Vol.2015-MUS-107 No.66
2015/5/24

フォルマント構造維持を規範とした音声信号混合法

大脇 渉[1,a)]　旭岡 舜介[1,b)]　高橋 弘太[1,c)]

概要：音声と BGM の同時再生時に音声が埋もれない，優先度付き音声信号混合法の，スマートミキサーでの新しい構成法を提案する．スマートミキサーは，入力信号を解析し処理する機構を音信号加算器自体に持たせる，新しい発想の音信号混合法である．従来法では，過度のスペクトル操作による出力信号の音質劣化が問題であった．そこで，母音認識に焦点を当て，スペクトル操作をフォルマント帯域に限定することで，聴きとりやすさと音質を両立できると考えた．本稿では，フォルマント構造維持を規範とし，音質劣化の改善を重視した新しいスマートミキサーを提案する．

1. はじめに

本稿では，優先度付き音声信号混合法 [1], [2] (以下，スマートミキサー) の，音声信号のフォルマント構造維持を規範とした新しいアルゴリズムを提案する．

スマートミキサーとは，入力信号をそれぞれ時間周波数平面 (以下，tf 平面) に展開し，非線形に重ね合わせることで出力信号を得る，新しい音信号混合法である．スマートミキサーの構成法として，音声と BGM とのミキシングにおいて，音声が埋もれない出力信号を得るアルゴリズムとして，優先度付き音声信号混合法が提案されてきた．

優先度付き音声信号混合法の従来法 [1], [2] では，調波構造の分布に着目し，音声信号の調波構造の強調を処理の主眼としていた．この手法では，過度の処理により音質が大幅に劣化する問題があった．そこで音質劣化の軽減策として，母音識別において重要となるフォルマント情報に着目し，処理領域をフォルマント帯域に制限する手法 [2] も検討された．しかし，音質劣化の軽減は特定の音源でのみ為されるに留まった．原因として，フォルマント帯域での処理領域の制限によって，強調する調波構造に沿った処理領域が断続的となった点にあると考えた．

そこで，処理領域をフォルマント周波数帯域とする，従来法よりも tf 平面上で滑らかなスペクトル操作が有効であると予想した．本稿では，フォルマント構造維持を規範とし，音質劣化の改善を重視した新しいスマートミキサーを提案する．

1 電気通信大学
UEC, Chofu, Tokyo 182–8585, Japan
a) wataru@uec.ac.jp
b) hinooka-shunsuke@uec.ac.jp
c) kota.t@uec.ac.jp

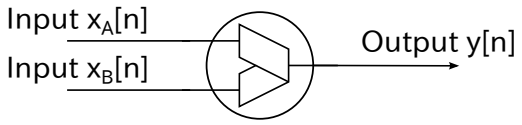


図 1　スマートミキサーを表す記号
Fig. 1　The symbol of the smart mixer.

2. スマートミキサー

従来の音信号のミキサー自体は，単純な加算器である．単純加算では，加算後のスペクトル上で，各入力信号のスペクトル遷移パターンが崩れうる．このとき，各入力信号の一部が聴き取れなくなる．そこで，プロのスタジオ技術者は，加算器の前後に，イコライザやコンプレッサなどの処理を追加することで，聴感上好ましい出力を得てきた．

一方，筆者らが提案するスマートミキサー [1], [2] では，入力信号を tf 平面に展開し，平面どうしの非線形な重ねあわせを行う．この原理により，従来のミキシング法を超えるきめ細かな処理を，手軽に実現できる．スマートミキサーを表す記号として，図 1 を用いている．

スマートミキサーの構成法として，音声と BGM とをミキシングしたとき，音声が埋もれない音信号混合法を検討してきた．また，入力信号間に優先度が存在することを前提としたアルゴリズムを，“優先度付き音信号混合法” として表記している．音声が埋もれない音信号混合法では，音声の優先度が高いものとみなして構成されている．tf 平面上の局所領域ごとに緻密な混合処理を行うことで，優先度の高い入力が埋もれず，容易に聴き取りが可能な混合出力を得る．本稿では，音声の特徴に着目し，音声に特化した構成法という意味合いで，“優先度付き音声信号混合法” と表記する．

© 2015 Information Processing Society of Japan

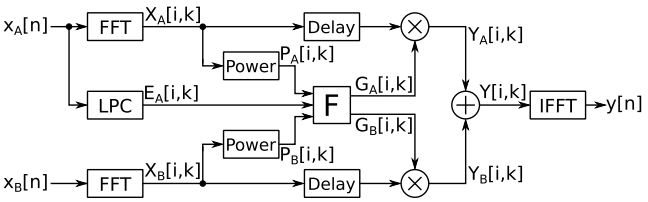


図 2 フォルマント構造維持を規範とした音声信号混合法 (提案法)

3. 線形予測符号を利用したフォルマント周波数推定

3.1 線形予測符号

線形予測符号 (Linear Predictive Coding，以下 LPC)[3], [4] は，過去の信号から現在の信号を予測する信号分析法である．LPC は以下の式 (1) で定義される．

$$y'_n = -\sum_{p=1}^{N_{\mathrm{LPC}}} a_p y_{n-p} \tag{1}$$

ここで，y_n，y'_n はそれぞれ時刻 n における，離散時間信号の実測値と予測値，a_p は LPC 係数，N_{LPC} は LPC 次数を表す．

3.2 最尤スペクトル推定法

LPC によって求められた LPC 係数 a_p と残差分散 $E(=\sigma^2)$ から，式 (2) に示す周波数応答を求めることで，音声のスペクトル包絡 $f(\omega)$ を近似する．$f(\omega)$ の極に対応する周波数を調べることで，フォルマント周波数を推定できる．

$$f(\omega) = \frac{\sigma^2}{2\pi}\frac{1}{|1+\sum_{p=1}^{N_{\mathrm{LPC}}} a_p e^{-jp\omega}|^2} \tag{2}$$

4. フォルマント構造維持を規範としたアルゴリズム

提案法のブロック図が，図 2 である．提案法の処理の概要を，以下に示す．

(1) 入力時間信号 $x_{\mathrm{A}}[n]$(優先音)，$x_{\mathrm{B}}[n]$(非優先音) を FFT し，時間周波数成分 $X_{\mathrm{A}}[i,k]$，$X_{\mathrm{B}}[i,k]$ に展開
(2) $X_{\mathrm{A}}[i,k]$, $X_{\mathrm{B}}[i,k]$ を，それぞれパワー $P_{\mathrm{A}}[i,k]$, $P_{\mathrm{B}}[i,k]$ に変換・平滑化
(3) $x_{\mathrm{A}}[n]$ に LPC を適用し，スペクトル包絡 $E_{\mathrm{A}}[i,k]$ を抽出
(4) $P_{\mathrm{A}}[i,k]$, $P_{\mathrm{B}}[i,k]$, $E_{\mathrm{A}}[i,k]$ より，ゲインマスク $G_{\mathrm{A}}[i,k]$, $G_{\mathrm{B}}[i,k]$ を決定・平滑化
(5) $X_{\mathrm{A}}[i,k]$，$X_{\mathrm{B}}[i,k]$ に $G_{\mathrm{A}}[i,k]$，$G_{\mathrm{B}}[i,k]$ をそれぞれ乗算し，処理後の入力信号 $Y_{\mathrm{A}}[i,k]$，$Y_{\mathrm{B}}[i,k]$ を生成
(6) $Y_{\mathrm{A}}[i,k]$ と $Y_{\mathrm{B}}[i,k]$ を加算し $Y[i,k]$ を生成
(7) $Y[i,k]$ を IFFT し，出力時間信号 $y[n]$ を出力

4.1 フォルマント帯域の決定

フォルマント帯域は，推定されたフォルマント周波数軌跡より決定する．フォルマント帯域幅 W_{fmt} に基づき，軌跡から周波数方向に上下 $\frac{W_{\mathrm{fmt}}-1}{2}$ ビンをフォルマント帯域とする．

4.2 ゲインマスク生成

本手法では，優先音および非優先音に対し振幅処理によるゲイン操作を行い，ゲインマスクを生成する．$D_{\mathrm{G}}[i,k]$ を式 (3) で定義する．ここで $P_{\mathrm{A}}[i,k]$，$P_{\mathrm{B}}[i,k]$ は優先音および非優先音のパワー，ATH$[k]$ は最小可聴閾値 [1]，$D_{\mathrm{fn}}[i,k]$ はフォルマント帯域判定とする．

$$D_{\mathrm{G}}[i,k] = \begin{cases} 1 & \left(\begin{array}{r} \mathrm{ATH}[k] \le P_{\mathrm{A}}[i,k] \\ \wedge \mathrm{ATH}[k] \le P_{\mathrm{B}}[i,k] \\ \wedge D_{\mathrm{fn}}[i,k] \end{array}\right) \\ 0 & \text{(otherwise)} \end{cases} \tag{3}$$

優先音へのゲインマスク $G_{\mathrm{A}}[i,k]$ は，時間フレーム毎の入力信号のパワー比率 $R_p[i] = \sum_{k=0}^{N_{\mathrm{FFT}}} P_{\mathrm{B}}[i,k] / \sum_{k=0}^{N_{\mathrm{FFT}}} P_{\mathrm{A}}[i,k]$，およびパラメータ $p_1 \sim p_4$ に基づき，式 (4) により決定される (N_{FFT} は FFT 点数)．ここで，$D_{\mathrm{G}}[i,k]$ が 1 の場合はゲインを増幅し，0 の場合はゲインを増幅しない．式 (4) における，$R_{\mathrm{P}}[i]$ と $G_{\mathrm{A}}[i,k]$ の関係を，図 3 に表す．

$$G_{\mathrm{A}}[i,k] = \begin{cases} \frac{p_3-1}{p_1-1}R_{\mathrm{P}}[i] + 1 - \frac{p_3-1}{p_1-1} & \left(\begin{array}{r} D_{\mathrm{G}}[i,k] \\ \wedge 1.0 \le R_{\mathrm{P}}[i] < p_1 \end{array}\right) \\ \frac{p_4-p_3}{p_2-p_1}R_{\mathrm{P}}[i] + p_3 - \frac{p_4-p_3}{p_2-p_1} & \left(\begin{array}{r} D_{\mathrm{G}}[i,k] \\ \wedge p_1 \le R_{\mathrm{P}}[i] < p_2 \end{array}\right) \\ p4 & \left(\begin{array}{r} D_{\mathrm{G}}[i,k] \\ \wedge p_2 \le R_{\mathrm{P}}[i] \end{array}\right) \\ 1.0 & \text{(otherwise)} \end{cases} \tag{4}$$

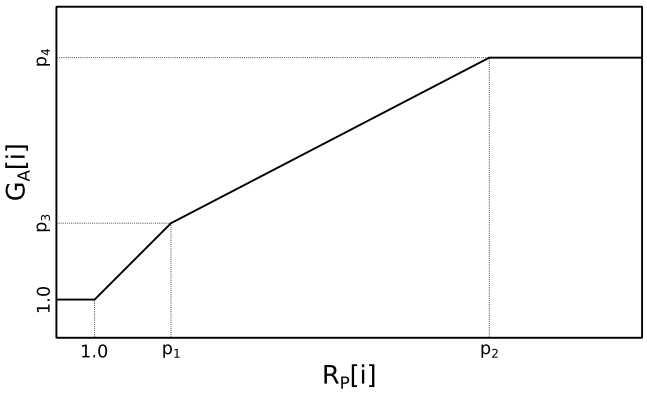


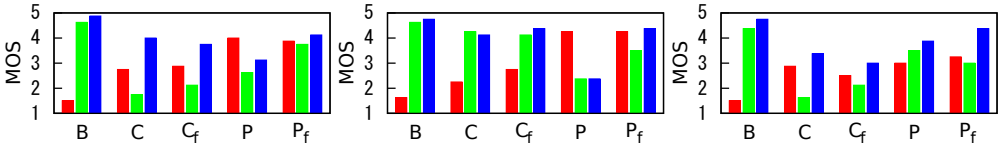
図 3 式 (4) における，$R_{\mathrm{P}}[i]$ と $G_{\mathrm{A}}[i,k]$ の関係

非優先音へのゲインマスク $G_{\mathrm{B}}[i,k]$ は，増幅処理後の優先音のパワーを非優先音のパワーが上回っている場合に，1 つ前の時間フレームにおける同じ周波数帯域のゲインマスク $G_{\mathrm{B}}[i-1,k]$ を，α 倍ずつ減衰させることで決定する．

© 2015 Information Processing Society of Japan

評価項目	比較対象	色
音声の聴き取りやすさ	元の音声	
音声の自然さ	元の音声	
音楽の自然さ	元の音楽	

表 1 評価項目



(a) S1 (b) S2 (c) S3

図 4 聴取実験による各評価対象ごとの平均 MOS 値

表 2 実験諸元

パラメータ		値
サンプリング周波数	F_s	44100 Hz
量子化ビット数	N_{bit}	16 bit
FFT 点数	N_{FFT}	1024 点
解析時窓関数	N_F	ハニング窓形 1024 点
合成時窓関数	N_I	ハニング窓形　256 点
フレームシフト	N_{SFT}	256 点
LPC 次数	N_{LPC}	12 次
フォルマント帯域幅	W_{fmt}	11 ビン (473.73 Hz)

もし処理後に倍率の下限値 L_α を下回る場合は，α を乗算しないこととする．また，$D_{\mathrm{G}}[i,k]=0$ の場合はゲインを 1 とする．

$$G_{\mathrm{B}}[i,k]=\begin{cases}\alpha G_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r}D_{\mathrm{G}}[i,k]\\ \wedge P_{\mathrm{A}}[i,k]<P_{\mathrm{B}}[i,k]\\ \wedge L_\alpha<\alpha G_{\mathrm{B}}[i-1,k]\end{array}\right)\\ G_{\mathrm{B}}[i-1,k] & \left(\begin{array}{r}D_{\mathrm{G}}[i,k]\\ \wedge P_{\mathrm{A}}[i,k]<P_{\mathrm{B}}[i,k]\\ \wedge L_\alpha\geq\alpha G_{\mathrm{B}}[i-1,k]\end{array}\right)\\ 1.0 & \text{(otherwise)}\end{cases} \tag{5}$$

本稿では，実験により各パラメータの値を，$p_1=3.0$，$p_2=100.0$，$p_3=3.0$，$p_4=6.0$，$\alpha=0.708$，$L_\alpha=0.01$ とした．

5. 聴取実験

実験諸元を表 2 に示す．評価対象は，表 3 に示す S1~S3 の 3 種類のセット音源を混合した，表 4 に示す 5 種類とした．表 1 に示す 3 つの評価項目について，5 段階の MOS 値 (1：Bad，2：Poor，3：Fair，4：Good，5：Excellent) による評価を行った．

入力音源では，すべてのセット音源について，優先音を音声，非優先音を音楽とした．音声には SRV-DB[5] の「発話のプロフェッショナルによる編集手帳（読売新聞）の読み上げ」のデータを，音楽には「RWC 研究用音楽データベース：ポピュラー音楽 [6]」のデータを用いた．今回の実験では，非優先音の音量を優先音の音量より，表 3 に示す音量差分あらかじめ大きく設定している．各セットをこの音量差で単純加算したとき，音声は音楽に埋もれ，聴き取りが困難である．ミキシングの出力信号において，スペクトル操作の影響をより顕著に現すためである．被験者は成人男性 8 名であり，全員ヘッドフォンによる聴取を行った．

表 3 音源の組み合わせ [5][6]

セット名	音声 (優先音)	音楽 (非優先音)	音量差
S1	女性音声	J-POP	12 dB
S2	男性音声	シティーロック	12 dB
S3	女性音声	ポップス	24 dB

表 4 評価対象

音源	備考
B	単純加算
C	スマートミキサー (従来法)
$\mathrm{C_f}$	従来法の処理範囲をフォルマント帯域に制限
P	提案法の処理範囲をフォルマント帯域外へ拡張
$\mathrm{P_f}$	フォルマント構造維持を規範とした手法 (提案法)

5.1 実験結果と考察

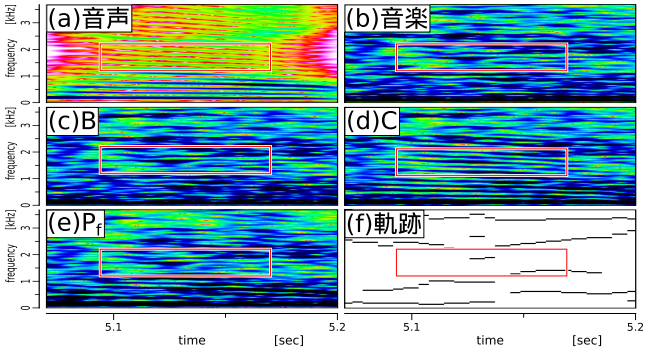


図 5 各信号の tf 平面 (a) 音声 (b) 音楽 (c) 音源 P (d) 音源 C (e) 音源 $\mathrm{P_f}$ (f) 音声のフォルマント軌跡

はじめに tf 平面上での観察結果について考察する．各信号の tf 平面を図 5 に示す．提案法の狙いは，フォルマント帯域外でのスペクトル操作を行わないことで，無用な音質劣化を軽減する点にある．したがって上記観点から各時間周波数平面を比較する．

枠で示す 1~2 kHz の領域は音声のパワーが小さく，主要なフォルマントが存在しない周波数帯域である．(d) の音源 C では，枠内の帯域においてもゲイン操作により音楽 (b) のスペクトルが抑制され，音声 (a) の調波構造が鮮明に現れている．一方 (e) の音源 $\mathrm{P_f}$ では，スペクトル操作を行っていない (c) の音源 B と同様に音楽 (b) の調波構造が現れている．この結果から，狙い通りにフォルマント帯域外におけるゲイン操作が適切に制限されたと考えられる．

次に聴取実験結果について考察する．図 4 に各評価対象ごとの平均 MOS 値を示す．フォルマント構造維持の有効性について議論するため，音源 P と音源 $\mathrm{P_f}$ を比較する．

© 2015 Information Processing Society of Japan

S1 と S2 においては，成分操作を行う範囲の制限により音声と音楽の両方の自然さが大きく改善され，かつ音声の聴き取りやすさが維持されていることが分かる．この結果より，音声信号混合処理におけるフォルマント構造維持が有効であると考えられる．

一方，S3 においては，音声の自然さの MOS 値が低下した．S3 は 3 種類の音源セットの中で最も音量差が大きく，音声の聴き取りやすさを維持するためにはフォルマント帯域成分を大幅に増幅する必要がある．そのため振幅増幅処理により音声全体のスペクトル構造が大きく崩れ，音質が劣化したと考えられる．

表 5 日本語子音の分類 [3]

調音位置		口唇		歯，歯茎		口蓋		声門
音源		有声	無声	有声	無声	有声	無声	無声
調音方式	摩擦音		f	z	s	ʒ	ʃ	h
	破擦音			dz	ts	dʒ	tʃ	
	破裂音	b	p	d	t	g	k	
	半母音	w		r		j		
	鼻音	m		n		ŋ		

各音源について，聴き取りづらい時区間での，tf 平面の観察を行った．観察により，子音『t』,『k』,『z』,『s』,『ts』の音節に該当することがわかった．表 5 に日本語子音の分類を示す．問題に該当する音節の子音は，摩擦音や破擦音，破裂音の子音に含まれていることが分かる．特に歯茎無声音は，すべての子音が該当している．したがって，これらの子音が現れた場合，現在の提案法の処理では埋もれを回避することができないと考えられる．

6. 結論

本稿では，フォルマント構造維持を規範とした，新しいスマートミキサーのアルゴリズムを提案した．

実音源での処理結果の観察により，提案法でフォルマント構造維持が実現できていることが確認できた．そして聴取実験により，従来法と比較して，音声の聴き取りやすさを維持したまま音質が向上していることから，提案法の有効性を確認した．

7. 今後の課題

今後は，音声の子音部に対する処理内容変更システムの導入，入力信号の音量差が大きい場合における処理方法の検討，および各入力信号に対するゲインマスク決定関数の改良を行う．またフォルマント推定法についても改良を行い，より高精度な周波数推定を実現することで，さらなる出力信号の音質向上を目指す．

謝辞 本研究は JSPS 科研費 20500111，23500147，26330188 の助成を受けたものです．

参考文献

[1] 大脇渉，有北知弘，宮地紘司，高橋弘太，“スマートミキサー〜時間周波数平面上での位相操作〜，” 信学技報，vol.112，no.478，pp.13-18，2013.
[2] 旭岡舜介，大脇渉，高橋弘太，“フォルマント構造維持を規範とした音声信号混合法，” 信学技報，vol. 114，no. 274，EA2014-27，pp. 21-26，2014 年 10 月.
[3] 古井貞熙，” 新音響・音声工学,” 近代科学社，2006.
[4] 板橋秀一編，“音声工学,” 森北出版株式会社，2005.
[5] “話速バリエーション型音声データベース:SRV-DB.” 〈http://www.it.ice.uec.ac.jp/SRV-DB/〉
[6] 後藤真孝，橋口博樹，西村拓一，岡隆一，“RWC 研究用音楽データベース：ポピュラー音楽データベースと著作権切れ音楽データベース,” Technical Report 103(2001-MUS-042)，Oct 2001.)

© 2015 Information Processing Society of Japan