拡散音場理論に基づく残響環境下音声認識の検討*

◎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、事前に伝達関数を測定して SS 法の引き去り係数を決定しておく必要があり、未知の環境で使いにくいという課題があった。

本稿では、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する。これによって、残響時間がわかっている場合に音声データから残響を除去できる。また残響時間が未知な場合でも、逐次的なアルゴリズムにより事前の学習なしに入力音声データのみから残響時間を推定し、未知の環境でも頑健に残響除去を行える。

2　拡散音場理論に基づく残響除去手法

2.1　残響時間が既知の場合

本節では室の残響時間 RT_0 が既知の場合に残響を除去する手法を示す。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 [3]。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。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。

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

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 $|Y_k[i-s]|^2$ に残響が重畳することによるエネルギー増加率 q をかけたものが、観測パワースペクトル $|X_k[i-s]|^2$ と等しいと近似する。

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

さらに $w[0]=1$ を仮定して、Eq. (3) が導かれる。

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

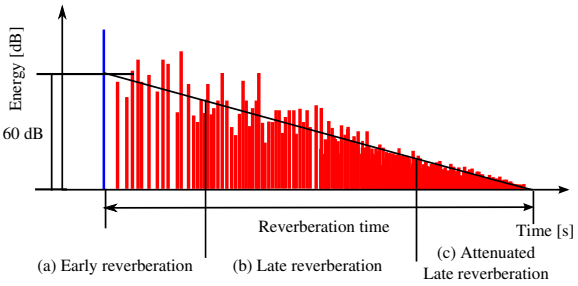

Fig. 1　Early and Late reverberation.

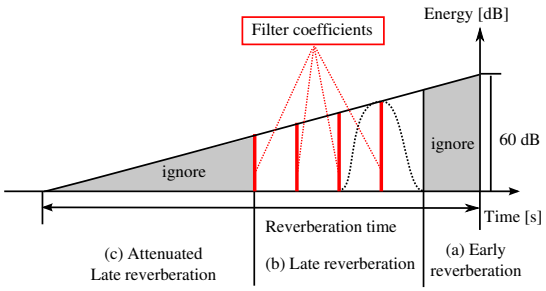

Fig. 2　How to define weight coefficients from an energy decay curve.

$q \simeq 1$ 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。

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) 十分にエネルギーが減衰した後期残響の 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。(a) はさまざまな要因が影響して複雑だが、(b)、(c) は拡散音場理論に基づき、時間 t[秒] に関する室の音響エネルギー密度の空間平均 $\bar{E}(t)$ を、Eq. (4) の指数関数の形に仮定できる。

$$\bar{E}(t) = \bar{E}(0)\exp\Big(-\frac{13.82}{RT_0}t\Big) \qquad (4)$$

音源・受音点間の伝達関数 h と受音点の音響エネルギー密度の時間変化 $E(t)$ は Eq. (5) の関係にある [4]。

$$E(t) = \frac{\int_t^\infty h^2(\tau)d\tau}{\int_0^\infty h^2(\tau)d\tau} \qquad (5)$$

$h^2(t)$ の減衰特性は $E(t)$ の時間微分で、$E(t)$ を $\bar{E}(t)$ とみなせば、これも同様に指数減衰を仮定できる。

これより、(a) は無視し、(b) は Eq. (4) の指数的減衰を仮定し、直接音より 30dB 減衰した (c) の部分は計算量を削減のため無視することで、Fig. 2 のように重み係数 $w[s]$(Eq. (6)) を決定できる。ここで f_s

*Speech recognition under reverberant environments based on the diffuse sound field theory.
by TACHIOKA, Yuuki, HANAZAWA, Toshiyuki, IWASAKI, Tomohiro (Mitsubishi Electric Co.)

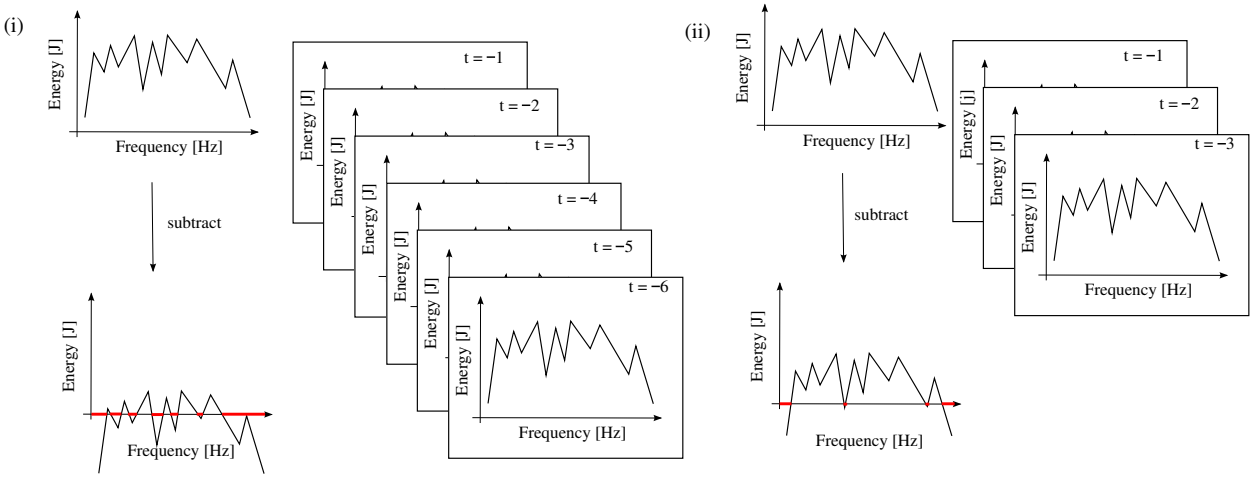

Fig. 3 Estimated power spectrum: (i) an estimated reverberation time (RT) is long. (ii) an estimated RT is short.

はサンプリング周波数 [Hz]、α はサブトラクト係数 ($\alpha > 0$)、D は後期残響域に遷移するのにかかるサンプル数である。

$$\begin{aligned} w[s] &= 0 && 1 \le s \le \frac{D}{fr} \\ &= \alpha \exp\Big(-\frac{13.82 fr}{RT_0 f_s} s\Big) && \frac{D}{fr} < s \le \frac{f_s}{fr}\frac{RT_0}{2} \\ &= 0 && \frac{f_s}{fr}\frac{RT_0}{2} < s \end{aligned} \quad (6)$$

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 ($0 < \beta < 1$) である。

$$|Y_k[i]|^2 > \beta|X_k[i]|^2 \quad (7)$$

さらに $\beta|X_k[i]|^2$ がノイズ区間における背景騒音よりも小さい場合はそれで置き換えることで、エネルギーの引きすぎによる認識性能の低下を抑えられる。

2.2 残響時間が未知な場合

本節では RT_0 が未知な場合に、フロアリングの考え方を利用しそれを推定する手法を示す。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. (8) のように発話区間 (開始フレーム番号 i_s, 終了フレーム番号 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。

$$r = \frac{cnt}{N \times (i_e - i_s + 1)} \quad (8)$$

このときすべてのフレームではなく、S/N の高い音声区間のフレームのみを対象とすることで r の分散を小さくし、残響時間推定の精度を向上させられる。

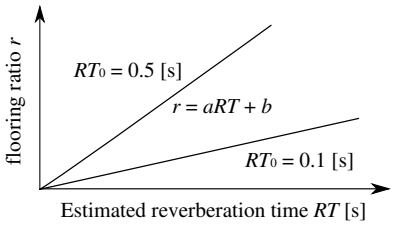

Fig. 4 Relationship between estimated RT and flooring ratio r. RT_0 is an actual RT measured at a receiving point.

ところで、観測された音の大半が残響成分である場合、Eqs. (2), (3) に示した q は室の残響時間 RT_0 に関して増加関数となる。(RT_0 が長いほど室に残留しているエネルギーが大きいため。) したがって、$q \simeq 1$ として SS を行うと、RT_0 が長いほど $1/q < 1$ であるので、Eq. (3) で求められる音源のスペクトル $|Y_k[i]|^2$ よりも引き去りすぎフロアリングしやすくなる。よってフロアリングしやすさの指標 a と RT_0 の間には正の相関があり、本稿では 1 次式 $RT_0 = \gamma a - c$ でモデル化した。ここで傾き γ と残響時間のオフセット c は事前に定めておく。

a は以下のようにして算出する。まず Fig. 4 のように、適当な残響時間 RT を仮定して、2.1 と同様に残響除去を行い、RT と r との関係を求める。一般に、Fig. 3 (i) のように RT が長いとフロアリングする部分が多く cnt および r が大きくなり、(ii) のように短く推定されると逆であるため、r は RT に関して単調増加関数となる。上述のとおり、RT_0 が長いほどフロアリングしやすいので、同じ RT に対しても r が大きくなる。そこでいくつかの RT を仮定して回帰直線 $r = aRT + b$ を求めることで、係数 a をフロアリングしやすさの指標として用いることができる。そこから $RT_0 = MAX(\gamma a - c, 0)$ と求められる。MAX は引数の 2 値の大きい方を返す関数である。

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、クリーン 4 種類 (clean1–4)、それに残響をそれぞれ畳み込んだ testa (office, elevator hall, in car, living room)、また別の残響をそれぞれ畳み込んだ testb (lounge, japanese room, meeting room, japanese bath) からなる。インパルス応答 2 乗積分法による 1 kHz の 1/3 オクターブバンド帯域の残響時間 (T_{20}) と残響減衰曲線を Table 1 と Fig. 5 に示す。

評価データは 16 kHz サンプリングとした。音声区間検出を自動で行うには、発話末尾のポーズが短かったので、ファイルの先頭の 0.3 秒を末尾に 5 回分 (1.5 秒) つけ、提供されているインパルス応答により評価データを作成した。音響モデルは音素片を用い、混合数は 8 とした。特徴ベクトルは MFCC (0–16 次) とその $\Delta, \Delta\Delta$ とした。音響モデルは地名、ATR 日本語音声データベースの発話から作成した。

3.2　パラメータ α, β の設定に関する検討

α を変化させたときの clean と rev (testa, testb の平均) の認識率を Fig. 6 に示す。D は 2000 ($RT <$ 0.25 [秒] では残響除去されない)、β は 0.05、RT_0 は、0.25, 0.5, 0.75 秒で固定とした。このように clean はあまり変化が見られなかったが、rev は若干の認識性能の差異が見られた。これより α は 5 とした。

β に関する検討結果を Fig. 7 に示す。$\beta \leq 0.05$ で clean, rev ともに認識性能が急激に低下するが、それ以上ではあまり変化がないので、β は 0.05 とした。

3.3　残響時間の自動推定結果

Figs. 6, 7 に示したように、clean は RT_0 が短いほど認識性能が高く、rev でも RT_0 が長すぎると性能が低下する。これから clean では残響除去があまり起こらず、最も残響時間の長い elevator hall で $RT_0 = 0.6$ 秒程度になるように $\gamma = 0.0055, c = 0.6$ 秒とした。

各条件における 1001 発話の平均より求めた RT_0 の推定結果を Table 2 に示す。回帰直線を求めるには残響時間を 0.25 秒より 0.05 秒ずつ増加させて r を求め、その変化量 Δr が 0.05 よりも大きい範囲を用

Table 1　RT_0 calculated by impulse responses (IRs).

condition	RT_0 [s]	condition	RT_0 [s]
office	0.22	lounge	0.36
elevator hall	1.16	japanese room	0.34
in car	0.09	meeting room	0.42
living room	0.77	japanese bath	0.70

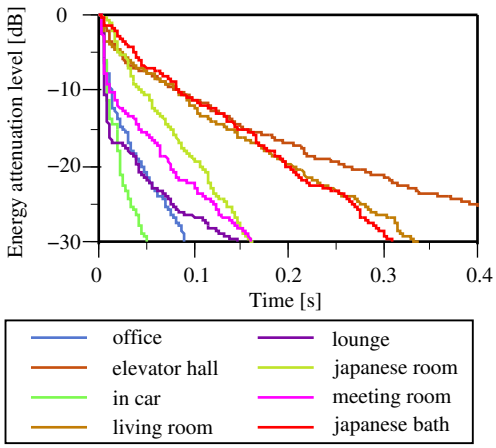

Fig. 5　Energy decay curves of testa and testb.

いた。今回は認識性能重視で γ, c を設定しているため絶対値は必ずしも一致していないが、相関係数は 0.99 と非常に高いので、γ, c の設定を変えることで一致させることも可能である。また「55619」と発話した clean2 と testa の elevator hall での推定残響時間 RT とフロアリング値 r、回帰直線を Fig. 8 に示す。Fig. 4 での仮定が成り立っていることがわかる。

3.4　認識率の検討

以上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。clean の認識率を Fig. 9 に示す。

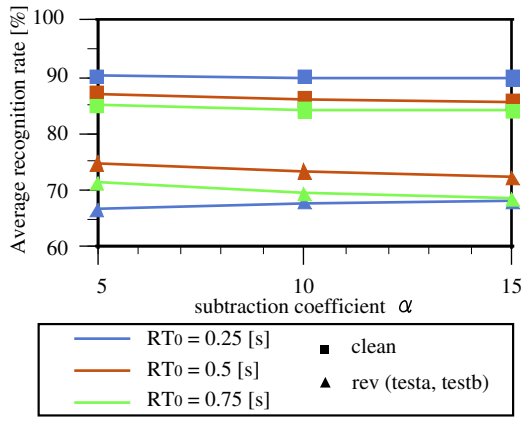

Fig. 6　Average recognition rate [%] with α.

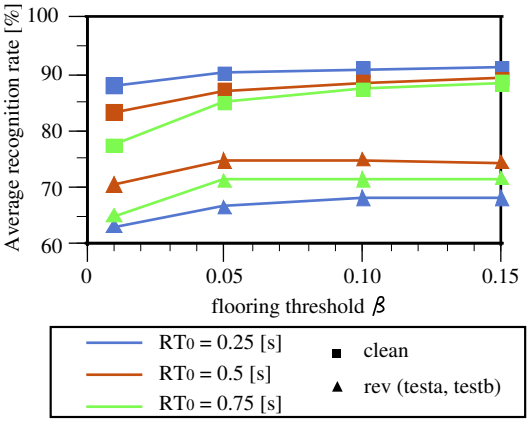

Fig. 7　Average recognition rate [%] with β.

Table 2　RT_0 estimated from speeches.

condition	RT_0[s]	condition	RT_0[s]	condition	RT_0[s]
clean1	0.23	office	0.27	lounge	0.37
clean2	0.24	elevator hall	0.59	japanese room	0.31
clean3	0.25	in car	0.24	meeting room	0.36
clean4	0.25	living room	0.48	japanese bath	0.43

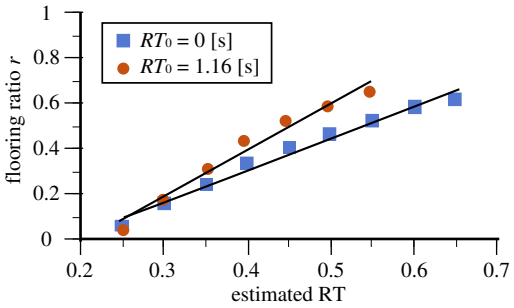

Fig. 8　Flooring ratio varying estimated RT [s] and calculated regression lines. "$RT_0 = 0$ [s]" is a clean speech "FAK_55619". "$RT_0 = 1.16$ [s]" is a reverberant speech convolved with an elevator hall IR.

"noderev" はクリーンな音響モデルで残響除去を行わないもの、"derev" は行ったものである。"revHMM" はクリーンに加え testb のインパルス応答を学習データに畳み込んだ音響モデルを用いたものである。"noderev", "derev" は "revHMM" より認識率が高い。

testa, testb の認識結果をそれぞれ Fig. 10、Fig. 11 に示す。testa の条件は clean と同様である。testb の "revHMM" に関しては testa のインパルス応答を畳み込んだ学習データより作成した。残響の長い elevator hall において "noderev" の認識性能の低下が顕著である。Fig. 5 に示したように office, elevator hall, in car, japanese room, japanese bath は指数減衰に近いため、"derev" と "revHMM" にあまり性能差はない。living room, lounge, meeting room は折れ曲がりが顕著で拡散音場のモデルに従っていないので、"revHMM"

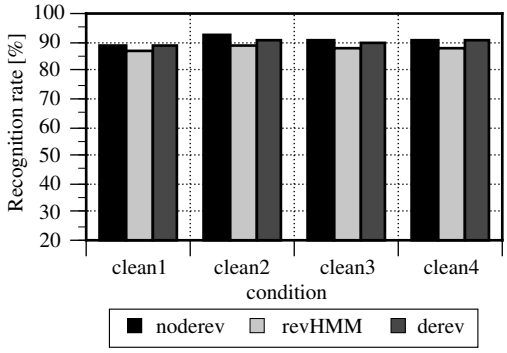

Fig. 9　Recognition rate [%] of clean speeches (clean). "noderev" and "derev" are clean HMMs without or with dereverberation. "revHMM" is an HMM made by speeches convolved with an IR of testb in addition to clean speeches.

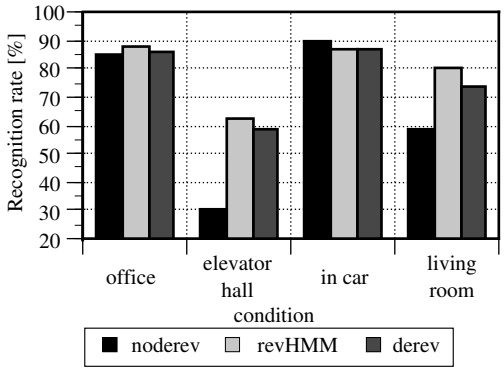

Fig. 10　Recognition rate [%] of reverberant speeches (testa).

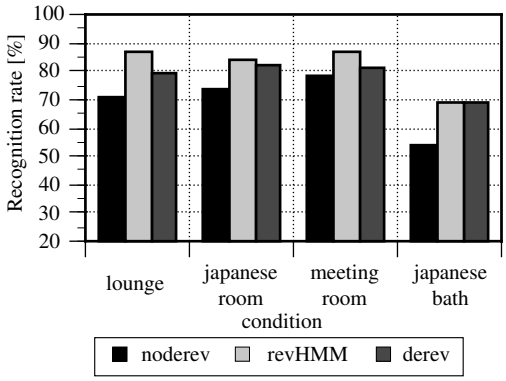

Fig. 11　Recognition rate [%] of reverberant speeches (testb). "revHMM" is an HMM made by reverberant speeches convolved with an IR of testa.

に比べて認識性能が低下している。平均の認識率は "clean" が 75.2 %、"revHMM" が 82.8 %、"derev" が 81.4 %であった。

4　まとめと今後の課題

拡散音場理論に基づく残響除去法を提案した。音声認識実験を行い、拡散音場を仮定できれば、クリーンな音響モデルでも残響を畳み込んだ音響モデルと同等の性能であることを示した。今後は、折れ曲がり残響となった際の認識性能の向上と、残響に加えて騒音を重畳した際の音声認識性能に関して検討する。

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] 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.

[4] 前川他, "建築・環境音響学," pp. 204–205, 共立出版, 2000.

[5] http://research.nii.ac.jp/src/list/index.html