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.

音響信号の非現実的ミックスの検出

薗田光太郎†

† 長崎大学 大学院 工学研究科
〒852–8521 長崎県長崎市文教町 1–14
E-mail: †sonoda@cis.nagasaki-u.ac.jp

あらまし　近年，音の模擬合成手法の発展により，人工的な模擬合成音は，現実の録音そのものであると聴覚的に認識されてもおかしくないほどとなりつつある．しかし，聴覚的には自然に感覚されても，現実の録音ではあり得ない波形に合成される場合もある．楽曲制作で一般的に用いられるマルチトラックレコーディングとミキシング過程では，異なる環境で収録された複数の音源を別の統一された環境下に配置合成される．既報では音源の再配置における位相パンニング処理の有無の検出手法について検討をおこない，左右チャネル間群遅延差の標準偏差が極端に低い場合に位相パンニング処理があったことを疑われることを示した．今回の報告ではマルチトラックミキシングを考慮した手法を検討する．

キーワード　非現実的ミックス，　なりすまし検知, ライブネス検知, ケプストラム

Study on detection of mixing unreality for PCM audio signal

Kotaro SONODA†

† Graduate School of Engineering, Nagasaki University
Bunkyo-machi 1–14, Nagasaki, Nagasaki 852-8521, Japan
E-mail: †sonoda@cis.nagasaki-u.ac.jp

Abstract　In recent years, due to the development of simulated sound synthesis techniques, artificial simulated sounds have become to be perceived as real actual recordings by human auditory. However, even if the artificial sound is perceived as natural or real by human auditory system, the waveform of sound may be synthesized as that cannot be recorded in a real recording. In practical music production, the multi-track recording and mixing process are commonly used and the multiple sound sources recorded in different environments are arranged and synthesized under another unified environment by the digital audio workstation. In the previous report, we have proposed the detection method of the presence or absence of the phase panning process in the given directional sound source, and showed that the phase panning process was suspected when the standard deviation of the group delay difference between the left and right channels (Inter-channel Group-Delay difference) was extremely low. In this report, we consider a method considering multi-track mixing.

Key words　Unreal mixing, Spoofing detection, Liveness detection, Cepstrum

1. はじめに

現代の音楽制作現場では，計算機の利用により簡単にデジタル音響加工を行うことが可能となった．商用の楽曲メディアでは，現実に直接録音された音に対して，残響処理や各種エフェクトを施し，さらにそれらを適切な時刻，到来方向に配置するなどを行うのが一般的になっている．また，深層学習の導入により，簡単に人間には現実の音と聞き分けることができない高品質な模擬音を作成し，音源とすることも可能である．このような音源・音場の模擬は，人間による聴取においては問題にならないが，真正性の担保が必要とされるシステムにおいては大きな問題となる可能性がある．一旦パッケージ化された信号についてその真正性を担保するには，真正の配信者が介在し，恣意的に電子透かしを埋め込んだり，フィンガープリントを記憶するなどによりそれらを照合することで解決できるが，パッケージされる前の直接録音されたものである真正性を担保することにはならない．事実，音声による個人認証・照合システム (ASV: Automatic speaker verification) で

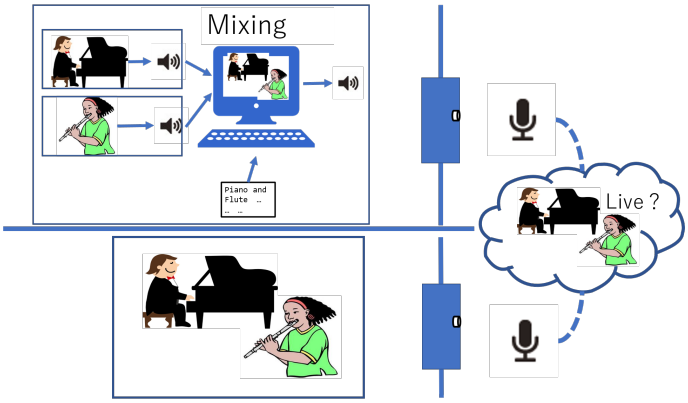

図 1　Anti-spoofing: 手元の録音はライブ音かなりすまし加工音か.

は，このような録音再生音（リプレイ音声）やディープフェイク音声などによる詐称がシステムの判断を誤らせる障害となることが取りざたされている，　音声以外でも例えば，工場の異常音を模擬した音が作成され，再生された場合，労働者に誤った判断をさせてしまう．

このような直接録音・現実音場の模擬加工の検知 (Anti-spoofing は，音声その他のバイオメトリクス認証においては生体検知 (Liveness detection / PAD: presentation attack detection [1]) として考えられ，ASV-PAD についてはその検知性能を競う ASVspoof challenge [2]〜[4] が開催されている．

2.　チャネル間群遅延差によるパンニング加工の検出．モノラル音の擬似ステレオ化

本章では，既報 [5] の IGDD (チャネル間群遅延差; Inter-channel Group-Delay Difference) について概要を述べる．いま，検出対象とする信号は 2ch ステレオ信号とする．この信号が，実在する音源に対してその場で録音された信号（ライブ信号）なのか，ある 1 チャネル音源信号をステレオパンニングで擬似ステレオ化された信号（加工信号）なのかを検出したい．既報では，加工信号ではチャネル間の群遅延差 Inter-channel Group-Delay Difference (以下 IGDD)　すなわち到来角度差が信号全体の中で変化しにくいと考えられることに着目した．対してライブ信号では，各チャネルの信号が録音時の音源ー各センサ間で決定され，チャネル間の到来角度差は加工音に比べ変化しやすいと考えられる．この考えに基づき，以下の式 (1) の指標をライブネス度（ライブ信号らしさ）とした．

$$\text{liveness}_{\text{IGDD}} = \text{stdv}_n \left[\text{ave}_k[\text{IGDD}(k,n)]\right] \tag{1}$$

新たにスタジオ録音したライブ信号 (A)，それをステレオパンニングした加工信号 (B)，パッケージ化されたクラシック音楽信号（ワントラックと想定）(C)，ポップス音楽信号（マルチトラックと想定）(D) について評価した結果，加工信号である (B)(D) の評価値は，ライブ信号である (A)(C) に比べ大幅に低い値をとることが確認された．

音声の生体検知性能を競う ASVSpoof における生体発話/スピーカ再生の識別においては，検出対象を 2ch ステレオ信号した場合に，マイク間（チャネル間）の到来時間差に基づく手法が多く提案されている．実発話の場合は，発話位置が口内で前後するため，左右チャネルへの到来時間の差が変化しやすく，対してスピーカ再生の場合には発音位置が変化しないためチャネル間の信号の差は時間差が優位をしめる．したがって，スピーカ再生のに対する録音信号で，チャネル間到来時間差を求める指標であるチャネル間 GCC(Generalized Cross-Correlation) [6] が高くなりやすいことを利用している．式 (2) に時間系列 $x(t)$ と $y(t+\tau)$ の $GCC(\tau)$ を示す．

$$GCC(\tau) = \mathcal{F}^{-1}\left[X(\omega)Y(\omega)^* \mathrm{e}^{j\tau\omega}\Phi(\omega)\right] \tag{2}$$

ここで，$X(\omega)$ および $Y(\omega)$ は $x(t)$ および $y(t)$ の Fourier 変換であり，$\Phi(\omega)$ は前 2 項で構成されるクロススペクトルに対する一般化周波数フィルタである．$\Phi(\omega)$ をクロススペクトル絶対振幅値とすれば式 (2) は白色化クロススペクトルの逆 Fourier 変換と考えられる (GCC-PHAT;GCC-PHAse Transform)．

さらに矢野・塩田らは，音声の生体検知でありながら，発話中に現れる無声区間の背景音の GCC に着目し，背景音においても，実録音（ライブ録音）に比べ大幅にスピーカ再生音の GCC が高くなることを利用し，識別性能を向上させている [7]．GCC は仮定する到来時間差 τ としたときの相関係数であり，我々のチャネル間群遅延差の散らばりが少ないことと同様の考え方に基づく．

しかし，これらの手法は，エコー・リバーブ環境下では真の到来時間差・群遅延差を計測できず，評価値の性能が悪い．

3.　マルチトラックミキシングの検出

ASVspoof で対象としているスピーカ再生音声は，実発話録音音声のスピーカ再生を再録音したものであり．再録音された信号は再生信号（実発話録音）と再生再録音時の背景音との合成および，再生再録音時の室内伝達関数との畳み込みと考えられる．元の再生信号じたいも実発話録音時の室内伝達関数との畳み込みであるため，1 個の信号全体の中で，録音環境（室内伝達関数）が変化することとなる．矢野・塩田らの無声区間 GCC に着目した手法は，無声区間では実発話録音時の環境が現れず，スピーカ再生環境のみとなるため性能向上したと考えられる．

一般の音響信号の場合，ライブ信号は，多数の方向から到来する実発音を一挙録音したものとする，一方で加工音は個々の楽器音の録音（マルチトラック）を加工合成（ミキシング）したものとする．ライブ信号がある特定の環境で一貫して録音されたものであるのに対し，加工音は録音環境の異なる音源が加算・時刻配置されたものである．

そこで，チャネル間ではなく，隣接する時間フレーム間での録音環境の急激な変化をとらえることとする．前節の到来方向も録音環境と考えることができるが，マルチトラックミキシングを捉えることはできない．ここでは，残響特性に着目することとする．

残響特性は，音源に対するフィルタ（畳み込み処理）であり，観測信号のケプストラム上では加算の関係になる．観測した

信号を $y(t)$ とすると，$y(t)$ は式 (3) のように音源信号 $s(t)$ と残響特性を $h(t)$ の畳み込み演算（積和演算）でモデル化でき，Fourier 変換ドメインのスペクトルでは，式 (4) のようにアダマール積となる．この両辺に対して対数をとったものは加算の関係となり，その Fourier 変換（すなわちケプストラム）でも式 (5) のように加算となる．

$$y(t) = s(t) * h(t) \tag{3}$$

$$Y(\omega) = S(\omega) \cdot H(\omega) \tag{4}$$

$$\mathcal{Y}(q) = \mathcal{S}(q) + \mathcal{H}(q) \tag{5}$$

ここで $Y(\omega), S(\omega), H(\omega)$ は角周波数 ω におけるスペクトル，$\mathcal{Y}(q), \mathcal{S}(q), \mathcal{H}(q)$ は quefrency:q におけるケプストルとする．

音源信号は逐時変化するが，残響特性は環境が変化しない限り大きくは変化しない [8] と考えられるので，仮に実録音時の残響特性のケプストラム $\mathcal{H}_{\varnothing}(q)$ を得ることができれば，観測信号のケプストラム $\mathcal{Y}_n(q)$ とのケフレンシードメインの相互相関 $\mathcal{C}_{\varnothing}(\tau)$ は，観測時の残響特性のケプストラム $\mathcal{H}(q)$ との相互相関を強く反映したものとなることが予想される．ただし，相互相関を求めるために，ケプストラムは位相特性を考慮しない振幅ケプストラムの実部とする．

フレーム n のケプストラムを $\mathcal{Y}_n(q)$ とすると，このフレームに関する正規化相互相関係数 $\mathcal{C}_{\varnothing}(n)$ は式 (6) で求められる．

$$\mathcal{C}_{\varnothing}(n) = \frac{\sum_q \mathcal{Y}_n(q)\mathcal{H}_{\varnothing}(q)}{\sqrt{\sum_q \mathcal{Y}_n(q)^2 \sum_q \mathcal{H}_{\varnothing}(q)^2}} \tag{6}$$

実録音を示す参照信号 $\mathcal{H}_0$ は，信号中で最もパワーの低いフレーム（無音区間）で得られる短ケフレンシーケプストラムを採用する．

$\mathcal{C}_{\varnothing}(n)$ がある閾値 $\rfloor_{\varnothing}$ を下回るとき，フレーム n は，実録音とは異なる環境のトラックがミックスされたことを意味すると考える．

4. 評価実験

評価音源として，新たにスタジオ録音したギター演奏#live，騒音データベース [9] から#env，時間方向に#live と#env を連結した#concat, #env の音源の途中に#live を加算した#mix, RWC 研究者用音楽データベース [10], [11] からポピュラー音楽#pops，ジャズ音楽#jazz，を用意した．いずれも 1 チャネル音源とし，フレーム内で量子化振幅値がすべて 0 となる区間は現れないものとする．

#live および #env　は加工されていない信号（ライブ信号）と考える．#concat および#mix は環境の異なるライブ信号を合成したものである．#pops はマルチトラックミキシングが行われており，#jazz は明記されていないが一般に一挙録音されたものと考える．

図 2 から図 7 に各信号の信号波形と対応する時刻のケプストラム相互相関係数の変化を示す．信号の合成のない#live(図 2)，#env(図 3) ではある一定の範囲内で変動しているが，ミキシングのあった#concat(図 4) および#mix(図 5) では連結

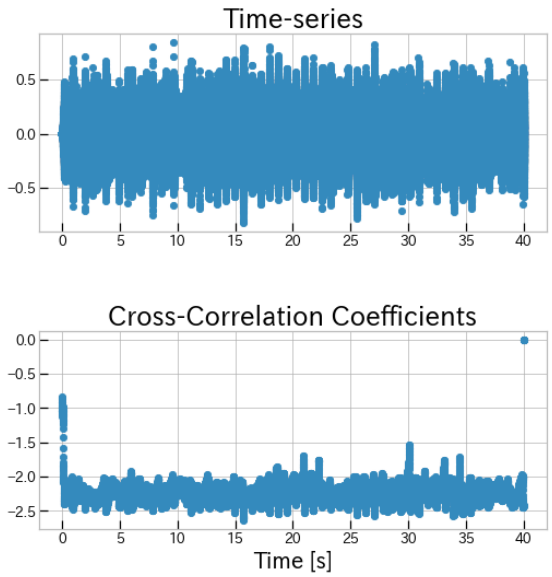

図 2　#live

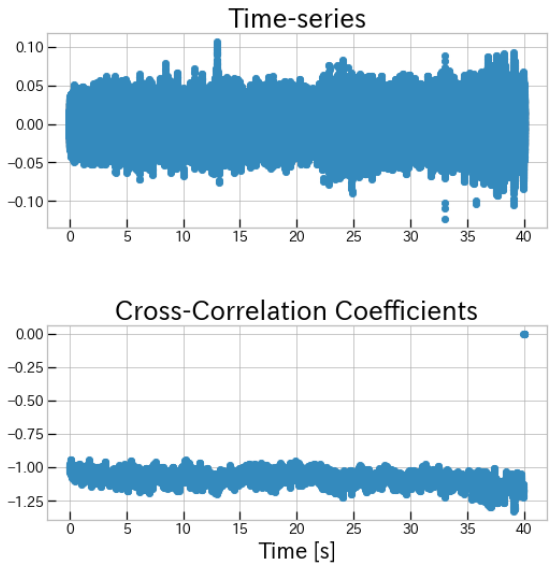

図 3　#enc

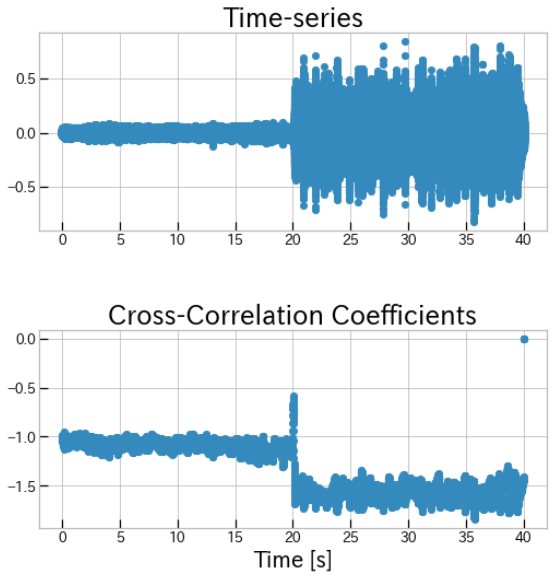

図 4　#concat

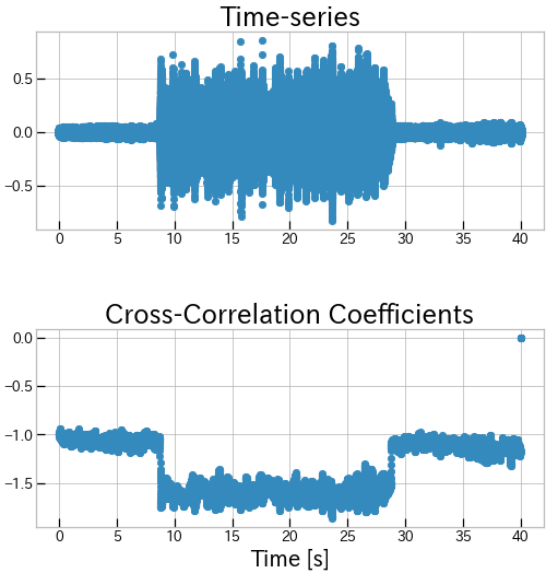

図 5　#mix

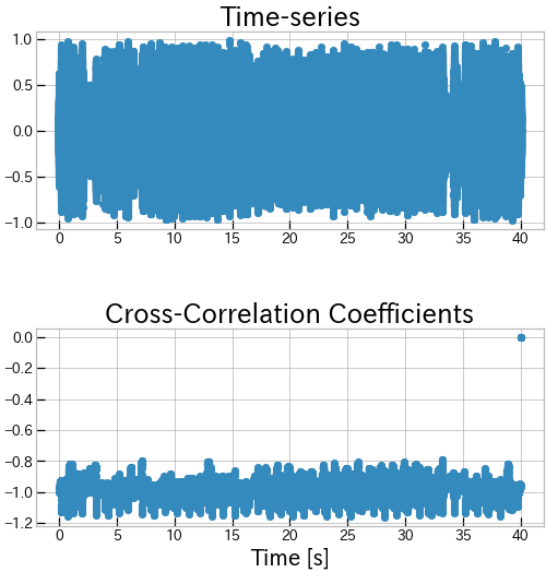

図 6　#pops

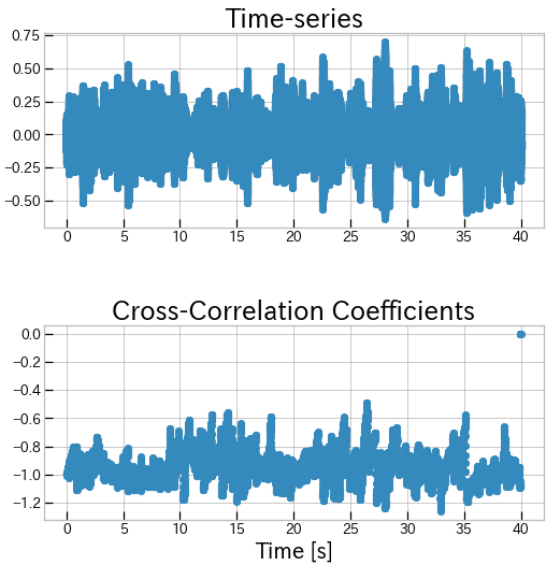

図 7　#jazz

されたフレームや合成がなされたフレームで大幅に値が変化していることが見てとれる．

5. 結　　論

本報告では，短ケフレンシー領域のケプストラムを録音環境と考え，信号中で最もパワーの低いフレームで得られる短ケフレンシーケプストラムを実録音環境とした上で，信号中の各フレームでのケプストラム相互相関係数の変化を計測し，急峻に変化するフレームを非現実的ミックスフレームとした．生収録が既知である音楽信号，環境騒音に対しては一定の範囲内での変動であるのに対し，ミキシングを行った信号では連結フレーム，または合成フレームにおいて大幅な変動があったことから，ケプストラムの相互相関により音源の背景に存在する音環境の移り変わりを捉えることができる可能性が示唆された．ただし，音源と環境の分離は不十分であり，音源の変化の影響も少なからずあることが考えられる．今後は，その精査と，ASVSpoof で利用されるリプレイアタック検知用評価音源データベースなど信頼できる評価音源に対する評価を行う．

文　　献

[1] ISO Central Secretary, “Information technology — biometric presentation attack detection,” Standard ISO/IEC 30107-1:2016, International Organization for Standardization, Geneva, CH, 2016. https://www.iso.org/standard/53227.html

[2] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, “Asvspoof: the automatic speaker verification spoofing and countermeasures challenge,” IEEE Journal on Selected Topics in Signal Processing, vol.11, no.4, pp.588–604, 2017.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” Interspeech 2015, pp.2037–2041, Sept. 2015.

[4] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.A. Lee, “The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” Interspeech 2017, pp.2–6, 2017.

[5] 黒田康弘，薗田光太郎，喜安千弥，“音のパンニング加工の検出に関する一検討，” 信学技報，vol.118，no.494，pp.320–327，March 2019. EMM2018-96.

[6] C.H. Knapp and G.C. Carter, “The generalized correlation method for estimation of time delay,” IEEE Transaction on Acoustics, Speech, and Signal Processing, vol.24, no.4, pp.320–327, Aug. 1976.

[7] 矢野凌也，塩田さやか，小野順貴，貴家仁志，“複数チャネル間の相互相関関数を用いた話者照合のためのなりすまし検出，” 日本音響学会 2018 年秋季研究発表会 講演論文集，pp.1335–1338，Sept. 2018.

[8] 榎本祐太，及川靖広，山崎芳男，“準同型処理を用いた室内伝達特性の抽出，” 日本音響学会 2014 年春季研究発表会 講演論文集，pp.741–742，March 2014.

[9] 社団法人 電子情報技術産業協会音声入出力方式標準化委員会，“電子協 騒音データベース JEIDA-NOISE”. http://www.sunrisemusic.co.jp/database/fl/noisedata01_fl.html

[10] 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，” 情報処理学会研究報告，vol.42，pp.35–42，oct 2001. https://ci.nii.ac.jp/naid/110002935763/

[11] 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：クラシック音楽データベースとジャズ音楽データベース，” 情報処理学会研究報告，vol.44，pp.25–32，feb 2002. https://ci.nii.ac.jp/naid/110002935774/