屋外用聴き取りにくさ計プロトタイプの改善

屋外用聴き取りにくさ計のプロトタイプの改善*

☆野口啓太, 小林洋介, △岸上順一 (室蘭工大), 栗栖清浩 (TOA)

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，コストがかかる。

我々はこれまでに，入力音声の特徴量としてMFCC(Mel Frequency Cepstrum Coefficients)を求め，主観品質である聴き取りにくさの指標LDR(Listening Difficulty Rating)[3] を機械学習手法の1つであるRF(Random Forest) [4] で作成した予測モデルを組み込んだ，聴き取りにくさ計測器の提案をした [5]。教師データとして主観評価値を直接用いて学習したモデルを使用したが，教師データが160音源と少ないデータ数であり，計測したLDRと予測したLDRのRMSE(Root Mean Square Error)が0.21と精度が低かった。

Deokgyu *et al.* は，MFCCを用いて，機械学習手法のLSTM(Long Short-Time Memory) アルゴリズムで客観評価値であるSTOI(Short Time Objective Intelligibility) [6] 予測モデルを作成し，評価した [7]。騒音と残響を考慮した音源に対し，RMSEが0.147と誤差が小さく，相関係数が0.576とやや相関があるモデルを作成した。

そこで，本稿では，より多くの教師データを得るために，MFCCを音響特徴量として客観評価値であるSTOIを予測するモデルと，STOIからLDRをマッピングするシグモイド関数を組み合わせたモデルを提案する。STOI予測モデルは，これまで取り組んでいる要因解析も可能なRFアルゴリズムを使用する。

2 提案システムの概要

Fig. 1に提案する計測器のフローを示す。この計測器は，コンピューティングプラットフォーム (NVIDIA, JETSON TX2) を用い，屋外に設置されている拡声器等から放送される音声を録音し，リアルタイムにLDR予測結果を表示する。そのためにまず，マイクロホンに入力された音源をオーディオインタフェースを通して，1 sec. ごとに録音し，フレーム長を100 msec. として12次元のMFCCとパワーおよび，deltaパラメータの合計26次元を算出する。MFCC計算終了後，客観指標であるSTOIを学習済みモデルを用いて予測する。最後に，予測したSTOIに対し，シグモイド関数を用いて主観評価値であるLDRへマッピングし，結果を表示する。このために，本稿では，音響特徴量としてMFCCを用いて，機械学習手法のRFでSTOI予測モデルの作成をし，予測したSTOIからLDRをマッピングするシグモイド関数を作成する。

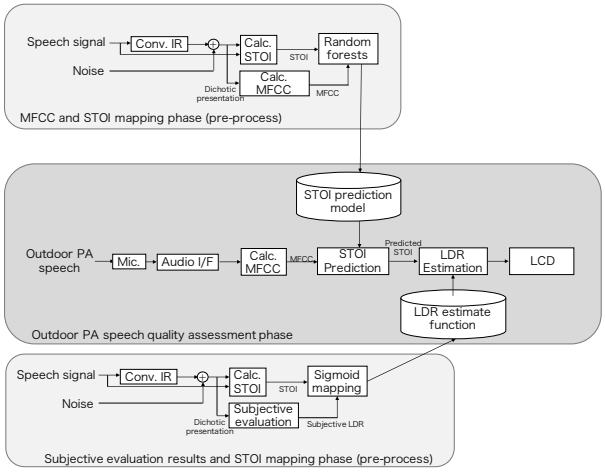

Fig. 1 Proposed LDR meter flow

3 客観評価モデルと主観評価

3.1 Short Time Objective Intelligibility

STOIは時間-周波数モデルに基づく了解度指標である。Fig. 2にSTOIの処理フローを示す。時間周波数分解は，原音と劣化音声の両方について10 kHzのサンプリングレートで，512サンプルまでゼロパディングしたハン窓付きフレーム分割を処理する。原音の最大振幅より40 dB低い無音区間を除去して再構成した信号に対し，1/3オクターブバンド分割を行い，原音 x と劣化音声 y の周波数包絡 $X_j(m)$ と $Y_j(m)$ を求める。

次に，原音と劣化音声の周波数包絡をセグメントしたフレームより長い区間 N で切り出し，$X_{j,m}$ と $Y_{j,m}$ を得る。さらに，劣化音声の周波数包絡 $y_j(m)$ を正規化し，了解度に強い影響を持たないレベル差を補正する。最後に，$\mathbf{x}_{j,m}$ と $\bar{\mathbf{y}}_{j,m}$ の同一バンド，同一フレームでの相関係数を求め，$d_{j,m}$ として，平均了解度指標値 d を得る。STOIの標準では，音声了解度へのマッピングを行うが，本稿ではLDRとのマッピングとした。

3.2 Listening Difficulty Rating

提案システムの評価指標は既報 [5] で用いたLDRを使用する。LDRは，式 (1) に示すように，Table 1の評価値 $L1$ から $L4$ の総集計数 T に対して，評価値 $L1$ 以外の割合で示す。

*Improvement of Listening Difficulty Rating Meter for Outdoor Public Address System. by NOGUCHI, Keita, KOBAYASHI, Yôsuke, KISHIGAMI, Jay(Muroran Institute of Technology) and KURISU, Kiyohiro (TOA Corporation)

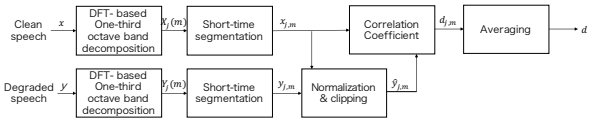

Fig. 2　Flowchart of STOI calculation

Table 1　Four-grade evaluation for LDR

Index	English	Japanese
L1	Not difficult	聴き取りにくくはない
L2	A little difficult	やや聴き取りにくい
L3	Fairly difficult	かなり聴き取りにくい
L4	Extremely difficult	非常に聴き取りにくい

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

4　音源の作成

4.1　屋外での音源収集

STOI 予測を行うため，機械学習を用いたが，学習を行う際に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音源も必要である。そこで，STOI 学習音源を作成するために，インパルス応答を Fig. 3 に示す室蘭工業大学 V/R 棟前広場で取得した。拡声のためのメガホン (TOA，ER-2830W) は，図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。この際に，メガホンからの出力は騒音レベル 92 dB で固定し，各計測地点での騒音レベルを記録した。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 5 分 20 秒間録音した。このインパルス応答と B.G.N. を使い，STOI 予測のモデルを学習した。

4.2　主観評価

LDR の主観評価音源は，Table 2 に示す条件で作成した 160 音を使用する。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ュミレータ (アコー，Type2128E) を用いて調整した。被験者には音量を変更しないように指示した。聴取者は各音源に対して Table 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。

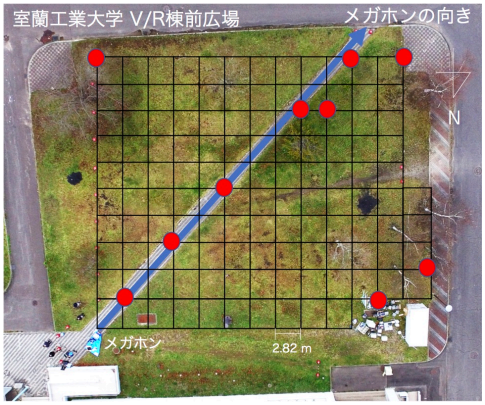

Fig. 3　IR measurement points in Muroran-IT: black crosses are IR measured point, red circles are used subjective evaluation

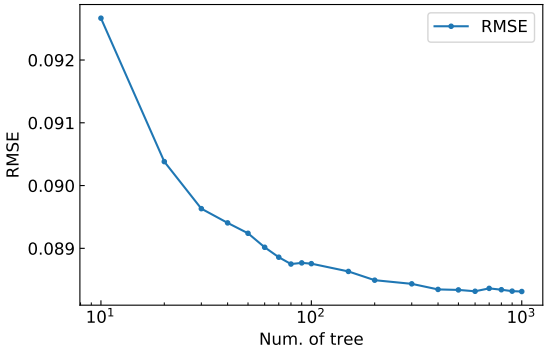

Fig. 4　Relationship between RMSE and num. of tree

実験の前に，聴取と回答の練習を兼ねて，STI(Speech Transmission Index) が 0.813 の地点と，STI が 0.830 の地点のインパルス応答と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考慮した。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。

4.3　STOI 予測モデルの評価音源

STOI 予測モデルの学習に用いる音源と評価音源は，Table 3 に示す条件で作成した。各音源に対し，1 sec. に切り出し，12 次元の MFCC とパワー，それらの delta パラメータを含む 26 次元の特徴量を 100 msec. ごとに求めて，STOI を予測する。よって，1 sec. の音源より 260 次元の特徴量が求まる。

5　モデルの作成

5.1　RF による STOI 予測

RF 予測モデルの精度を向上するために，決定木の数を多くする必要がある。しかし，ハンドヘルド可能な計測器に組み込むことを考慮し，動作が軽い予測

Table 2　Subjective evaluation signal generate conditions

IR	5 points (STI: 0.596 ∼ 0.732), 5 points (Straight line with PA system)
Sentences	ATR 503 P.B. sentence A，F，G sets (150 sentences)
Speaker	Two male and two female speakers
Speech levels	50 dB ∼ 80 dB per 10 dB
Noise level	40 dB
Sampling rate	48 kHz
Signal length	Sentence length + transition interval
Num. of generated speech signal	160

Table 3　Training and test signal generate conditions for STOI prediction model

	Training speech signal	Test speech signal
Recorded IR (Using section 4.2)	7 points	3 points
Recorded IR (Not Using section 4.2)	94 points (About 70% of recorded IR)	41 points (About 30% of recorded IR)
speaker	A male and a female (ECL0001，ECL1003)	A male and a female (ECL0002，ECL1004)
Noise	Recorded in Muroran-IT，and 6 sounds from JEIDA-NOISE (inside the station，highway，junction，crowd， train (conventional line)，air conditioner (large))	
SNR	5 dB∼45 dB per 10 dB	0 dB∼50 dB per 10 dB
Num. of speech signal (divided into 1 sec.)	99,658	53,264

モデルを用いる必要がある。そこで，決定木の数を調整し，実測の STOI と予測した STOI 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 刻みと，150，200 から 1000 まで 100 刻みとし，モデルの精度を比較した。

Fig. 4 に 4.3 節で作成したテスト音源に対してのモデルの精度を示す。決定木の数が 100 以上の時，RMSE の減少率が飽和したため，計測器に組み込むモデルは，計算速度を考慮し決定木の数を 100 とした。Fig. 5 に決定木の数を 100 としたモデルを用いて，テスト音源に対してのモデルの性能を示す。予測した STOI は，Equal rate に漸近することが示された。この時，RMSE は 0.0888 となり，テスト音源に対して予測モデルの精度が高いことが示された。また，先行研究 [7] と比較し，RMSE が 0.0582 低く，モデルの精度が向上した。

Fig. 6 に 4.2 節で作成した主観評価音源に対してのモデルの性能を示す。Fig. 5 と比較し，予測した STOI は実測の STOI と比べ高い値を予測する傾向があることが分かる。また，RMSE が 0.1603 となり，テスト音源より 0.0715 低くなった。学習音源とテスト音源は，予測モデルを汎化させるために騒音を 7 種類用いたが，主観評価音源は室蘭工大で録音した暗騒音のみ用いたため，精度の差が生じたと考えられる。

5.2　STOI から LDR へのマッピング

客観評価指標である STOI から主観評価指標である LDR への回帰は，一般線型モデルであるシグモイド関数でフィットさせた。LDR は，4.2 節で聴取実験をした結果を使用した。

式 (2) にシグモイド関数を示し，Fig. 7 に実測の STOI と実測の LDR のマッピングを示す。実測の LDR と，実測の STOI をシグモイド関数を用いてマッピングした予測 LDR との RMSE が 0.0996 と誤差が小さく，相関係数が 0.9665 と強い相関があることが示された。これは，主観評価値である LDR を客観評価値である STOI で説明することが可能であることを示している。しかし，4.2 節で作成した音源は，IR を畳み込み，騒音を加算した音源であり，実際の屋外拡声器から放送した音源で検討する必要がある。

$$f(d) = \frac{0.9870}{1 + \exp\left(25.9877 \times d - 18.9425\right)} \qquad (2)$$

5.3　結果と考察

4.2 節で作成した主観評価音源に対し，MFCC を音響特徴量として，学習済み STOI 予測モデルを用いて予測し，LDR へマッピングして評価を行った。Fig. 8 に実測の LDR と，STOI を予測し LDR へマッピングした予測値を示す。Fig. 8 より，実測の LDR は 0.00 から 1.00 の範囲であるが，予測した LDR は 0.80 から 1.00 の範囲であることから，LDR 予測が失敗していることを示している。Table 4 に既報 [5] と本稿における RMSE と相関係数を示す。既報 [5] と比較し，RMSE が 0.29 大きくなり，相関係数が 0.17 低くなり，予測精度が低くなったことを示している。

5.2 節で作成したシグモイド関数は，STOI が 0.60 から 1.00 の範囲で急激に変化量が大きくなり，STOI 予測値の誤差に大きく精度が左右されるため，STOI 予測モデルの精度をより高める必要がある。STOI を計算する際の DFT フレーム幅と，RF 学習モデルを作成する際の MFCC のフレーム幅が異なるため，フレーム幅を合わせることが必要である。

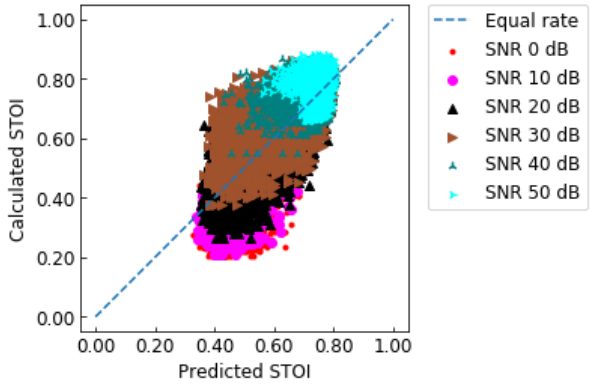

Fig. 5　Relationship between calculated STOI and measured LDR using section 4.3 speech signal

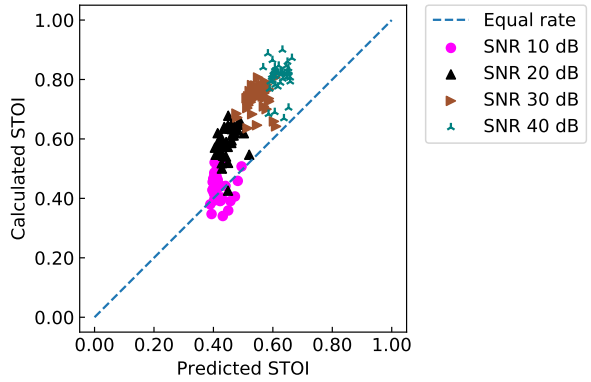

Fig. 6　Relationship between calculated STOI and predicted STOI using section 4.2 speech signal

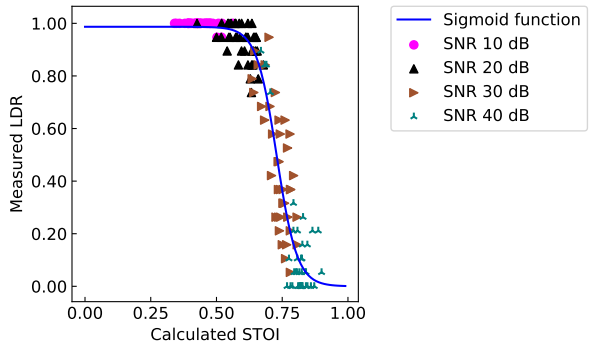

Fig. 7　Relationship between calculated STOI and measured LDR

6　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改善として，MFCC を特徴量とし，RF による STOI 予測モデルを用い，LDR へ回帰することを提案した。その結果，MFCC を特徴量として STOI を予測するモデルは，RMSE が 0.0888 と誤差が小さく，精度が高いことが示された。また，STOI を LDR への回帰曲線は，相関係数が 0.9665 と強い相関があることが示された。しかし，予測した STOI から LDR へマッピングしたところ，実測の LDR を予測した LDR

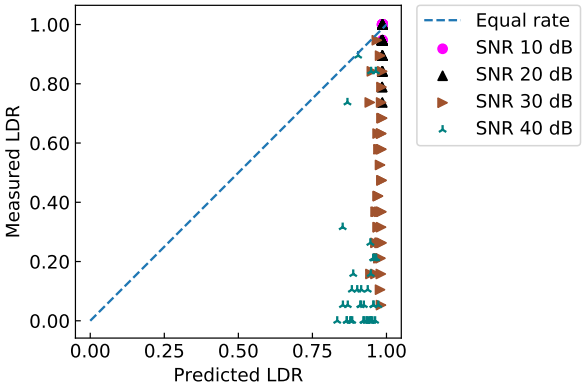

Fig. 8　Relationship between mesured LDR and predicted LDR

Table 4　LDR prediction results

	RMSE	Corr. coef.
Previous research [5]	0.21	0.84
STOI model & sigmoid function	0.4966	0.6672

で RMSE が 0.4966 と誤差が大きいことが示された。今後は，さらなる STOI 予測モデルの精度の向上を検討する。

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 回会合 (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] M. Morimoto *et al.*, J. Acoust. Soc. Am. vol. 116(3), pp.1607–1613 (2004)

[4] Leo Breiman, Machine Learning vol. 45(1) pp.5-32 (2001)

[5] 野口他, 春季音響学会, pp.659–660 (2018)

[6] Cees H. Taal *et al.*, IEEE Trans. Audio., vol. 19(7), pp.2125–2136 (2011)

[7] Deokgyu YUN *et al.*, IEICE TRANS. INF. & SYST., vol. E101-D(4), pp.1207–1208 (2018)