残響と騒音が存在する環境で最適な音響モデルの作成法*

◎太刀岡勇気, 花沢利行, 岩崎知弘 (三菱電機・情報総研)

1 はじめに

残響の長い環境で音声認識を行う場合、認識性能が大きく低下する。対策として認識に用いる音響モデルを作成する際に、残響を加えたデータを学習に用いることで認識性能が向上することが知られている [1]。しかし残響環境下もしくは騒音環境下の一方での音響モデルの学習データの作成法は様々検討されているものの、残響と騒音が同時に重畳した場合についての検討は見られない。

我々は、先に拡散音場理論に基づき残響時間をパラメータとする SS 法の引き去り係数の決定方法を提案し、残響と騒音が同時に存在する環境で残響除去により認識性能が向上することを示した [2]。その際には、残響を重畳したデータで学習した音響モデルを合わせて用いることで、認識性能のさらなる向上が図れることがわかった。本報では、残響に加え、種々の S/N で騒音が重畳したデータで音響モデルを学習し、認識性能を評価することで、学習データへの騒音重畳の効果を検討する。

2 音響モデルの作成法

2.1 学習データ

学習データは Fig. 1 に示すように、残響・騒音のないドライソースに騒音を重畳したもの (Anechoic データ) とドライソースにインパルス応答を畳み込み騒音を重畳したもの (Reverberant データ) の 2 種類作成した。Reverberant データの作成に用いたインパルス応答は、評価データの作成に用いたものを使用した。その際、評価がクローズにならないように、評価データ作成用のインパルス応答を 2 つのテストセット test1, test2 に分類し、test1 の実験を行う際に用いる音響モデルは test2 のインパルス応答を畳み込んだ音声から学習し、逆も同様にした。

学習データに加える騒音は、電子協の騒音データベース (JEIDA-NOISE) の騒音 4 種類 (ブース、交差点、板金工場、幹線道路) である。S/N は低 (12, 18)・中 (18, 24)・高 (24, 30)・クリーン (inf) の 4 水準とした。Fig. 2 に示すように、Anechoic データ 4 種類 (低・中・高・なし) ×Reverberant データ 5 種類 (低・中・高・クリーン・なし) の組み合わせの 20 種類の学習データからそれぞれ音響モデルを作成した。ただし、予備実験の結果からドライソースは学習データに必ず含めた。音素片を基本単位とする混合数 8 の混合分布とし、特徴ベクトルは MFCC (0–16 次) とその $\Delta, \Delta\Delta$ を用いた。

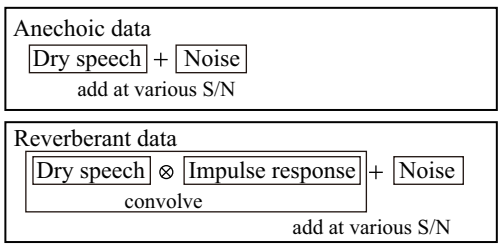

Fig. 1 Procedure of constructing anechoic and reverberant data.

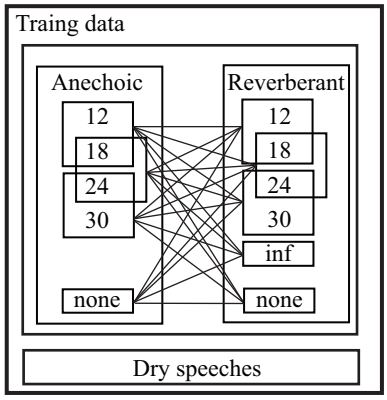

Fig. 2 How to select training data from anechoic and reverberant speeches at various S/N.

2.2 音響モデルの表記法

音響モデルは Anechoic データの S/N を N の後に、Reverberant データの S/N を R の後に表記する。たとえば、ドライソースとそれに S/N 12, 18 dB で騒音を加えたデータ、S/N 18, 24 dB で騒音と残響を加えたデータを音響モデルの学習に用いた場合は、N-12-18_R-18-24 と表記する。データを使っていない場合には none と表記する。たとえば上記のデータで Reverberant データを学習に使わなかった場合には、N-12-18_R-none となる。

3 認識実験による検証

3.1 実験条件

評価環境には残響下音声認識評価用のデータベース CENSREC-4 [3] を用いた。2.1 で述べた通り、8 環境を test1 (Office, Elevator hall, In car, Living room)

*Appropriate acoustic model for reverberant and noisy environments.
by TACHIOKA, Yuuki, HANAZAWA, Toshiyuki, IWASAKI, Tomohiro (Mitsubishi Electric Corp.)

と test2 (Lounge, Japanese (style) room, Meeting room, Japanese (style) bath)) に分割した。それぞれの環境のインパルス応答と騒音が用意されている。

評価音声は電子協 日本語共通音声データ (JEIDA-JCSD (B-set)) の 100 地名を用いた。それぞれの環境のインパルス応答を畳み込み、騒音を S/N 20, 25, 30 dB で重畳して評価データを作成した。認識対象の語彙は評価に用いた 100 地名を含む 676 地名とした。前後に 2 秒のポーズを付加して、音声区間検出は自動で行った。

3.2 実験結果 (残響環境がオープンな音響モデルを用いた場合)

S/N 20 dB の評価データに対する 8 環境平均の認識率を、音響モデルごとに比較する。

3.2.1 比較した音響モデル

今回作成した音響モデルは、学習データの騒音の有・無、残響の有・無により 4 パターンに分類できる。

(i) 騒音・残響をともに含まないモデル (N-none_R-none)、(ii) 騒音のみ含むもの (N-18-24_R-none)、(iii) 残響のみ含むもの (N-none_R-inf)、(iv) 残響・騒音ともに含むもの (N-18-24_R-inf と N-none_R-18-24) の 5 種類の音響モデルによる認識結果を比較した。

(i),(ii),(iii) は作成法が 1 通りしかないが、(iv) に関しては 2 通りの作成法が考えられる。騒音と残響の両方を同時に重畳すると、残響環境数 × 騒音環境数の学習データを作成する必要がある。一方で、残響ありデータ (騒音なし) と騒音ありデータ (残響なし) を両方用いることも考えられる。後者によって前者と同等の認識率向上効果が得られれば、学習データ数が残響環境数+騒音環境数となるため前者より少なく効率的である。よって (iv) は上述の 2 つの方法を比較した。Table 1 に用いた上記 5 種類の音響モデルをまとめた。ここでは学習時の S/N は 18-24 としている。

3.2.2 騒音、残響それぞれを重畳した効果

結果を Fig. 3 に示す。クリーンモデル (i) の認識率が他のものと比べて著しく低くなっている。騒音環境のみを学習データに含む (ii) は、(i) より認識率が向上している。これは学習データに重畳した騒音によって音響モデルの各ガウス分布の分散が広がり、残響によって他音素の影響で特徴量があいまいになった際も

Table 1 Conditions of acoustic models.

name	rev	noise	note
N-none_R-none	no	no	
N-18-24_R-none	yes	no	
N-none_R-inf	no	yes	
N-18-24_R-inf	yes	yes	anechoic (noisy) & reverberant (clean)
N-none_R-18-24	yes	yes	reverberant (noisy)

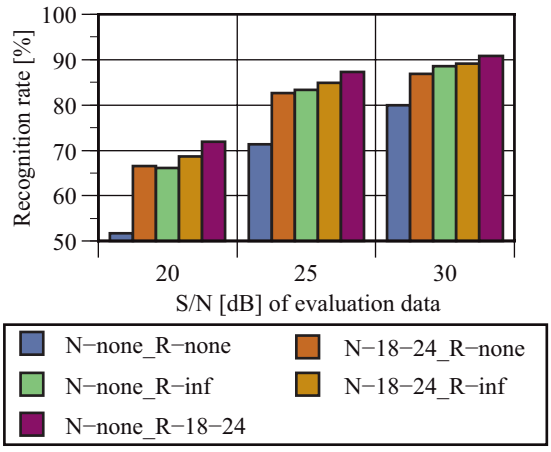

Fig. 3 Recognition rate [%]. The recognition rate is the average rate in 8 environments. Testing set is independent of training set of each HMM (open).

認識率の低下が抑えられたためと思われる。

(iii) の残響環境を学習データに含んだ場合は、騒音を重畳しなくても認識率が向上している。特に評価データの S/N が高い場合にその傾向があり、残響が認識率低下の主要因になっているためと思われる。

3.2.3 残響と騒音をともに重畳した効果

(iv) の残響と騒音をともに学習データに含むモデルは、(i),(ii),(iii) に比べてどの評価データの S/N でも認識率が向上している。

3.2.1 に述べた (iv) の 2 つの作成法の比較では、残響と騒音を同時に重畳した学習データより作成したモデルの性能のほうが、それぞれ重畳した学習データより作成したモデルよりも高かった。これより残響と騒音による認識率の低下は双方が関連しており、同時に扱う必要があることがわかる。これらの傾向は評価データの S/N によらず同様であった。

3.2.4 評価環境ごとの認識率

Fig. 3 と同じ音響モデルでの、評価環境ごとの認識率を Fig. 4 に示す。全体的な傾向は平均で見た場合と同じである。“Office”, “In car” において残響重畳の効果があまり見られないのを除けば、残響の重畳は認識性能の向上に大きな効果があることがわかる。また、残響と騒音を同時に重畳したデータの認識率が環境によらず最もよい。

3.2.5 学習データの S/N が認識性能に及ぼす影響

騒音と残響を同時に重畳したモデルの学習時の S/N が認識性能に及ぼす影響を検討するため、Reverberant データの S/N を変化させた場合の認識率を Fig. 5 に示す。モデルは Anechoic データを含まないものとし (N-none)、評価データの S/N 別にまとめた。これらの場合は Reverberant データの S/N による認識率

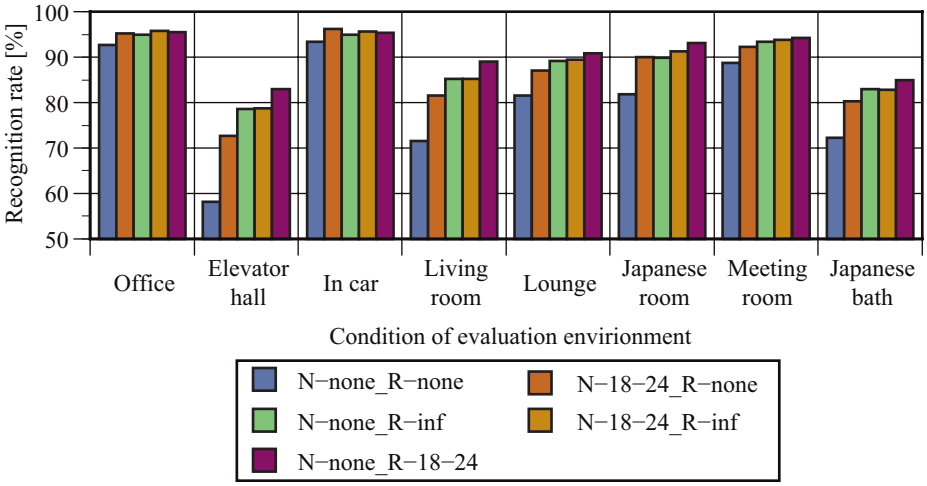

Fig. 4　Recognition rate [%] (S/N 30 dB) of each evaluation environment.

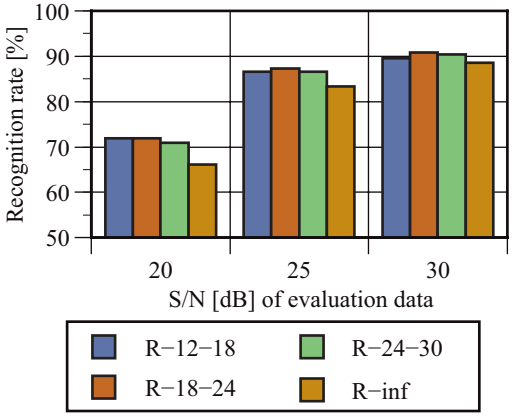

Fig. 5　Recognition rate [%]. The recognition rate is the average rate of HMM which is N-none* in 8 environments. Testing set is independent of training set of each HMM (open).

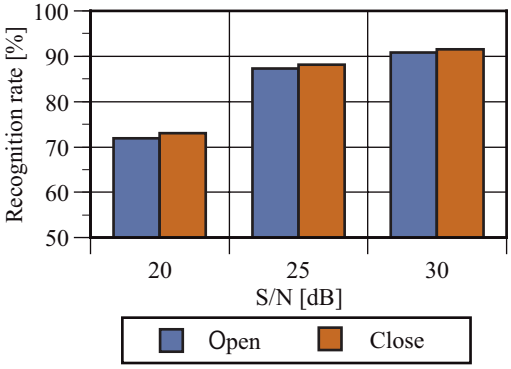

Fig. 6　Recognition rate [%] of N-none_R-12-18. Testing set is independent of training set (open) and included by training set (close) .

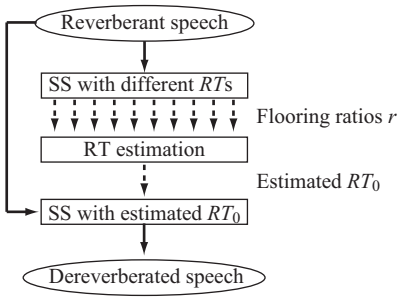

Fig. 7　Block diagram of the proposed method. (line: signal, dotted line: parameter)

の差異はさほど見られず、学習時の S/N が認識時とミスマッチであっても認識率に大きな影響を与えないことがわかる。

評価データの S/N が 30 dB の場合には、残響のみを加えたもの (R-inf) が R-12-18, R-18-24, R-24-30 の残響と騒音を同時に加えたものと同程度の認識率を示している。評価データの S/N が 20 dB の場合には、残響のみよりも残響と騒音を同時に加えたもののほうが認識率が高くなっており、評価データの S/N の低下によって残響に加えて騒音も認識率低下の主要因になっていることがわかる。

3.3　実験結果 (残響環境がクローズな音響モデルを用いた場合)

音響モデルの残響環境がクローズになるようにした場合の認識率を、Fig. 6 に示す。オープンとクローズの認識率を比較すると、認識率の差異は最大 1.2 %であり、学習時と認識時の残響環境が異なっていても認識率の低下は小さいことがわかった。

3.4　実験結果 (残響除去を行った場合)

文献 [2] に示す残響除去法を用いて認識率を検討した。残響除去法の概要図を Fig. 7 に示す。本残響除去法は、拡散音場理論に基づき複数の残響時間を仮定して、SS (Spectral Subtraction) 法により残響成分の引き去りを行うことで、音声データから認識環境の残響時間を推定する。求まった残響時間により SS 法の引き去り係数を分析フレームごとに決定することで、残響成分の除去を行う。

前節までの検討で認識率の高かった N-none_R-18-24 と、残響環境を学習データに含まない N-18-24_R-

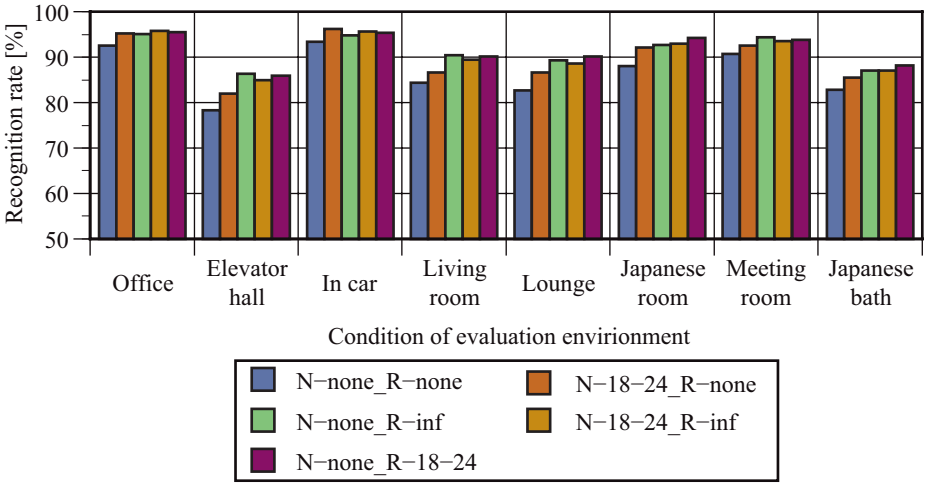

Fig. 9　Recognition rate [%] (S/N 30 dB). Speech data is dereverberated.

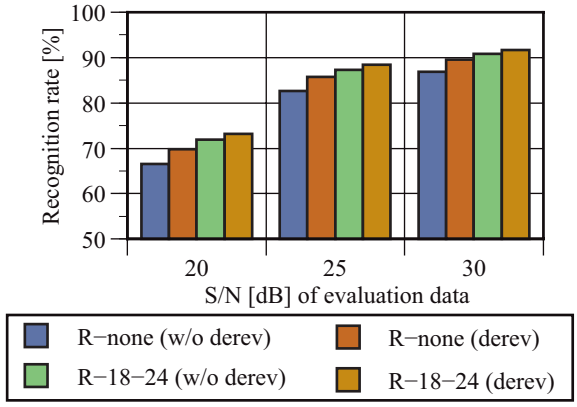

Fig. 8　Recognition rate [%] of N-18-24_R-none and N-none_R-18-24 with and without dereverberation method.

none を比較したものを Fig. 8 に示す。残響除去により音響モデルにかかわらず認識率が向上し、残響の影響を考慮した音響モデルを用いることで最高の認識率を得ることができている。これは残響除去が初期残響域を無視した簡易なものであり、残響成分を完全に除去できていないためと考えられる。なお音響モデルをクローズな環境で学習した残響除去しないものの認識率を、オープンな環境で学習した音響モデルで残響除去したものの認識率が上回っている。

残響除去を行った場合の S/N 30 dB の各環境、各音響モデルの認識率を Fig. 9 に示す。クリーンモデル等、残響環境を学習データに含んでいないモデルだけでなくすべてのモデルで、多様な環境で認識率向上効果があることがわかる。

4　まとめ

残響と騒音が重畳した場合に最適な音響モデルの作成法を検討し、以下の 3 点が明らかとなった。

第 1 に残響を加えた学習データにより認識率が向上した。評価データの S/N が低い場合、残響と騒音を同時に学習データに加えたモデルは、残響のみを加えたモデルより高い認識率が得られた。また残響畳み込みと騒音重畳を別々に行った学習データを用いる (N-18-24_R-inf) よりも、残響畳み込みデータのそれぞれに騒音を重畳した学習データを用いる (N-none_R-18-24) ほうがよいことがわかった。この際、学習時と認識時の S/N が異なっていても効果があった。

第 2 にクローズな残響環境で学習した音響モデルを用いても認識率は 1%程度しか改善せず、音響モデル学習時の残響環境と認識時の環境は一致していなくても効果があることがわかった。

第 3 に残響除去法を合わせて用いることで、認識率が向上し、オープンな残響環境で学習した音響モデルを用いても、クローズな音響モデル以上の認識率を得ることができた。

以上より、残響に加えて騒音のある環境では、認識の際の残響環境や騒音の S/N と一致していなくても、残響と騒音を同時に加えた学習データから音響モデルを作成しておくことが有効であることがわかった。また残響除去法の有効性も示された。

参考文献

[1] 西亀他, “複数残響特性下の音声を単一モデル学習に用いた未知残響環境に頑健な音声認識の検討,” 信学技報 (SP2008-8), pp. 43–48, 2008. 5.
[2] 太刀岡他, “拡散音場理論に基づく残響環境下音声認識,” 信学技報 (SP2010-4), pp. 19–24, 2010. 5.
[3] http://research.nii.ac.jp/src/list/index.html