社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2002-65(2002-07)

定常部抑圧による音声明瞭度改善のための前処理
―異なる残響環境下について―

程島 奈緒† 井上 豪† 荒井隆行† 木下慶介† 楠本亜希子‡

† 上智大学理工学部電気電子工学科 〒102-8554 東京都千代田区紀尾井町 7-1

‡ Department of Veterans Affairs, Portland VA Medical Center, OR 97207, USA

E-mail: † n-hodosh@splab.ee.sophia.ac.jp

あらまし 残響環境下では，音声が聞き取りづらくなる場合がある．我々は以前の研究 (荒井他, 2001; Arai et al., 2002)で，直前の残響成分が現在の音声信号をマスクするために音声明瞭度が減少すると考え，実験によりそれを確認した．そこではマスキングの影響を抑えるためにエネルギは比較的大きいが，音声知覚にはそれほど重要ではないとされる定常部を抑圧する処理を提案した．処理の効果は残響時間によって異なることから，本研究では定常部抑圧処理と残響時間の関連を調べるために人工的に作成した数種類の残響を用いて聴取実験を行った．結果からいくつかの残響条件で処理による改善が見られ，荒井らの手法は，残響環境下において音声明瞭度を改善するための前処理として有効であると確認することができた．

キーワード 音声強調，残響，音声明瞭度，マスキング，定常部抑圧

Suppressing steady-state portions of speech for improving intelligibility as pre-processing
-Under various reverberant environments-

Nao HODOSHIMA† Tsuyoshi INOUE† Takayuki ARAI† Keisuke KINOSHITA† and Akiko KUSUMOTO‡

† Department of Electrical and Electronics Eng., Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan

‡ Department of Veterans Affairs, Portland VA Medical Center, OR 97207,USA

E-mail: † n-hodosh@splab.ee.sophia.ac.jp

Abstract In previous studies (Arai et al., 2001; Arai et al., 2002), we hypothesized that segments of an acoustic signal are masked by reverberation components of previous segments, degrading speech intelligibility. To reduce masking influences, we suppressed steady-state portions having more energy, but which are less crucial for speech perception. We have presently conducted a perceptual test with a set of artificial reverberations to explore the relationship between steady-state suppression and several reverberation conditions. The results indicated clear improvements for some reverberation conditions. We certified that Arai's technique was an effective pre-processing method for improving speech intelligibility under reverberant conditions.

Keyword speech enhancement, reverberation, speech intelligibility, masking , steady-state suppression

1. はじめに

大きなホールにおいては，しばしば講演の内容を理解することが難しい場合がある．これは様々な位相遅れと振幅をもった反射音が重なり合ってできる残響の影響である．残響の尾は後続の音声信号に影響を与えるため，直前の残響成分が現在の音声信号をマスクし，音声の明瞭度が減少すると考えられる[1][2]．よってエネルギの大きい定常部のような音声区間に後続する音素は，その音素に重畳される前の音声区間に対する残響成分のエネルギ量が多くなるため，マスキングの影響をより多く受けると考えられる．

単音節明瞭度の実験の結果から，音声の遷移部は音声知覚に関して非常に重要な役割を果たしているとされている[3]．また，自動音声認識において音声の定常部は遷移部と比較するとそれほど重要ではないことが分かっている[4]．デルタ処理とよばれるケプストラムに対する動的特徴量[3]や RelAtive SpecTrAl (RASTA) 処理[4]は音声の遷移部を強調した特徴量であり，自動音声認識の分野において認識率の増加に貢献している．

残響環境下において音声の明瞭度を改善する方法は，主に pre-processing と post-processing の 2 種類のアプローチに分けることができる．pre-processing は音声

に残響が付加される前，すなわちマイクロフォンとスピーカの間で音声処理を行なう(例えば[1][2][5][6][10]). post-processing では音声が室内に放射され，残響が付加された後で dereverberation を行なう(例えば[6]-[9]). 荒井らは，pre-processing としてエネルギは比較的大きいが音声認識にはあまり重要ではないとされる音声の定常部を抑圧する処理を行ない，残響によるマスキングの影響を軽減させている[1][2]. 一方，音声の変調スペクトルを変化させる変調フィルタリングも，音声の明瞭度を改善する手法として用いられている(例えば[5]-[7]). 楠本らは，pre-processing として音声知覚に重要とされる変調周波数帯 4Hz を強調する処理を行なった[5]. 荒井ら[1][2]，楠本ら[5]共に音声明瞭度の改善を示唆する結果を得ている.

Langhans らは pre-processing, post-processing の両方に，残響によって減少した音声の変調指数を人工的に増加させる理論的な IMTF (Inverse Modulation Transfer Function)フィルタを適用した[6]. Avendano らは同じく post-processing として，音声の変調指数を人工的に増加させているが，理論的な IMTF フィルタを求める代わりに彼らのトレーニングデータから計算された IMTF フィルタを用いた[7].

本研究の最終目的は，任意のホールに対して音声明瞭度の低下を抑える最適な pre-processing を提供することである. そのためには，残響時間と処理の効果の関連を知る必要がある. 我々はこの関係を調べるため，残響時間を変化させた数種類のインパルス応答と，[5]の変調フィルタ処理を用いて聴取実験を以前行なっている[10]. 実験結果から変調フィルタリングの効果は残響時間によって異なり，特定の条件下では明瞭度の低下を抑えることができると確認された. [10]では変調フィルタ処理に関して実験を行っているのに対し，本論文では定常部抑圧処理を様々な残響条件において調べることを目的とした. そこで，残響時間を変化させた数種類のインパルス応答と[1][2]で用いた定常部抑圧処理を用いて聴取実験を行った.

2. 聴取実験

2.1. 残響時間

我々は様々な残響条件を実現するため，[10]にならい数種類のインパルス応答を人工的に作成した. まず理想的な室内において測定されるインパルス応答 h_o は，式(2.1)のように時定数 τ_o をもつ包絡成分 $\exp(-t/\tau_o)$ と，定常的なノイズであるキャリア成分 $w(t)$ との掛算で近似する事が出来る.

$$h_o(t) = e^{-t/\tau_o} w(t) \qquad (2.1)$$

式(2.1)より、τ_o の値を変えることで，任意の残響時間を持つインパルス応答を作成することができる. 式(2.1)から，新たに作成されるインパルス応答 h_n は包絡の時定数を τ_n とすると式(2.2)のように表され，さらに h_o を用いて表すと式(2.3)のように書きかえることができる.

$$h_n(t) = e^{-t/\tau_n} w(t) \qquad (2.2)$$

$$h_n(t) = e^{-t/\tau} h_o(t) \quad \left(1/\tau = 1/\tau_n - 1/\tau_o\right) \qquad (2.3)$$

式(2.3)から，h_o の残響時間を基準に，残響時間を 0.1 秒ずつ変化させるような包絡を h_o にかけ，計 4 種類のインパルス応答を作成した. 表 1 に，本論文で使用したインパルス応答の残響時間を示す(h_o は表 1 で h3 に対応する). ここで，本論文で用いた h_o は東大和市大ホール（反射板なし）で測定されたインパルス応答である. 本論文において，残響時間はインパルス応答の減衰曲線が定常状態から 60dB 減少する時間である T_{60} を用いた. ただし，減衰曲線は Schroeder が提案した方法に基づいて計算した[11]. その際，dB 値に変換した減衰曲線 $y(t)$ を式(2.4)に示す.

$$y(t) = 10\log_{10} N\int_t^{\infty} h_n^{\,2}(x)dx \qquad (2.4)$$

ここで N は h_n の単位周波数あたりのパワーである.

表 1 本論文で使用した残響条件.

インパルス応答	h1	h2	h3	h4	h5
残響時間(s)	0.9	1.0	1.1	1.2	1.3

2.2. 定常部抑圧処理

定常部抑圧処理は，[1][2]と同じ処理を用いた. まず，音声信号を 1/3-oct に帯域分割し，各帯域において包絡を抽出した. 次にダウンサンプリング(M=160)された包絡成分の対数をとり，前後計 5 点の回帰係数を計算し，帯域に渡って回帰係数の 2 乗平均 (以下では D と

表 2 実験で使用した CV(子音-母音)の分類.

	Stop+V	Fricative+V	Affricate+V	Nasal+V
Voiceless Consonant+V	/pa/ /ta/ /ka/ /pi/ /ki/	/sa/ /ʃa/ /ha/ /ʃi/ /hi/	/tʃa/ /tʃi/	
Voiced Consonant+V	/ba//da/ /ga/ /bi/ /gi/		/dza/ /dʒa/ /dʒi/	/ma/ /na/ /mi/ /ni/

する)を求めた．ここで，D は Furui にならい音声のスペクトル遷移を表すパラメータを表す [3]．元の標本化周波数に戻した後，D が一定の閾値より小さい箇所を定常部とし，定常部とみなした箇所は元の波形の振幅を 40%に抑圧した．

2.3. 刺激

刺激は，日本語の単音節 CV (子音–母音)をターゲットとし，日本語のキャリアセンテンス「題目としては__といいます」に挿入した．V として/a/, /i/を，C として/p/, /t/, /k/, /b/, /d/, /g/, /s/, /ʃ/, /h/, /tʃ/, /dz/, /dʒ/, /m/, /n/の 14 種類を用いた．実験で使用した 24 種類の CV を表 2 に示す．各刺激は ATR 研究用日本語音声データベース (話者：MAU, 40 才男性)を用いた．ターゲットは単音節データを使用したのに対し，キャリアセンテンスは文章データの中から 2 文を選択し，その一部ずつを使用した（前半部と後半部で異なる文を使用した)．全ての音声サンプルで，直前の音節からのターゲットに対するマスキング量の統制をとるため，ターゲットの母音の開始位置はその直前のキャリアセンテンスの終了部から 150ms に統一した．また，ターゲットの実効値は同じ母音の CV セットごとに正規化した．

刺激音はオリジナルの音声信号に残響を畳み込んだ刺激セット(org_ rev)と，定常部抑圧処理を行った後に残響を畳み込んだ刺激セット(proc_rev)の 2 種類を用意した．

2.4. 被験者

被験者は日本語を話す健聴者 24 人(男性 14 人，女性 10 人，年齢 18-26 才)であった．

2.5. 手順

実験の指示は防音室内のコンピュータの画面上で行なった．刺激音の提示はヘッドフォン(STAX SR-303)を用い，被験者ごとに適した音圧レベルに調整した．実験の形式に慣れさせるため，本番前に被験者に対して練習を行った．実験の手順は，各試行においてまず刺激音を一度だけ提示し，提示終了後画面上に実験で使用した 24 種類の CV を選択肢としてかなで表示した．被験者には，画面上の選択肢を強制的に一つマウスでクリックさせ回答させた．選択が終わると，次の刺激が自動的に提示された．各被験者に対して，計 240 刺激(残響 5 種類×24 単音節×処理 2 種類)をランダムに並べて提示した．

3. 実験結果

各残響条件，処理条件における子音の正解率の平均値を表 3，図 1 に示す．ただし母音の正解率は，いずれの条件においても 100%であった．以下では外れ値である(正解率の平均値から大きくかけ離れた) 2 名を除き，22 人を分析の対象とした．繰り返しのある 2×5 の分散分析を行なったところ，処理による主効果($p<.001$)，残響による主効果($p<.001$)が有意であった．処理条件間での t 検定では，h1～h4 条件において有意差が得られた (h1: $p=.049$; h2: $p=.026$; h3: $p=.003$; h4: $p<.001$).

表 3 各条件における子音の正解率.

	h1	h2	h3	h4	h5
org_rev (%)	66.5	63.5	61.4	55.1	58.1
proc_rev (%)	73.1	68.3	67.4	64.2	58.5

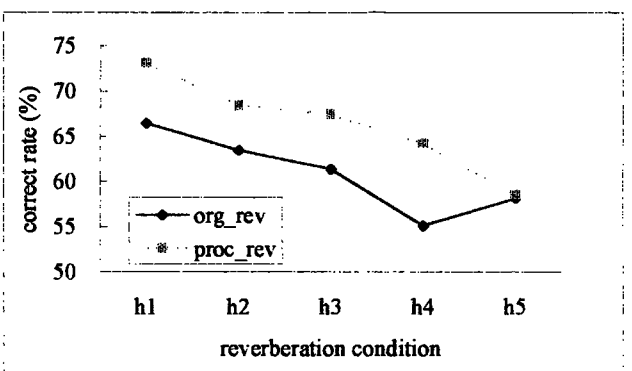

図 1 各条件における子音の正解率.

4. 考察

残響条件における主効果は有意であり，さらに主効果の成分を多項式展開すると一次成分で有意であったため，正解率は処理によらず残響時間が長くなるにつれて単調減少することが確認された．実験の結果から，全ての残響条件において proc_rev の方が正解率は高く，さらに残響時間が 0.9～1.2 秒では proc_rev の方が org_rev よりも正解率が有意に高いことが示された．残

響時間が 1.2 秒において最も処理による改善が得られたことから，この残響条件においては直前の残響成分によるターゲットへのマスキング量を最も抑えることができたと考えられる．また Proc_rev と Org_rev の正解率の間には，残響時間が短くなるにつれて差がみられなくなると予想されるが，本実験で用いた条件の中で残響時間が最も短い 0.9 秒においても処理による改善が得られていることから，残響時間が h1 より短い場合においても処理による効果のある条件が存在するものと考えられる．以上より，定常部抑圧は前処理として音声明瞭度の改善にとって有効であり，残響時間によって処理の効果が異なることが確認された．

次に，定常部抑圧処理を行なうことによって音声の明瞭度が改善した理由を，以下に検討をする．定常部抑圧処理は，残響によるマスキングの影響を抑えることを目的としている[1][2]．マスキング量の変化を調べるため，図 2 に音声サンプル「題目としてはざといいます」の残響付加前の時間波形(org)，定常部抑圧処理を行った後の時間波形(proc)，残響付加後の時間波形(org_rev)，定常部抑圧処理を行った後，残響を付加した時間波形(proc_rev)を示す．ここでは，残響条件として h4 を用いた．org と proc を比較すると，処理を行なうことによって音声の定常部が抑圧されている様子が分かる．一方 proc_rev では，org_rev に比べてターゲットに付加される直前の音声信号に対する残響成分によるマスキング量が減少する [1][2]ことから，音声明瞭度が改善したものと考えられる．

変調フィルタリングが本質的には音声の定常部抑圧と類似した処理をしているため[1][2]，我々は定常部抑圧処理による効果を，音声の変調スペクトルの変化として捉えることができると考えた．そこで，定常部抑圧処理を行う前後の変調スペクトルを求め，両者の比較を行った．図 3 に org の変調スペクトル(ms_org)と，proc の変調スペクトル(ms_proc)を示す．本論文においては，変調スペクトルは帯域制限された音声サンプルの時間包絡を周波数分析することで計算した．図 3 では，まず音声サンプルを[12]にならい帯域通過フィルタによって 4 帯域 (band 1: 0 – 800 Hz, band 2: 800 – 1600 Hz, band 3: 1600 – 3200 Hz, band 4: 3200 – 8000 Hz) に分割してから変調スペクトルを計算した．各帯域において org, proc 共に 24 種類の音声サンプルに対する変調スペクトルの平均を求め，その平均値に対し変調周波数軸上において前後 1 点ずつ計 3 点の移動平均による平滑化を行った．図 3 から全ての帯域において，定常部抑圧処理を加えることにより変調周波数 4Hz 付近と 10Hz 以上で変調指数が上昇している．通常音声明瞭性に寄与している変調周波数は 1-16Hz 付

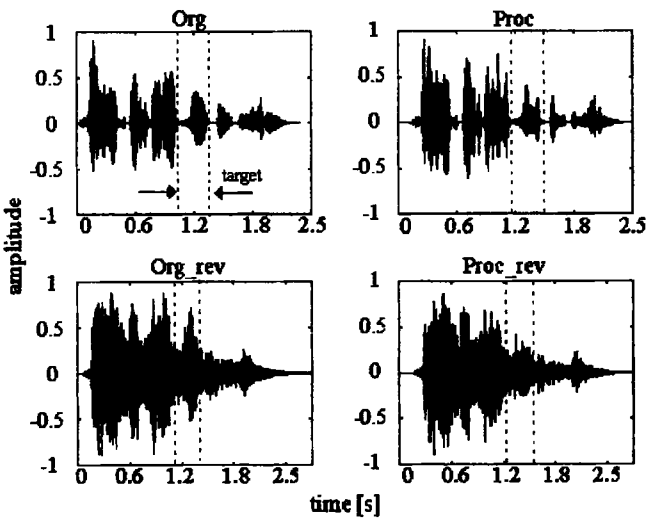

図 2 実験で使用した音声サンプル「題目としてはざといいます」における，残響付加前の時間波形(org)，定常部抑圧処理を行った後の時間波形(proc)，残響付加後の時間波形(org_rev, 残響時間=1.2 s)，定常部抑圧処理を行った後，残響を付加した時間波形(proc_rev,残響時間=1.2s) ．ただし破線に挟まれた区間は，ターゲットの区間を示す．

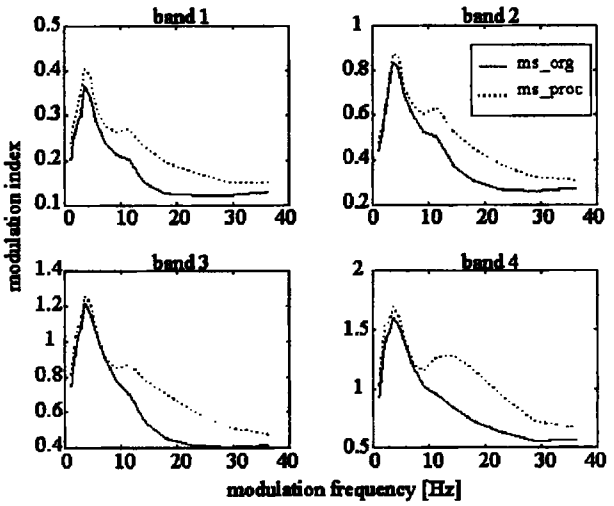

図 3 各帯域における org の変調スペクトル (ms_org) と proc の変調スペクトル (ms_proc)．直線は ms_org を，破線は ms_proc を示す．

近であるといわれ，特に変調スペクトルが最大値に近づく 4 Hz 付近が最も重要であるとされる[13][14]．しかし音声に残響が付加されると，変調スペクトルのピークが低い変調周波数領域に移り，変調指数も減少する[15]ことから，変調スペクトルと音声明瞭性の間には相関関係がある．変調周波数 4Hz 付近は音声における一秒間に出現する音節の割合に対応していると考え

られ[16]，一方変調周波数 10Hz 以上は一秒間に出現する音素の割合に対応していると考えられることから，定常部抑圧処理を行なうことによって音節や音素の遷移部が強調されたといえる．よって音声明瞭度にとって重要である変調周波数帯域の変調指数があらかじめ強調されたことにより，残響の影響によって変調指数が減少することを抑える事ができたと考えられる．

5. おわりに

本論文は，荒井らによって提案された手法[1][2]に基づき，音声の定常部を抑圧することで直前の残響成分による現在の音声信号へのマスキングの影響を軽減し，残響環境下において音声明瞭度の低下を抑えることを目的とした．定常部抑圧処理を様々な残響条件において調べるため，残響時間を人工的に細かく変化させた複数のインパルス応答を用いて聴取実験を行なった．結果から，定常部抑圧処理の効果は残響時間によって異なり，残響時間が 0.9～1.2 秒において処理による改善が得られた（特に残響時間が 1.2 秒において最も改善が得られた）．以上より荒井らの手法[1][2]は，残響環境下において音声明瞭度を改善するための前処理として有効であることが確認された．

6. 謝辞

インパルス応答のデータを提供して頂いた，東京大学の橘秀樹先生，上野佳奈子さん，横山栄さんに心から感謝申しあげます．また実験に参加して頂いた全ての被験者の方々に感謝いたします．

文　献

[1] 荒井隆行，木下慶介，程島奈緒，楠本亜希子，喜田村朋子，“音声の定常部抑圧の残響に対する効果，” 日本音響学会秋季研究発表会講演論文集，1，pp. 449-450, 2001.

[2] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, “Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments,” Acoustical Science and Technology, 23, 2002.

[3] S. Furui, “On the role of spectral transition for speech perception,” J. Acoust. Soc. Am., 80(4), pp. 1016-1025, 1986.

[4] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech Audio Process., 2, pp. 578-589, 1999.

[5] A. Kusumoto, T. Arai, M. Takahashi and Y. Murahara, “Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired,” Proc. IEEE ICASSP, pp. 933-936, 2000.

[6] T. Langhans and H. W. Strube, “Speech enhancement by nonlinear multiband envelope filtering,” Proc. IEEE ICASSP, pp. 156-159,1982.

[7] C. Avendano and H. Hermansky, “Study on the dereverberation of speech based on temporal envelope filtering,” Proc. ICSLP. pp. 889-892,1996.

[8] H. Wang and F. Itakura, “An approach of dereverberation using multi-microphone sub-band envelope estimation,” Proc. IEEE ICASSP. pp. 953-956, 1991.

[9] T. Yamada, S. Nakamura and K. Shikano, “Hands-free speech recognition based on 3-D viterbi search using a microphone array,” Proc IEEE ICASSP, pp. 245-248, 1998.

[10] N. Hodoshima, T. Arai and A. Kusumoto, “Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments,” Proc. Forum Acusticum Sevilla, 2002.

[11] M. R. Schroeder, “New method of measuring reverberation time,” J. Acoust. Soc. Am., 37, pp. 409-412, 1965.

[12] T. Arai and S. Greenberg, “Speech intelligibility in the presence of cross-channel spectral asynchrony,” Proc. IEEE ICASSP, pp.933-936, 1998.

[13] R. Drullman, J. M. Festen and R. Plomp, “Effect of temporal envelope smearing on speech reception,” J. Acoust. Soc. Am., 95(2), pp. 1053-1064, 1994.

[14] T. Arai, M. Pavel, H. Hermansky and C. Avendano, “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” J. Acoust. Soc. Am., 105(5), pp. 2783-2791, 1999.

[15] T. Houtgast and H. J. M. Steeneken, “A review of MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” J. Acoust. Soc. Am., 77(6), pp. 1069-1077, 1985.

[16] S. Greenberg, “Understanding speech understanding – Towards a unified theory of speech perception,” Proc. ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, pp. 1-8, 1996.

盛光印刷所