Vol.2010-SLP-83 No.9
2010/10/29

室内音響指標に基づく残響下音声認識性能の計測，評価，保証

西　浦　敬　信[†1]

本稿では，室内音響指標を用いた残響下における頑健な音声認識性能の計測，評価，保証を検討する．室内音響指標の中でも，特に初期反射音と後続残響音の関係を表す音響パラメータに着目し，事前に様々な環境で複数箇所計測したインパルス応答を基に算出した音響パラメータと，実際の音声認識性能に基づき回帰分析を行うことで，残響下における頑健な音声認識性能の計測，評価，保証を試みる．評価実験の結果，従来の残響時間に基づく手法と比較して提案手法は，より頑健に残響下音声認識性能を計測，評価，保証できることを確認した．

Estimation, Evaluation and Guarantee of the Reverberant Speech Recognition Performance based on Room Acoustic Parameters

Takanobu Nishiura[†1]

We study on estimation, evaluation and guarantee of the reverberant speech recognition performance based on room acoustic parameters in this paper. We first designed the suitable reverberation criteria with the relation between room acoustic parameters and speech recognition performance. We then estimated, evaluated and guaranteed the speech recognition performance based on our designed reverberation criteria. As a result of evaluation experiments, we could confirm that the recognition performance could be accurately and robustly estimated, evaluated and guaranteed with proposed criteria.

†1 立命館大学 情報理工学部
College of Information Science and Engineering, Ritsumeikan University

1. はじめに

近年，雑音残響下における音声認識性能の向上に関する研究に触発され，実環境における音声認識性能の計測，評価，保証に関する研究が高い注目を浴びている．事前に各環境での音声認識性能の劣化を計測し，結果を音声認識システムの前処理等に反映させることで，各環境に適した音声認識製品の実現が可能となる．また音声認識性能の評価や保証が実現できれば，ユーザの動きが伴う遠隔発話音声認識に対しても有効的であり，音声認識性能の評価が低い場合は受音機への接近を促し，評価が高い場合は離反を許容することが可能となる．

残響環境下に対する音声認識性能については，1965 年に M. R. Schroeder によって提案された残響時間測定法[1] に基づいて音声認識性能の評価，保証が行われている[2]．また残響時間に加えて入出力間距離に基づいて音声認識性能を評価する手法[2] も提案されている．しかし，残響時間は同一室内で固有の値とされているが，仮定音場と実環境との差異から残響時間だけでなく他の残響特性も変化することから，残響時間のみで音声認識性能を計測，評価，保証することは困難である．また入出力間距離に基づいて音声認識性能を評価する手法についても，入出力間距離の把握が困難な条件においては音声認識性能の計測，評価，保証が困難である．

そこで本稿では，直接音対間接音比と音声認識性能の関係を用いて音声認識性能を計測可能な残響指標 RSR-D_n(Reverberant Speech Recognition criteria with D_n) を提案し，実残響下における音声認識性能の高精度な計測，評価，保証を試みる．

2. 音声認識性能計測のための従来の残響指標

2.1 残響時間 (T_{60})

残響時間 (T_{60})[3] は室内音場を評価する基本的な概念であり響きの長さを表す．室内に放射した音が平衡状態に達した後，音を停止し，その後の残響エネルギー密度が音源停止直前のエネルギー密度に比べて 100 万分の 1（-60 dB）になるまでの時間を表したものである．残響理論では室内で拡散音場を仮定しているため，吸音材料をどの位置に配置してもその効果は変化せず，音源位置によって残響時間が変わらないと定義されている．

残響時間は現在の音声認識の残響指標として積極的に利用されているが，完全拡散音場を仮定していることから，残響時間のみで音声認識性能を計測，評価，保証することに限界があることが問題視されている．

Vol.2010-SLP-83 No.9
2010/10/29

3. 音声認識性能の頑健な計測，評価，保証のための残響指標 RSR-D_n の提案

3.1 室内音響指標

音声認識性能を残響に対して頑健に計測，評価，保証するために本研究では室内音響指標[4] に着目した．ISO3382 Annex A で提案されている室内音響指標は残響時間を補う残響尺度として，音の初期部分の減衰状態を表現可能である．この室内音響指標の中で,「音の了解性」に最も関連性がある「初期反射音と後続残響音のバランス」に着目し，音声認識システムの整合性を検証する．

3.1.1 Definition(D 値)

「初期反射音と後続残響音のバランス」を構成する要素として，C 値 (Clarity)，D 値 (Definition) と Ts(Centre time) の 3 つが存在する．C 値と D 値は可逆変換可能な指標であり，かつ D 値は音声の明瞭性を表現可能な指標として提案されていることから，本研究では D 値に注目する．D 値は系のインパルス応答を基に式 (1) より算出され，直接音と初期反射音のエネルギーに対する直接音と全ての反射音のエネルギー比を示す．

$$D_n = \int_0^n h^2(t)dt / \int_0^\infty h^2(t)dt. \quad (1)$$

ここで $h(t)$ はインパルス応答を，n は初期反射音と後続残響音の境界時間を示す．直接音と初期反射音のエネルギーが大きいほど D 値は向上し，後続残響のエネルギーが大きいほど低下する．

3.2 残響指標 RSR-D_n

前述の D 値と残響下音声認識性能の関係を明らかにした上で，従来の残響指標である残響時間を基に，D 値と残響下音声認識性能間の相関関係を基に回帰分析を行い，残響下音声認識性能の計測，評価，保証のための残響指標 RSR-D_n(Reverberant Speech Recognition criteria with D_n) の策定を試みる．

3.2.1 残響指標 RSR-D_n 策定アルゴリズム

音声認識性能を計測，評価，保証するための残響指標 RSR-D_n の策定アルゴリズムを図 1 の上部に示す．

Step.1 インパルス応答計測

各環境でインパルス応答を数 10〜数 100 箇所にて計測し，さらに残響時間を算出する．残響時間は同一室内では固有の値をもつため，計測したインパルス応答の全てから残響時間を算出する必要は無く，数箇所のインパルス応答から算出した残響時間の平均値を各環境の残響時間とすることが一般的である．

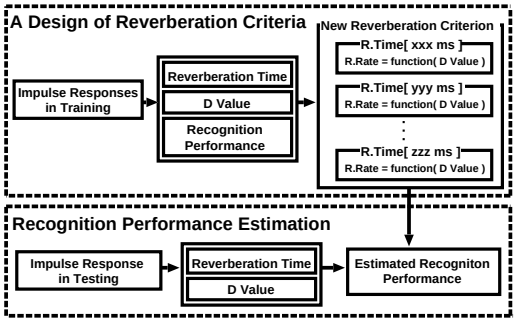


図 1 提案手法の概要
Fig. 1 Overview of the proposed method.

Step.2 D 値の算出

Step.1 で計測した各インパルス応答に対して式 (1) に基づいて D 値を算出する．また初期反射音と後続残響の境界時間を表す n は，音声認識性能と D 値の最大相関値を示すように設定する必要がある．そこで最適な境界時間 n の値を **4.1.1** 節で実験的に検討し，その結果を基に D_n を算出する．

Step.3 音声認識性能の算出

Step.1 で計測した各インパルス応答と学習データとして予め用意した音声ソースを畳み込み，音声認識エンジンを用いて音声認識性能を確認する．

Step.4 回帰分析・回帰曲線の算出

Step.2 と **Step.3** で各インパルス応答から算出した D 値と音声認識性能を基に回帰分析を行う．回帰分析に基づいて算出する回帰曲線は 1 次関数，2 次関数とする．各回帰曲線の定義式と推定パラメータを表 1 に示す．1 次関数，2 次関数で回帰分析を行う際に用いる係数予測方法は，最小二乗法を用いる．

3.3 残響下音声認識性能の計測，評価，保証

策定した残響指標 RSR-D_n に基づく音声認識性能の計測，評価，保証アルゴリズムを図 1 の下部に示す．音声認識性能を計測する系で測定したインパルス応答に基づいて残響時間

Vol.2010-SLP-83 No.9
2010/10/29

表 1 回帰曲線と推定パラメータ
Table 1 Regression curve and parameters to estimate

回帰曲線	$y = ax + b$	$y = ax^2 + b$
推定パラメータ	a, b	

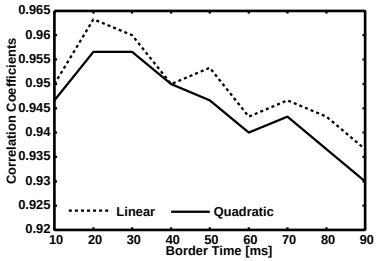


図 2 各回帰曲線の相関係数と境界時間 n の関係
Fig. 2 The relation between correlation coefficient in each regression curve and border time n

と D 値を算出する．ここで同一残響時間の指標が存在しない場合，近接の残響時間の指標を線形補間する．そして同一残響時間における残響指標 RSR-D_n と D 値から音声認識性能の評価，保証を試みる．

4. 性能評価実験

まず各環境において算出した D 値と音声認識性能に基づいて回帰分析を行い，残響指標 RSR-D_n を策定する．策定した RSR-D_n と性能評価，保証を行う系のインパルス応答を基に，残響下音声認識性能の評価，保証を行う．なお音声認識性能は特徴量や言語・音響モデルなどに依存するため，残響尺度策定と音声認識性能の計測，評価，保証における認識条件を統一させる必要がある．

4.1 実 験 条 件

室内音響指標と残響下音声認識性能の関係を分析するために表 2**(A)** に示す 9 つの学習環境にて計 732 箇所のインパルス応答を計測した．表中の RIRs（Room Impulse Responses）は，計測したインパルス応答数を示す．なお表 2 に示す環境は，様々な残響環境を想定して，残響時間が異なる環境でインパルス応答を計測した．

4.1.1 残響指標 RSR-D_n のための最適境界時間の検討

式 (1) における n は，初期反射音と後続残響音の境界時間を示し，D 値を算出する際に

表 2 実験条件
Table 2 Experimental conditions

(A) Environments in training	Soundproof room (T_{60}=100 ms, 72 RIRs) Japanese style room (T_{60}=400 ms, 72 RIRs) Laboratory (T_{60}=450 ms, 72 RIRs) Conference room (T_{60}=600 ms, 120 RIRs) Living room (T_{60}=600 ms, 72 RIRs) Corridor (T_{60}=600 ms, 120 RIRs) Bath room (T_{60}=650 ms, 28 RIRs) Elevator hall (T_{60}=850 ms, 120 RIRs) Standard stairs (T_{60}=850 ms, 56 RIRs)
(B) Environments to calculate a suitable n	Japanese style room (T_{60}=400 ms, 72 RIRs) Conference room (T_{60}=600 ms, 120 RIRs) Standard stairs (T_{60}=850 ms, 56 RIRs)
(C) Environments to design RSR-D_n	Japanese style room (T_{60}=400 ms, 72 RIRs) Conference room (T_{60}=600 ms, 120 RIRs) Standard stairs (T_{60}=850 ms, 56 RIRs)
(D) Environments in testing	Laboratory (T_{60}=450 ms, 72 RIRs) Bath room (T_{60}=650 ms, 28 RIRs) Elevator hall (T_{60}=850 ms, 120 RIRs)
Measured distance	100〜5,000 mm
Speech	ATR phoneme balance 216 words[5)] 7 female and 7 male speakers
Decoder	Julius[6)]
HMM	IPA monophone model (Gender-dependent)
Feature vectors	12 MFCC + 12 ΔMFCC + 1 ΔPower
Frame length	25 ms (Hamming window)
Frame interval	10 ms

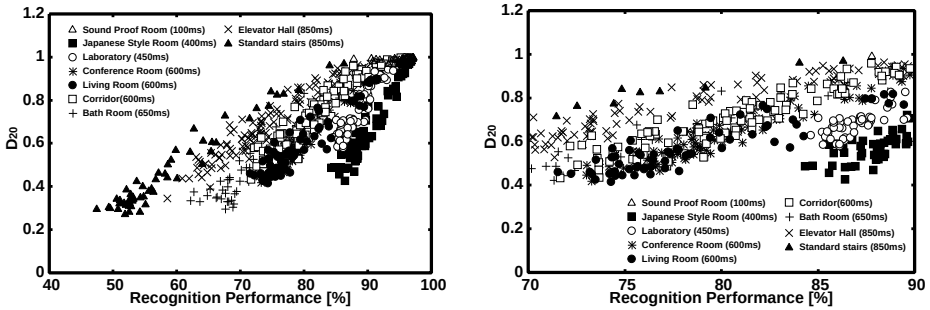
適切な値を設定する必要がある．そこで音声認識性能と D 値の間で高い相関を示す境界時間 n を検討するために，表 2**(B)** に示す残響時間が異なる 3 環境で評価実験を行った．実験方法は **3.2.1** 節の分析アルゴリズムと同様である．また境界時間 n は 10〜90 ms の 10 ms 間隔に設定し，D 値を算出する．そして境界時間 n ごとに算出した D 値と音声認識性能との関係を回帰分析し，3 環境の相関係数の平均値を各回帰曲線ごとに算出した．

初期反射音と後続残響音の境界時間 n と各回帰曲線の相関係数の関係を図 2 に示す．1，2 次関数共に境界時間 n が 20 ms で最も高い相関係数を示し，以降は減少傾向にあることを確認した．従って，今回の表 2**(B)** に示す 3 環境における評価実験結果では残響指標 RSR-D_n のための境界時間 n は 20 ms が最適であることを確認した．本稿では，最も高い相関係数

Vol.2010-SLP-83 No.9
2010/10/29

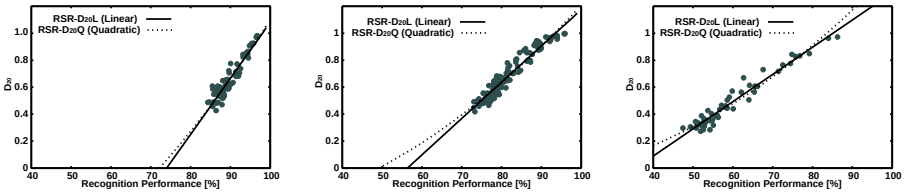
表 3 相関係数
Table 3 Correlation coefficients

	RSR-D_{20}L (Linear)	RSR-D_{20}Q (Quadratic)
T_{60}=400 ms	0.937	**0.939**
T_{60}=600 ms	**0.966**	0.963
T_{60}=850 ms	**0.977**	0.972



(a) 全体図 (Overall) (b) 拡大図 (Closeup)

図 3 D_{20} と音声認識性能の関係
Fig. 3 The relation between D_{20} and speech recognition performance



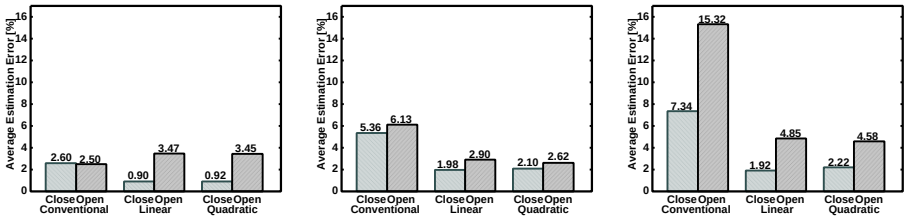
(a) 和室 (T_{60}=400 ms) (b) 会議室 (T_{60}=600 ms) (c) 階段 (T_{60}=850 ms)

図 4 RSR-D_{20} と音声認識性能の関係
Fig. 4 The relation between RSR-D_{20} and speech recognition performance

であった n=20 ms を採用して D 値 (D_{20}) および RSR-D_{20} を算出する．

4.2 残響指標 RSR-D_{20} の策定

表 2**(A)** に示す 9 つの学習環境における D_{20} と音声認識性能の関係を図 3(a) に，拡大図を図 3(b) に示す．そしてこの 9 環境の中から表 2**(C)** に示す残響時間が異なる 3 環境につ



(a) 研究室 (b) 浴室 (c) エレベータホール

図 5 平均誤差
Fig. 5 Average error

表 4 標準偏差
Table 4 Standard deviation

	Conventional Method		RSR-D_{20}L (Linear)		RSR-D_{20}Q (Quadratic)	
	Close	Open	Close	Open	Close	Open
T_{60}=450 ms	3.10	3.26	1.10	3.62	1.13	3.60
T_{60}=650 ms	6.92	7.18	2.46	3.49	2.59	3.14
T_{60}=850 ms	8.80	17.64	2.41	5.35	2.81	5.23

いて回帰分析を行った結果を図 4 に，3 環境に対する各回帰曲線の相関係数を表 3 に示す．また，音声認識性能と D_{20} の関係を 1 次関数で回帰分析した結果を RSR-D_{20}L(Linear)，2 次関数で回帰分析した結果を RSR-D_{20}Q(Quadratic) と表している．

結果より，会議室 (T_{60}=600 ms) と階段 (T_{60}=850 ms) における両関数の相関係数が 0.96 を上回り，非常に高精度で近似可能であった．また和室 (T_{60}=400 ms) における両関数の相関係数も 0.93 を上回っており，全体的に安定した回帰分析が可能であった．この結果から D_{20} と音声認識性能の関係を 1 次，2 次関数で回帰分析した RSR-D_{20}L，RSR-D_{20}Q ともに有力な残響指標であることを確認した．

4.3 残響下音声認識性能の計測，評価，保証の検討

策定した音声認識指標の有効性を検証するために音声認識性能の計測，評価，保証実験を行う．実験は表 2**(D)** に示す残響時間が異なる 3 つのテスト環境の下で，音声認識性能の評価，保証を行う．また各環境の精度を比較するために，環境クローズテストおよび環境オープンテストを行う．環境クローズテストでは，環境が既知という条件で，学習時と同一環境の回帰曲線から音声認識性能を評価，保証する．一方，環境オープンテストでは，環境が未

知という条件で，学習時と残響時間が同一でかつ異なる環境の回帰曲線から音声認識性能を評価，保証する．精度評価には回帰曲線から算出した音声認識性能の評価値とテストデータの真値との差を示す平均誤差を用いた．

なお提案手法との比較のために残響時間のみを用いた従来の音声認識性能評価も併せて行った．従来法は表 2**(D)** に示す 3 つのテスト環境の残響時間を基に，各環境に対する音声認識性能の平均値に基づいて音声認識性能を評価，保証した．

図 5 に各環境の環境オープンテストおよび環境クローズテスト結果を，表 4 に各テストの標準偏差を示す．高残響環境では RSR-D_n を用いた場合，平均誤差と標準偏差が従来法と比較して全体的に改善し，高精度に音声認識性能を評価，保証できた．また残響時間のみを用いても十分に評価可能な低残響環境についても，同程度の精度を確認できた．そして環境オープンテストにおいて RSR-D_nQ の平均誤差と標準偏差ともに RSR-D_nL の結果よりも改善でき，高精度な音声認識性能の評価ができた．したがって音声認識性能と D_{20} の関係を 2 次関数で回帰分析した残響指標 RSR-D_{20}Q が残響下音声認識性能の評価，保証指標として最適であると考えられる．

4.4 考　　察

4.4.1 RSR-D_{20} の環境変化に対する頑健性

策定した RSR-D_{20} の環境変化に対する頑健性について考察する．表 2**(A)** に示す 9 つの学習環境における D_{20} と音声認識性能の関係を示した図 3(b) の残響時間が 600 ms の環境 (会議室，リビング，廊下) より，同一残響時間または近傍の残響時間をもつ環境における計測値の分布が類似していることがわかる．残響時間が 400〜450 ms の和室と研究室，850 ms のエレベータホールと階段においても同様の傾向が確認できる．このことから近傍の残響時間であれば異なる環境の RSR-D_{20} を用いても音声認識性能を頑健に評価，保証できると考えられる．

5. 高精度な音声認識性能の計測，評価，保証を目指して

本稿では提案した残響指標 RSR-D_n を入出力間距離や発話位置と壁からの距離情報を未知として策定した．もし入出力間距離と音声認識性能，または壁からの距離と音声認識性能に相関があれば, これらの関係と RSR-D_n を組み合わせることで音声認識性能の評価，保証精度の向上が期待できる．そこで RSR-D_n の拡張を目指して認識器と発話者の位置関係と音声認識性能の関係を詳細に分析する．

表 5　各環境・発話位置におけるインパルス応答の D_{20}
Table 5　D_{20} of measured impulse response in each environment

	壁に対する発話位置	
残響時間	近接	遠隔
T_{60} = 450 ms	0.48	0.63
T_{60} = 850 ms	0.72	0.67

5.1 発話位置に対する伝達特性の変化

はじめに発話位置によって入出力間の伝達特性がどのように変化するのかを明らかにする．具体的には発話者と認識器を壁に接近させた場合と離反させた場合のインパルス応答を計測する．ここで実際に表 2**(A)** に示す環境の中から和室（T_{60}=450 ms）とエレベータホール（T_{60}=850 ms）で計測したインパルス応答 (入出力間距離：50 cm) を用いて入出力間の伝達特性の変化を調査する．また計測において壁と発話位置からの距離は和室の場合は 25 cm，132 cm，エレベータホールの場合は 25 cm，300 cm とした．表 5 に計測した各インパルス応答の D_{20} を示す．結果より，残響時間が短い環境では壁に接近させた場合は壁から離れた場合よりも D_{20} が減少し，後続残響量が多いことが確認できた．それに対して残響時間が長い環境においては壁から離れて発話すると後続残響量が増加した．このように壁に対する発話位置が同じでも初期反射音と後続残響のエネルギーの割合は部屋の残響時間によって大きく異なることが確認できた．

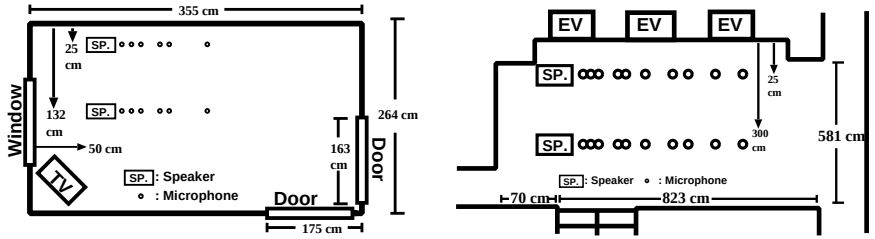
5.2 発話位置に対する音声認識性能の変化

ここでは発話位置および入出力間距離と音声認識性能の関係について調査する．図 6 に和室とエレベータホールでの収録図を示す．また図 7 に各残響環境で計測したインパルス応答を用いて算出した音声認識性能の結果を示す．結果から，どの環境においても入出力間距離が長いと音声認識性能が低下することが確認できた．そして残響時間が短い環境においては壁から離れて発話すると音声認識性能が向上した．一方，残響時間が長い環境においては，壁に接近して発話すると音声認識性能が向上することを確認した．

5.3 考　　察

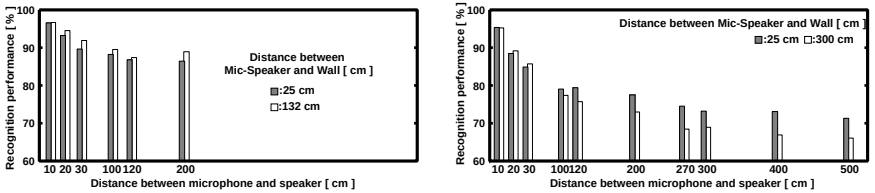
残響時間が短い和室（T_{60}=450 ms）では壁から離れて発話すると音声認識性能が向上し，残響時間が長いエレベータホール（T_{60}=850 ms）では，壁に接近して発話すると音声認識性能が向上した．この結果は 5.1 節で分析した伝達特性の入出力間の伝達特性の初期反射音と後続残響エネルギーの比率を表現する D_{20} の結果と類似していることがわかる．つまり D_{20} が低下（後続残響のエネルギーが増加）することにより音声認識性能が低下することが

Vol.2010-SLP-83 No.9
2010/10/29



和室 (T_{60}=450 ms)　　エレベータホール (T_{60}=850 ms)

図 6　実験配置図
Fig. 6　Placement of microphone and speaker



和室 (T_{60}=450 ms)　　エレベータホール (T_{60}=850 ms)

図 7　音声認識性能結果
Fig. 7　Reverberant speech recognition performance

確認できた．また本実験結果から各環境の残響時間と入出力間距離や壁からの距離などの発話位置に基づいて音声認識性能を評価，保証できる可能性がある．そしてこれらの指標と本稿で提案した残響指標 RSR-D_n を組み合わせることにより, 音声認識性能評価，保証精度の向上が期待できると考えられる．

6. おわりに

本稿では，残響指標 RSR-D_n を提案し，音声認識性能の高精度な計測，評価，保証を試みた．その結果，提案した残響指標は，高精度な音声認識性能の計測，評価，保証が行えることを確認した．今後は発話者と認識器との距離情報と残響指標を組み合わせた高精度な音声認識性能の計測，評価，保証手法の検討や MTF(Modulation Transfer Function)[7] などの周波数指標も含めた音声認識に適した残響指標の確立を目指す．また PESQ を利用した認識性能計測，評価，保証法[8] を残響環境下へ拡張する手法などを検討し，雑音と残響が混在した環境における音声認識性能の計測，評価，保証に取り組む計画である．

謝辞　本研究の一部は科研費による研究助成を受けた．また，社団法人 情報処理学会音声言語情報処理研究会 雑音下音声認識評価ワーキンググループの諸氏，および立命館大学理工学研究科の福森隆寛氏に感謝する．

参 考 文 献

1) M. R. Schroeder, “New Method of Measuring Reverberation Time,” JASA, Vol. 37, pp. 409-412, 1965.
2) Rico Petrick, Xugang Lu, Masashi Unoki, Masato Akagi, and Ruediger Hoffmann, “Robust Front End Processing for Speech Recognition in Reverberant Environments: Utilization of Speech Characteristics,” Proc. Interspeech2008, pp. 658-661, Brisbane, Australia, Sept. 2008.
3) 日本音響学会, “新版音響用語辞典,” コロナ社, 2003.
4) ISO3382:Acoustics-Measurement of the reverberation time of rooms with reference to other accoustical parameters. Internatinal Organization for Standardization, 1997.
5) K. Takeda, Y. Sagisaka, and S. Katagiri, “Acoustic-Phonetic Labels in a Japanese Speech Database,” Proc. European Conference on Speech Technology, vol. 2, pp. 13-16, Oct. 1987.
6) A. Lee, T. Kawahara, and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine,” In Proc. European Conf. on Speech Communication and Technology, pp. 1691-1694, 2001.
7) T. Houtgast, H. J. M. Steeneken, and R. Plomp, “Predicting speech intelligibility in room acoustics,” Acustica, vol. 46, pp. 60-72, 1980.
8) T. Yamada, M. Kumakura, N. Kitawaki, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” IEEE Trans. on ASLP, Vol. 14, No. 6, pp. 2006-2013, Nov. 2006.