音声定常部抑圧処理のリアルタイム化のためのアルゴリズムの検討
～残響環境下における音声明瞭度改善に向けて～

山辺 祐史[†]　　荒井 隆行[†]　　安 啓一[†]　　栗栖 清浩[‡]

[†]上智大学大学院　理工学研究科　理工学専攻　情報学領域
[‡]TOA 株式会社

Abstract

コンサートホールや講堂, 地下鉄のホームなどの残響環境下において, 音声の明瞭性が低下し, 語音が聞き取りづらくなることがある. その原因に一つとして, 先行する音素による残響の尾が, 後続の音素に影響を与える overlap-masking が挙げられている[1, 2]. この問題に対し, 荒井ら[3, 4] は, overlap-masking を軽減するための前処理として, 音声の定常部を抑圧し音声明瞭度の低下を防ぐ処理（定常部抑圧処理）を提案した. また, 定常部抑圧処理が特定の残響環境下で音声明瞭度を改善することも報告されている[5].

この一連の研究は, 定常部抑圧処理を DSP 上でリアルタイム動作させ, 公共空間における拡声システム（PA システム）に導入することを最終目標としている. 本論文では, リアルタイムで定常部抑圧処理を実現するため, 音声の定常部における母音性に着目した特徴量を用いる簡易的な手法を提案する. 特徴量の計算方法は以下の通りである. (A) 線形予測法によって, スペクトル包絡を 20 ms のフレーム毎に計算する. (B) 50～3000 Hz 間に存在するフォルマントのピークのうち, ピーク値が高い上位 2 つを検出する. (C) 検出されたピークのピーク値の和を取り, 特徴量とする. この手法により, 日本語単音節において母音部を検出できることが確認された.

1. はじめに

1.1. 残響と定常部抑圧処理について

残響とは, 音源から発生した音が壁や天井に反射しながら減衰することで, 音の響きが残る現象である. 残響は音楽に余韻を加えるので, コンサートホールなどでは適度な残響が求められる. 一方, 演説を聞く際は, 長い残響によって音声の明瞭性が下がってしまう. その原因の一つとして, overlap-masking が挙げられている[1, 2]. 図 1 は音声(ワタクシ)が overlap-masking の影響を受けている例である. 音声の出力レベルによってマスキング量は変化するが, 先行する音素から, 後続の音素へのマスキングが発生していることがわかる. また, エネルギーの大きい音素から小さい音素へのマスキングが起きている様子も観察できる. この現象によって残響環境下では音声明瞭度が低下する.

荒井ら[3,4]は, 残響環境下での聞き取り改善のための信号処理として, 定常部抑圧処理を提案している.

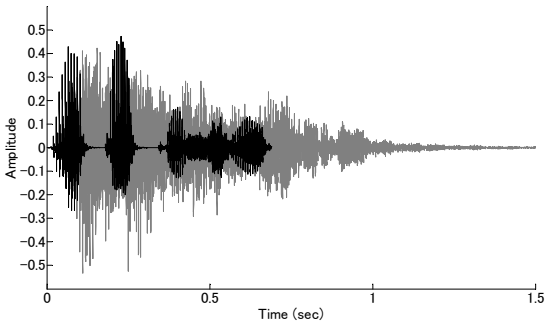

図 1　原音声(濃色)と残響のかかった音声(淡色)

一般に, 先行する音素のエネルギーが大きいほど, 後続の音素へのマスキング量は増加する. そのため, この処理では音声の中で比較的エネルギーが高いが, 人間の音声知覚にはあまり重要でない[6]定常部（主に母音部）の振幅を抑圧することで overlap-masking の影響を軽減し, 音声明瞭度の改善を図る. またこの処理は前処理であり, スピーカから音が放射される前に入力音声に対して信号処理を施し, 残響にロバストな音声として拡声することを目的としている.

1.2. フィルタバンクによる定常部抑圧処理(FB 法)

図 2 に定常部抑圧処理[3, 4]のブロックダイアグラムを示す. 最初に入力信号は臨界帯域を模擬した 1/3 オクターブごとのバンドパスフィルタ群に通される. 帯域毎に振幅の時間包絡を抽出し, ダウンサンプリングした後振幅包絡の対数をとる. 10 ms 毎に前後 2 点, 計 5 点にわたって回帰係数をとり, Furui [7]にならいスペクトル遷移を表す D を計算する. D が閾値以下であれば定常部とみなし, 元の音声信号の振幅を設定した抑圧率で抑圧する. 本処理における抑圧率とは, 定常部の振幅を元の振幅の何%にするかという変数であり, 図 2 の weighting function の 0.4 に相当する. なお, 抑圧率 40%において聞こえの改善が得られている[3, 4]. また, weighting function の値が変化する際に直線傾斜を持たせることで, 抑圧の開始・終了の際に急激な振幅の変化が起きることを防いでいる. フィルタバンクによる定常部抑圧処理について, 多くの聴取実験が行われている. Hodoshima *et al.* [8]は, 上智大学 10 号館講堂(822 座席, 残響時間約 1 秒)において若年健聴者 24 名に対し日本語単音節による聴取実験を行い, 定常部抑圧処理によって残響環境下での聞こえが有意に改善することを示した(図 3). また, Miyauchi *et al.* [8]は, 上智大学 10 号館講堂のインパルス応答を用いて残響環境下での聴取実験を行い, 高齢難聴者に

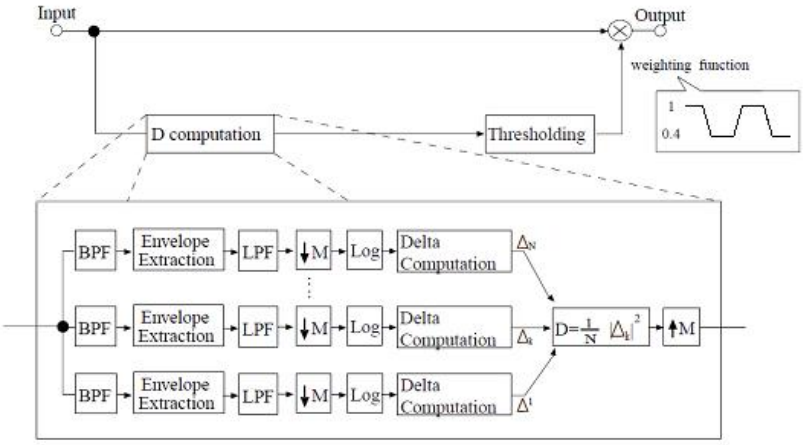

図 2　定常部抑圧処理のブロックダイアグラム([5]より改変)

おいても残響環境下での明瞭度が改善することを確認した（図 3）．これらの結果から，定常部抑圧処理により, 公共空間において年齢を問わず聞こえを改善することが期待できる．

1.3. 定常部抑圧処理のリアルタイム化

次のステップとして, 公共空間の音響システムに定常部抑圧処理を組み込むため, DSP によるリアルタイム化の取り組みが行われてきた．荒井ら[3, 4]のフィルタバンクによる手法は, 録音済みの音声全体を読み込んでから処理を行うため, リアルタイム処理には不向きである．そのため後藤ら[10]は，FFT ケプストラムの変化により定常部を検出する手法(FFT 法)を開発し，DSP への実装を行った．しかし，サンプリング周波数が 8 kHz と低く, 音声の情報を劣化させていたため，明瞭度の向上が得られなかった[11]．また，入出力を含めた遅延が 120 ms と大きく，実環境で使用した際, 音と講演者の口がずれて違和感が起こることも考えられる上，発話者にとっても，PA システムを通した音声が遅れて聞こえることで, 発話しづらくなることが考えられる．そこで Takahashi *et al.* [12]は，連続する 2 フレーム間のエネルギー比による簡易的手法を提案し，48 kHz のサンプリング周波数で 30 ms の低遅延を実現した．一方，音声定常部におけるエネルギーの時間変化の少なさに着目した処理を行っているため, 話速の変化によって定常部を正しく判定できなくなる問題も考えられた．

これらの経緯を踏まえ, 本研究では音声の母音性に着目した特徴量に基づく手法を提案する．

2.　提案法

2.1. 特徴量の設定

音声において, エネルギーが大きい定常部が主に母音部に現れる事に着目し, 音声の母音性を示す特徴量 α' を提案する. 母音性の強い部分の振幅を抑圧することで, 定常部抑圧処理を簡易的に再現することを目標としている．α' は，録画した動画への字幕付与を補助するための, 自動音声区間検出のために提案されたパラメータ α [13]がベースとなっている．α' は，α に対して①リアルタイム処理に適したアルゴリズム, ②より高い精度で母音を検出すること, の 2 点に注目した改良を行ったものである．

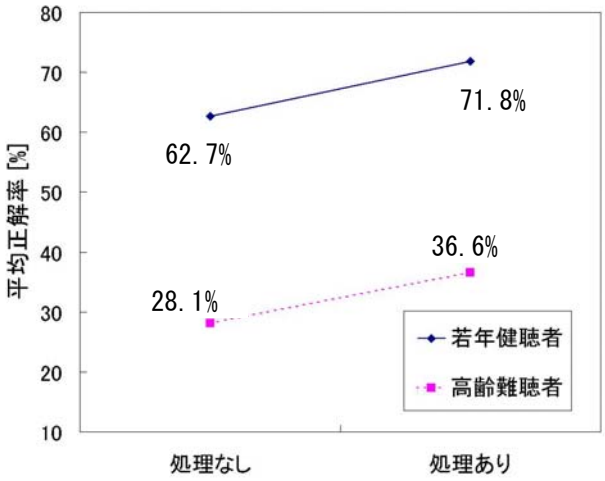

図 3　定常部抑圧処理による聞こえの改善([7, 8]より改変)

2.2. アルゴリズム

図 4 に，線形予測法(LPC)による定常部抑圧処理のブロックダイアグラムを示す. 処理の流れは以下の通りである．

1. サンプリング周波数 16 kHz で標本化された入力信号に，前後サンプルでの差分を取ることで +6dB/octave の高域強調を施す.
2. フレーム長 20 ms でフレーム分けし，ハニング窓による窓かけを行う．オーバーラップは行わない.
3. フレーム毎に RMS を計算する.
4. 各フレームにおいて 15 次の LPC の算出と 512 点の高速フーリエ変換を行い, LPC スペクトル包絡を得る.
5. スペクトル包絡の 50Hz～3kHz 間で，1 階微分を取り，1 階微分が正から負に変わる点をピークとして検出する.

提案法の α' の計算手法は, 字幕付与向け自動音声区間検出アルゴリズム[13]と以下の点が異なる．

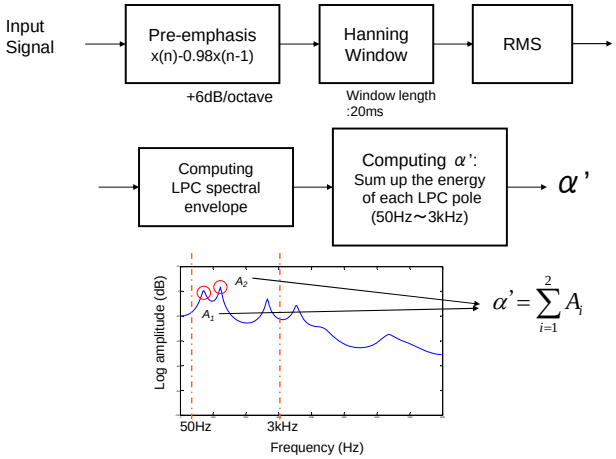

図 4　提案法のブロックダイアグラム

1. 得られたピークのうち，全てのピークではなく，第 1・第 2 ピークの対数振幅を足し合わせ，特徴量α'とする(/s/等の摩擦音で現れる多量の小さなピークを検出する事を避けるため)
2. LPC の多項式の根を直接計算せず，スペクトルの 1 階微分から計算する（DSP での処理時間を考慮）
3. 摩擦音等で，50 Hz～3 kHz 間にピークが存在しない場合，α'=0 とするのではなく 100 Hz 付近における対数振幅を代入する(中央値の計算を正確にするため)
4. その他パラメータの変更
   (ア) サンプリング周波数を 12 kHz から 16 kHz に変更
   (イ) LPC と FFT の次数の変更(LPC:12→15 次，FFT:256→512 点)

提案法では，母音が声道の共振により生成されていることを利用して母音部を判別する．上記の周波数分析により，母音における声道の共振（フォルマント）の存在を検出しているため，母音部のフレームでは子音部に比べてα'の値が大きくなる．α'の値に対して閾値処理を行い，α'の値が閾値以上となるフレームは母音部であると判定し，元の振幅の 40%として出力する抑圧処理を行った．なお，閾値は，単音節の発話区間全体の中央値とした．発話区間の推定には，RMS による閾値処理を利用した．この閾値の設定法は，音声全体が読み込まれている前提のためリアルタイム処理に向いていないが，本論文では提案法の有効性を検証するため，理想的な閾値の設定法を用いた．また，先行研究同様，抑圧の開始・終了の際に急激な振幅の変化を防ぐため 10 ms の直線傾斜をつけた．

提案法では，アルゴリズム上の遅延は 1 フレーム分(20 ms)となる．そのため FFT に基づく手法[10]に比べ，リアルタイム動作させた際に，聴取者に対しては口の動きと音声とのずれが少なくなること，また発話者に対しては発話と聴覚へのフィードバックのずれが低減されることから，実環境での使用に適していると考えられる．本論文では上記のアルゴリズムを MATLAB で実装し，処理の妥当性を検討する．

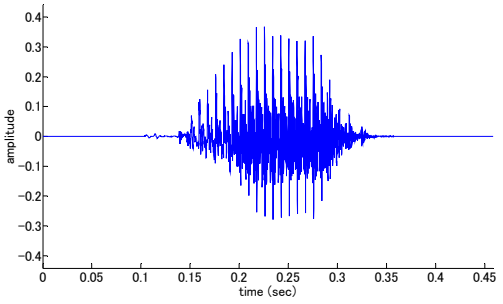

図 5　原音声の時間波形

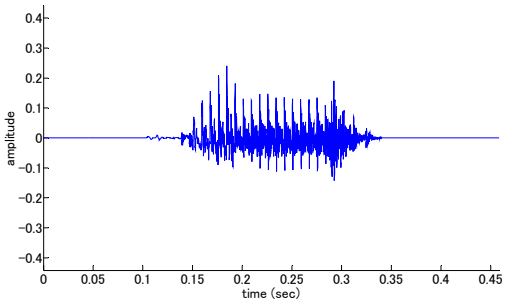

図 6　提案法の時間波形

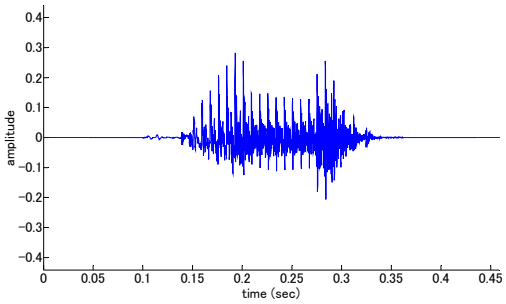

図 7　FB 法の時間波形

3.　処理の検討・考察

3.1. 母音の検出

音声「が」の原音声を図 5，提案法を施した時間波形を図 6 に示す．提案法がエネルギーの大きい母音部を検出し，抑圧している様子が確認できる．また，先行研究で用いられた代表的な 14 種類の子音（/p/, /t/, /k/, /b/, /d/, /g/, /s/, /h/, /ʃ/, /tʃ/, /dʒ/, /dz/, /m/, /n/)と母音「あ」からなる日本語単音節に関しても，同様に提案法による母音部の検出と抑圧が確認できた．これにより，特

表 1　FB 法と提案法の一致率

単音節	p+a	d+a	ʃ+a	tʃ+a	dʒ+a	m+a	平均
一致率	72.7%	69.1%	92.0%	83.1%	77.9%	77.3%	78.7%

徴量α'に，子音部と母音部の特徴が反映されていることがわかった．しかし/ʃ/において，子音の一部が 1 フレーム分だけ母音と判定され抑圧された．これは，そのフレームにおいて/ʃ/のスペクトルが 1437 Hz 付近にピークを 1 つ持っていた事による. 一般に母音の第一フォルマントは 1 kHz 以上にならないので, 一つ目のピークが 1 kHz 以上の場合は, その点をα'の計算に含めないことでこのような問題を回避することが出来ると考えられる．

3.2. フィルタバンクによる手法との比較

フィルタバンクによる手法(FB 法) [3, 4]は，残響環境下での聴取実験により, 明瞭度を改善する事がわかっている．そのため，FB 法と同様の抑圧結果が得られれば，提案法でも明瞭度の改善が期待できる．図 7 は音声「が」に FB 法を施した時間波形である．提案法（図 6）と比較すると，提案法の方が母音のより長い範囲を抑圧している事がわかる．これは，FB 法がスペクトル遷移が定常な部分を抑圧しているのに対し, 提案法では母音性の共振がある部分を抑圧しているためと考えられる．そこで，FB 法の抑圧箇所を基準とし, 提案法の抑圧箇所との一致率を求め表 1 に示した．比較に用いる子音は Takahashi *et al.* [12]と同様とし，表 1 に示したものを使用した．平均の一致率は 78.67%であり，FFT 法(72.17%) [13]，エネルギー比による手法(80.20%) [13]とほぼ同等であった．提案法では, FFT 法やエネルギー比による手法に比べ子音部を抑圧する事は少ないが，FB 法に比べ母音部の長い範囲を抑圧する傾向にあったため，8 割程度の一致率になったと考えられる. 今後は単音節だけでなく単語などの音節数の多い音声でも比較を行い, 処理の検討を行っていきたい.

3.3. リアルタイム化に向けての課題

前述の通り, 提案法によって母音部と子音部を判定し，簡易的に定常部抑圧処理を実現する事ができた. しかし, 今回は単音節全体のα'の中央値を閾値としたため, 閾値の設定方法がリアルタイム動作に不向きである．今後は①日本語の平均的なモーラ長である 160 ms 程度の時間毎にα'の中央値を計算し閾値を更新していく, ②α'の時間変化が少ない定常的な部分を検出し，かつα'が閾値を越えている部分を抑圧する，などのリアルタイム処理が可能な方法を検討する必要がある．

4.　まとめ

本論文では, 線形予測法を用いたリアルタイム処理向け定常部抑圧処理の簡易アルゴリズムを提案した. また，提案法を MATLAB 上で実装し，日本語単音節での母音部検出を確認した. 今後は閾値設定方法の改善や単音節以外でのパラメータ調整などを行っていきたい．また，先行研究同様 Texas Instruments 社製 TMS320C6713 DSK への実装も進めていきたい.

謝辞

本研究は, 文部科学省私立大学学術研究高度化推進事業上智大学オープン・リサーチ・センター「人間情報科学研究プロジェクト」の支援を受けて行われた.

参考文献

[1] R. H. Bolt and A. D. MacDonald, “Theory of Speech masking by reverberation,” *J. Acoust. Soc. Am.*, **21**(6), 577-580, (1949).

[2] A. K. Nabelek, T. R. Letowski, and F. M. Tucker, “Reverberant overlap- and self-masking in consonant identification,” *J. Acoust. Soc. Am.* **86**(4), 1259-1265, (1989).

[3] 荒井隆行, 木下慶介, 程島奈緒, 楠本亜希子, “音声の定常部抑圧の残響に対する効果,” 日本音響学会講演論文集, 449-450 (2001.9).

[4] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, “Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments,” *Acoust. Sci. & Tech.*, **23**, 229-232 (2002).

[5] N. Hodoshima, T. Arai, A. Kusumoto and K. Kinoshita, “Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments,” *J. Acoust. Soc. Am.*, **119**(6), 4055-4064, (2006).

[6] W. Strange, J. Jenkins, and T. Johnson, “Dynamic specification of coarticulated vowels,” *J. Acoust. Soc. Am.*, **74**(3), 695-705, (1983).

[7] S. Furui, “On the role of spectral transition for speech perception,” *J. Acoust. Soc. Am.*, **80**, 1016-1025 (1986).

[8] N. Hodoshima, T. Goto, N. Ohata, T. Inoue and T. Arai, “The effect of pre-processing for improving speech intelligibility in the Sophia University lecture hall,” *Proc. of International Congress on Acoustics*, **3**, 2389-2392, (2004).

[9] Y. Miyauchi, N. Hodoshima, K. yasu, N.Hayashi, T. Arai and M. Shindo, “A preprocessing technique for improving speech intelligibility in reverberant environments: The effect of steady-state suppression on elderly people,” *Proc. Interspeech*, 2769-2772, (2005).

[10] 後藤崇公，荒井隆行，安啓一，“定常部抑圧処理のリアルタイム化に向けて DSP による開発,” 第 7 回 DSPS 教育者会議予稿集, 91-94, (2005)

[11] K. Takahashi, T. Goto, F. Tadokoro, K. Yasu, T. Arai, “Implementation of steady-state suppression using a digital signal processor for real-time processing in an actual hall,” *IEICE Technical Report*, **105**(685), 25-30,　(2006)

[12] K. Takahashi, K. Yasu, N. Hodoshima, T. Arai and K. Kurisu, “Enhancing speech in reverberation by steady-state suppression,” *Proc. Int. Congr. Acoustics*, (2007).

[13] 藤樫佑樹，古賀綾子，荒井隆行，金寺登，吉井順子, “字幕付与システムを目的とした線形予測に基づく音声端点検出,” 日本音響学会講演論文集，33-34 (2005.9).

[14] 高橋慶, “DSP への実装を目的とした簡易式定常部判定法の検証,” 上智大学修士論文, (2008).