音声認識による言語特徴を用いた屋外拡声器品質計測器の構築

代表研究者　　小　林　洋　介　　室蘭工業大学　大学院工学研究科　助教
研究分担者　　太　田　健　吾　　阿南工業高等専門学校　講師

1　はじめに

2011 年 3 月に発生した東日本大震災では，約 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。本研究では，複数被験者による音声の聴き取り易さ指標である了解度[3] または関連する指標である主観的聴き取りにくさ[4]を予測する品質予測モデルを開発する。これは単純な音の大きさを計測する騒音計といった既存の計測器と異なり，音の物理量だけではなく，人工知能的側面を持つ音声認識結果を組み込む必要があるため，以下の 3 点に取り組んだ。

1. 了解度計(聴き取りにくさ計）に組み込む音声認識システムの最適化
2. 予測モデルの教師データとなる了解度(聴き取りにくさ）の主観評価実験
3. 音声認識を利用した音声了解度(聴き取りにくさ）予測モデルのプロトタイプ作成

本研究の計画段階では音声の品質として，有意味単語の正答率である了解度の主観評価値を重視していたが，後半では，より屋外拡声器の品質評価に適していると考えられ，機器メーカーでも利用が始まっている主観的聴き取りにくさ[4]の予測問題にシフトした。最終的にプロトタイプを開発したシステムは，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である。この計測器は，入力音声の特徴量として MFCC(Mel Frequency Cepstrum Coefficients）を求め，音声区間判別 VAD(Voice Activity Detection）を行い，主観品質である聴き取りにくさの指標 LDR(Listening Difficulty Rate)[4] を機械学習法の 1 つである RF(Random Forest）で作成した予測モデルの結果を表示する。

本報告執筆時点では，最終的な人間の評価と提案計測器との比較が終了していないため，実用上の限界性能が確定していないが，この計測器は本研究における最終的な利用形態に限りなく近い形で実際に動作するシステムである。

2　音声認識システムと了解度

2.1　概要

拡声音声のみから了解度を予測するためには，音響伝搬系による劣化を観測信号から求める必要がある。音声認識を用いた予測の基礎的な検討では，音響系の適応を行っておらず適応範囲に限界があった[8]。そこで，本稿では屋外拡声系に適応するため，拡声器のインパルス応答を畳み込んだ音声で学習した音声認識システムを構築し，その出力値を利用した了解度予測モデルを作成し，主観評価結果と比較する。

2.2　音源の設定と主観評価値

評価音声は，親密度別単語了解度試験用音声データセット 2007(FW07)[3] の高親密度語女性 1 話者分用いた。FW07 の音源に，東日本大震災後に仙台市若林区荒浜小学校周辺で収集されたスピーカによる TSP 信号を 10 地点分畳み込み評価音源を作成した。以後，各地点を p01～p10 とする。また， 実際の拡声音声システムでは，無線通信等で電話帯域に帯域制限されて伝送されているため，原音声を電話帯域に帯域制限した後に伝送系のインパルス応答を計算機上で畳み込んだのちに，音量のゲイン調整を行った。評価は予測モデルの作成用データとテスト用データに分けるため 2 セット行った。Set 1 は被験者 22 名分，Set 2 は Set 1 と異なる評価リストを用いた被験者 13 名分であり，被験者のほとんどは 10 代後半の高専生である。評価結果は Table 1 に音声認識結果と共に示す。評価音源の設定と主観評価の詳細は本研究の呼び検討であった文献[9]を参照されたい。

2.3　音源の設定と主観評価値

屋外拡声音声の特性を考慮した音声認識器を構築するために，Julius 音響モデルの環境適応を行う。ベースラインの音響モデルとして，Julius のディクテーションキットに付属する GMM 版音響モデルを用いる。このモデルは，ASJ-JNAS コーパス 86 時間分から学習された性別非依存 triphone モデルである(3 状態 LR 型対角共分散 HMM)。特徴量は MFCC12 次元とそれらの 1 次差分およびエ ネルギーの 1 次差分の計 25 次元を用い，ケプストラム平均正規化を適用している。ベースラインの音響モデルに対し，評価音源と同様に屋外拡声系のインパルス応答を畳み込んだ音素バランス文を適応データとして，MAP 推定に基づく環境適応を行う。適応データには，ASJ-JIPDEC[11] に収録さ れた 12 名の話者(男性 6 名，女性 6 名）による ATR 音素バランス 503 文に対し，前述した 10 地点において測定されたインパルス応答を畳み込んだものを用いる。

デコーダには Julius rev.4.3.1 を用い，音節タイプライタ用文法を用いて連続音節認識を行い，音素単位の認識率で評価した。評価単語に用いた FW07 の認識率を正答率 Corr. と挿入誤差を考慮した Acc. で評価する。屋外拡声器への適応を行った音声認識システム(Adap.）と適応を行っていないベースライン(Base.）の評価結果と主観評価のセット別の平均了解度（Intell.）を Table 1 に示す。結果より，Corr.は適応により下がっているが，より実態に近い認識率指標である Acc. は拡声器への適応で大幅に改善している。適応モデルの Acc. と主観評価による了解度の相関係数を求めると，0.41 と若干の相関はあるが，人間の了解度評価による知覚モデルとの差は未だ大きく，今後の改良が必要である。

Table 1Speech intelligibility and recognition results.

	Intell.		Base.(%)		Adap.(%)	
	Set1	Set2	Corr	Acc.	Corr	Acc.
p01	92.3	90.9	48.8	-4.6	41.3	32.5
p02	78.8	74.1	44.0	0.6	38.1	27.8
p03	96.9	92.9	46.4	-13.4	38.7	31.1
p04	88.0	78.4	47.7	-5.3	39.4	28.6
p05	91.1	83.2	46.3	-6.9	37.9	29.7
p06	90.4	91.8	46.9	-14.4	37.8	26.4
p07	91.5	82.5	46.1	-11.9	36.4	28.4
p08	94.2	80.5	46.4	-10.6	36.5	25.7
p09	86.9	90.6	45.2	-8.1	37.8	29.4
p10	85.0	87.0	49.1	-4.5	37.7	28.0

Table 2Speech intelligibility and recognition results.

	Intell.(%)	Base.(%)	Adap.(%)
p01	91.6	-4.6	32.5
p02	76.5	0.6	27.8
p03	94.9	-13.4	31.1
p04	83.2	-5.3	28.6
p05	87.2	-6.9	29.7
p06	91.1	-14.4	26.4
p07	87.0	-11.9	28.4
p08	87.3	-10.6	25.7
p09	88.8	-8.1	29.4
p10	86.0	-4.5	28.0

2.4　音声認識と機械学習による了解度予測モデル

音声認識の出力結果の相関係数が 0.5 未満であり，現状の認識結果のみでは了解度の予測は困難であると考え，我々が以前提案した被験者の反応を模擬する判別器を機械学習で作成した了解度予測システムの特徴

量として用いることを検討する。特徴量に認識結果を用いる方法は幾つかあるが，先行研究[8] と同様に認識に用いた対数尤度スコアを利用する。これは，認識率である Corr. と Acc. は正答のテキストが必要な指標であるため，観測された音声信号のみでは求められないためである。

2.5 音源の設定と主観評価値

了解度予測には我々がこれまでに提案している，ノンブラインド音響特徴量を用い被験者の反応を模擬する手法をベースに音声認識部を付加する。予測法のフローを Fig. 1 に示す。この手法は学習用の主観評価を複数被験者に対して行い，その正答と誤答のレスポンスを模擬する判別器を被験者の数だけ作成し，主観評価と同様の平均処理によって予測値を得る。判別器は SVM(Support vector machine)[12]とし，ハイパーパラメータを学習データで最適化する。個々の判別器に用いるノンブラインド音響特徴量は，ITU-T P.563 勧告[13] の内部特徴量 12 次元と拡声器のインパルス応答を適応した音声認識システムの尤度スコアの計 13 次元を用いた。これまでの検討より，被験者数の多い主観評価 Set1 で学習した方が予測精度が高くなることがわかっているため，本稿では特徴量に音声認識に用いた対数尤度を加えない場合とを比較する。比較は予測値と主観評価値との相関係数と平均二乗誤差 RMSE を指標に用いる。

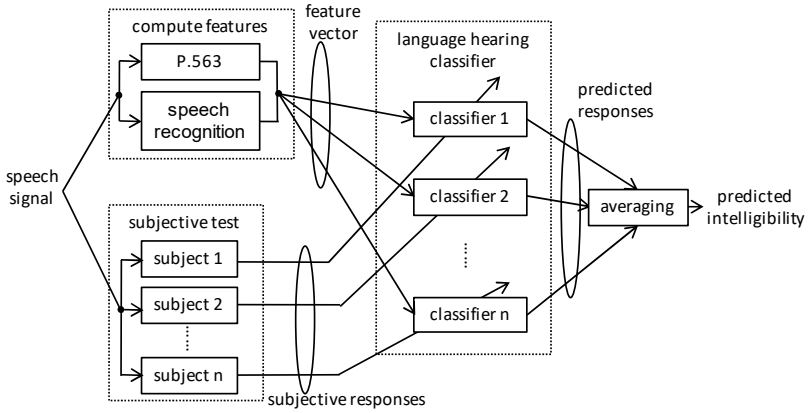

Fig. 1 Intelligibility prediction flow.

2.6 予測結果と考察

予測実験の結果を Table 3 に示す。表の Conventional は音声認識による尤度スコアを用いていない場合で前報告[9]の値であり，Proposed は音声認識を組み込んだ提案法である。学習したデータ（主観評価 Set 1 の予測）の結果を Train に，学習していないテストデータ（主観評価 Set 2 の予測）の結果を Test に示す。結果より，提案法は Train データの相関と RMSE がわずかに悪化しているものの，テストデータの相関が 0.547 から 0.724 と向上し，RMSE が 0.005 ポイント下がった。この結果は，新しく加えた音声認識結果による効果である。Fig.2 に提案法の主観値と予測値をインパルス応答ごとにプロットした。図より，学習したデータはほぼ対角線上にあり，テストデータも概ね対角線に沿うようにマップされており，本予測法のロバストさがわかる。

Table 3 Intelligibility prediction results.

	Conventional		Proposed	
	Train	Test	Train	Test
Correlatio	0.965	0.547	0.940	0.724
RMSE	0.023	0.041	0.040	0.035

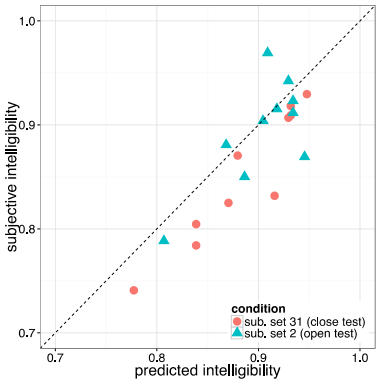

Fig. 2 Subjective vs. predicted intelligibility by test conditions. Doted line is equal rate.

3　聴き取りにくさ計の開発

3.1　概要

前節で開発したモデルは，MFCC による音声認識の音響モデルと言語モデル，P.563 による音声品質モデルの３つを混合し，多数の SVM を構築するという点で計算コストがかかっていた。このため，最終目的のハンドヘルド計測器とするには課題が多い。加えて，東日本大震災後の屋外拡声器のインパルス応答を用いた正確な音声了解度を用いたが，学習データの収集に行う心理実験としては，実験規模（１被験者あたり数時間）の割にデータが集まりにくいという問題を抱えていた。

そこで，全体の計算と主観評価過程の簡略化を行う。特に近年大きく発展してきた，シングルボードコンピュータでも運用可能な軽量なシステムとすることを目標とした。このため，前節のシステムで効果があったと考えられた音声認識の音響モデルに利用される入力音声の特徴量として MFCC(Mel Frequency Cepstrum Coefficients) の時系列信号のみに特徴量を絞り込み，計算コストが線形時間で設計しやすい機械学習法の 1 つである RF(Random Forest) を利用する。また，了解度よりも拡声器など品質評価指標として利用されている主観品質である聴き取りにくさの指標 LDR(Listening Difficulty Rate)[4] を用いた“聴き取りにくさ計” とすることにした。LDR は，1 つの文章を用いた評価も可能であるため，被験者の拘束時間を短くすることが可能である。さらに，リアルタイム動作時には，今解析している信号が音声かどうかの判別が必要である。そこで，VAD(Voice Activity Detection：音声区間判別) も RF で実装し，音声区間の聴き取りにくさを求めることとした。

3.2　システムフロー

Fig.3 に提案する計測器のフローを示す。この計測器は，2 台のシングルボードコンピュータを用い，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。

実機システムには，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 AD 変換し，Board 1 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 とパワーおよび，delta パラメータを算出する。次に，Board 2 で学習済みモデルを使用して，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。開発したシステムの外観を Fig.4 に示す。

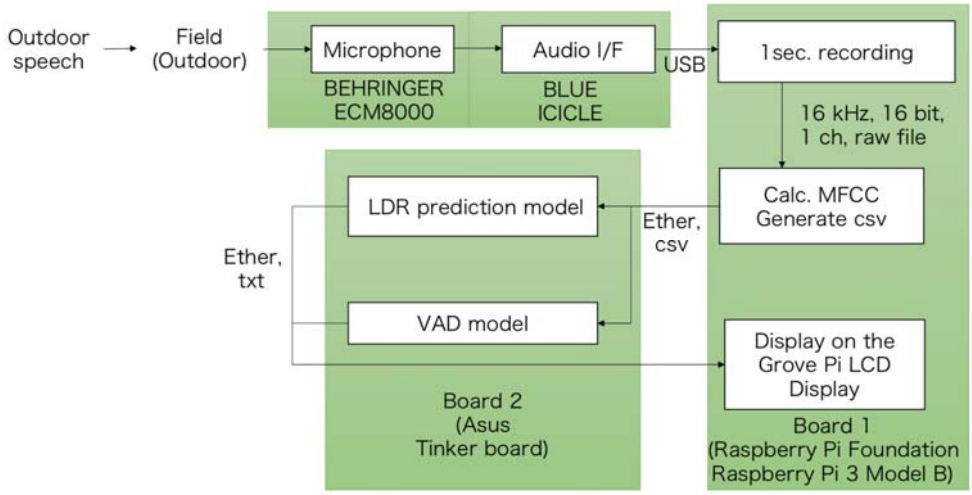

Fig. 3 Proposed LDR meter flow

Fig. 3 Proposed LDR meter

3.3 屋外での音源収集

VAD と LDR 予測を行うため，機械学習を用いた。機械学習を行う際，多数の音源が必要である。また，本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，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 Fig.5 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, 本助成で整備）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 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 地点で録音した。このインパルス応答と B.G.N. および，先行研究［5］の音源を使い，VAD と LDR 予測のモデルを学習した。

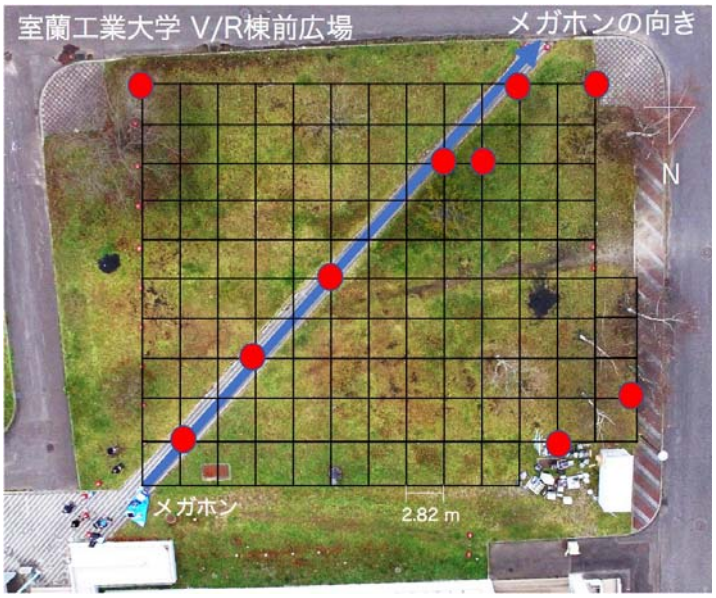

Fig.5 IR measurement points in Muroran-IT: black crosses are IR measured point, red circles are used subjective evaluation.

3.4 RF による VAD

VAD の学習音源は，Table 4 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，フレーム長 100 msec. としてパワーと MFCC を 12 次元，およびそれらのデルタパラメータとして全 26 次元を利用した。音声区間および非音声区間として，Table 4 に示す VAD の総音源数 3593038 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。delta パラメータの重要度が 0.015 以下と低く，delta を除いた 13 特徴量の重要度の積算が 0.927 であったため，delta 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。

Table 5 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。本計測器の録音は 1 sec. ごとに行うが，MFCC のフレーム長が 100 msec. であるため，1 sec. の音源では 10 回 MFCC を求め，それぞれ音声区間判別を行う。

Table 4 Test signal generate conditions

用途	VAD		LDR	
	非音声区間	音声区間	主観評価	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 から 80 dB を 10 dB 間隔		
騒音レベル	40 から 80 dB まで 1 dB 間隔	40 dB, 45 dB, 50 dB	40 dB	
Sampling rate	16 kHz		48 kHz	16 kHz
音源長	0.1 sec.		文章長+遷移区間	1.0 sec.
総音源数	510860 (騒音数 × 騒音レベル × 1780 (ノイズを切り出す数))	3082178 (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

Table 5 VAD results

真のクラス＼予測クラス	音声	非音声
音声	924059	811
非音声	3840	149202

3.5 音源の設定と主観評価値

LD の主観評価音源は，Table 4 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 19 人(男性 18 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「聴き取りにくくはない」，「やや聴き取りにくい」，

「かなり聴き取りにくい」，「非常に聴き取りにくい」の 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 で回答した。実験の前に，聴取と回答の練習を兼ねて，STI 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。

3.6 RF による LDR 予測結果

前節の主観評価音源を 1 sec. ごとに切り出し，delta を含む 26 次元の特徴量を 100 ms ごとに求めて，LDR 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。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 [5] 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，5.1 節の主観評価音源数は 744 であり，先行研究 [5] の音源数は 392 である。LDR 予測モデルは，これらを合わせた 1136 音源のうち，約 70%をランダムに選びトレーニングデータとしてモデルの学習に，残りの約 30%を予測精度のオープンテストに用いた。シングルボードコンピュータに実装すること，音源数が 1136 と少なく，RF による VAD モデルより決定木の深さが浅くなることを考慮し，RF の決定木の数を 1000 とした。

Fig.6 にオープンテストの主観評価による LDR(Subjective LDR) と予測した LDR(Predicted LDR) を示す。5.1 節の主観評価音源でのテストデータに対して，誤差が 0.65 と大きく外れるデータが存在するが，先行研究 [5] のテストデータに対しては RMSE が 0.15 と主観評価による LDR に漸近する。Table 6 に先行研究 [5] と本稿における RMSE，相関係数，決定係数を示す。先行研究 [5] と，本稿で決定係数が低かった先行研究 [5] の音源を比較し，決定係数が男声と比較すると 0.11 低いが，女声と比較すると 0.13 高い。また相関係数が先行研究 [5] の音源に対して 0.88，5.1 節の主観評価音源に対して 0.84 と強い相関がある。よって本計測器で実用できる可能性があることを示している。

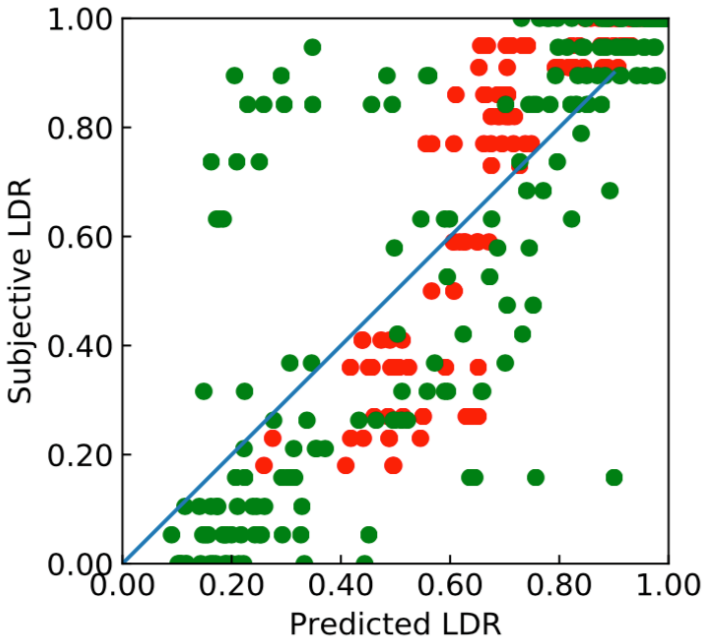

Fig. 6 Relationship between subjective and objective LDR: red dots are evaluated at previous reserch[5], green dots are evaluated at section 5.1, blue line is equal line.

4　まとめ

本研究では，以下の 2 つの手法による屋外拡声器の品質評価モデルを開発した。

1. 人間の音声品質聴取と多数の判別器による主観評価を模擬した了解度予測モデル
2. 音声認識に利用される MFCC を利用した聴き取りにくさ計のプロトタイプ

2. の聴き取りにくさ計は，実際に動く実システムとなったため，最終的な実用化に向けてより詳細な検討を加えていく。特に現時点では実現できていない，人間の主観評価との同時実験による予測精度評価を今後行って行く。そして，最終的な屋内外の拡声システムの設計に利用可能な計測器としての実用化を目指す。

5　謝辞

公益財団法人 電気通信普及財団による本助成により，研究が大きく進行したことに感謝いたします。また，本研究の進行に必要な有益な助言とデータを提供をいただいた TOA 株式会社栗栖清浩様，異動直後に本研究を実施する環境を整えてくださいました室蘭工業大学の岸上順一教授，聴き取りにくさ計の実装に協力いただいた室蘭工業大学学生の野口啓太君，そして実験に協力していただいた被験者各位に感謝いたします。

【参考文献】

[1] 内閣府，“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報告，” 2011.
[2] 日本音響学会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，“災害等非常時屋外拡声システム性能確保のための規準案（第 1 版），” 2015.
[3] 近藤公久，天野成昭，坂本修一，鈴木陽一，“親密度別単語了解度試験用音声データセット 2007(FW07) の作成，” 電子情報通信学会技術研究報告 TL2007-62, pp.43-48, Jan. 2008.
[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] 栗栖清浩, “変動量解析による屋外拡声音声の明瞭性評価,” 聴覚研究会資料, H-2015-74, pp.405-411 (2015)
[6] IEC 60268-16, “Sound system equipment-Part 16：Objective rating of speech intelligibility by speech transmission index,” IEC 60268-16:2011, 2011.
[7] 佐藤逸人, 崔正烈, 坂本修一, 鈴木陽一, 森本政之, 青木雅彦, 小池宏寿, 高島和博, 鶴秀生, 光枝太一, “音声了解度による屋外拡声システムの評価-総務省平成 23 年度 3 次補正予算による技術開発- ,” 日本音響学会 2013 年秋季研究発表会講演論文集, pp. 1533-1536 (2013).
[8] 栗栖清浩, 川島佑亮, 安啓一, 荒井隆行, “音声認識技術を用いた明瞭性評価の試みー屋外拡声 音の「聴き取りにくさ」と Julius 尤度の関係ー,” 日本音響学会 2014 年秋季研究発表会講演論文集, pp. 1087-1088 (2014).
[9] 小林洋介, 近藤和弘 “判別器を用いた屋外拡声音声了解度の予測法,” 電子情報通信学会技術研究報告 EA, 302, Vol. 115, pp. 43-48 (2015).
[10] julius: http://julius.osdn.jp/
[11] 磯健一, 渡辺隆夫, 桑原尚夫 “音声データベース用文セットの設計,” 日本音響学会 1988 年春季研究発表会講演論文集, pp. 89-90 (1988).
[12] V. Vapnik, “The Nature of Statistical Learning Theory; Statistics for Engineering and Infor-mation Science,” Springer, 1995.
[13] ITU-T, “Single-ended method for objective speech quality assessment in narrow-band tele-phony applications, ITU-T P.563,” March 2004.

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
機械学習と音声認識による拡声音声品質予測	情報処理学会研究報告, Vol.2016-MUS-111	2016 年 5 月 21-22 日
Speech intelligibility prediction method using machine learning for outdoor public address systems	5th Joint Meeting Acoustical Society of America and Acoustical Society of Japan	2016 年 11 月 28 日 12 月 2 日
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	2017 年 10 月 28-29 日
Listening Difficulty Meter Using the ITU-T P.563 Feature Set for Public-Address System	Joint Seminar on Environmental Science and Disaster Mitigation Research 2018	2018 年 3 月 2 日
屋外用聴き取りにくさ計のプロトタイプ開発	日本音響学会 2018 年春季研究大会	2018 年 3 月 2 日
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	情報処理学会研究報告, Vol.2016-MUS-119	2018 年 6 月 15-16 日 (発表予定)