音環境センサネットワークを活用した「よく聴こえる」拡声システム

Easy listening system using sound field feedback sensors

2171011

研究代表者　　室蘭工業大学　　助　教　　小　林　洋　介

［研究の目的］

図１に示す拡声装置は，音声がそのまま伝わるという直感的な動作をする機器であり，駅や空港，学校といった公的な空間を中心にあらゆる場所に設置されている。しかし，直感的に利用できるが故に最適とは程遠い利用による音割れや相手を意識しないボソボソとした発話により聴こえにくく，聴き返しができないため，聴くこと自体を諦めてしまう人も多い。極端に言えば，最適に運用されない拡声装置は誰にも聴かれることがない騒音発生装置であり，市民生活の QOL を低下させている。

東日本大震災では，20 % の市民が聞き取りにくかったと回答している [1]。一方で，北海道地区の事例ではあるが，年齢が高くなるほど，拡声装置の情報をもとに避難を実施している [2]。拡声装置が聴き取り易くなる事は命にも関わる重要な問題である。

本研究では，この問題を解決するために，Internet of things（IoT）技術を利用して，小型のワンボードコンピュータとマイクロホンからなる音環境センサを複数用いて，拡声フィールドの実際の聴取感のフィードバックを検討する。特に音声の聴取感は主観評価による心理実験が必要であるため，計測された物理量と，心理実験値を定式化し，計測システムに組み込む必要がある。このフィードバックを受けて，拡声装置自身が自律的に聴き取りやすいように拡声内容を調整する事で，人間と機械が調和した豊かな音環境構築に貢献する。

図１　屋外拡声器の例

［提案計測システムのフロー］

図２に提案システムのフローを示す。このシステムは，フィールドの音環境を計測する図３のクライアントが５台と計測結果を集約し，聴取実験値に変換するサーバからなる。クライア

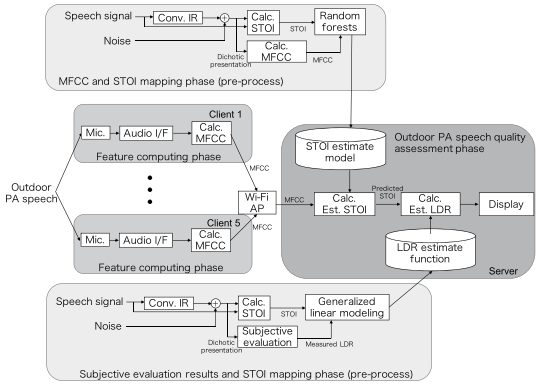

図２　提案システムのフロー

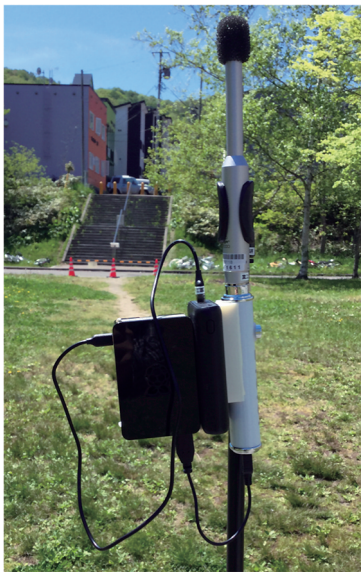

図3　クライアントの外観

ントは Raspberry Pi 3 model B を利用し，オーディオインタフェース（BLUE, ICICLE）に接続したマイクロホン（BEHRINGER, ECM8000）への入力を 100 msec. にセグメント分割して音響特徴量の Mel frequency cepstrum coefficients（MFCC）とパワーの 13 次元およびその Δ パラメータ 13 次元の合計 26 次元を求めた。

セグメントごとの MFCC の計算が終了次第，計算した 26 次元の MFCC ベクトルをサーバに転送する。サーバでは，収集した MFCC を用いて音声の明瞭性/了解性指標として近年広く使われ始めた Short time objective intelligibility（STOI）[3] の推定値を求める。次に主観的な音声の聴き取りにくさの心理実験値である Listening difficulty rate（LDR）[4] の主観評価値と STOI の関数を利用して LDR 推定値を得る。このため，事前に計測したインパルス応答を畳み込んで再現した拡声音声のシミュレーション音声から求めた STOI と MFCC との関係を機械学習アルゴリズムの一つである Random Forest（RF）[5] でモデリングした。LDR と STOI のマッピング関数は一般化線形モデルを用いた。最後に LDR の予測値を画面表示する。

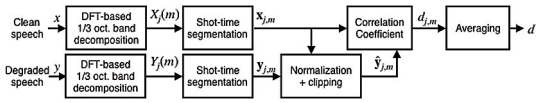

図4　STOI の計算フロー

［Short time objective intelligibility］

C. H. Taal et al. が提案した STOI は，時間周波数モデルに基づく知覚的歪みをモデル化した了解度指標である。STOI の処理フローを図 4 に示す。

STOI の時間周波数分解は，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 50 % オーバーラップで処理する。クリーン音声の最大振幅より 40 dB 低い無音領域フレームを除去して再構成した信号に対し，中心周波数が 150 Hz から約 4.3 kHz までの 15 帯域の 1/3 オクターブバンド分析で周波数エンベロープを求め，時間－周波数ユニットとして，クリーン音声 x と雑音加算音声 y のエンベロープ $X_j(m)$ と $Y_j(m)$ を求める。

次にクリーン音声と雑音加算音声の周波数エンベロープ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 $Y_{j,m}$ を求める。この後に $Y_{j,m}$ を正規化し，了解度に強い影響を持たないグローバルなレベル差を補正した $\hat{Y}_{j,m}$ を得る。

最後に $X_{j,m}$ と $\hat{Y}_{j,m}$ 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，平均して了解度指標値 d を得る。STOI の標準ではこの後音声の了解度へのマッピングを行うが，本項では LDR とのマッピングとした。

［Listening difficulty rate］

Morimoto et al. が提案した主観評価指標 LDR は，音声了解度は 100 % に飽和するが，まだ聴き取りにくいと感じる領域の評価に利用される主観評価指標である [4]。

表 1　LDR の主観評価指標

index	説明文
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

被験者に提示した音声単語の正答率である音声了解度はシステムの評価によく利用されている。しかし，被験者が評価単語を能動的に傾聴するため，劣悪な音声サンプルであっても正答率が高くなる傾向にある。突然の放送に耳を傾けるような受動的な音声聴取が想定される屋外拡声システムは，放送冒頭からの十分な傾聴は期待できず，受動的に聴き始めるため，了解性が高いことに加え，聴き取りにくくはないことが重要である。LDR の主観評価は，被験者に雑音加算音声を評価音として提示し，表 1 の指標値から一つを選択させる。評価音ごとに全被験者の結果から，「聴きとりにくい」という表記のある L2–4 の割合を算出し，指標としている。

本実験では日本語音素バランス文 160 文に実験フィールド 10 箇所のインパルス応答を畳み込み屋外騒音を加算して再現したシミュレーション音源を利用して LDR の主観評価を被験者 25 名に対して行った。

[STOI 推定モデル]

図 2 のサーバ内の MFCC から LDR を推定する処理は RF モデルを用いた。LDR の評価音源長は最長で 11 sec. を超えたため，100 msec. ごとに計算した MFCC を統合し，3120 次元の入力を持つ RF モデルとした。RF は複数の決定木を統合する学習アルゴリズムであり，個々の決定木の数と決定木の深さを指定できる。図 5 に決定木の深さ 50–200 とした際の Root mean squared error（RMSE）の比較結果を示す。結果と計算速度を考慮して深さ 75 の決定木数 100 とした。

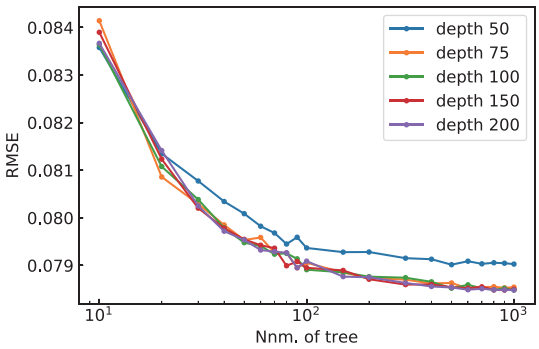

図 5　STOI 予測時の決定木深さと RMSE

[LDR 推定モデル]

STOI と LDR のマッピングは図 6 となり，一般化線形モデルでシグモイド関数をフィットさせて得た関数は以下の式で，決定係数は 0.93 で，RMSE は 0.102 であった。

$$f(\mathrm{d}) = \frac{0.9888}{1+\exp(25.5809\times \mathrm{d} - 18.6406)}$$

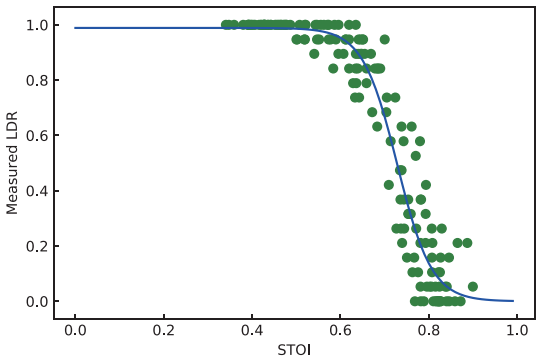

図 6　STOI と LDR の関係

[屋外での計測条件]

2018 年 6 月 23 日土曜日（天候：晴）に室蘭工業大学の V/R 棟前の中庭で実機の動作テストを行った。フィールドの様子を図 7 に示す。実験は建物による反射を意図して V/R 棟に向かってメガホン（TOA, ER-2830W）を設置し，サーバから音声放送できるように接続した。音量はメガホンの前 2 m に設置した騒音計で 80,85,90 dB とし，音素バランス文を 50 文づつ放送した。クライアントの設置場所は前年度の

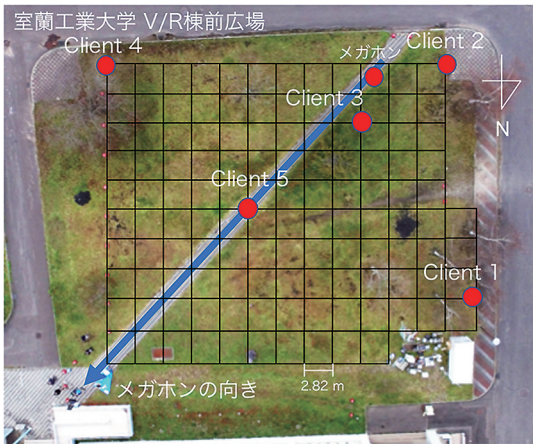

図 7　屋外実験のフィールド

予備実験の際に収集したインパルス応答と LDR の主観評価結果から選択した。サーバとクライアントは Wi-Fi の屋外で利用できる周波数帯である 2.4 GHz 帯で接続した。

［実験結果］

表 2 に 85 dB 放送時の推定 STOI と推定 LDR の平均値をそれぞれ示す。結果より，メガホンにもっとも近い Client 3 の聴き取りにくさが 0.575 と最も低くこのフィールドの中では最も聴きとりやすいことがわかる。次は直線上にある Client 5 の 0.717 であり，距離は近いが真横にあってメガホンの指向性の影響を受ける Client 2 が 0.825, 遠方の Client 1,5 は 0.9 以上とほとんど聴き取れないことが明らかである。

図 8 と 9 に，Client 3 と 1 の推定 LDR の時間分布を示す。結果より，Client 3 は放送のタイミングで LDR が低い谷を深く形成するのに対し，Client 1 は常に推定 LDR が高く，ほとんど音声が到来していないことがわかる。

表 2　Client ごとの平均 STOI 推定値と LDR 推定値（85 dB 放送時）

Client	Est. STOI	Est. LDR
1	0.555	0.961
2	0.638	0.825
3	0.698	0.575
4	0.592	0.925
5	0.676	0.717

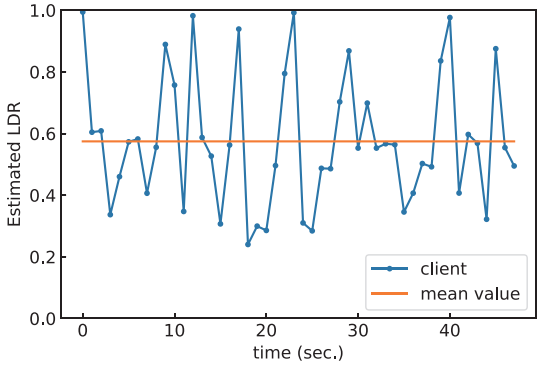

図 8　Client 3 の推定 LDR 分布

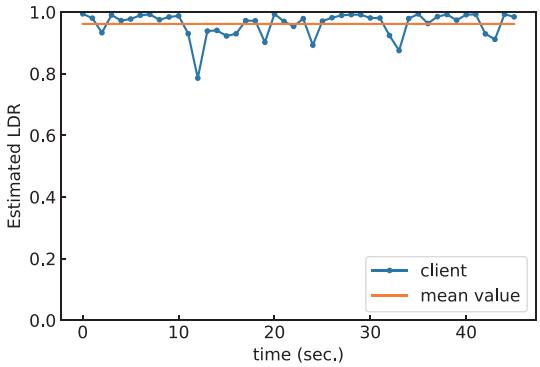

図 9　Client 1 の推定 LDR 分布

［今後の課題］

環境をフィードバックするセンサのプロトタイプは完成し，フィールド内の聴き取りにくさがフィードバックされるようになった。しかし，拡声音声との連携部まではまだ完成していない。並行して開発したより聴こえやすい放送システム（成果発表 [4]）は，入力音声を認識し，自然言語処理技術による構文解析を行い，音声合成して聴きとりやすくする。しかし，近年の合成音声技術では自然度の高い音声が作れるようになったものの，まだまだ自然音声の品質に到達していなく，不明瞭でも自然音声の方が高品質であった。このため，より高品質な音声合成の実装と並行して長期的にセンシングシステムと連携した総合的な拡声音声システムを構築して行きたい。

［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］ 田中岳, 渡部靖憲, 中津川誠, “2011 年東北地方太平洋沖地震津波に対する北海道沿岸域住民の避難行動調査”, 土木学会論文集 B2（海岸工学）, 69 巻, 1 号, (頁 48～63), 2013 年

［3］ Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, “An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech”, IEEE Trans. Audio., Speech, Language Processing, vol. 19, no. 7, pp. 2125-2136, Sep. 2011.

［4］ M. Morimoto, H. Sato and M. Kobayashi, “Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces,” J. Acoust. Soc. Am., Vol. 116 (3), pp. 1607-1613, 2004.

［5］ Leo Breiman. “Random Forest”, Machine Learning 45 (1) : 5-32, 2001.

［成果の発表, 論文等］

［1］ 児島悠太, 小林洋介, “音環境センシングシステムの開発”, 電子情報通信学会技術研究報告, vol. 117, No. 138, EA2017-22, pp. 125-130, 2017 年 7 月 20-21 日

［2］ 児島悠太, 小林洋介, “音環境センサネットワークの基礎検討”, 日本音響学会 2017 年秋季研究大会講演論文集, 1-P-40（Poster), pp.749-750, 2017 年 9 月 25-27 日

［3］ 野口啓太, 小林洋介, 岸上順一, 栗栖清浩, “屋外録音音声の聴き取りにくさ予測に用いる機械学習手法の比較”, 平成 29 年度電気・情報関係学会北海道支部連合大会講演論文集, 143, pp. 210-211, 2017 年 10 月 28-29 日

［4］ 赤泊寛和, 石川耕輔, 太田健吾, 小林洋介, 岸上順一, “HMM 音声合成と発話構文解析を利用した「よく聴こえる」拡声システム”, 情報処理学会研究報告, Vol. 2018-MUS-119 No. 51, 4 pages, 2018 年 6 月 16-17 日

［5］ 小林洋介, 野口啓太, 栗栖清浩, “拡声音声への環境情報フィードバックシステムの基礎検討”, 聴覚研究会資料, Vol. 48, No. 6, pp. 591-596, 2018 年 8 月 23-24 日