瞬時振幅を利用したブラインド残響音声回復の可能性の検討

柴野 洋平 (510045)

北陸先端科学技術大学院大学 情報科学研究科

2007 年 2 月 8 日

キーワード: 残響音声回復，キャリア，瞬時振幅，瞬時位相.

実環境の雑音や残響は観測される音声信号に歪みを与える．特に残響環境下では，壁からの反射音が直接音に重なり歪みを与え，音声の明瞭度を著しく低下させる．そのため音声認識システムや，拡声音声通話，補聴システムなどでは性能が低下する．これらの影響を取り除き，残響音声を回復させることが求められる．残響の影響は室内の構造や室温の変化とともに変動する可能性があるため，逐次，室内インパルス応答を測定することなく，適応的に残響音声を回復する必要がある．先行研究では残響音声を帯域分割し，帯域内で伝達系の測定を必要としない MTF に基づくパワーエンベロープ逆フィルタ処理を用いて音声のパワーエンベロープを回復し，残響環境下から推定された基本周波数 (F0) と有声無声区間の情報と群遅延を用いることで，自然性の高い音声のキャリアを生成している．この手法は，残響環境からブラインドで音声を回復できる．しかし，残響環境下では F0 と有声無声区間の情報の推定精度の低下が懸念されるため，これらの手法では十分な回復効果を得られない可能性がある．

本研究の目的は，F0 を用いることなく，残響音声回復の可能性を示すことである．改良モデルでは残響音声を帯域分割し，先行研究で用いられたパワーエンベロープ逆フィルタ処理を用いて瞬時振幅を回復する．回復した瞬時振幅から位相を推定し，帯域内の信号を回復し，最終的にそれらを足し合わせ音声を回復する．

本研究では，まず瞬時振幅から残響の影響が取り除かれたものとして，元の信号の瞬時振幅から瞬時位相を推定し，元の信号を合成する方法について検討する．フィルタバンクによって帯域分割した信号 $x_n(t)$ を AM-PM 信号として考える．

$$x_n(t) = s_{x,n}(t)\cos(\omega_{c,n}t + p_{x,n}(t)), \tag{1}$$

$s_{x,n}(t)$ は瞬時振幅，$p_{x,n}(t)$ は瞬時位相である．$\omega_{c,n}$ はフィルタバンクの中心周波数，n はチャネル番号である．ここで，瞬時振幅とフィルタバンクの中心周波数は既知であるから，残る未知の項である元の信号の瞬時位相が分かれば，元の信号を復元することができる．しかし，1 つの式に対して未知の項が 2 つ存在しており，このままでは式を解くことができない．また，既知の項である瞬時振幅や中心周波数と未知の項である瞬時位相は一般に

独立で，既知の項から未知の項を推定することも難しい．本研究では，瞬時振幅と瞬時位相を関係付けるための条件を示し，元の瞬時振幅から信号を合成する方法について検討する．

周波数領域では最小位相信号であれば振幅スペクトルから位相スペクトルを求めることができる．周波数領域の実部と虚部に Hilbert 変換の関係があり，実部が正になっている信号は最小位相信号である．時間領域と周波数領域のアナロジーから，時間領域でも瞬時振幅から瞬時位相を推定できる．帯域分割した信号 $x_n(t)$ とその Hilbert 変換から，解析信号 $x_{a,n}(t)$ を下式のように定義する．

$$\begin{aligned} x_{a,n}(t) &= x_n(t) + j\mathrm{Hilbert}[x_n(t)] \\ &= s_{x,n}(t)\cos(\omega_{c,n}t + p_{x,n}(t)) + js_{x,n}(t)\sin(\omega_{c,n}tp_{x,n}(t)) \\ &= s_{x,n}(t)\exp(jp_{x,n}(t))\exp(j\omega_{c,n}t) \end{aligned} \tag{2}$$

今，簡略のため $s_{x,n}(t)\exp(jp_{x,n}(t))$ を中心に検討する．この実部と虚部は Hilbert 変換の関係にある．また，$x_{a,n}(t)$ はフーリエ変換すると片側が 0 となる．この対数をとると

$$\log[x_{a,n}(t)] = \log[s_{x,n}(t)] + j(p_{x,n}(t)). \tag{3}$$

ここまでの式の展開は，周波数領域と時間領域が入れ替わっただけで同じである．本研究の着想点は，もし対数瞬時振幅と瞬時位相が Hilbet 変換の関係にある場合，瞬時振幅から信号を復元できるということである．その場合は，下式により対数瞬時振幅の Hilbert 変換から瞬時位相を求めることができる．

$$\hat{p}_{x,n}(t) = \mathrm{Hilbert}(\log[s_{x,n}(t)]). \tag{4}$$

一般に帯域分割した信号は正の値のみもつことはない．このままでは，瞬時振幅から瞬時位相を推定できず，元の信号を回復することはできない．そこで，信号を正の部分と負の部分に分割して，負の部分についてはマイナスをかけることにする．正の部分と負の部分に分割した信号からは，半波整流と同様にエンベロープを求めることができる．それぞれのエンベロープの残響の影響を取り除き，回復瞬時振幅を求め，正の部分と負の部分信号を回復する．最終的に，正の部分とマイナスをかけた負の部分は足し合わされ，帯域内の信号は回復される．ATR 音声データベースの 10 話者（MAU，MHT，MNM，MTM，MTT，FAF，FFS，FKN，FSU，FYN），10 単語（相変わらず，季節，新聞，冗談，中間，滑らか，施す，間に合う，楽観，わがまま），合計 100 単語を対象にシミュレーションを行った．元の音声と回復音声を LSD と SNR を用いて比較した．結果は，平均 LSD 0.1dB，平均 SNR 24.0 dB と音声を回復できていることが分かった．すなわち，元の帯域分割した信号の瞬時振幅が分かれば，瞬時位相を回復でき，帯域内の元の信号を回復し，音声を回復できることが分かった．

本研究ではまず難しい問題である瞬時振幅から位相を推定し，信号を復元する方法について検討した．この問題は 1 つの式に未知の項が 2 つあり，また既知の項である瞬時振幅

と未知の項である瞬時振幅は一般に独立であり非常に難しい．しかし，本論文ではこれを，元の信号が正であるという条件下であれば，対数瞬時振幅の Hilbert 変換から瞬時位相を推定し，信号を完全に復元できるということを示した．この方法を音声に適用するために，音声を帯域分割した信号を正の部分と負の部分に分割し，それぞれの瞬時振幅から帯域内の信号を復元させ，音声を復元するシミュレーションを行った．シミュレーションは ATR 音声データベースの 100 単語を対象に行われた．その結果，復元音声は平均 LSD 0.1 dB，平均 SNR 24.0 dB であり，瞬時振幅から音声を復元することができた．これにより，残響音声を帯域分割した信号の瞬時振幅が回復した場合，残響音声を回復することができる．瞬時振幅の回復については，戸井らの方法をそのまま用いて残響音声を帯域分割した信号の瞬時振幅を回復することはできないと考えられるが，今後はパワーエンベロープ逆フィルタ処理を低域のみに対応させたり，瞬時振幅の周波数成分が低域のみ存在するように残響音声の帯域幅を決めることで，瞬時振幅は回復する可能性がある．瞬時振幅が回復すれば，残響音声は回復することができる．したがって，F0 を用いることなく，残響音声が回復する可能性がある．今後は，残響信号の瞬時振幅を回復させ．回復した瞬時振幅から信号を復元するシミュレーションを行う．これにより，本手法を総合的に評価することができる．音声認識システムや，拡声系音声通話，補聴システムでの性能の向上が期待できる．