残響環境下における音声の明瞭度低下を抑える前処理技術

Pre-processing technique to prevent degradation of speech intelligibility in reverberant environments

荒井 隆行, 程島 奈緒, 後藤 崇公, 井上 豪, 大畑 典子,
木下 慶介[*1], 喜田村 朋子[*2], 楠本 亜希子[*3]
Takayuki Arai, Nao Hodoshima, Takahito Goto, Tsuyoshi Inoue, Noriko Ohata,
Keisuke Kinoshita[*1], Tomoko Kitamura[*2] and Akiko Kusumoto[*3]

上智大学理工学部電気・電子工学科
Sophia University, Department of Electrical and Electronics Engineering

*1 現在は、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
NTT Communication Science Laboratories, Speech Open Laboratory

*2 現在は、ワイデックス株式会社, 技術部技術管理グループ
WIDEX Co., Ltd., Engineering Department

*3 現在は、Portland VA Medical Center,
National Center for Rehabilitative Auditory Research

内容概要　残響環境下において音声明瞭度が低下する主要因の一つである overlap-masking に対し、音声の定常部を抑圧することでそのマスキング量を減らし、明瞭度の低下を防ぐ前処理技術を以前から提案している(荒井他, 2001; Arai et al., 2002)。聴取実験の結果、予めコンピュータ上で残響がたたみ込まれた音声では、特定の残響時間において有効であることが示されている。本稿では、その実験結果に加え、新たに実際の残響環境下における聴取実験を行った。それにより、Arai et al.の定常部抑圧処理は、残響環境下において音声明瞭度を保持するための前処理として有効であることを改めて確認した。また、実時間処理に適したアルゴリズムについても検討した。

1. まえがき

残響の多い環境では、一般に音声の聞き取りが悪くなることが知られている。ごく短い初期反射音は先行音効果によって直接音と聴覚的に結合する結果、音声の聞き取りを助けるが(e.g., Haas, 1972)、直接音より数十 ms 以降の残響音は音声明瞭度を下げる (Nabelek and Pickett, 1974)。多目的ホールでは音楽や講演会など複数の目的のために同じホールが使われる。大きなホールになると音楽には豊かさが増す反面、音声明瞭度の劣化は顕著となる。

残響の存在はしばしば大きな問題となる。それは、特に聴覚障害者やお年寄りなどにとって深刻な問題である(Nabelek and Mason, 1981)。また、自分の母語以外の言語を使った音声コミュニケーションにおいても、また残響は好ましくない (Takata and Nabelek, 1990)。語学の聞き取り試験で同じ音声が違う残響環境で再生されれば、受験者にとって不利益が生じることにもなりかねない。

残響の影響による音声明瞭度低下に対する対策としては、建築音響的なアプローチを含め様々なものが存在する。電気音響的なアプローチには、1）聞き手側に何らかの処理や装置を必要とするものと、2）音声を流す側に何らかの処理や装置を設けるものに大別される。

1）には、マイクロフォンアレイにより空間情

Trans. Tech. Comm. Psychol. Physiol. Acoust., The Acoustical Society of Japan, Vol. 33, No.5, H-2003-59
Trans. Tech. Comm. Architectural Acoust., The Acoustical Society of Japan, AA2003-21

報を利用するものや、1つのマイクロフォンによって信号の時間包絡を操作するといったdereverberation などの後処理(post-processing)が含まれる(Kaneda and Ohga, 1986; Flanagan et al., 1991; Neely and Allen, 1979; Miyoshi and Kaneda, 1988)。この場合、聞き手は何らかの形で処理装置を携帯し、またイヤフォンなどを用いて処理された音を聞く必要がある。

一方、2）は前処理(pre-processing)であり、音声が室内に放射される前に、何らかの強調処理が音声に施されることになる。つまり、マイクロフォンからスピーカまでの間のPA (public access)機器の中に処理装置が組み込まれることになる。処理の対象となる音声は、話者が使うマイクロフォンから直接得られた、まだ残響の掛かっていない音声であるので、音声の特徴に基づいた強調処理をするのであれば、なおさら処理がしやすいというのが、1）とは大きく違う点である。

Langhans and Strube (1982)はこのpre-processingに関する試みを以前行っているが、そこでは音声の各帯域信号の時間包絡に対してフィルタ処理が行われている。音声信号は、ある音源が調音器官によって変調を受けたときの出力信号と見なすことができ、また音声の時間包絡はその変調の様子を表すものと見なすことができることから、Langhans and Strubeの処理は、変調フィルタリングと呼ぶことができる。Langhans and Strubeの報告によると、変調フィルタリングをpre-processingに用いた際の効果は確認されなかった。

ところで、我々はLanghans and Strube (1982)と同様、pre-processingとしての変調フィルタリングに注目し、いくつかの実験を繰り返してきた(Kusumoto et al., 2000; Kitamura et al., 2000; Hodoshima et al., 2002a)。それらは、残響が低域通過型の変調度伝達関数 MTF (modulation transfer function) (Houtgast and Steeneken, 1985)を有していることに基づいている。音声信号に対し時間包絡の周波数分析を行うことによってその変調スペクトルを見てみると、その周波数特性は通常、4 Hz付近にピークを持つが、これは音声の音節に基づく音響的な時間特性を反映している(Duquesnoy and Plomp, 1980; Arai et al., 1999)。残響環境下では、音声の変調スペクトルはそのピークがより低域に下がると共にその変調指数も小さくなる。それを予め強調するためにIMTF (inverse MTF)の特性を持つ変調フィルタを用いるというものである。実験の結果から、将来性のある改善傾向を確認している。

従来の変調フィルタリングでは、時間包絡に対して線形フィルタリングが施されていた。一方、より直接的に変調フィルタリングを実現するため、我々はその後、時間包絡に非線形処理を施す定常部抑圧処理を提案し、ある残響条件下では残響による明瞭度低下を防げることを実証した(荒井ら, 2001; Arai et al., 2002)。この定常部抑圧処理は、残響によるoverlap maskingを減らすことを目的とした処理である。

残響が音声の明瞭度を下げる要因としては、self-maskingとoverlap-maskingの2つがあるとされている(Bolt and MacDonald, 1949; Nabelek and Robinette, 1978; Nabelek et al., 1989)。一つ目のself-maskingはそれぞれの音素内でマスキングが起こる結果、音素自身が変形するもので、音の立ち上がりや立ち下がりといった遷移部が特になまる。もう一方のoverlap-maskingは、先行する音素に伴う残響が後続する音素をマスクするというもので、特に先行する音素のエネルギが大きく後続する音素のエネルギが小さい場合、その影響は大きくなると考えられる。

このようにoverlap-maskingは音声明瞭度を下げる主な要因であると考えられるが、このoverlap-maskingを減らすためには、適当に原音声

を間引くことが考えられる。しかし、音声とは無関係に、機械的に間引いてしまったのでは、逆に音声情報が失われてしまい、結果として間引くことがかえって明瞭度低下を招いてしまう。

そこで荒井ら(2001)および Arai et al. (2002)は、音声信号のうち"定常部"を間引く（抑圧する）ことを考えた。このような処理を定常部抑圧処理と呼んでいる。音声の定常部は典型的には母音部の中央（音節核）であり、そのエネルギは大きいものの、音声としての情報量は少ない。一方、音声の遷移部は、明瞭度実験の結果からも音声知覚に関して非常に重要な役割を果たしていることがわかっている(Furui, 1986)。

音声の母音定常部は一般にエネルギが大きいことが多いので、それに後続する遷移部やエネルギの小さい子音そのものは overlap-masking の影響をまともに受けやすい。そこで定常部抑圧処理を施すと、音声情報の損失は最小限に抑えながら、overlap-masking による遷移部へのマスキング量をなるべく減らすことが可能となる。実際は、音声情報の損失と overlap-masking のマスキング量低下との間にはトレードオフが存在し、両者のバランスが取れるような最適なパラメータ値を探ることになる。

本稿では、残響環境下において明瞭度の低下を抑える前処理技術としての定常部抑圧処理を改めて紹介し、さらにその実用的応用について考える。特に、実際の残響環境下において行われた実験についても触れ、実時間処理応用をも視野に入れて議論する。

2. 定常部抑圧処理

2.1 フィルタバンクを用いた手法

荒井ら(2001)および Arai et al. (2002)によって提案された定常部抑圧処理では、次のような信号処理が行われている。まず、音声信号を 1/3-oct に帯域分割し、各帯域において時間包絡を抽出する。次に、時間包絡を 100 Hz にダウンサンプリングし、その対数軌跡から前後 2 点計 5 点に対する回帰係数をサンプルごとに計算する。すべての帯域に渡って、その回帰係数の 2 乗平均 (以下では D とする)を求める。ここで，D は Furui (1986)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。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では元の波形の振幅を抑圧する。抑圧する際、振幅を 0%に抑圧すると不自然性が高まることがあるので、現在は 40%程度に抑圧することにしている。また、振幅を抑圧する際、急激に振幅が変化しないよう、実際にはエッジに傾きを持たせている。

2.2 FFT を用いた手法

高速フーリエ変換(FFT)を用いると、フレーム処理に基づく、より実時間処理向きのアルゴリズムが実現できる（図 1）。

まず、音声信号に対し 20 ms のハニング窓によってフレーム分けを施す。このとき、フレームシフトは 10 ms（50%オーバラップ）とする。各フレームにつき対数スペクトルを計算し、その逆 FFT によりケプストラム係数を求める。そのうち 16 次までの係数について、各係数の時間軌跡に対し前後 2 点計 5 点の回帰係数をサンプルごとに計算する。16 次までに渡って、その回帰係数の 2 乗平均 D を求め、D が一定の閾値より小さい箇所を

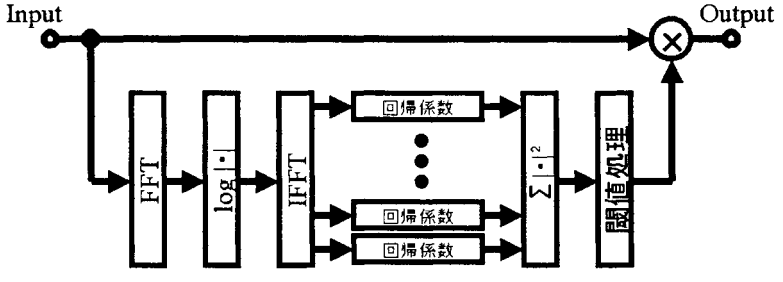

図 1：FFT を用いた定常部抑圧処理のブロックダイアグラム

定常部とし、定常部とみなした箇所では元の波形の振幅を抑圧する。

3. 聴取実験

3.1 実験室環境における実験

ここでは、Hodoshima et al. (2002b)ならびに程島ら(2003a)、Hodoshima et al. (2003b)によって行われた聴取実験についてを紹介する。残響環境は、コンピュータ上で音声信号と残響のインパルス応答をたたみ込むことによって実現した。なお、使用したインパルス応答は、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（残響時間が 1.1s）を基に、それを人工的に加工することによって残響時間 0.4 秒から 1.3 秒までの範囲に変化させたものである。

3.1.1 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 子音を用いた。結局、実験では 24 種類の CV を使用した。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。

刺激音はオリジナルの音声信号に残響をたたみ込んだ刺激セット（処理なし）と、定常部抑圧処理を行った後に残響をたたみ込んだ刺激セット（処理あり）の 2 種類を用意した。

3.1.2 被験者

日本語を母語とする健聴者 44 名(残響時間が短いセットに対して 22 名、長いセットに対して 22 名)であった。

3.1.3 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。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 を選択肢としてカナで表示した。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。

表 1 ：実験室環境における単音節明瞭度試験の結果（残響時間が短いセットに対する子音正解率）

残響時間(s)	0.4	0.6	0.8	0.9	1.0
処理なし(%)	90.7	84.3	73.9	70.1	69.5
処理あり(%)	92.8	86.6	82.4	79.2	73.7

表 2 ：実験室環境における単音節明瞭度試験の結果（残響時間が長いセットに対する子音正解率）

残響時間(s)	0.9	1.0	1.1	1.2	1.3
処理なし(%)	68.3	63.5	61.4	55.1	58.1
処理あり(%)	73.1	68.3	67.4	64.2	58.5

3.1.4 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 1（残響時間の短いセット）と表 2（残響時間の長いセット）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。処理による主効果はいずれも有意($p<.001$)であった。処理条件間での t 検定の結果、表 1 では残響時間が 0.8, 0.9, 1.0 秒において処理ありのほうが、表 2 では残響時間が 0.9, 1.0, 1.1, 1.2 秒において処理ありのほうが有意に正解率が高かった。

3.1.5 考察

これらの実験の結果から、全ての残響条件において処理ありの方が正解率は高く、さらに残響時間が 0.8～1.2 秒では処理の効果が確認された。

3.2 実際の残響環境における実験

実験室環境で効果を示した定常部抑圧処理を、実際の残響環境下においてもその効果を確認するために大学の講堂にて実験を行った。実験は単音節明瞭度試験と、文の書き取り試験を行った。

3.2.1 刺激

単音節明瞭度試験では第3.1節の刺激のうち、母音が/a/のもの（14単音節、キャリア文付き、処理あり/なし）を用いた。文了解度試験では、NTT-AT音素バランス1000文から20文を用いた。

3.2.2 被験者

日本語を母語とする健聴者24名であった。

3.2.3 手順

実験は上智大学構内で一番大きな収容人数（822名）を持つ10号館講堂で行った。壇上にスピーカを設置し、PCから予め準備された刺激音を再生した。被験者は講堂正面の後方のブロックに配置した。始めに指示を与えた後、テスト用の刺激文を用いて被験者全員が問題なく聞き取れる程度の音量に出力を調整した。

単音節明瞭度試験では、28刺激（14単音節のそれぞれについて処理あり、処理なし）を2回の計56刺激をランダムに並べ替え提示した。各試行において刺激音を一度だけ提示し、回答を14単音節のリストから1つ強制的に選んで用紙に書いてもらった。次の刺激提示までは5秒とした。

文了解度試験では、24名の被験者をグループA（13名）とグループB（11名）に分け、各グループごとに実験を行った。各グループでは、異なる20文（処理ありの10文と処理なしの10文）をランダムに並べ替えて提示した。また、グループAで処理あり（なし）であった10文は、グループBで処理なし（あり）になるように組み合わせることによって、バランスをとった。各試行において刺激音は2度、20秒間隔をあけて提示し、回答をカナで用紙に書いてもらった。

3.2.4 実験結果

単音節明瞭度試験では子音の正解率を比較した結果、処理あり（69.3%）のほうが処理なし（62.7%）よりも正解率が高くなった。

文了解度試験では、書き取られた文を処理ありと処理なしで比較した。その結果、処理ありと処理なしではともにモーラごとの正解率が95%以上と高く、その差はほとんど観測されなかった。

3.2.5 考察

単音節明瞭度試験では実験室環境のdiotic受聴の場合と同じ刺激を用いたが、両耳（dichotic）環境においてもその効果を確認できた。

文の書き取りでは文脈情報を利用できることから、多少の聞き取りづらさが存在しても特に健聴者の場合には問題ない。今回用いた刺激文は、比較的平易で、訓練を受けたアナウンサがゆっくりと明瞭に発話したもので、また、残響時間もそれほど長くない環境で、かつ直接音のエネルギも強かったことが、そもそもの了解度が高かった要因として考えられる。しかし、より劣悪な残響環境下で、親密度の低い語が存在したり自然発話音声にみられるように話速が速かったり不明瞭な音声になると、本手法の効果が現れてくるものと予想される。このことは、お年寄りや聴覚障害者に対してはなおさらのことであろう。

4. おわりに

残響環境下における音声の明瞭度低下を抑える前処理技術として、特に定常部抑圧処理(荒井ら, 2001; Arai et al., 2002)の有効性を検討した。実験室環境での実験に加え、実環境における実験を新たに行った結果、定常部抑圧処理がoverlap-maskingの影響を軽減し、残響環境下における前処理として有効であることが改めて示された。今後はさらに実用化を進めていきたい。

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。

参考文献

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

[2] 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, 449-450, 2001.

[3] 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(4):229-232, 2002.

[4] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," *ICSLP*, 889-892,1996.

[5] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, 21, 577-580, 1949.

[6] A. J. Duquesnoy and R. Plomp, "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," *J. Acoust. Soc. Am.*, 68(2):537-544, 1980.

[7] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West and M. M. Sondhi, "Autodirective microphone systems," *Acoustica*, 73 (2):58-71, 1991.

[8] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, 80(4):1016-1025, 1986.

[9] H. Haas, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc.*, 20:145-159, 1972.

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

[11] N. Hodoshima, T. Inoue, T. Arai and A. Kusumoto, "Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments," *China-Japan Joint Conference on Acoustics*, 199-202, 2002.

[12] 程島奈緒, 荒井隆行, 井上豪, 木下慶介, 楠本亜希子, "小・中規模ホール環境を想定した定常部抑圧による拡声音声の明瞭度改善," 日本音響学会春季研究発表会講演論文集, 1073-1074, 2003.

[13] N. Hodoshima, T. Arai, T. Inoue, K. Kinoshita and A. Kusumoto, "Improving speech intelligibility by steady-state suppression as pre-processing in small to medium sized halls," *Eurospeech*, 2003.

[14] 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):1069-1077, 1985.

[15] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech and Signal Process.*, ASSP-34(6):1391-1400, 1986.

[16] T. Kitamura, K. Kinoshita, T. Arai, A. Kusumoto and Y. Murahara, "Designing modulation filters for improving speech intelligibility in reverberant environments," *ICSLP*, 3:586-589, 2000.

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

[18] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," *IEEE ICASSP*, pp. 156-159,1982.

[19] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech and Signal Process.*, 36(2):145-152, 1988.

[20] A. K. Nabelek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation," J. Speech and Hearing Res., 17:724-739, 1974.

[21] A. K. Nabelek and L. Robinette, "Influence of precedence effect on word identification by normally hearing and hearing-impaired subjects," *J. Acoust. Soc. Am.*, 63:187-194, 1978.

[22] A. K. Nabelek, T. R. Letowski and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, 86:1259-1265, 1989.

[23] A. K. Nabelek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *J. Speech and Hearing*, 24:375-383, 1981.

[24] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, 66(1): 165-169, 1979.

[25] Y. Takata and A. K. Nabelek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *J. Acoust. Soc. Am.*, 88:663-666, 1990.