論　文

室内音響指標を用いた残響指標 RSR-D_n に基づく残響下音声認識性能の予測

福森　隆寛[†]　　森勢　将雅[††]　　西浦　敬信[††]　　山下　洋一[††]

Performance Estimation of Reverberant Speech Recognition Based on Reverberant Criteria RSR-D_n with Acoustic Parameters

Takahiro FUKUMORI[†], Masanori MORISE[††], Takanobu NISHIURA[††], and Yoichi YAMASHITA[††]

あらまし　近年，雑音及び残響下における音声認識手法に関する研究が盛んに行われている．それに伴い雑音環境下で音声認識性能を頑健に予測可能な指標も多数提案されている．一方，残響環境下における音声認識性能の有力な予測指標は提案されておらず，残響下音声認識性能の頑健な予測指標の策定は急務である．これまでに残響下音声認識性能の優劣を判別する残響指標として同一室内で固有の値となる残響時間が提案されているが，仮定する拡散音場と実環境との差異から他の残響特性が変化することにより同一環境でも計測箇所によって音声認識性能が変動する．そのため残響時間は音声認識の難しさを表す指標として不十分であることが問題視されている．そこで本論文では，ISO3382 Annex A で提案されている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予測法を提案する．提案法では初期反射音と後続残響音の関係を表す室内音響指標の中でも特に Definition（D 値）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もとに算出した D 値と音声認識性能の関係を一次直線や二次曲線で近似することで残響指標 RSR-D_n を策定する．策定した残響指標 RSR-D_n と性能予測を行う残響環境の発話位置におけるインパルス応答をもとに残響下音声認識性能の予測を試みる．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残響指標 RSR-D_n は，より頑健に残響下音声認識性能を予測できることを確認した．

キーワード　残響下音声認識，性能予測，室内音響指標，残響時間，残響指標

1. まえがき

近年，情報機器の急速な発展に伴い入力インタフェースが複雑化している．現在の基本となる入力インタフェースはキーボードとマウスである．しかし手足が不自由な身体障害者や情報機器に不慣れな高齢者には使用が困難であるのが現状である．万人が使い勝手の良い入力インタフェースとしてマイクロホンを意識せずに音声入力ができるハンズフリー音声インタフェースの実現に高い注目が集まっている．

しかしながら，ハンズフリー音声インタフェースはマイクロホンを装着しない上，入出力間距離も 10〜500 cm と様々な状況を想定しているため実環境下で使用者がマイクロホンから離れて発話した際に，雑音や室内残響等の混入により音声認識性能が著しく低下するという問題がある．この実環境下で音声認識性能を向上させる耐雑音対策としては，複数のマイクロホンを使用して指向性を形成するマイクロホンアレーの利用 [1]，雑音スペクトルを受音信号から減算するスペクトルサブトラクション [2] や雑音信号を含めて音響モデルの学習 [3]・適応 [4] を行う手法など様々な試みが行われている．また耐残響対策としてもケプストラム係数の平均正規化法である CMN [5]，空間伝達特性の逆フィルタを用いて残響を除去する手法 [6] や空間伝達特性を含めて音響モデルを学習 [7]・適応 [8] する手法などが提案されている．

これに対し，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と比例して，実環境における音声

[†] 立命館大学大学院理工学研究科，草津市
Graduate School of Science and Engineering, Ritsumeikan University, 1–1–1 Nojihigashi, Kusatsu-shi, 525–8577 Japan
[††] 立命館大学情報理工学部，草津市
College of Information Science and Engineering, Ritsumeikan University, 1–1–1 Nojihigashi, Kusatsu-shi, 525–8577 Japan

認識性能の予測に関する研究に高い注目が集まっている．例えば音声認識システム搭載のハンズフリーホンのような携帯機器は，場所にとらわれることなく様々な環境で使用されている．そこで事前に各環境が与える音声認識性能の劣化を予測し，その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が可能となる．また音声認識性能の予測はユーザの動きが伴う遠隔発話音声認識に対しても有効であり，音声認識性能予測値が低い場合は受音器への接近を促し，予測値が高い場合は離反を許容することが可能となる．雑音環境下ではこれまで信号対雑音比を示す SNR（Signal to Noise Ratio）によって音声認識性能を予測する研究が一般的であったが，2006 年山田らによって原信号と劣化信号に基づいて品質を予測する PESQ（Perceputual Evaluation of Speech Quality）を利用した研究 [9] が提案され雑音下における音声認識性能の予測精度は飛躍的に向上した．一方，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 [10] に基づいて音声認識性能の評価が行われている [11]．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 [11] も提案されている．ところが残響時間については同一室内で固有の値となるが，仮定する拡散音場と実環境との差異から他の残響特性が変化し同一環境でも計測箇所によって音声認識性能が変動することから，音声認識の難しさを表現する指標としては不十分であると考えられる．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予測が困難である．

そこで本論文では，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予測可能な残響指標 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予測を試みる．具体的には，最初に初期・後続反射音が音声認識性能に与える影響を分析し，残響下音声認識に利用可能な初期反射音と除去すべき後続残響音との分離時間を明らかにする．その上で，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もとに算出した室内音響指標と音声認識性能の関係に基づき残響指標 RSR-D_n の策定を試みる．更に策定した残響指標 RSR-D_n と音声認識性能予測を行う残響環境の発話位置におけるインパルス応答をもとに残響下音声認識性能の予測精度を検証する．

2. 音声認識性能予測のための従来の残響指標

残響時間（T_{60}）[12] 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．また残響時間は M.R. Schroeder によって二乗積分法に基づく残響測定法 [10] が提案され，系の残響曲線はインパルス応答 $h(x)$ を用いて式 (1) に基づき容易に算出できるようになった．

$$<Sd^2(t)> = N \int_t^{\infty} h^2(x)dx, \qquad (1)$$

ここで N は単位周波数当りのパワー，$<Sd^2(t)>$ は残響曲線を表す．これまで残響曲線は入力信号をランダム雑音として長時間かつ複数回観測した信号から集合平均を利用して算出したのに対して M.R. Schroeder はインパルス応答 $h(t)$ のみから集合平均を利用せずに残響曲線を算出する手法を提案した．残響時間は算出した残響曲線に基づき 60 dB 減衰するまでの時間となるが，計測したインパルス応答の後続部分は暗騒音に埋没し，実際に残響エネルギー密度が 60 dB 減衰する時間を算出することは困難である．この問題に対して，通常は初期部分を回帰した直線が 60 dB 減衰するまでの時間を残響時間とすることが一般的である．

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，仮定する拡散音場と実際の環境との差異から他の残響特性が変化し，同一環境でも計測箇所によって音声認識性能が変動する．そのため固有の値をとる残響時間のみで音声認識の難しさを表現することに限界があると考えられる．

ここで残響時間と音声認識性能の関係を調査するために表 1 に示す残響時間が異なる 3 環境にて評価実験を行った．まず各環境で数十〜数百系のインパルス応答を計測し，各インパルス応答とクリーン音声を畳み込んで音声認識エンジンを用いて音声認識性能を算出した．なお表中の RIRs（Room Impulse Responses）は，計測したインパルス応答数を示す．ここで実験結果を図 1 に示す．図中の線は各残響環境の音声認識性

表 1 残響時間と音声認識性能の関係調査のための実験条件
Table 1 Experimental conditions for investigating the relation between reverberation time and recognition performance.

Environments	Laboratory (T_{60}=450 ms, 72 RIRs) Conference room (T_{60}=600 ms, 120 RIRs) Elevator hall (T_{60}=850 ms, 120 RIRs)
Distance between mic. and sp.	100～5,000 mm
Speech	ATR phoneme balance 216 words [16] 7 female and 7 male speakers
Decoder	Julius [17]
HMM	IPA monophone model (Gender-dependent)
Feature vectors	12 orders MFCC+ 12 orders ΔMFCC+ 1 order ΔPower
Frame length	25 ms (Hamming window)
Frame interval	10 ms

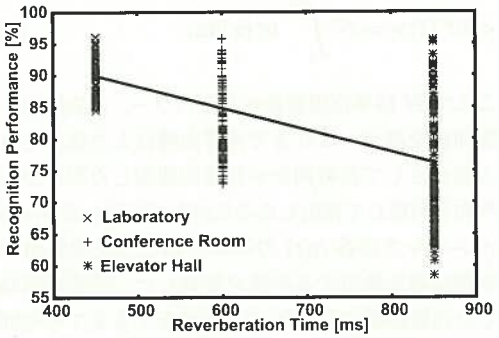


図 1 残響時間と音声認識性能の関係
Fig. 1 The relation between reverberation time and recognition performance.

能の平均を表す．図 1 の結果から長い残響時間ほど音声認識性能の平均が低下し，分散が上昇していることが確認できた．このことから残響時間のみを用いて音声認識性能を予測することは低残響環境では比較的容易であるが，高残響環境では困難であると予測できる．

3. 頑健な音声認識性能予測のための残響指標 RSR-D_n の提案

3.1 音声認識における初期反射音と後続反射音の影響

前章において同一環境でも計測箇所によって音声認識性能が変動することから，同一室内で固有の値となる残響時間では音声認識性能の予測が困難であることを述べた．そこで本節では音声認識に影響を与える残響特性を明らかにするために，音声認識性能の著しい

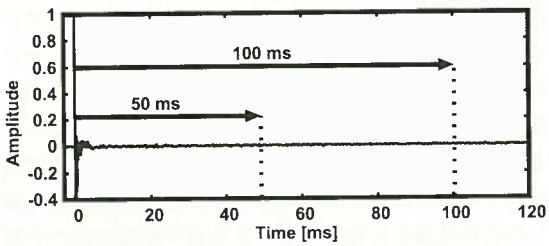


図 2 直接音からのインパルス応答長
Fig. 2 Length of impulse response from a direct sound.

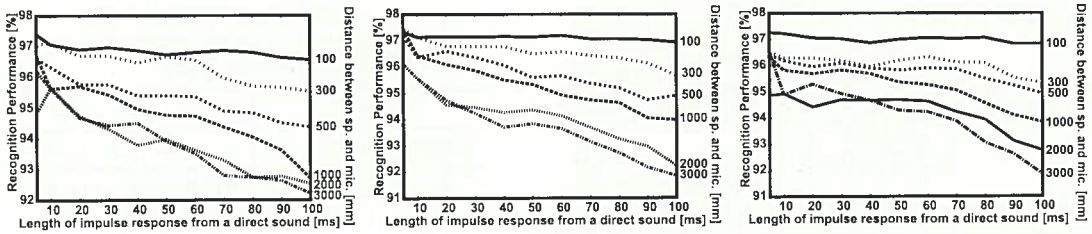
表 2 反射音と音声認識性能の関係調査のための実験条件
Table 2 Experimental conditions for investigating the relation between reflection and speech recognition performance.

Environments	Laboratory (T_{60}=450 ms, 72 RIRs) Corridor (T_{60}=600 ms, 120 RIRs) Elevator hall (T_{60}=850 ms, 120 RIRs)
Distance between mic. and sp.	100～3,000 mm
Length of impulse response	5 ms, 10～100 ms at 10 ms intervals

低下が顕著に確認できる反射継続時間と音声認識性能の関係について調査する．

音声認識性能と反射音の関係を調査する方法として，TSP (Time Stretched Pulse) 信号 [14] を用いて系のインパルス応答を計測し，図 2 及び表 2 の実験条件に示す範囲に基づいて初期反射時間分だけインパルス応答を切り出した上で音声ドライソースと畳み込むことで，初期反射音の継続時間と音声認識性能との関係を調査する．なおハース効果 [12] に基づき本実験では直接音から最長 100 ms までの反射音を調査する．

図 3 に初期反射音の継続時間と音声認識性能の関係を示す．音声認識性能は，マイクロホンとスピーカ間の距離が 500～1,000 mm となる系を境界として低下する傾向が確認できた．更に，同一残響時間でも音声認識性能に差異があることや，20～30 ms 程度より後続の反射音，特に 60 ms 程度より後続の反射音は音声認識性能を大きく低下させる要因であることが確認できた．また図 3 (c) におけるマイクロホンとスピーカ間の距離が 300 mm の結果では，直接音からのインパルス応答長が 10～80 ms において音声認識性能はほぼ同程度であるため，本実験において最長 80 ms までの反射音を含むインパルス応答を用いても音声認識性能は低下せず，直接音から 60 ms 以降の後続の反射音が音声認識性能の劣化原因とならない環境が存在するこ



(a) 研究室　(b) 廊下　(c) エレベータホール

図 3　音声認識性能と初期反射音の関係（マイクと壁の距離：250 mm）
Fig. 3　The relation between speech recognition performance and early reflections. (Distance between a mic. and a wall: 250 mm)

とも確認できた．この結果から音声認識性能の予測指標として，従来の残響時間では高精度な音声認識性能の予測が困難であることを再確認した．

そこで本論文では，音声認識が著しく低下するまでの初期反射音の継続時間に基づき初期部分の反射音エネルギーと後続部分の反射音エネルギーの割合に着目する．この着目点に対して室内音響指標（ISO3382）[15] の導入を念頭に残響下音声認識のための残響指標の策定を試みる．

3.2　A 値（反射音の総合振幅）

計測したインパルス応答の反射エネルギーを表現する尺度としてよく利用されるのが直接音に対する反射音の総合振幅を表す A 値 [13] である．A 値は式 (2) のように定義される．

$$A = \sqrt{\int_{\epsilon}^{\infty} h^2(t)dt \Big/ \int_{0}^{\epsilon} h^2(t)dt}, \qquad (2)$$

ここで $h(t)$ はインパルス応答を表す．また ϵ は直接音の持続時間を示し，インパルス応答の場合 3〜5 ms となる．A 値は受音信号における反射音エネルギーに対する直接音エネルギー比であり，同一室内でも各受音点により大きく異なる．音源に近接して受聴すると反射音に比べて直接音のエネルギーが高くなるため，A 値が低下するのに対して，遠方から受聴すると反射音のエネルギーが大きくなり，A 値は上昇する．しかし A 値では系の初期反射音と後続残響のどちらのエネルギーが大きいのかを判断できないため音声認識性能を著しく低下させる後続残響エネルギーを表現することが困難である．したがって反射エネルギーの中で音声認識性能に影響する成分を明確に示すことができず，A 値に基づいて音声認識性能を予測することは困難であると考えられる．

3.3　室内音響指標

ISO3382 Annex A で提案されている室内音響指標 [15] 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するために 1997 年に提案され，建築音響学の分野ではよく用いられている指標の一つである．この室内音響指標は以下の四つから構成される．

（1）　音圧レベル
（2）　残響時間
（3）　初期反射音と後続残響音のバランス
（4）　両耳パラメータ

この中で音の了解性に最も関連性がある「(3) 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．

3.3.1　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）の三つが存在する．C 値と D 値は可逆変換可能な指標であり，かつ D 値は音声の明りょう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もとに式 (3) 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すべての反射音のエネルギー比を示す．

$$D_n = \int_{0}^{n} h^2(t)dt \Big/ \int_{0}^{\infty} h^2(t)dt, \qquad (3)$$

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 値は向上を示し，後続残響のエネルギーが大きいほど低下する．D 値は計測したインパルス応答から音声認識性能に影響を与える初期反射音と後続残響音の割合を表現できることから，音声認識性能に与える劣化の度合を表現するパラメータと

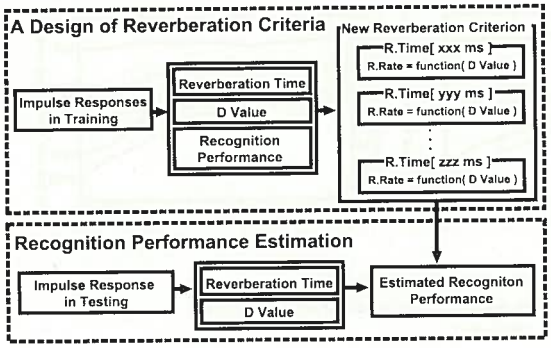


図 4 提案手法の概要
Fig. 4 Overview of the proposed method.

なる可能性がある．

3.4 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，それぞれの相関関係について曲線近似し，残響下音声認識性能予測のための残響指標 RSR-D_n（Reverberant Speech Recognition criteria with D_n）の策定を試みる．

3.4.1 残響指標 RSR-D_n 策定アルゴリズム

音声認識性能を予測するための残響指標 RSR-D_n の策定アルゴリズムを図 4 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十～数百箇所にて計測する．その際，式 (1) に基づいて算出した残響曲線から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すべてから残響時間を算出する必要はなく，数箇所のインパルス応答から算出した残響時間の平均を各環境の残響時間とすることが一般的である．

(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 (3) 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 を **4.1.1** で実験的に検討し，その結果をもとに 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，学習データとしてあらかじ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算出する．

(Step.4) 近似曲線の算出

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もとに近似曲線を算出する．算出する近似曲線は一次直線，二次曲線とする．各近似曲線の定義式を表 3 に示す．なお本論文では一つの D 値から音声認識性能が一意に定まるように二次曲線の一次項を省略している．一次直線，二次曲線で分析を行う際に用いる係数予測方法は，最小二乗法 [18] を用いる．最小二乗法は，予測値と測定値の残差の二乗和が最小となるようにモデルパラメータを決定する方法である．

表 3 近似曲線と音声認識性能予測値
Table 3 Regression curve and estimation value of recognition performance.

近似曲線	$y = ax + b$	$y = ax^2 + b$
音声認識性能予測値 ($\hat{x}$)	$\hat{x} = \frac{y-b}{a}$	$\hat{x} = \sqrt{\frac{y-b}{a}}$

※ a, b：近似係数，x：音声認識性能，y：D 値

3.5 残響下音声認識性能の予測

策定した残響指標 RSR-D_n に基づく音声認識性能の予測アルゴリズムを図 4 の下部に示す．音声認識性能を予測する系で計測したインパルス応答に基づいて残響時間と 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予測を試みる．

4. 性能評価実験

まず各環境において算出した D 値と音声認識性能の関係について曲線近似して残響指標 RSR-D_n を策定する．そして策定した RSR-D_n と性能予測を行う系のインパルス応答をもとに，残響下音声認識性能の予測を行う．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予測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 4 (A) に示す九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．なお表 4 に示す環境は，様々な残響環境を想定するために，残響時間が異なる環境でインパルス応答を計測した．また各残響環境の中でも各系の D 値の分散が大きくなるように 10～500 cm の入出力間距離及び正背左右の放射面の条件で計測を行った．そして計測したインパルス応答をもとに，室内音響指標と音声認識性能の関係について曲線近似する．

表 4 実験条件
Table 4 Experimental conditions.

(A) Environments in training	Soundproof room (T_{60}=100 ms, 72 RIRs) Jpn. 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	Jpn. 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	Jpn. 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 open test	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

4.1.1 残響指標 RSR-D_n のための最適境界時間の検討

式 (3) 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 値の算出において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 4 (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 **3.4.1** の分析アルゴリズムと同様である．また D 値は境界時間 n を 10～90 ms の 10 ms 間隔に設定して算出する．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曲線近似し，3 環境の相関係数の平均を各近似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 と各近似曲線の相関係数の関係を図 5 に示す．一，二次曲線ともに境界時間 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．したがって，今回の表 4 (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 が最適であることが分かった．

本論文では，最も高い相関係数を確認した n=20 ms を採用して D 値（D_{20}）及び RSR-D_{20} を算出した．

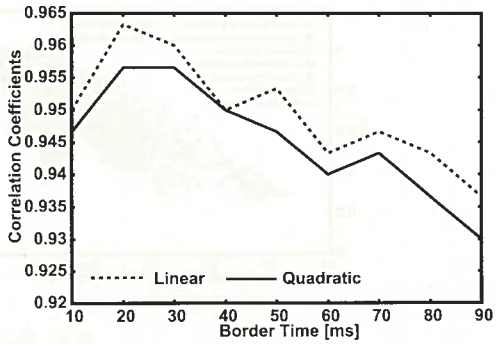


図 5 各近似曲線の相関係数と境界時間 n の関係
Fig. 5 The relation between correlation coefficient in each regression curve and border time n.

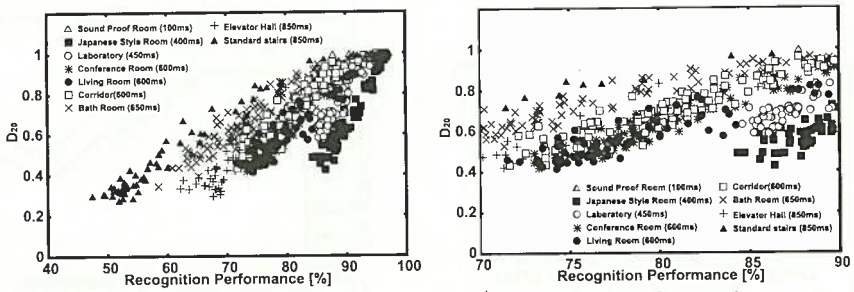
4.2 残響指標 RSR-D_{20} の策定

表 4 (A) に示す九つの学習環境における D_{20} と音声認識性能の関係を図 6 (a) に，拡大図を図 6 (b) に示す．そしてこの 9 環境の中から表 4 (C) に示す残響時間が異なる 3 環境について曲線近似を行った結果を図 7 に，3 環境に対する各近似曲線の相関係数を表 5 に示す．また音声認識性能と D_{20} の関係を一次曲線で近似した結果を RSR-D_{20}L (Linear)，二次曲線で近似した結果を 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）における両曲線の相関係数が 0.96 を上回り，高精度に近似可能であった．また和室（T_{60}=400 ms）における両曲線の相関係数も 0.93 を上回っており，全体的に高精度な曲線近似が可能であった．この結果から D_{20} と音声認識性能の関係を一次，二次曲線で近似した 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予測

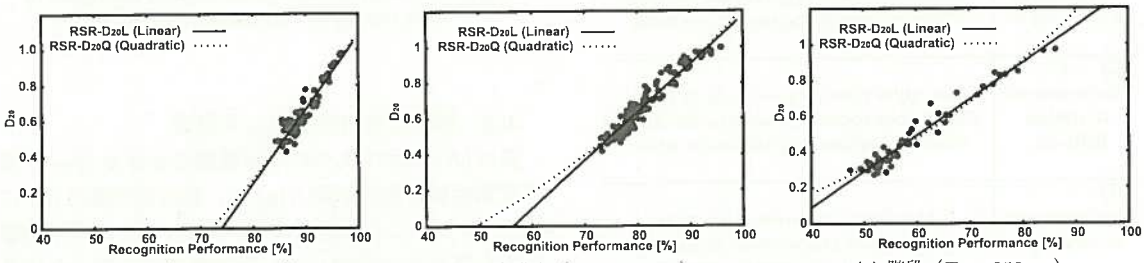
策定した音声認識指標の有効性を検証するために音声認識性能予測実験を行う．各環境の予測精度を比較するために，環境クローズテスト及び環境オープンテストを行う．環境クローズテストでは，環境が既知という条件で，学習時と同一環境の RSR-D_{20} から音声認識性能を予測する．本論文では表 4 (C) に示す 3 環境において策定した RSR-D_{20} を用いて同一環境の音声認識性能の予測を試みる．一方，環境オープンテストでは，環境が未知という条件で，学習時と残響時間は近いが環境が異なる RSR-D_{20} から音声認識性能を予測する．本論文では表 4 (C) に示す 3 環境のイン



(a) 全体図（Overall）　(b) 拡大図（Closeup）

図 6　D_{20} と音声認識性能の関係

Fig. 6　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 7　RSR-D_{20} と音声認識性能の関係

Fig. 7　The relation between RSR-D_{20} and speech recognition performance.

表 5　相関係数

Table 5　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

表 6　標準偏差

Table 6　Standard deviation.

	Conventional Method		RSR-D_{20}L (Linear)		RSR-D_{20}Q (Quadratic)	
	Close	Open	Close	Open	Close	Open
T_{60}=400 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

パルス応答に基づいて策定した RSR-D_{20} を用いて，表 4 (D) に示す 3 環境の音声認識性能の予測を試みる．予測精度評価には RSR-D_{20} から算出した音声認識性能の予測値とテストデータの真値との差を示す平均予測誤差を用いた．

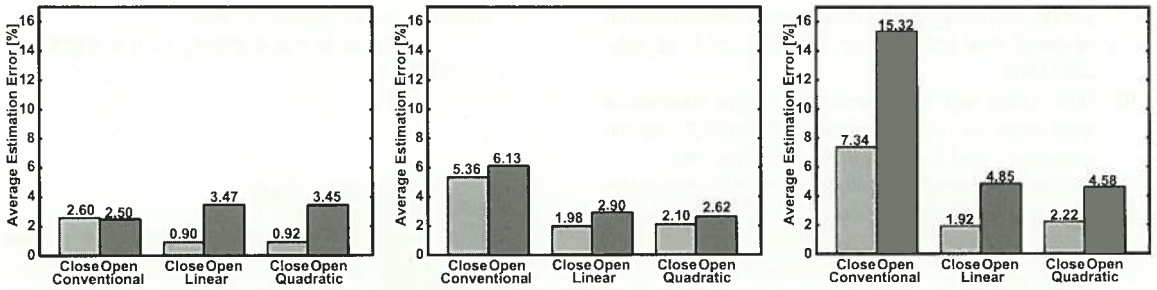
なお提案手法との比較のために残響時間のみを用いた従来の音声認識性能予測も併せて行った．従来法は表 4 (C) に示す三つのテスト環境の残響時間をもとに，各環境に対する音声認識性能の平均に基づいて音声認識性能を予測した．

図 8 に各環境の環境クローズテスト及び環境オープンテスト結果を，表 6 に各テストの標準偏差を示す．高残響環境では RSR-D_{20} を用いた場合，平均性能予測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予測できた．また残響時間のみを用いても十分に予測可能な低残響環境についても，同程度の予測精度を確認できた．そして環境オープンテストにおいて RSR-D_{20}Q の平均性能予測誤差と標準偏差ともに RSR-D_{20}L の結果よりも改善でき，高精度な音声認識性能の予測ができた．したがって音声認識性能と D_{20} の関係を二次曲線で近似した残響指標 RSR-D_{20}Q が残響下音声認識性能の予測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 4 (A) に示す九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 6 (b) の残響時間が 600 ms の環境（会議室，リビング，廊下）よ



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

図 8　平均予測誤差
Fig. 8　Average estimation error.

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分かる．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予測できると考えられる．

4.4.2　RSR-D_n に基づく室内音響指標の評価

残響指標 RSR-D_n の策定によって室内音響指標の D 値から音声認識性能が予測できる上に音声認識性能から D 値を予測することも可能となった．そこで目標の音声認識性能を達成するため D 値を算出することで，その D 値を満たす発話位置等も推定することが可能となり，自律移動型音声対話ロボット等の発展にも大きく貢献できると考えられる．

4.4.3　三次曲線近似による RSR-D_{20} の検討

本論文では音声認識性能と D 値の関係を一次，二次曲線に基づき近似することで RSR-D_n を策定したが，更に三次曲線（$y = ax^3 + b$，x：音声認識性能，y：D 値，a, b：係数）を利用した近似も検討した．表 4 (C) に示す残響時間が異なる環境で RSR-D_n を策定した結果，各環境の相関係数が和室では 0.941，会議室では 0.959，階段では 0.960 となり，RSR-D_{20}L と RSR-D_{20}Q とほぼ同等の性能を達成した．これにより残響指標策定において高次数の曲線で近似する必要はなく RSR-D_{20}L や RSR-D_{20}Q を用いることで十分な性能が期待できると考えられる．

5. むすび

実環境下における音声認識ではマイクロホンから離れた地点で発話すると壁や床からの反射音の影響で音声認識性能が低下する．しかし，残響環境下に対する頑健な音声認識のための残響指標は存在せず，残響下の音声認識性能の予測が困難であるという問題があった．これまでに音声認識の難しさを判別する残響尺度として同一室内で固有の値をとる残響時間（T_{60}）が利用されている．ところが仮定音場と実環境との差異から他の残響特性が変化し，同一環境でも計測箇所によって音声認識性能が変動することから，残響時間のみで音声認識性能を予測することは困難であった．そこで本論文では，音声認識性能を残響に対して頑健かつ簡便に予測できる残響指標 RSR-D_n を提案し，音声認識性能の高精度な予測を試みた．その結果，各環境の D_{20} と音声認識性能の関係を二次曲線で近似した残響指標 RSR-D_{20}Q によって高精度な音声認識性能の予測が行えることを確認した．今後は MTF (Modulation Transfer Function) [19] 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予測法 [9] 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予測指標の策定に取り組む計画である．

謝辞　本研究の一部はグローバル COE，科研費 17700216 と 20700169 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に感謝する．

文　献

[1] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am., vol.78, no.5, pp.1508–1518, Nov. 1985.

[2] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," Proc. IEEE, vol.60, no.8, pp.926–935, Aug. 1972.

[3] M.J.F. Gales and S.J. Young, "An improved ap-

proach to the hidden Markov model decomposition of speech and noise," Proc. ICASSP, vol.1, pp.233–236, 1992.
[4] H.M. Cung and Y. Normandin, "Noise adaptation algorithms for robust speech recognition," Speech Commun., vol.12, no.3, pp.267–276, July 1993.
[5] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Signal Process. Society, vol.29, no.2, pp.254–272, April 1981.
[6] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Trans. Acoust. Speech Signal Process., vol.36, no.2, pp.145–152, 1988.
[7] 清水秦博，梶田将司，武田一哉，板倉文忠，"空間音響特性を考慮したスペースダイバシチ型音声認識，" 信学論（D-II），vol.J83-D-II, no.11, pp.2448–2456, Nov. 2000.
[8] T. Takiguchi, M. Nishimura, and Y. Ariki, "Acoustic model adaptation using first-order linear prediction for reverberant speech," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.908–914, March 2006.
[9] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," IEEE Trans. Audio, Speech, and Language Processing, vol.14, no.6, pp.2006–2013, Nov. 2006.
[10] M.R. Schroeder, "New method of measuring reverberation time," J. Acoust. Soc. Am., vol.37, pp.409–412, 1965.
[11] R. Petrick, X. Lu, M. Unoki, M. Akagi, and R. Hoffmann, "Robust front end processing for speech recognition in reverberant environments: Utilization of speech characteristics," Proc. Interspeech2008, pp.658–661, Brisbane, Australia, Sept. 2008.
[12] 日本音響学会，新版音響用語辞典，コロナ社，2003.
[13] H. Kuttruff, Room Acoustics, Spon Press, 2000.
[14] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," J. Acoust. Soc. Am., vol.97, no.2, pp.1119–1123, 1995.
[15] ISO3382: Acoustics-measurement of the reverberation time of rooms with reference to other accoustical parameters, Internatinal Organization for Standardization, 1997.
[16] 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.
[17] A. Lee, T. Kawahara, and K. Shikano, "Julius — An open source real-time large vocabulary recognition engine," Proc. European Conf. on Speech Communication and Technology, pp.1691–1694, 2001.
[18] 田中敏幸，数値計算法基礎，コロナ社，2006.
[19] T. Houtgast, H.J.M. Steeneken, and R. Plomp, "Predicting speech intelligibility in room acoustics," Acustica, vol.46, pp.60–72, 1980.

（平成 22 年 7 月 2 日受付，11 月 8 日再受付）

福森　隆寛

平 22 立命館大・情報理工・メディア情報卒．同年 4 月同大大学院理工学研究科博士課程前期課程入学，現在に至る．音響信号処理の研究に従事．日本音響学会，情報処理学会各会員．

森勢　将雅　（正員）

平 16 和歌山大・システム工・デザイン情報卒．平 18 同大大学院システム研究科博士前期課程了．同年 4 月より日本学術振興会特別研究員（DC1）．平 20 同大大学院博士後期課程了．同年 4 月より関西学院大・理工学・博士研究員．平 21 立命館大・情報理工・助教，現在に至る．博士（工学）．音声・音響信号処理，インタフェース設計及び聴覚情報処理の研究に従事．平 18 電気通信普及財団賞．日本音響学会，情報処理学会，日本バーチャルリアリティ学会各会員．

西浦　敬信　（正員）

平 9 奈良高専・専攻科・電子情報卒．平 11 奈良先端大学大学院情報科学研究科博士前期課程了．平 13 同大博士後期課程了．同年和歌山大・シス工・助手．平 16 立命館大・情報理工・助教授．平 19 同准教授，現在に至る．博士（工学）．音響信号処理，主として音環境の解析・理解・再現・生成に関する研究に従事．平 13 電気通信普及財団賞，平 13 ATR 発明・論文表彰．平 21 日本バーチャルリアリティ学会論文賞．日本音響学会，情報処理学会，日本騒音制御工学会，日本バーチャルリアリティ学会各会員．

山下　洋一　（正員）

昭 57 阪大・工・電子卒．昭 59 同大大学院修士課程了．同年阪大・産研・文部技官，平 5 同助手，平 6 同講師，平 9 立命館大・理工・助教授，平 13 同教授，平 16 同大・情報理工・教授，現在に至る．博士（工学）．音声情報処理に関する研究に従事．日本音響学会，情報処理学会，人工知能学会，ISCA，IEEE 各会員．