Vol.2016-MUS-111 No.43
2016/5/22

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

小林 洋介[1,†1,a)]　太田 健吾[2]　近藤 和弘[3]

概要：屋外拡声器からの音声の聴こえにくさを予測するシステムを構築した．提案システムは拡声音声に適応した音声認識システムの出力とブラインド信号処理による音響特徴量を用いて「聴き取れる」か「聴き取れない」かの判別モデルを被験者ごとに機械学習によって構築し，その出力値を主観評価と同様に平均し了解度予測値を求める．このシステムの予測誤差は構築した条件において主観評価値に対して 4%以下と高い精度を持つことを確認した．

1. はじめに

東日本大震災では，避難放送などの屋外拡声器の音声が聴き取れなかった事例 [1] があったため，その改善のための技術規準ができた [2]．技術基準ではシステムを，音源系，信号伝送系，音響出力系，音響伝搬系の 4 段階ごとに技術基準を設定している．最終出力である伝搬系では，聴取位置での拡声音の品質に問題が無いかが性能確認規準になっている．拡声音の品質としては，評価地点の音圧レベルと明瞭性の聴取試験が挙げられており，主観評価実験である了解度試験の実施が推奨されている．

しかし，主観評価実験では評価コストが膨大になるあらゆる環境での評価は困難である．これまでに拡声系のインパルス応答より計算した STI（Speech Transmission Index）[3] などの伝送指標を利用した了解度予測の可能性が示されている [4]．一方で，生活空間でのインパルス応答の取得は突発的な外乱による測定値のバラつきが発生しやすく困難であるとの指摘もある．

この問題を解決するために，本稿では実際に拡声された音声のみから了解度を予測する手法を提案する．具体的には，拡声音声に適応した音声認識システムの出力とブラインド信号処理による音響特徴量を用いて「聴き取れる」か「聴き取れない」かの判別モデルを被験者ごとに機械学習によって構築し，その出力値を主観評価と同様に平均し了解度予測値を求める手法である．本稿では，提案法の構築に用いた主観評価結果と Julius[5] の音声認識結果を述べたのちに，提案予測法の概要と提案法を用いた予測結果について述べる．

[1] 都城工業高等専門学校
Yoshio, Miyakonojo, Miyazaki 885–8567, Japan
[2] 阿南工業高等専門学校
Aoki Minobayashi, Anan, Tokushima 774–0017, Japan
[3] 山形大学
Jonan, Yonezawa,Yamagata 992–8510, Japan
[†1] 現在，室蘭工業大学
Presently with Mizumoto, Muroran, Hokkaido 050–8585, Japan
[a)] yosuke_kobayashi@m.ieice.org

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 [6]．特に屋外拡声器品質で議論されるロングパスエコーによる妨害は，音声に対する音声の妨害であるため，ブラインド信号処理によるノイズリダクション等で前提とする音声と雑音の統計量の異なりといった物理的な特性を利用できない．しかし，人間は伝送系に関する知識が特になくともこれまでの経験に基づいて音声品質の主観評価を行うことが可能である．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)[7] の高親密度語女性 1 話者分を用いた．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別報 [8] を参照されたい．

2.3 音声認識システムの構築と認識率

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[9] 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．

2.4 主観評価条件間の分析

評価セット間での了解度のピアソン相関係数は 0.518 で，インパルス応答ごとの結果間の平均二乗誤差（RMSE）は 0.070 であった．これは物理条件が一致し，評価単語が異なる場合に 7%の誤差で再現可能であることを意味する．このときの評価セット間の関係を図 1 に示す．テストセット 2 の方が了解度が高くなっている地点があり，先に述べた傾向差のある評価地点と一致する p05，p07 及び p08 と p04 である．了解度予測を高精度に行うためには，この傾向差を再現できるようにする必要がある．テストセット間

表 1 Speech intelligibility and recognition results.

	Intell.		Base.(%)		Adap.(%)	
	Set1	Set2	Corr.	Acc.	Corr.	Acc.
p01	0.92	0.91	48.8	-4.6	41.3	32.5
p02	0.79	0.74	44.0	0.6	38.1	27.8
p03	0.97	0.93	46.4	-13.4	38.7	31.1
p04	0.88	0.78	47.7	-5.3	39.4	28.6
p05	0.91	0.83	46.3	-6.9	37.9	29.7
p06	0.90	0.92	46.9	-14.4	37.8	26.4
p07	0.92	0.83	46.1	-11.9	36.4	28.4
p08	0.94	0.80	46.4	-10.6	36.5	25.7
p09	0.87	0.91	45.2	-8.1	37.8	29.4
p10	0.85	0.87	49.1	-4.5	37.7	28.0

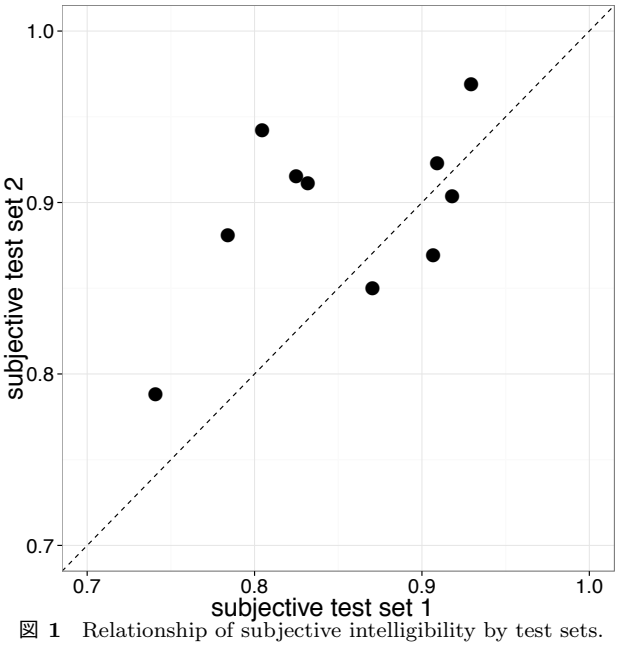

図 1 Relationship of subjective intelligibility by test sets.

の RMSE 値は人間による別の評価環境を評価した際の誤差量であるため，機械で予測した場合にこれ以下の誤差であれば，人間による評価と同等であるといえる．よって，テストセット 1 とテストセット 2 間の相関係数 0.518 と RMSE7.0%を予測実験の目標値と設定する．

3. 音声認識と機械学習による了解度予測

3.1 概要

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 [6] 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．

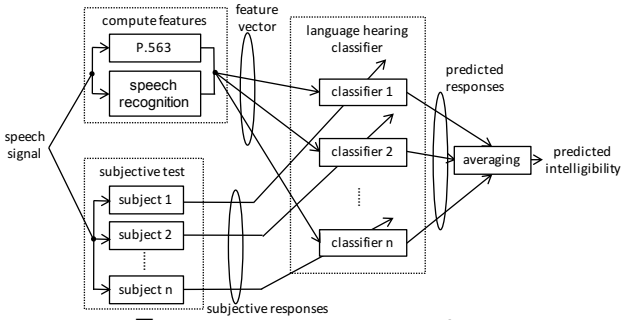

図 2　Intelligibility prediction flow.

3.2　提案する予測システムと特徴量

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 [8] をベースに音声認識部を付加する．予測法のフローを Fig. 2 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)[10] 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 [11] 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．

3.3　予測結果と考察

予測実験の結果を Table 2 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 [8] 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. 3 に Fig. 4 従来法の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，どちらも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いる．しかし，提案法の方で加えた音声認識の尤度スコアの差はあまり明確ではない．これは従来法の段階で RMSE 自体は十分に小さかったためである．

表 2　Intelligibility prediction results.

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

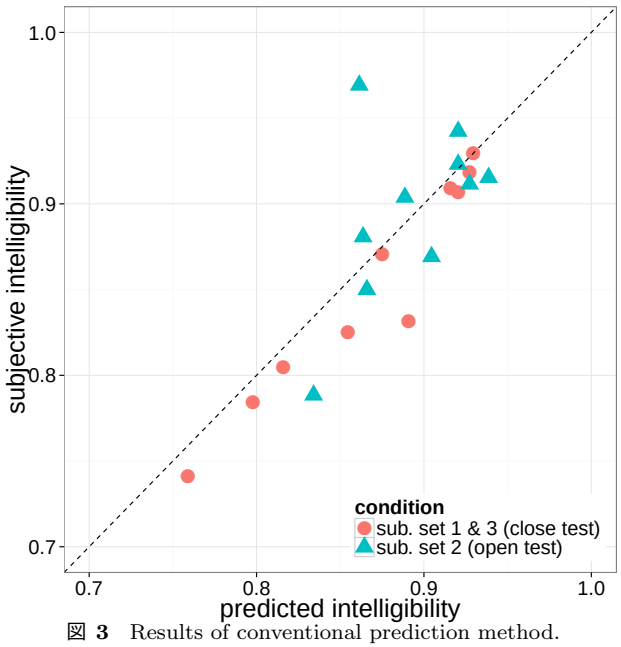

図 3　Results of conventional prediction method.

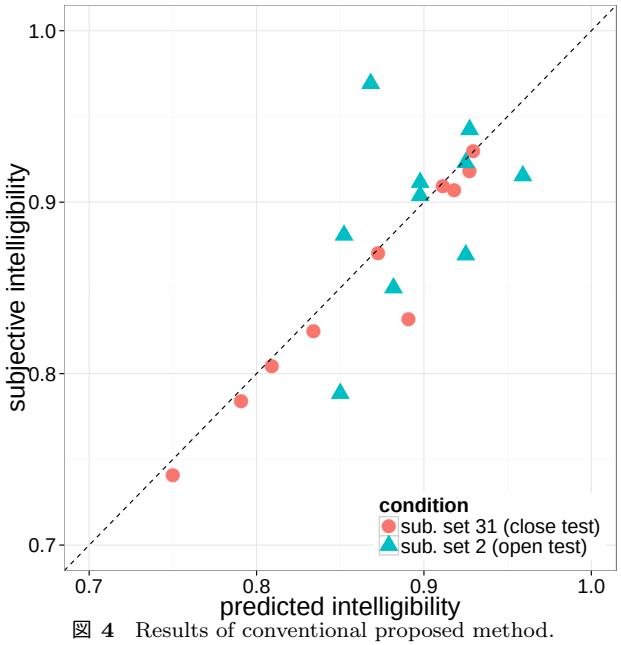

図 4　Results of conventional proposed method.

4.　まとめと今後の課題

我々が提案してきた屋外拡声器の了解度予測法に拡声音声で適応した音声認識を組み込み，計算上で利用する尤度スコアを利用して予測精度を向上させた．今後は音声認識部へ適応させたインパルス応答と主観評価のインパルス応答を変えるなど提案システムのロバスト化へ取り組むとともに，学習させる主観評価データ数の増加や，文章音声へ

の対応を検討していく．

謝辞 本研究を遂行するにあたり，拡声器のインパルス応答を提供していただいた東北大学電気通信研究所の坂本修一准教授に感謝する．また，本研究の一部は一般財団法人カワイサウンド技術・音楽振興財団第 32 回サウンド技術振興部門助成，一般財団法人人工知能研究振興財団平成 27 年度研究助成，公益財団法人電気通信普及財団平成 27 年度助成，東北大学電気通信研究所共同研究プロジェクト（H26/A14）及び国立高専機構平成 27 年度研究プロジェクト経費助成の助成を受けて実施した．

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] IEC: IEC60268-16 Sound system equipment — Part 16: Objective rating of speech intelligibility by speech transmission index (2011).
[4] 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).
[5] julius: http://julius.osdn.jp/.
[6] 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—屋外拡声音の「聴き取りにくさ」と Julius 尤度の関係—，日本音響学会 2014 年秋季研究発表会講演論文集，pp. 1087–1088 (2014).
[7] 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 TL，432, Vol. 107, pp. 43–48 (2007).
[8] 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).
[9] 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).
[10] Vapnik, V.: *The Nature of Statistical Learning Theory; Statistics for Engineering and Information Science*, Springer (1995).
[11] ITU-T P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications (2004).