

成分位相の制御により声の肌触りを変える

河原 英 紀^{†1} ロイ・パターソン^{†2} 森 勢 将 雅^{†3}
坂 野 秀 樹^{†4} 津 崎 実^{†5} 高 橋 徹^{†6}
西 村 竜 一^{†1} 入 野 俊 夫^{†1}

成分の位相は音色に影響を与えないという常識は、第一次近似としては正しい。しかし、ニアフィールドリスニングやヘッドフォンでの受聴等の状況では、成分の相対位相を操作することにより、音の『肌触り』とでも表現すべき属性のバリエーションを元の音色に加えることができる。ここでは、TANDEM-STRAIGHT の拡張機能として実現した、成分の振幅・周波数・位相制御インタフェースを紹介する。

Voice texture manipulation through constituent component phase control

HIDEKI KAWAHARA,^{†1} ROY D. PATTERSON,^{†2} MASANORI MORISE,^{†3}
HIDEKI BANNO,^{†4} MINORU TSUZAKI,^{†5} TORU TAKAHASHI,^{†6}
RYUICHI NISIMURA^{†1} and TOSHIO IRINO^{†1}

Common understanding, “phase deaf,” in our auditory perception is valid as first order approximation. However, it is possible to add delicate timbre, voice texture in other words, by manipulating relative phase of constituent components, when listening using headphones or under so-called near field listening conditions. This article introduces a specialized interface to control amplitude, frequency and phase in re-synthesis stage of TANDEM-STRAIGHT system as an interesting extension.

1. はじめに

含まれる成分の相対位相により、知覚される音色が影響を受けることは、専門家の間では広く知られている^{1)~6)}。この効果は、例えば、合成音声特有の『バズ感』を軽減するための手段として、音声分析変換合成システム STRAIGHT⁷⁾ で利用されている。しかし、一般には「位相は音色に影響しない」ことが常識となっており、他での応用例は少ない⁸⁾。

この常識には、十分な理由がある。教室等の広い室内では、複数の音の伝播経路を経た音の干渉により、聴取位置での信号に含まれる成分の位相は、ほぼランダム化されている。そのため、元の信号の相対位相は破壊されてしまい、この効果を感じることはできない。リビング等での音楽鑑賞でも、同様に成分位相はランダム化されてしまう。

しかし、ヘッドフォンやニアフィールドリスニング環境での受聴の普及は、この状況を変えつつある。それらの状況では、成分の相対位相による音色の変化は、明瞭に知覚される。ここでは、位相制御された音声を用いて、今までの常識を覆す体験をしてもらうためのシステムについて紹介する。

デモシステムは、STRAIGHT を再定式化した TANDEM-STRAIGHT⁹⁾ の合成部分を拡張して実現されている。具体的には、パルスと雑音による混合駆動音源のパルスの代わりに、AM と FM を含む複合正弦波の各成分の振幅・周波数・位相を明示的に指定することで、直ちに音声再合成される。このデモシステムで同一の滑らかな時間-周波数表現を用いて成分

^{†1} 和歌山大学システム工学部

Faculty of Systems Engineering, Wakayama University

^{†2} ケンブリッジ大学 聴覚神経基盤研究センタ

CNBH, Cambridge University

^{†3} 立命館大学情報理工学部

College of Information Science and Engineering, Ritsumeikan University

^{†4} 名城大学理工学部

Department of Information Engineering, Meijo University

^{†5} 京都市立芸術大学

Kyoto City University of Arts

^{†6} 京都大学大学院情報学研究科

Graduate School of Informatics, Kyoto University

の振幅を設定しても、相対位相を、cos 位相, sin 位相, 交替位相²⁾, ランダム位相, Schroeder 位相¹⁰⁾, 最小位相⁷⁾などに設定することにより、知覚される音色にはかなりの違いが生じる。しかも、それらの間を連続的にモーフィングすることも可能である¹¹⁾。来場者の方々には、このデモシステムで、音の『肌触り』とでも表現すべき音色の新たな属性のバリエーションを体験して頂きたい。

2. TANDEM-STRAIGHT の拡張

まず、TANDEM-STRAIGHT の概要を説明し、次に、本資料のポイントである合成部分の拡張方法について説明する。

2.1 TANDEM-STRAIGHT

TANDEM-STRAIGHT は、基本的には VOCODER¹²⁾である。音声は、駆動情報とフィルタ情報に分離され、再合成される。駆動情報は、有声音の主要な成分である周期成分と、それ以外の非周期成分とから構成されている。

TANDEM-STRAIGHT では、二つの処理を組み合わせることで、駆動情報の周期成分による干渉を排除した滑らかなフィルタ特性を抽出している。最初の処理では、基本周期の 1/2 の間隔を隔てて計算されたパワースペクトルの平均が求められる。この処理により、周期性に起因する時間方向の周期的変動が消去される¹³⁾。次の処理では、周波数方向で F0 に適応した平滑化と特性の補償が行われる。この処理により、残っていた周波数方向の周期的変動が消去される。consistent sampling¹⁴⁾ の考え方に基づいて、調波位置でのレベルを保存するための補償処理を対数スペクトル上で行うことにより、常に非負の値を取る滑らかな時間周波数表現である STRAIGHT スペクトル $P_{TST}(\omega, t)$ が求められる。なお、 $P_{TST}(\omega, t)$ は、パワーの次元を持つ。

2.2 周期成分の拡張：振幅の制御

元のシステムの再合成部分では、フィルタの最小位相インパルス応答を求めていた。拡張したデモシステムでは、この STRAIGHT スペクトルを用いて、再合成音声の周期成分の振幅を設定する。ここでは、簡単のため、非周期成分を考慮せず、 $P_{TST}(\omega, t)$ は、周期成分のパワースペクトル（包絡）のみを表していると仮定しておく。

再合成音 $x_S(t)$ は、次式で表される。

$$x_S(t) = \sum_{k=1}^N a_k(t) \sin \left(\int_0^t \omega_k(\lambda) d\lambda + \varphi_k(t) \right) \quad (1)$$

ここで、 N は成分の個数を表す。 $\omega_k(t) = 2\pi f_k(t)$ は、 k 番目の成分の瞬時角周波数を表しており、調波複合音の場合には、基本（瞬時）周波数 $f_0(t)$ の k 倍となる。また、 $a_k(t)$ は k 番目の調波成分の瞬時振幅であり、次式により決められる。

$$a_k(t) = \sqrt{P(k\omega_0(t), t)} \quad (2)$$

2.3 周期成分の拡張：位相の制御

式 1 の位相項 $\varphi_k(t)$ は、cos 位相, sin 位相, 交替位相, ランダム位相の場合には、以下に示すように、時間に依存しない定数項になる。

- **cos 位相**： $\varphi_k(t) = \pi/2$
- **sin 位相**： $\varphi_k(t) = 0$
- **交替位相**： k の偶奇に応じて $\varphi_k(t) = \pi/2$ と $\varphi_k(t) = 0$ を交互に設定
- **ランダム位相**： $\varphi_k(t)$ の値を $(0, 2\pi)$ の区間で一様分布する乱数を用いて設定

これらは、定数項であるため、初期位相として設定すればよい。

Schroeder 位相は、文献 10) で与えられた方法により設定される。Schroeder 位相は、成分の振幅が与えられたときに、合成された波形のピークファクタ（波形の最大値と最小値の差）を最小化する。具体的には、次式により n 番目の成分の位相 ϕ_n を決定する。

$$\phi_n = \phi_1 - 2\pi \sum_{l=1}^n (n-l)p_l \quad (3)$$

$$p_l = \frac{a_l^2}{\sum_{k=1}^N a_k^2} \quad (4)$$

ここで用いられている a_k は、式 2 で設定されたものである。

最小位相は、TANDEM-STRAIGHT とその前身である STRAIGHT で用いられている。各成分の位相 $\varphi_k(t)$ は、以下のように STRAIGHT スペクトル $P_{TST}(\omega, t)$ から求めた cepstrum である $C(q, t)$ を用いて複素 cepstrum である $c(q, t)$ を作り、その逆 Fourier 変換の虚部として求められる。

$$\varphi_k(t) = \Im(S_{mp}(k\omega_0, t)) \quad (5)$$

$$S_{mp}(\omega, t) = \frac{1}{4\pi} \int_{-\infty}^{\infty} c(q, t) e^{j\omega q} dq \quad (6)$$

$$c(q, t) = \begin{cases} 2C(q, t) & (q > 0) \\ C(q, t) & (q = 0) \\ 0 & (q < 0) \end{cases} \quad (7)$$

$$C(q, t) = \int_{-\infty}^{\infty} \log(P_{TST}(\omega, t)) e^{-j\omega q} d\omega \quad (8)$$

ここで q は, quefrequency を表し, 時間の次元を持つ. なお, 式 6 の係数の分母が 2π ではなく 4π となっているのは, 平方根を求めるためである. ($\exp(S_{mp}(\omega, t))$ が, $P_{TST}(\omega, t)$ に対応する最小位相インパルス応答)

2.4 周期成分の拡張：両耳聴

成分の個別制御は, もう一つの自由度を提供する. 両耳への成分毎の割当である. 脳内での左右からの情報の統合を前提とすれば, 偶数次成分と奇数次成分を左右に振り分けることで, 単耳聴では不可能であった周波数分解能の劣化を回避できる可能性がある. あるいは, 左右で知覚される基本周波数を同一にするために, 隣接する 2 つの調波成分を単位として, その添字の偶奇性にもとづいて左右に振り分ける等, より複雑な設定を用いることもできる.

2.5 非調波複合音への拡張

拡張には, 式 1 の成分の周波数 $\omega_k(t)$ の設定において $k\omega_0(t)$ という制約を外す. この場合, $P_{TST}(\omega, t)$ をフィルタ特性を設定するものと解釈するか, 信号のパワーの分布を設定するものと解釈するかによって, 以下のように $a_k(t)$ の定義が異なる.

- フィルタ特性と解釈する場合.

$$a_k(t) = \sqrt{P(\omega_k(t), t)} \quad (9)$$

- パワー分布を設定するものと解釈する場合.

$$a_k(t) = \arg \min_{a_k} \int |R(\omega, t)|^2 d\omega \quad (10)$$

被積分関数 $R(\omega, t)$ は, 次式で定義される.

$$\int_0^\omega \sum_{k=1}^N \delta(\omega - \omega_k) a_k^2 - P_{TST}(\lambda, t) d\lambda \quad (11)$$

ここで, $\delta(\omega)$ はデルタ関数を表す.

2.6 実装

デモシステムは, TANDEM-STRAIGHT と同様に, 科学技術計算用の環境である Matlab を用いて実装されている. TANDEM-STRAIGHT の既定値では, 分析は 5 ms 周期で更新されるフレーム毎に行われる. 成分の瞬時周波数 $\omega_k(t)$ と位相 $\varphi_k(t)$ は, フレーム時刻における値に基づいて, 標本化周期毎に線形補間された値が求められる. 調波複合音の場合には, 線形補間された $\omega_0(t)$ を用いて, $k\omega_0(t)$ として k 番目の成分の瞬時周波数が求められる. 式 1 の積分には, Matlab の累積和関数 (cumsum) を用いた.

3. 数値例

ここでは, 幾つかの条件で, システムの動作例を示す. 図 1 に, 成分が等振幅の場合の波形の例を示す.

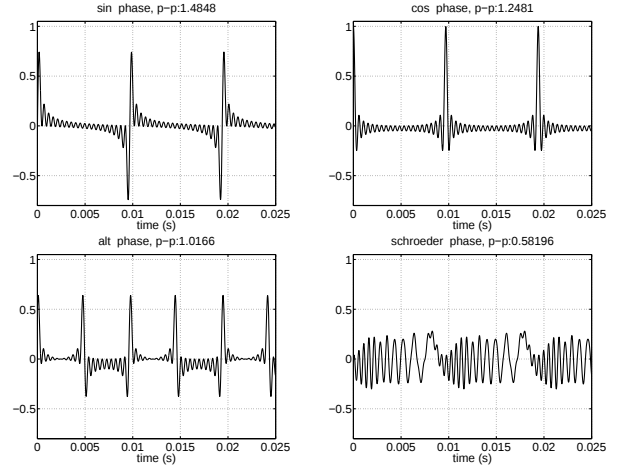


図 1 初期位相による波形の違い

Fig. 1 Waveform with different initial phase conditions

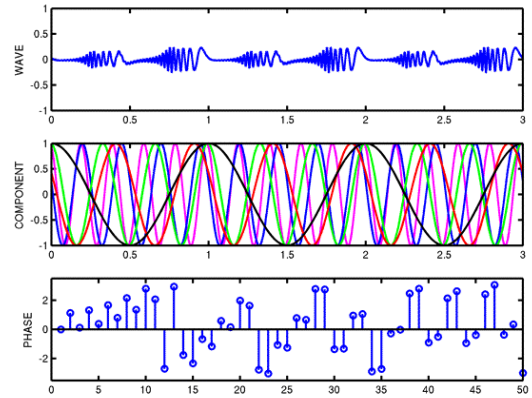


図 2 Schroeder 位相から交替位相へのモーフィング

Fig. 2 Snapshot of morphing from Schroeder phase to alternating phase

成分の個数は, この例では $N = 20$ とした. プロットは左右上下の順に, sin 位相, cos 位相, 交替位相, Schroeder 位相の場合を示す. それぞれの図の上には, ピークファクタとして p-p 値が付記されている. sin 位相が最大のピークファクタを示し, Schroeder 位相が最小となっている. 両者の比は 2.55 倍になる.

図 2 に, Schroeder 位相から交替位相へのモーフィングの途中の状態を示す. ここでは, 50 個の成分を用いている. 図 2 の最上段は, 合成された波形, 中段は各成分, 最下段は, それぞれの成分の位相を表している. (試聴できるデモムービーが用意されている.)

図 3 に, デモシステムにより実際の音声进行分析し再合成した波形の一部を示す. 音声は, 男性の発声した日本語母音連鎖/aiueo/であり, 標本化周波数は 22,050 Hz である. 図 3 の中のプロットは, 左右上中下の順に, sin 位相, cos 位相, 交替位相, Schroeder 位相, 最小位相, 原音声波形を示す. これらの信号のパワーは, 同一になるように正規化されている. 最大

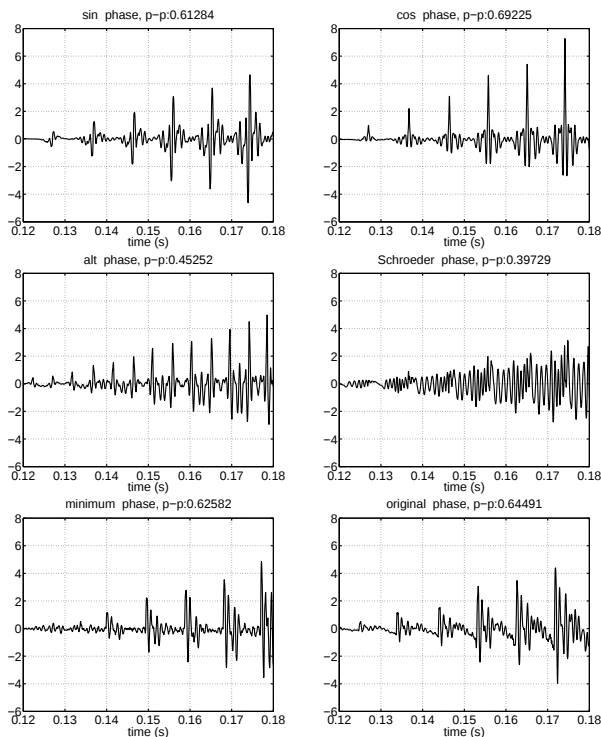


図3 初期位相による波形の違い (音声)

Fig. 3 Waveform with different initial phase conditions

のピークファクタは、cos 位相の場合に得られており、最小は等振幅の場合と同様に Schroeder 位相である。最大と最小のピークファクタの比は、1.74 倍となった。これらを比較すると、原音声は最小位相に近い性質を持つことが分かる。実際、原音声の波形では、高域では主要なエネルギーが声門閉止において供給され、低域では、声門を通過する体積流波形の影響が大きいことが知られている。

4. ま と め

これまで注目されることの少なかった、波形の相対位相による『音の肌触り』の操作を体験できるデモシステムを作成した。システムは音声分析変換合成システム TANDEM-STRAIGHT の合成部分の拡張として実装した。会場では、様々な音響トランスデューサーによる違いも併せて体験してもらう予定である。

謝辞 本研究の一部は、科学研究費補助金基盤(A)19200017 および科学技術振興財団の CrestMuse プロジェクトによる支援を受けた。

参 考 文 献

- 1) Plomp, R. and Steeneken, H. J. M.: Effect of phase on the timbre of complex tones, *J. Acoust. Soc. Am.*, Vol.46, No.2B, pp.409–421 (1969).

- 2) Patterson, R. D.: A pulse ribbon model of monaural phase perception, *J. Acoust. Soc. Am.*, Vol.82, No.5, pp.1560–1586 (1987).
- 3) 河原英紀, 津崎 実, Patterson, R.D.: オールパスフィルタの位相操作による時間構造制御とその知覚への影響について, 日本音響学会聴覚研究会資料, Vol.H-96–79 (1996).
- 4) Skoglund, J. and Kleijn, W. B.: On time-frequency masking in voiced speech, *IEEE Trans. Speech and Audio Processing*, Vol. 8, No.4, pp.361–369 (2000).
- 5) Uppenkamp, S., Fobel, S. and Patterson, R.D.: The effect of temporal asymmetry on the detection and perception of short chirp, *Hearing Research*, Vol.158, No.1-2, pp.71–83 (2001).
- 6) 坂野秀樹, 武田一哉, 板倉文忠: 群遅延に基づく位相の聴覚的距離尺度, 信学技報・音声, Vol.SP2001, No.64, pp.23–29 (2000).
- 7) Kawahara, H., Masuda, I. and de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Communication*, Vol.27, No.3-4, pp.187–207 (1999).
- 8) 松原貴司, 森勢将雅, 西浦敬信: 高品質音声合成における有声音の位相特性が知覚に与える影響, 日本音響学会聴覚研究会資料, Vol.40, No.8, pp.659–664 (2010).
- 9) Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation, *Proc. ICASSP 2008*, Las Vegas, pp.3933–3936 (2008).
- 10) Schroeder, M.R.: Synthesis of low-peak-factor signals and binary sequences with low autocorrelation, *IEEE Trans. Information Theory*, Vol.16, No.1, pp.85–89 (1970).
- 11) Kawahara, H., Banno, H., Irino, H. and Zolfaghari, P.: Algorithm AMALGAM: morphing waveform based methods, sinusoidal models and STRAIGHT, *Proc. ICASSP 2004*, Montreal, pp.13–16 (2004).
- 12) Dudley, H.: Remaking speech, *J. Acoust. Soc. Am.*, Vol.11, No.2, pp.169–177 (1939).
- 13) 森勢将雅, 高橋 徹, 河原英紀, 入野俊夫: 窓関数による分析時刻の影響を受けにくい周期信号のパワースペクトル推定法, 電子情報通信学会論文誌D, Vol.J 90-D, No.12, pp.3265–3267 (2007).
- 14) Unser, M.: Sampling – 50 years after Shannon, *Proceedings of the IEEE*, Vol.88, No.4, pp.569–587 (2000).