1}{1 + \exp(a - b \times d)} \tag{2}$$

表 4 に決定した回帰係数を示す。図 5 ～ 7 に客観指標の実測値と LDR のマッピングした結果，及び回帰モデルを示す。3 指標ともモデル式から大きく外れる音源が無いことが示された。表 5 に客観評価指標ごとに実測の LDR と予測した LDR の相関係数を SNR 別に示す。実測の LDR と STOI から予測した LDR の相関係数は最小で 0.36 と他指標より低いことが示された。SNR が 0, 30, 40 dB の場合，実測の LDR と SIMI から予測した LDR の相関係数が 0.60 から 0.76 と最も高く，SNR が 10, 20 dB のとき，実測の LDR と ESTOI から予測した LDR の相関係数が 0.64 から 0.90 と最も高い事が示された。

4.3　LDR 予測システム全体の性能評価 (Model 1 and Model 2)

MFCC を入力特徴量として中間特徴量を予測するモデル (Model 1) と中間特徴量から主観評価値を予測するモデル (Model 2) 全体で指標間の比較を行う。図 8 ～ 10 に実測の LDR と予測した LDR を示す。STOI と ESTOI は Equal rate に漸近する傾向にあるが，SIMI は予測値が 0.45 から 1.00 の範囲となり，予測値に偏りが生じている。表 6 に客観評価指標ごとの予測 LDR と実測の LDR の RMSE を SNR 別に示す。SNR 別に結果を比較すると 0, 10, 20 dB の場合，SIMI を用いた場合の RMSE が 0.12 以下と他指標と比べて良い。一方で，それよりも SNR が高い際は，ESTOI を用いた場合で RMSE が 0.14 と 0.22 と 3 指標の中では良い結果となった。

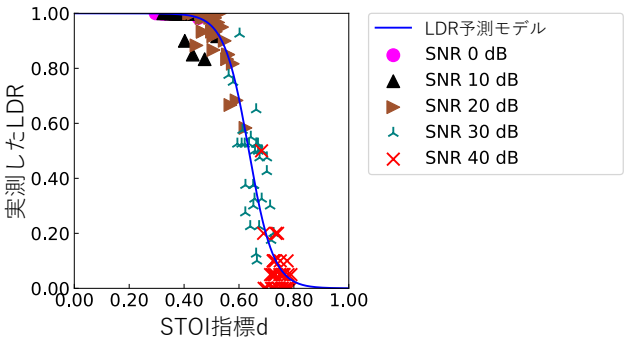

図 5　表 2 の主観評価音源を用いた STOI と LDR

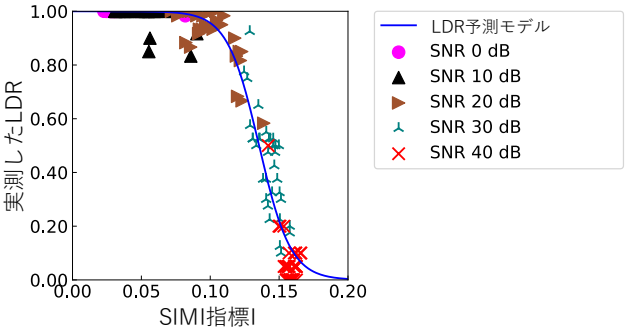

図 6　表 2 の主観評価音源を用いた SIMI と LDR

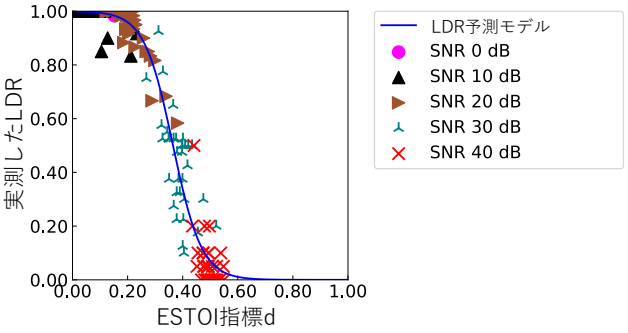

図 7　表 2 の主観評価音源を用いた ESTOI と LDR

4.4　考察

システム全体の評価において，SNR が 20 dB 以下と相対的に小さい場合は SIMI を用いた場合が精度が良く，SNR が 30 dB 以上と相対的に大きい場合は ESTOI を用いた場

© 2019 Information Processing Society of Japan

Vol.2019-MUS-123 No.53
Vol.2019-SLP-127 No.53
2019/6/23

表 5　図 1 B から C への相関係数

SNR (dB)	0	10	20	30	40
STOI	0.36	0.50	0.89	0.56	0.51
SIMI	**0.60**	0.58	0.86	**0.76**	**0.73**
ESTOI	0.53	**0.64**	**0.90**	0.67	0.60

表 6　図 1 A から C への RMSE

SNR (dB)	0	10	20	30	40
STOI	0.22	0.18	0.20	0.16	0.30
SIMI	**0.12**	**0.07**	**0.07**	0.37	0.60
ESTOI	0.25	0.21	0.24	**0.14**	**0.22**

合が精度が良い。この傾向は，SIMI 提案時の雑音環境下の了解度推定性能が STOI より高い結果 [4] と同傾向である。また，本稿で比較した STOI-type の指標はどれもノイズリダクションを施した音声を対象としており，ノイズが抑圧された場合の SNR に近い，相対的に高 SNR の場合，ノイズリダクション音声の了解度推定精度が最も高い ESTOI [5] が高精度になったと考えられる。以上のように，SNR により予測精度が異なるため，拡声音の SNR に応じたモデル切り替えや複数モデルの結果統合により，精度を向上すると考えられる。

5. まとめ

屋外拡声システムによる拡声音の LDR を予測するシステムの改善のため，中間特徴量として STOI, SIMI, ESTOI を用いた予測モデルを提案し，それらの性能を比較した。その結果，相対的に SNR の低い場合は SIMI が，SNR が高い場合には ESTOI を用いた予測モデルの精度が良いことが示された。今後は，実際に屋外拡声音を使用して評価する。

謝辞　本研究の一部は JSPS 科研費 (19K15146, 16K21584)，(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。

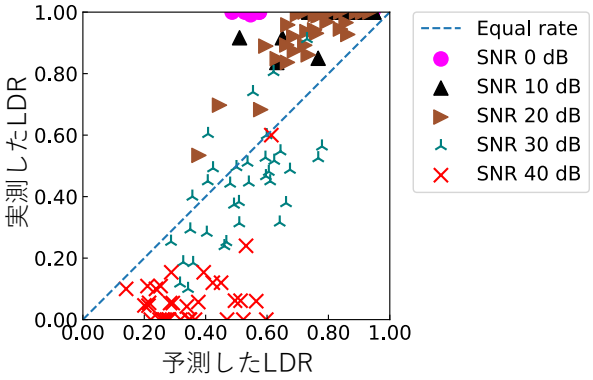

図 8　STOI を用いた LDR 予測値と実測の LDR

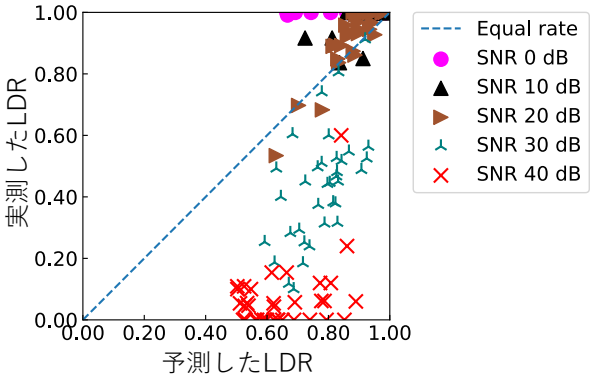

図 9　SIMI を用いた LDR 予測値と実測の LDR

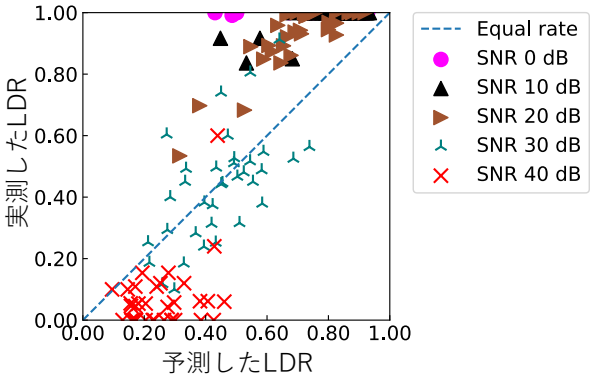

図 10　ESTOI を用いた LDR 予測値と実測の LDR

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] Cees H. Taal and Richard C. Hendriks. “An Algorithm for Intelligibility Pre- diction of Time-Frequency Weighted Noisy Speech,” IEEE Trans. Audio., Speech, Language Processing”. *IEEE Trans. Audio.*, pp. 2125–2136, 2011.

[4] Jesper Jensen and Cees H. Taal. “Speech Intelligibility Prediction Based on Mutual Information”. *IEEE Trans. Audio.*, pp. 430–440, 2014.

[5] Jesper Jensen and Cees H. Taal. “An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers”. *IEEE Trans. Audio.*, pp. 2009–2022, 2016.

[6] 小林洋介. “雑音下音声了解度と客観的音声了解度指標との関係”. *IEICE Technical Report*, pp. 1–6, 2018.

[7] Leo Breiman. “Random Forest”. *Machine Learning*, pp. 5–411, 2001.

[8] 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.

[9] 野口啓太, 小林洋介, 岸上順一, 栗栖清浩. “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”. 研究報告音楽情報科学 (MUS), pp. 1–4, 2018.

[10] Deokgyu YUN, Hannah LEE, and Seung Ho CHOI. “A Deep Learning-Based Approach to Non-Intrusive Objective Speech Intelligibility Estimation”. *IEICE TRANS. INF. SYST*, pp. 1207–1208, 2018.

© 2019 Information Processing Society of Japan