拡散音場理論に基づく残響環境下音声認識の検討
– 騒音環境下での評価 – *

◎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ことが知られている。対策手法がいくつか提案されているが、組み込み向け用途では計算量の少ない手法が求められる。計算量の少ない方式としてはSpectral Subtraction(SS) 法 [1] により残響成分を取り除く手法 [2] が有効だが、事前に伝達関数から SS 法の引き去り係数を決定しておく必要があり、未知の環境で精度が落ちるという問題があった。

我々は、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、残響時間が既知または未知いずれの場合においても頑健に残響除去を行えることを示した [3]。しかしながらこの評価はクリーンデータのみであり、騒音が重畳した場合の有効性は検討していない。またさまざまな騒音環境下で残響除去法の効果を検証した論文はあまりない。そこで本報では残響と騒音が重畳した場合の認識性能について検討する。

2 拡散音場理論に基づく残響除去手法 (騒音がある場合)

2.1 残響成分の引き去り法

反射音と直接音はインコヒーレントであるためエネルギーの加算が許される [4]。ゆえに残響のある環境では、観測された音のパワースペクトル $|X_k[i]|^2$ は s フレーム前の音源のパワースペクトル $|Y_k[i-s]|^2$ と現在のそれ $|Y_k[i]|^2$ と Eq. (1) のように重み付き和の関係にあると考えられる。既報 [3] に加え、背景騒音のパワースペクトル $|N_k|^2$ を考慮した。$|N_k|^2$ は定常的であると仮定する。スペクトルは、あるフレーム幅、フレームシフト fr をもって短時間フーリエ変換を行い求める。

$$|X_k[i]|^2 = \sum_{s=0}^{i} w[s] \cdot |Y_k[i-s]|^2 + |N_k|^2 \qquad (1)$$

ここで i は現在のフレーム番号、k はフーリエ変換の次元 $(0 \le k \le N-1)$、$w[s]$ は重み係数 $(0 \le s \le i)$ である。音源のスペクトルは未知であるので、過去の音源のパワースペクトル $|Y_k[i-s]|^2$ に残響が重畳

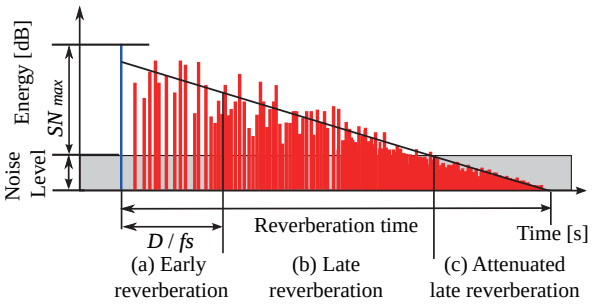

Fig. 1 Early and Late reverberation under noisy environments. (blue line : direct sound, red line : reflected sound)

することによるエネルギー増加率 q をかけて背景騒音のパワースペクトル $|N_k|^2$ を加算したものが観測パワースペクトル $|X_k[i-s]|^2$ と等しいと近似する。

$$q|Y_k[i-s]|^2 + |N_k|^2 \approx |X_k[i-s]|^2 \qquad (2)$$

さらに $w[0]=1$ を仮定して、Eq. (3) が導かれる。

$$|Y_k[i]|^2 + |N_k|^2 = |X_k[i]|^2 - \frac{1}{q}\sum_{s=1}^{i} w[s] \cdot \left(|X_k[i-s]|^2 - |N_k|^2\right) \qquad (3)$$

$q \simeq 1$ と考えると Eq. (1) と (3) の $w[s]$ $(1 \le s \le i)$ が同一視できるので、重み係数 $w[s]$ を決定すれば、Eq. (3) にしたがい SS 法により残響除去ができる。ここで $|N_k|^2$ は発話前の数フレームの平均から求めることとする。理想的には $|Y_k[i]|^2$ を求めたいが、背景騒音成分は定常ではないので SS によって背景騒音を除去することはできない。そして除去できなかった成分が認識率に悪影響を与えることになる。そこで本報では背景騒音は除去せずに残響のみを除去することを考える。つまり $|Y_k[i]|^2$ ではなく、Eq. (3) の左辺 $|Y_k[i]|^2 + |N_k|^2$ を求めることとする。

残響は減衰の程度に応じて、Fig. 1 のように、(a) 反射音が疎な初期残響と (b) 密な後期残響、(c) 背景騒音に隠された後期残響の 3 段階に分けられる。音声認識に悪影響を与えるのは主に (b) である。室の音響エネルギー密度の空間平均 $\bar{E}(t)$ は、複雑な (a) を除き Eq. (4) のように表せる。(b) は拡散音場理論に基づき時間 t[秒] に関する指数関数の形に仮定でき、

---

* A study on speech recognition under reverberant environments based on the diffuse sound field theory.
– Evaluation under noisy environments –
by TACHIOKA, Yuuki, HANAZAWA, Toshiyuki, IWASAKI, Tomohiro (Mitsubishi Electric Corp.)

(c) では背景騒音のレベルと一致するためである。

$$\begin{aligned}\bar{E}(t) &= \bar{E}(0)\exp\Big(-\frac{13.82}{RT_0}t\Big) \quad for(b)\\ &= \sum_{k=0}^{N-1}|N_k|^2 \qquad for(c) \end{aligned} \tag{4}$$

これより、(a) は無視し、(b) は Eq. (4) の指数減衰を仮定し、直接音より SN_{max} だけ減衰した (c) の部分は背景騒音と等しいとすることで、Eq. (5) の重み係数 $w[s]$ を決定できる。ただし SN_{max} は 1 発話中の最良の S/N で、30 dB を超えないとした。ここで f_s はサンプリング周波数 [Hz]、α はサブトラクト係数 ($\alpha \geq 0$)、D は後期残響域に遷移するのにかかるサンプル数、RT_0 は室の残響時間である。σ は Eq. (6) に示す引き去りエネルギーの正規化係数である。SN_{max} の違いによって引き去り量に差が出てしまうが、音声の指数減衰を仮定して 30 dB 未満の場合には σ を乗じることで引き去り量を正規化することができる。

$$\begin{aligned} w[s] &= 0 \qquad 1 \leq s \leq \frac{D}{fr}\\ &= \sigma\alpha\exp\Big(-\frac{13.82fr}{RT_0 f_s}s\Big)\\ &\qquad \frac{D}{fr} < s \leq \frac{f_s}{fr}\frac{SN_{max}RT_0}{60}\\ &= 0 \qquad \frac{f_s}{fr}\frac{SN_{max}RT_0}{60} < s \end{aligned} \tag{5}$$

$$\begin{aligned}\sigma &= \frac{\int_0^{\frac{30RT}{60}}\exp(-t)dt}{\int_0^{\frac{SN_{max}RT}{60}}\exp(-t)dt}\\ &= \frac{1-\exp(-30RT/60)}{1-\exp(-SN_{max}RT/60)}\end{aligned} \tag{6}$$

作成された $w[s]$ を用いて SS 法により残響成分の除去を行う。一般に SS 法では、引き去り後のパワースペクトル $|Y_k[i]|^2$ が小さくなりすぎないように、Eq. (7) の条件を満たすようにする。満たさない場合は $|Y_k[i]|^2$ を $\beta|X_k[i]|^2$ で置き換える。これをフロアリングと呼ぶ。β はフロアリングの閾値 ($0 \leq \beta \leq 1$) である。

$$|Y_k[i]|^2 > \beta|X_k[i]|^2 \tag{7}$$

発話直後の騒音区間では、音声の存在したスペクトル帯域のみ SS されることによって騒音のスペクトルが歪み、認識性能に悪影響をもたらす。また残響成分は指数的に減衰するため、直接音成分が大きい場合には SS を行わないほうが認識性能が良くなる。音声/非音声、直接音/残響音を判定すると、残響を最も引き去りたい Fig. 1(b) に対応しているのは、音声で残響音である場合である。そこでそのように判定された場合には、残響除去の効果を得るために β を小さく設定し、それ以外の場合には β を大きく設定する。これにより、もとの音声データのスペクトルに認識性能に悪影響を与える歪みを生じさせなくできる。

その判定は過去の騒音の平均スペクトル $|N_k|$ と現在の音声スペクトル $|X_k[i]|$ の比較により行う。判定の基準には S/N やそれらの相関係数を利用することができるが、ここでは S/N を用いる。S/N の水準を 4 段階設け、現フレームの S/N(SN) が mp 以上では直接音が優位、cp 付近では背景騒音と音声が混在、np 以下では背景騒音である可能性が高いとする。$cp < SN \leq mp$ となるフレームが音声でかつ残響音である可能性が高く、そこでの β が小さくなるように Eq. (8) のとおり設定した。

$$\begin{aligned}\beta &= 1 \qquad SN > mp\\ &= \frac{1-\beta_{min}}{mp-cp}(SN-cp)+\beta_{min}\\ &\qquad cp < SN \leq mp\\ &= \frac{\beta_{min}-1}{cp-np}(SN-np)+1 \quad np < SN \leq cp\\ &= 1 \qquad SN \leq np\end{aligned} \tag{8}$$

2.2 残響時間の推定法

Eq. (5) 中の RT_0 の推定法は既報 [3] と同様である。残響成分を除去した際に、Eq. (7) の条件を満たさずフロアリングした数 cnt を観測する。これを Eq. (9) のように発話区間 (開始, 終了フレーム番号 i_s, i_e) の時間・周波数平面での時系列パワースペクトルの総要素数 $N \times (i_e - i_s + 1)$ で除した割合を r とする。

$$r = \frac{cnt}{N \times (i_e - i_s + 1)} \tag{9}$$

このときすべてのフレームではなく、$cp \leq SN < mp$ のフレームのみを対象とすることで背景騒音の影響を小さくし、残響時間推定の精度を向上させられる。

r から導かれるフロアリングしやすさの指標 a と事前に定めた傾き γ、残響時間のオフセット c より、残響時間 RT_0 は Eq. (10) で求められる。

$$RT_0 = MAX(\gamma a - c, 0) \tag{10}$$

3 認識実験による検証

3.1 実験の条件

認識実験には残響下音声認識評価用のデータベース CENSREC-4 [5] を用いた。発話内容は 1–7 桁の連続数字であり、8 種類の異なる環境を模擬してある。騒音下での評価用データは、4 種類の残響と騒音 (office, in car, lounge, meeting room) を付加した testc が用意されている。本報では残りの 4 環境 (elevator hall, living room, japanese room, japanese bath) でも同

Table 1　Parameter list.

name	value	Eq.	name	value	Eq.
α	5	(5)	cp	2	(8)
D	1500	(5)	np	1	(8)
β	0.05	(7)	γ	0.0035	(10)
mp	$SN_{max}-5$	(8)	c	6	(10)

様に評価データを作成した。本報では前者を testc_1、後者を testc_2 とする。S/N は 20, 25, 30 dB とした。

音響モデルは音素片を用い、混合数は 8 とし、騒音重畳型 HMM と残響畳み込み型 HMM(“revHMM”) の 2 種類作成した。騒音重畳型 HMM は地名と ATR 日本語音声データベースのドライ音声と、電子協のデータベースの騒音 (自動車の走行騒音、ブースの騒音等) を S/N 12, 18 dB で加えた音声より学習した。“revHMM” はそれに加え、インパルス応答を畳み込んだクリーン音声から学習した。testc_1 の実験を行う際は testc_2 のインパルス応答を畳み込んだ音声から学習した。これを互いに行い、実験がオープンになるようにした。さらにインパルス応答を畳み込みかつ上記騒音を 12, 18 dB で重畳した音声から学習した “revHMM2” も作成した。特徴ベクトルは MFCC (0–16 次) とその $\Delta, \Delta\Delta$ を用いた。認識実験のパラメータは Table 1 のとおり設定した。

3.2　残響時間の自動推定結果

残響時間 RT_0 を各評価環境における 1001 発話から推定した結果の平均と実際の残響時間を Fig. 2 に示す。それらの相関係数は S/N 20, 25, 30 dB のときそれぞれ 0.11, 0.53, 0.73 であり、S/N が低いときには残響時間の推定が難しくなることがわかる。

3.3　認識実験の結果

3.1 の設定で認識実験を行った。認識率は発話単位で算出した。“noderev” は騒音重畳型 HMM で残響除去を行わないもの、“derev” は行ったものである。“revHMM” と “revHMM2” は残響畳み込み型 HMM で残響除去を行わなかったものである。

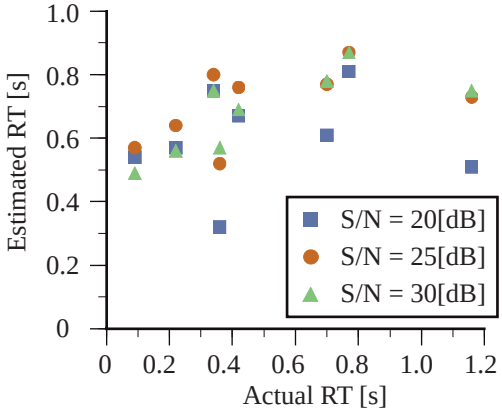

Fig. 2　Actual and estimated reverberation time (RT) [s] (S/Ns are 20, 25 and 30 dB).

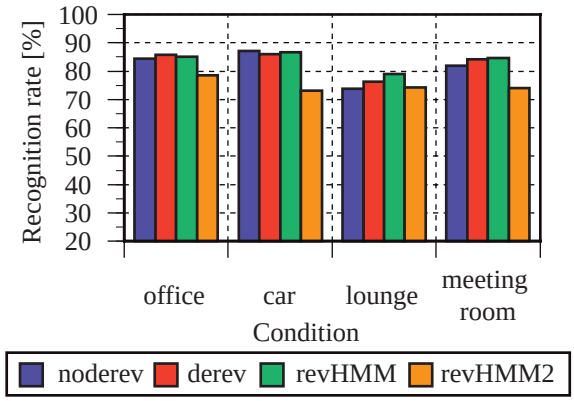

Fig. 3　Recognition rate [%] of testc_1. “noderev” and “derev” show the results without and with dereverberation respectively when HMMs are trained by clean and noisy anechoic speeches. “revHMM” shows the result without dereverberation when an HMM is trained by clean reverberant speeches convolved with impulse responses (IRs) of testc_2 in addition to anechoic speeches. “revHMM2” shows the result when an HMM is trained by clean and noisy reverberant speeches (S/N 30 dB)

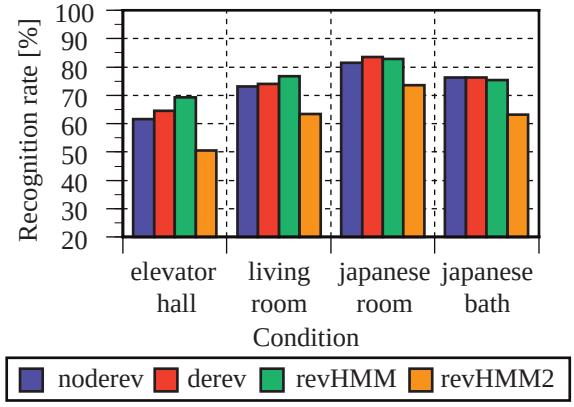

Fig. 4　Recognition rate [%] of testc_2. “revHMM” shows the result when an HMM is trained by speeches convolved with IRs of testc_1. (S/N 30 dB)

S/N 30 dB の testc_1、testc_2 の認識率を Figs. 3, 4 に示す。“noderev” と “derev” の比較では、残響が短い car 以外は認識率が 0.7–4.7 %向上し、本手法の効果が確認できた。“noderev” では評価音声がクリーンな場合 [3] よりも性能が良くなる場合がある。これは語中または語尾の無音区間の残響が騒音で隠されることにより、音響モデルとのミスマッチが低減したためと考えられる。次に全方式を比較すると “revHMM” が最も良い。これは評価音声の S/N が高く、騒音の影響が小さかったためと考えられる。一方 “revHMM2” は逆に認識率が低下した。これは残響を畳み込んだ音声は見かけ上パワーが増加し、適切な S/N で学習データを作成できなかったためと考えられる。

S/N 20 dB の testc_1、testc_2 の認識率を Figs. 5,

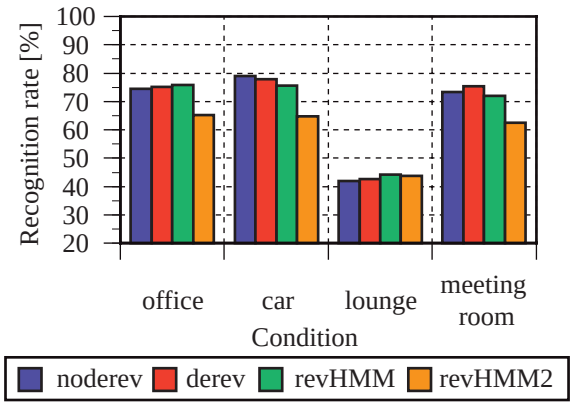

Fig. 5 Recognition rate [%] testc$_1$. (S/N 20 dB)

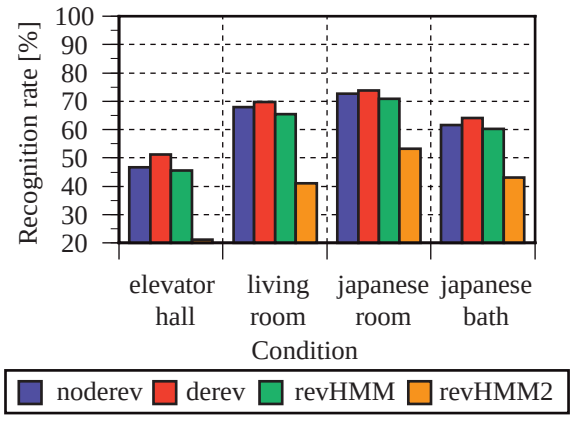

Fig. 6 Recognition rate [%] of testc$_2$. (S/N 20 dB)

6 に示す。“noderev” と “derev” は “revHMM” より office, lounge 以外で認識率が高い。また “derev” は “noderev” より car を除き認識率が向上している。ここでは、“revHMM” の認識率が “noderev” より低い。これは残響を畳み込んだ音声を学習する際に騒音を重畳していないためと考えられるが、騒音を加えた “revHMM2” はさらに性能が低い。

elevator hall での残響除去する前後の音声のスペクトログラムを Fig. 7 に示す。 騒音が主体的な低周波数域に大きな影響を与えることなく、高周波数域の音声の残響成分を取り除いていることがわかる。

以上に S/N 25 dB の場合を加えた認識率を Table 2 にまとめる。total は全評価環境での平均であり、total2 は car を除いた平均である。“revHMM” は S/N 30 dB の場合には最もよいが、20, 25 dB で性能が悪化している。“derev” は S/N 25 dB 以下では “revHMM” を上回り、全 S/N の平均では最も性能が高い。

Table 2 Average recognition rate of each S/Ns. “total” is all average of that when S/Ns are 20, 25 and 30 dB. “total2” is total average without car.

S/N	noderev	derev	revHMM	revHMM2
20	64.7	66.3	63.8	49.3
25	75.7	77.4	75.7	63.9
30	77.5	78.8	80.0	68.9
total	72.6	74.2	73.2	60.7
total2	70.4	72.2	71.2	58.5

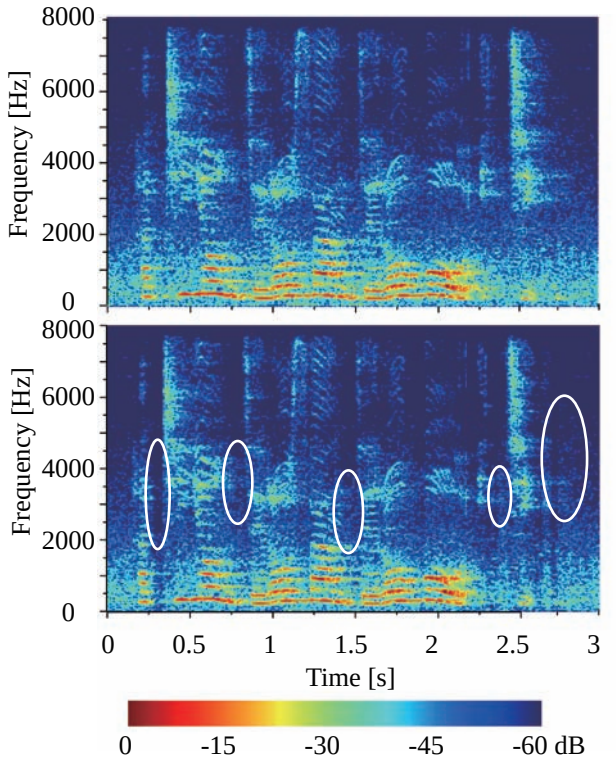

Fig. 7 Spectrogram of original (upper) and dereverberated (lower) speeches. (FAK_84Z3Z51A, elevator hall, S/N 20 dB)

4 まとめ

残響と騒音が重畳した場合の残響除去法による音声認識性能を検討した。

クリーンな場合には音声認識に悪影響を与えていた語中または語尾の無音区間の残響成分が騒音によってマスキングされるため、騒音を重畳しただけの音響モデルでもクリーンな評価データよりも認識性能が向上する場合があった。

残響を畳み込んだ音響モデルは学習時の S/N が低いと認識性能が大きく低下するが、クリーンな音声だけから学習すると S/N が低い場合に騒音を重畳しただけのモデルよりも認識性能が低下した。

本報で検討した残響除去手法は残響時間の推定精度は S/N とともに悪化するものの、残響除去により認識性能が向上し手法の有効性を確認した。

参考文献

[1] S. F. Boll, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-27-2, pp. 113–120, 1979. 4.
[2] 馬場他, 音講論 (秋), pp. 17–18, 2004. 9.
[3] 太刀岡他, 音講論 (秋), pp. 35–38, 2009. 9.
[4] H. Kuttruff, “室内音響学 – 建築の響きとその理論 –,” p. 89, 茅ヶ崎出版社, 2003.
[5] http://research.nii.ac.jp/src/list/index.html