遠隔マイクで集音した音声の認識精度を向上させる残響抑圧技術

Dereverberation Method to Improve Accuracy of Recognition of Spoken Language Collected by Distant Microphone

籠嶋　岳彦 KAGOSHIMA Takehiko　　金　　宜鉉 KIM Uihyun　　赤嶺　政巳 AKAMINE Masami

口元から離れたマイク（以下，遠隔マイクと呼ぶ）を用いた自由な話し言葉の音声認識では，遠隔マイクの集音で生じる残響により，音声認識精度が低下するという問題がある。

東芝は，この問題に対処するため，残響成分を抑圧する信号処理手法として，RWPE（Recursive Weighted Prediction Error）を開発した。RWPEでは，単一の遠隔マイクで集音するときに生じる残響の物理現象に合致した残響モデルを採用するとともに，再帰型の残響抑圧フィルターが不安定化する問題を回避した。残響のシミュレーション音声を用いた評価実験により，RWPEは，従来手法であるWPE（Weighted Prediction Error）に比べて，単一の遠隔マイクを用いる場合に音声認識精度が高いことを確認した。また，演算量の削減により，パソコン（PC）に実装しても，CPU能力の５％程度の演算量でリアルタイム処理できることが分かった。

A speech recognition system for spoken language collected by a single distant microphone is expected to be applicable to the preparation of minutes of meetings. To further expand the applicability of such a system, improvement of recognition accuracy is required. However, the degradation of recognition accuracy due to reverberation when using a single distant microphone is a crucial issue.

Toshiba Corporation has now developed a dereverberation method based on recursive weighted prediction error (RWPE) through the application of a moving average (MA) model of reverberation received by a single distant microphone, as well as an effective means of avoiding instability of the infinite impulse response (IIR) filter employed in this method. Experiments using samples of spoken language generated with simulation of the reverberation occurring when utilizing a single distant microphone have confirmed that our newly developed method based on RWPE offers superior recognition accuracy compared with conventional methods based on weighted prediction error (WPE). This method makes it possible to perform real-time processing on a PC by reducing the amount of computation required, utilizing only about 5% of the processing capacity of the central processing unit (CPU).

1. まえがき

音声認識技術は，深層学習の導入を契機とした，近年の音声認識性能の向上により，応用分野が急速に拡大している。スマートフォンの音声検索など，口元のマイクに向かって単語や短文を発声する用途では，既に十分な音声認識精度が得られており，多くのユーザーに利用されている。また，音声検索タスクを据え置き型の端末で行うスマートスピーカーも，実用化されている。

遠隔マイクを用いた自由な話し言葉の音声認識は，会議の議事録作成などへの応用が期待されている。今後，更に音声認識技術の応用分野を拡大していくためには，遠隔マイクを用いた音声認識の精度向上が不可欠である。遠隔マイクの音声は，口元のマイクで集音した音声に比べて，雑音と残響の影響が大きいため，音声認識精度が低下する。残響とは，壁・床・天井などで複雑に反射した音声が，音源から直接マイクに届く音声（直接音）よりも遅れて集音され，直接音と混合される現象である。特に室内で利用する場合は，たとえ静かな環境であったとしても，残響による影響が避けられない。会話の聞き取りに支障がない程度の家庭や会議室の残響でも，現状の音声認識技術では認識精度を低下させる要因となっている。

これまでに，集音した音声から残響成分を抑圧して聞き取りを容易にする，様々な信号処理手法が提案されている。例えば，部屋の環境などの情報を必要とせずに，オンラインで残響成分を推定して抑圧する手法のWPE（Weighted Prediction Error）[(1)]が知られている。WPEは，マイクで集音された過去の観測信号から現在の残響成分を予測するフィルターをオンラインで推定することによって，残響成分を抑圧する手法である。WPEが前提としている残響の自己

回帰（AR）モデルは，複数のマイクを用いる場合には遠隔マイクの残響を矛盾なく表現できるものの，単一のマイクを用いる場合には実際の物理現象と残響モデルとの間に差異が生じる。

この物理現象と残響モデルのミスマッチの問題を解決し，単一の遠隔マイクで集音された音声の認識精度を向上させるために，東芝は，新たな残響抑圧手法としてRWPE（Recursive Weighted Prediction Error）を開発した[2]。RWPEは，遠隔マイクにおける残響の物理現象と矛盾しない移動平均（MA）モデルに基づいて，過去の出力信号から現在の残響成分を予測し抑圧する手法である。RWPEでは，残響抑圧フィルターが再帰型となるため，推定したフィルター係数が不安定となった場合に出力が発散するという問題がある。そこで，この問題を回避するフィルター係数の推定方法と安定化手法を考案した。また，実用化に向けて，低遅延及び低演算量の実装についても検討し，残響に伴う音声認識精度の評価実験を行って，開発した手法の有効性を確認した。

ここでは，今回開発したRWPEの概要と，残響のシミュレーション音声を用いた，音声認識精度の評価実験の結果について述べる。

2. RWPE

2.1 残響モデル

短時間フーリエ変換（STFT：Short-Time Fourier Transformation）領域における，開発した手法の観測信号を，式(1)で表されるMAモデルで定義する。

$$\boldsymbol{x}_{n,f} = \boldsymbol{s}_{n,f} + \sum_{k=0}^{K-1} \boldsymbol{c}_{k,f} \boldsymbol{s}_{n-D-k,f} \tag{1}$$

ここで，$\boldsymbol{x}_{n,f}$及び$\boldsymbol{s}_{n,f}$は観測信号と音源信号のSTFTスペクトルを，$\boldsymbol{c}_{k,f}$は残響フィルター係数を，n,fはSTFTの時間フレーム番号と周波数ビン番号を，K,Dは残響フィルタータップ数と後部残響遅延フレーム数を，それぞれ表している。式(1)の右辺の第1項が直接音を，第2項が残響成分を表している。室内インパルス応答によって規定される残響フィルター係数と，過去の音源信号の畳み込みによって残響成分が表されることから，遠隔マイクで集音した音声の残響モデルと一致している。

一方，従来手法のWPEの残響モデルは，式(2)のARモデルで表される。

$$\boldsymbol{x}_{n,f} = \boldsymbol{s}_{n,f} + \sum_{k=0}^{K-1} \boldsymbol{c}_{k,f} \boldsymbol{x}_{n-D-k,f} \tag{2}$$

ARモデルでは，残響成分が観測信号とフィルター係数の畳み込みで表される。これは，マイクで集音した音声がスピーカーから再生され，室内で反響して再度マイクに集音される場合の残響と考えることができる。放送設備（PA：Public Address）を利用する場合に生じる残響という意味で，以後これをPA残響と呼ぶ。

2.2 残響フィルターの推定

開発した手法の残響モデルに基づく，残響フィルター係数の推定方法について述べる。以下の処理は，周波数ビンごとに独立に行われるため，fは省略する。音源$\boldsymbol{s}_n$を推定音源信号$\boldsymbol{y}_n$で置き換え，畳み込みをベクトルの内積で表現すると，式(1)より式(3)を得る。

$$\boldsymbol{y}_n = \mathbf{x}_n - \bar{\boldsymbol{c}}^H \bar{\boldsymbol{y}}_{n-D} \tag{3}$$

ここで，$\bar{\boldsymbol{c}}$と$\bar{\boldsymbol{y}}$は，それぞれ$\boldsymbol{c}_k$と$\boldsymbol{y}_n$を要素とするK次元のベクトル，Hは複素共役転置である。$\boldsymbol{y}_n$が時変ガウス分布に従うと仮定し，式(4)で表される対数尤度（ゆうど）関数を最大化するフィルター係数$\bar{\boldsymbol{c}}$を求める。

$$p(\bar{\boldsymbol{c}}, \bar{\boldsymbol{\sigma}}) = \sum_n \log N(\boldsymbol{y}_n = \boldsymbol{x}_n - \bar{\boldsymbol{c}}^H \bar{\boldsymbol{y}}_{n-D}; 0, \bar{\boldsymbol{\sigma}}) \tag{4}$$

$$= -\sum_n \frac{\left| \boldsymbol{x}_n - \bar{\boldsymbol{c}}^H \bar{\boldsymbol{y}}_{n-D} \right|^2}{\sigma_n^2} - \sum_n \log \sigma_n^2 \tag{5}$$

ここで，Nはガウス分布を表している。また，$\bar{\boldsymbol{\sigma}}$は各フレームの出力の分散を表すベクトルであり，$\bar{\boldsymbol{\sigma}} = \{\sigma_n^2\} = \{E(\boldsymbol{y}_n \boldsymbol{y}_n^*)\}$で表される。ただし，*は複素共役を示す。

$p(\bar{\boldsymbol{c}}, \bar{\boldsymbol{\sigma}})$を最大化する$\bar{\boldsymbol{c}}$及び$\bar{\boldsymbol{\sigma}}$を求めるアルゴリズムは，以下のとおりである。

まず，初期化を行う。

$$i = 1,\ \sigma_{n,i-1}^2 = \left| \boldsymbol{x}_n \right|^2,\ \boldsymbol{y}_{n,i-1} = \boldsymbol{x}_r \tag{6}$$

次に，収束するまで以下のステップを繰り返す。

$$\left(\sum_n \frac{\bar{\boldsymbol{y}}_{n-D,i-1} \bar{\boldsymbol{y}}_{n-D,i-1}^H}{\sigma_{n,i-1}^2} \right)^+ \sum_n \frac{\bar{\boldsymbol{y}}_{n-D,i-1} \boldsymbol{x}_n^*}{\sigma_{n,i-1}^2} \rightarrow \bar{\boldsymbol{c}}_i \tag{7}$$

$$\boldsymbol{x}_n - \bar{\boldsymbol{c}}_i^H \bar{\boldsymbol{y}}_{n-D,\,i-1} \rightarrow \boldsymbol{y}_{n,i} \tag{8}$$

$$\left| \boldsymbol{y}_{n,i} \right|^2 \rightarrow \sigma_{n,i}^2 \tag{9}$$

$$i+1 \rightarrow i \tag{10}$$

ここで，iは繰り返しのインデックスを，$^{+}$はムーア・ペンローズの疑似逆行列を表している。式(3)の残響抑圧フィルターは，不安定になる可能性がある再帰型フィルターであるが，式(8)では$\boldsymbol{y}_{-n-D,i}$の代わりに直前のイテレーションの$\boldsymbol{y}_{-n-D,i-1}$を近似的に用いることで再帰を回避し，不安定化の問題を解決している。収束した時点での$\boldsymbol{y}_{n,i}$が，残響成分を抑圧した音声スペクトルの出力であり，これを短時間フーリエ逆変換することで，出力波形を得る。

2.3 低遅延の実装

2.2節で述べたアルゴリズムでは，観測信号の全区間を用いて最適化を行うため，リアルタイム性が要求される応用分野には適用できない。そこで，短時間のブロック単位でフィルター係数を逐次的に更新することにより，遅延を低減する手法を開発した。1ブロックの長さをBフレーム（$B \leq D$）とする。ブロック長Bが小さくなると，ブロック内のデータだけでは信頼できるフィルター係数を求めることが難しいため，指数移動平均を用いた平滑化を導入する。b番目のブロックにおける係数更新のアルゴリズムは，2.2節で述べたアルゴリズムにおいてフィルター係数を更新する式(7)を，式(11)，(12)，(13)で置き換えたものとなる。

$$\alpha \boldsymbol{R}_{b\text{-}1} + (1-\alpha)\sum_{n=Bb}^{Bb+B-1} \frac{\bar{\boldsymbol{y}}_{n-D}\bar{\boldsymbol{y}}_{n-D}^{H}}{\sigma_n^2} \rightarrow \boldsymbol{R}_{b,i} \tag{11}$$

$$\alpha \bar{\boldsymbol{r}}_{b\text{-}1} + (1-\alpha)\sum_{n=Bb}^{Bb+B-1} \frac{\bar{\boldsymbol{y}}_{n-D}\boldsymbol{x}_n^{*}}{\sigma_n^2} \rightarrow \bar{\boldsymbol{r}}_{b,i} \tag{12}$$

$$\boldsymbol{R}_{b,i}{}^{+}\bar{\boldsymbol{r}}_{b,i} \rightarrow \bar{\boldsymbol{c}}_{i,} \tag{13}$$

ここで，行列$\boldsymbol{R}_0$及びベクトル$\boldsymbol{r}_{-0}$の全要素は0で初期化する。また，α ($0 \leq \alpha < 1$)は平滑化を制御するパラメーターである。当該ブロックの最適化が収束した後に，$\boldsymbol{R}_{b,i} \rightarrow \boldsymbol{R}_b$，$\boldsymbol{r}_{-b,i} \rightarrow \boldsymbol{r}_{-b}$により更新して，次のブロックの処理に移行する。

2.4 演算量の削減

2.3節で述べたリアルタイム処理を適用するためには，低遅延とともに低演算量が要求される。2.3節のアルゴリズムでは，行列更新の式(11)及び連立方程式の解法の式(13)が演算量の多くを占める。そこで，演算量を削減するため，行列$\boldsymbol{R}_{b,i}$を式(14)の三角行列$\hat{\boldsymbol{R}}_{b,i}$で近似する[3]。

$$\hat{R}_{b,i}(k,\ m) = \begin{cases} R_{b,i}(k,\ m) & (k \leq m) \\ 0 & (k > m) \end{cases} \tag{14}$$

行列$\boldsymbol{R}_{b,i}$は，STFTスペクトルの重み付き自己相関行列とみなすことができる。このとき，フレーム間でのスペクトルの相関が小さい場合は，対角要素が支配的になることから，近似誤差は比較的小さくなる。そこで，$\hat{\boldsymbol{R}}_{b,i}$を用いると，式(13)はガウスの消去法の後退代入を表す式(15)で置き換えられ，演算量が大幅に削減される。

$$\bar{\boldsymbol{c}}_i(k) = \frac{\bar{\boldsymbol{r}}_{b,i}(k) - \sum_{m=k+1}^{K} \boldsymbol{R}_{b,i}(k,m)\bar{\boldsymbol{c}}_i(m)}{\boldsymbol{R}_{b,i}(k,k)} \tag{15}$$

3. 評価実験

3.1 残響モデルの評価

残響モデルと音声認識精度の関係を把握するため，残響シミュレーションにより生成した音声を用いた評価実験を行った。音源として，日本語話し言葉コーパス[4]から男女5名ずつの音声資料を選択して連結し，25分間のクリーン音声（16 kHzサンプリング）を作成した。シミュレーションに用いた室内インパルス応答h_kは，Reverb Challenge(注1)[5]で配布されているデータ（Room 2, Far, Angle A, Channel 1）を用いた。残響時間（T60）は0.5秒，音源とマイク間の距離は2 mである。インパルス応答長Lは，残響フィルターのタップ数と対応するように，0.18秒相当で打ち切った。これを用いて，式(16)と式(17)で表される2種類の残響を模擬するフィルターを作成した。

$$H_{\mathrm{MA}}(z) = \sum_{k=0}^{L-1} h_k z^{-k} \tag{16}$$

$$H_{\mathrm{AR}}(z) = 1/\left(1 - g z^{-d} H_{\mathrm{MA}}(z)\right) \tag{17}$$

$H_{\mathrm{MA}}(z)$と$H_{\mathrm{AR}}(z)$は，それぞれ遠隔マイク残響とPA残響を模擬するフィルターである。ここで，dはPAのスピーカーとマイクの距離によって決まる遅延を表し，2 mに相当する94に設定した。また，gは，PAのボリュームに相当するパラメーターであり，ハウリングが発生しない範囲で最大の値に調整した。

音声認識の精度評価には，当社で開発した音声認識エンジンを用いた。音響モデルは，隠れ層4層（各層1,024ノード）のDNN（Deep Neural Network）-HMM（Hidden Markov Model）であり，言語モデルは，隠れ層1層（512ノード）のニューラルネットワークモデルである。音源として用

（注1） 残響音声の音声認識精度を競うイベント。

いたクリーン音声の文字誤り率（CER：Character Error Rate）は11.7％であった。

残響抑圧に用いたSTFTの窓関数は，512サンプルのハニング窓で，フレームシフトは128サンプルとした。従来手法のWPEと，開発した手法のRWPEにおいて，D=4，K=20とした。バッチ処理の実装では，繰り返し回数は3回に固定した。また，低遅延処理の実装ではB=4，α=0.95に設定し，繰り返し回数は1回とした。

処理なしの残響シミュレーション音声と，バッチ処理及び低遅延処理を行った残響抑圧音声のCER評価結果を，**図1**に示す。開発した手法と従来手法のどちらも，残響成分の抑圧によってCER低減効果が見られた。また，バッチ処理と比較して，低遅延処理でもほぼ同等かそれ以上のCER低減効果が得られた。従来手法との比較では，遠隔マイク残響に対しては開発した手法が優位，PA残響に対しては従来手法が優位となった。すなわち，残響の物理現象に合致した残響抑圧手法を適用することで，より良好なCERが得られた。これらの結果から，単一の遠隔マイクによる音声認識タスクでは，従来手法よりも開発した手法の方が音声認識精度が高いと考えられる。

3.2　実用性の評価

次に，実際に遠隔マイクで集音された音声データを用いて，残響抑圧による音声認識精度の向上効果と演算量の関係を把握するため，開発した手法の実用性を評価した。

3.1節の実験と同じ音源を用いて，スピーカーから再生した音声を，0.5 m，1.0 m，1.5 mの距離のマイクでそれぞれ収録した音声データ（16 kHzサンプリング，16ビットPCM（Pulse Code Modulation））を実験に用いた。音声認識のCERの評価には，3.1節の実験と同じ音声認識エンジンを用いた。音声認識精度の向上効果は，式⑱の誤り削減率（RERR：Relative Error Reduction Rate）で測定した。

$$RERR(\%)=\frac{CER_{\mathrm{org}}-CER_{\mathrm{derev}}}{CER_{\mathrm{org}}}\times 100 \qquad ⑱$$

ここで，CER_{org}及びCER_{derev}は，残響抑圧前後のCERをそれぞれ表している。

残響抑圧手法として，開発した手法の低遅延処理を用いることで演算量削減の有無を切り替えて，演算量と音声認識精度の向上効果を測定した。プログラムはPC上のC++言語で実装した。開発環境は，CPUがIntel®Core™ i5-4300Mプロセッサー（動作周波数：2.60 GHz），メモリーがRAM 4 Gバイト，OS（基本ソフトウェア）が64ビット版Windows 7 Professional Service Pack 1である。演算量は，式⑲で定義される実時間係数（RTF：Real Time Factor）で測定した。

$$RTF=（演算時間）/（処理した音声の長さ） \qquad ⑲$$

残響抑圧に用いたSTFTの窓関数は512サンプルのハニング窓，フレームシフトは128サンプルとした。パラメーターはD=4，K=20，B=1，α=0.999，繰り返し回数は1回とした。また，周波数ビンの中で，低域の128個だけに残響抑圧処理を適用した。残響抑圧処理のアルゴリズム遅延は，STFTによる遅延を含めて32 msであった。

評価結果を**表1**に示す。演算量削減により6倍の高速化を達成し，十分にリアルタイム処理が可能であることが分かった。また，演算量削減のための近似導入によるRERRの低下は1ポイント以内で，いずれの距離でも10％以上のRERRが得られた。更に，実収録音声でも，音声認識誤りの削減効果を確認した。

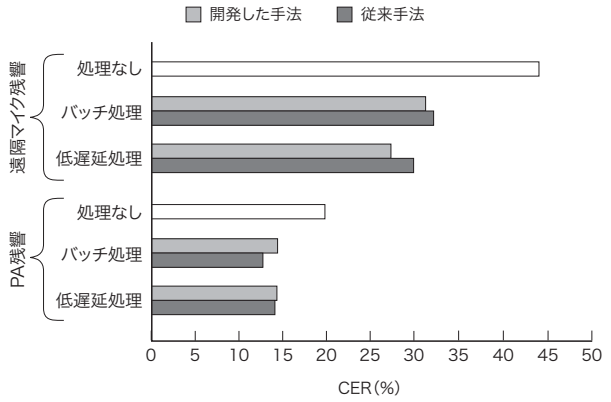

図1．開発した手法と従来手法におけるCERの比較

遠隔マイク残響に対しては，従来手法よりも開発した手法の方が良好なCERが得られ，音声認識精度が高いことが分かった。

Comparison of character error rates (CERs) of conventional and newly developed speech dereverberation methods

表1．開発した手法での，演算量の削減有無によるRTFとRERRの関係

Results of measurement of real-time factors (RTFs) and relative error reduction rates (RERRs) of newly developed method with and without reduction in amount of computation

演算量削減	RTF	RERR（%）		
		0.5 m	1.0 m	1.5 m
なし	0.30	11.38	16.84	11.38
あり	0.05	10.90	15.86	10.52

4. あとがき

遠隔マイクで集音した音声の認識精度を向上させるための残響抑圧手法として，RWPEを開発した。開発した手法により，単一の遠隔マイクによる音声認識タスクにおいて，従来手法と比較して音声認識精度が高いことをシミュレーション実験で確認した。また，低遅延実装と演算量削減により，PCによるリアルタイム処理もできることを検証した。

文　献

⑴ Nakatani, T. et al. Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction. IEEE Transactions on Audio, Speech, and Language Processing. 2010, 18, 7, p.1717–1731.

⑵ Kagoshima, T. et al. Speech dereverberation based on recursive weighted prediction error. 信学技報. 2018, 117, 515, p.367–372.

⑶ 金 宣鉉，籠嶋岳彦.“線形予測に基づく低遅延残響除去処理の高速化”. 日本音響学会2018年春季研究発表会講演論文集. 埼玉，2018-03, 日本音響学会. 2018，p.587–588.

⑷ Maekawa, K. et al. "Spontaneous speech corpus of Japanese". Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC) 2000, Athens, Greece, 2000-05, ELRA. 2000, p.947–952.

⑸ REVERB challenge organizers, "Materials for challenge participants". REVERB challenge. <https://reverb2014.dereverberation.com/download.html>, (accessed 2018-05-06).

籠嶋　岳彦　KAGOSHIMA Takehiko, Ph.D.
研究開発本部 研究開発センター
メディア AI ラボラトリー
博士（工学）　電子情報通信学会・日本音響学会会員
Media AI Lab.

金　　宣鉉　KIM Uihyun, Ph.D.
研究開発本部 研究開発センター
メディア AI ラボラトリー
博士（情報学）　日本音響学会会員
Media AI Lab.

赤嶺　政巳　AKAMINE Masami, D.Eng.
研究開発本部 研究開発センター
メディア AI ラボラトリー
工博　IEEE・電子情報通信学会・日本音響学会会員
Media AI Lab.