調波構造に基づく音声信号のブラインド残響除去*

○中谷 智広　　三好 正人
(日本電信電話株式会社 NTT コミュニケーション科学基礎研究所 音声オープンラボ)

1　はじめに

一般に、遠隔マイクで録音された音声信号には残響が含まれており、音声認識の性能劣化原因となる。特に残響時間が 0.5 秒以上になると、残響を含む音声信号を学習に用いても認識率の改善に限界があることが報告されている [1]。これに対処する一つの方法は、音声認識の前処理として残響除去を行うことである。本稿では、単一マイクで録音された音声の残響除去のために調波構造を用いる方法を提案し、実験によりその有効性を示す。

2　調波構造に基づく残響除去法

2.1　音声のモデル

実環境では、音源信号 $X(\omega)$ は残響を伴ってマイクに収音される。この残響を含む観測信号 Y は、X と伝達関数 H の積で表現することができ (式 (1))、H はさらに直接音成分 D と残響成分 R の二つの関数に分けることができる (式 (2))。前者は X から直接音 DX へ、後者は残響 RX への伝達関数である。

$$\begin{aligned} Y(\omega) &= H(\omega)X(\omega), && (1)\\ &= D(\omega)X(\omega) + R(\omega)X(\omega), && (2)\\ &= H'(\omega)X'(\omega), && (3)\\ X'(\omega) &= Y(\omega)/H'(\omega). && (4) \end{aligned}$$

本稿では直接音 $X'(= DX)$ を残響除去の目的信号とする。直接音 X' は、Y から RX を削除するか、もしくは、伝達関数 $H'(= H/D)$ を推定することができれば、Y に逆伝達関数 $1/H'$ を乗ずることで求めることができる (式 (3), (4))。

音声の場合、X は調波成分 X_h と雑音成分 X_n からなると仮定でき (式 (5))、この信号に伝達関数 H を乗じた信号として観測される (式 (6))。また、この観測信号 Y は、調波成分の直接音 DX_h とその他の成分の和とも解釈できる (式 (7))。

$$\begin{aligned} X &= X_h + X_n, && (5)\\ Y &= H(X_h + X_n), && (6)\\ &= DX_h + (RX_h + HX_n). && (7) \end{aligned}$$

このうち、直接音 DX_h は、観測信号 Y から、その基本周波数 (F_0) の整数倍に位置する各高調波の位相と強度を抽出することで近似的に求めることができる [2]。F_0 は時々刻々変動するのに対し、残響には変動以前の F_0 に起因する成分も多く含まれている。このため、F_0 とともに変動する調波成分を抽出する操作（調波構造フィルタ）により残響を低減することができる。調波構造フィルタによる直接音 DX_h の近似音 $\hat{X}'_h$ には、各高調波と同じ周波数に重畳している調波成分の残響の一部 $\hat{R}X_h$ とその他の雑音成分の一部 $\hat{N}$ が残存するのみである[1]。これを以下のようにモデル化する。

$$\hat{X}'_h = DX_h + (\hat{R}X_h + \hat{N}). \quad (8)$$

ここで、$\hat{X}'_h$ の 近似誤差はすべて $\hat{R}X_h + \hat{N}$ に含まれているとみなしている。

2.2　残響除去の原理

$\mathcal{O}(\hat{R}) = (D + \hat{R})/H$ を “残響除去オペレータ” と定義する。$\mathcal{O}(\hat{R})$ に Y を乗ずることで得られる信号 $DX + \hat{R}X$ が、残響が抑制された信号となるからである。

$$\mathcal{O}(\hat{R})Y = DX + \hat{R}X. \quad (9)$$

すなわち、$\mathcal{O}(\hat{R})Y$ は直接音 DX と一部の残響 $\hat{R}X$ から構成される信号となる。式 (2) に含まれている残りの残響 $(R - \hat{R})X$ は、残響除去オペレータによって取り除かれる。

残響除去オペレータ $\mathcal{O}(\hat{R})$ を求めるために、調波構造フィルタによる直接音の近似信号 $\hat{X}'_h$ を利用する。まず、複数の観測信号 Y に対して個別に $\hat{X}'_h$ を求める。次に、$\hat{X}'_h$ と Y の各組に対して H' の逆伝達関数の初期推定値 $1/\hat{H}'(= \hat{X}'_h/Y)$ を計算する。最後に、この初期推定値の平均 $E(1/\hat{H}')$ を計算することで、$\mathcal{O}(\hat{R})$ を近似的に求める。

$E(1/\hat{H}')$ が $\mathcal{O}(\hat{R})$ の良い近似となることは、$E(\hat{X}'_h/Y)$ に式 (6), (8) を代入することで、以下のように示すことができる。

$$E(\frac{1}{\hat{H}'}) = \mathcal{O}(\hat{R})E(\frac{1}{1+\frac{X_n}{X_h}}) + E(\frac{1}{1+\frac{Y-\hat{N}}{\hat{N}}}). \quad (10)$$

ここで、複素関数 $f(z) = 1/(1+z)$ は、z の偏角が一様分布で、z の偏角と $|z|$ が独立であり、かつ $z \neq -1$ と仮定すると、留数定理を用いて $E(f(z)) = P(|z| < 1)$ を示すことができる（$P(\cdot)$ は確率、証明略）。また、式 (10) 中の $\hat{N}$ は調波構造フィルタが抽出する調波成分に重畳した雑音成分であるため、十分に長い分析フレームを用いれば、その大きさは $Y - \hat{N}$ と比べて小さな値になると考えられる。これらの性質を用いると上式は以下のように近似できる。

$$E(\frac{1}{\hat{H}'(\omega)}) \simeq \mathcal{O}(\hat{R}(\omega))P(|X_n(\omega)| < |X_h(\omega)|). \quad (11)$$

つまり、逆伝達関数の平均値は、残響除去オペレータに対して、音声の調波成分が雑音成分よりも大きくなる確率（0〜1 の間の実数）を乗じた値となる。

*Harmonic structure based speech signal dereverberation, by Nakatani, T., and Miyoshi, M. (NTT Corporation)

[1] 単純化のため $\hat{R}$ は一定の変換として表現している。厳密には、$\hat{X}'_h$ に含まれる残響は X' の時間構造に依存して変動する。

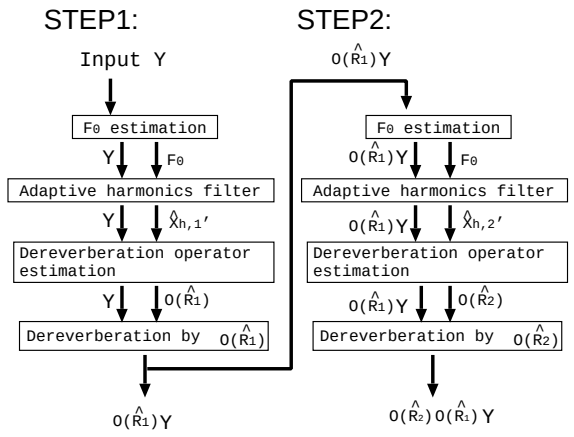

図 1: 残響除去フロー

音声では、通常、高域ほど雑音成分の影響が強くなる。このため、$P(|X_n(\omega)| < |X_h(\omega)|)$ は ω が大きくなるほど小さくなり、それに応じて逆伝達関数の平均値も減衰する。この減衰特性は、音声に関する $P(|X_n(\omega)| < |X_h(\omega)|)$ の平均的な特性を用いれば補正することは可能と考えられる。以下、本稿ではこうして求められる逆伝達関数の平均値を残響除去オペレータの推定値として扱う。

2.3 残響除去フロー

残響除去処理は図 1 に示した 2 段で構成される。

1. STEP1 では、まず、観測信号 Y から F_0 を求めたのち、調波構造フィルタを用いて Y に含まれる調波成分 $\hat{X}'_{h,1}$ を抽出する。次に、複数の観測信号 Y に対する $\hat{X}'_{h,1}/Y$ の平均から残響除去フィルタ $\mathcal{O}(\hat{R}_1)$ を推定する。最後に、Y に $\mathcal{O}(\hat{R}_1)$ を乗じて残響除去した信号を得る。
2. STEP2 は、STEP1 で残響除去した信号 $\mathcal{O}(\hat{R}_1)Y$ を入力として用いる以外は STEP1 と同じ処理を行う。$\mathcal{O}(\hat{R}_1)Y$ を用いることで、式 (8) に含まれる残響成分 $\hat{R}_2 X_{h,2}$ が抑制されるので、さらなる残響除去効果を達成できると期待される。

なお、各ステップで、式 (10) における $1/\hat{H}'(\omega)$ の平均は、振幅スペクトル $|\hat{X}'_h(\omega)|$ の重み付きで計算する。これにより、雑音成分の影響を抑制しつつ占有的な調波成分の影響を強調することができ、より正確に残響除去オペレータを求めることができる。

2.4 残響に頑健な F_0 推定法

提案法を用いて効果的な残響除去を行うためには正確な F_0 推定が重要である。しかし、残響の長い環境では、変動前の F_0 に起因する残響が変動後の F_0 の推定精度を劣化させる。この影響を低減するために、我々は、継続音を抑制する単純なフィルタを設計し、STEP1 の F_0 推定の前処理として用いた。継続音抑制フィルタは、振幅スペクトル $|Y(\omega, n)|$ の時系列に適用する。この継続音抑制フィルタによる F_0 推定性能の向上は、予備実験により確認されている。

また、残響除去オペレータ自身も、F_0 推定の前処理として有効に機能する。この処理は STEP2 で F_0 推定を $\mathcal{O}(\hat{R}_1)Y$ に対して適用するという形で実施される。この結果、STEP2 では、STEP1 よりもさらに正確に F_0 推定ができるようになる。

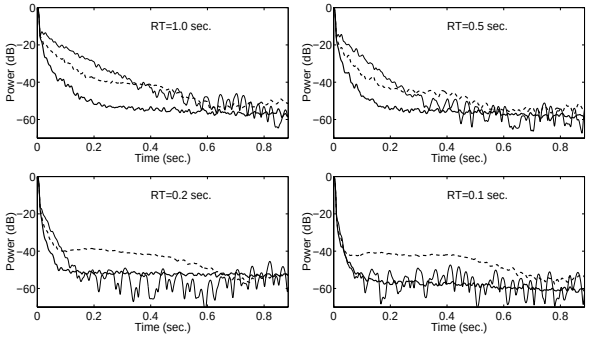

図 2: 残響時間 (RT) の異なるインパルス応答 (細実線)、および残響除去後の各インパルス応答 (男性：点線、女性：太実線) の残響曲線

3 残響曲線による評価

提案法の性能を残響曲線 [3] で評価した。ATR 単語 DB から男女一名（MAU と FKM、12kHz 標本化）の各 5240 単語を音源信号 X とし、可変残響室で測定した 4 種類のインパルス応答（残響時間が 0.1, 0.2, 0.5, 1.0 秒）を畳み込んで観測信号 Y を合成した。

図 2 は、残響除去処理を行う前後の残響曲線を示している。女性話者の場合、残響時間 (RT) が 0.1 秒より長い時には効果的に残響を抑制できている。しかし、RT$=0.1$ の時は残響の抑制効果は認められず、この実験で除去できる残響時間の最小限界を示していると思われる。一方、男性話者の場合も RT>0.1 の時はインパルス直後の残響は効果的に抑制できている。これは目的音の明瞭度が改善することを意味する [3]。例えば、RT$=0.2$ の時、Time>0.1 の残響は増えてはいるが、よりパワーの強い Time<0.1 の残響が抑制されているので、全体として明瞭度は向上する。

4 まとめ

本稿では、単一チャンネル音声のブラインド残響除去のために調波構造に基づく方法を提案した。残響を含んだ音声から抽出される調波成分を直接音の近似とみなし、そこから得られる逆伝達関数の平均値を計算することで、残響除去オペレータが求められることを示した。さらに提案法にとって重要な残響に頑健な F_0 推定法を提案した。残響時間が 0.1 秒より長い音声の残響が効果的に抑制されることを、ATR 単語 DB(5240 語) を用いて示した。提案法で処理した音声サンプルはインターネット上で試聴していただくことができる [4]。今後の課題は、男性話者に対する残響除去性能の向上、残響除去オペレータの計算に必要なデータサイズの縮小、適応処理への適用などがあげられる。

日頃から有益な議論をいただいている NTT コミュニケーション科学基礎研究所の方々、可変残響室のインパルス応答測定および継続音抑制フィルタの性能評価をしていただいた京都大学吉田尚史氏に感謝する。

参考文献

[1] 馬場他, 音講論, pp. 27–28, 秋田, 9 月, 2002.
[2] 中谷他, *ICASSP-2003* to appear, Apr., 2003.
[3] Yegnanarayana, B. et al., *JASA*, vol. 58, pp. 853–857, Oct. 1975.
[4] http://www.kecl.ntt.co.jp/icl/signal/nakatani/sound-demos/dm/derev-demos.html