瞬時周波数に基づく非混合度評価指標を導入したステレオ音源分離

Stereo audio source separation with disjointness map based on instantaneous frequency

1w143114-4　蛭間 涼　　指導教員　及川 靖広　教授
HIRUMA Atsushi　　Prof. OIKAWA Yasuhiro

概要：ステレオ音源分離はステレオ音源から目的の音源のみを強調，抑圧するといった，音源の編集を行う際に必要な信号処理である．ただし，一般的にステレオ音楽信号はモノラル音源をチャンネルごとにミキシングしたものであり，各音源のチャンネル間に位相差がなく，位相情報を用いた手法による分離を行うことは難しい．しかしながら，スペクトログラムの位相にも各音源に関する情報が含まれており，これを考慮することで分離精度の向上が期待できる．本研究ではステレオ音源分離に位相情報を考慮するために位相情報を用いて時間周波数ビンの非混合度を評価する指標を提案する．位相差ではなく瞬時周波数を用いることで，スペクトログラムの構造情報を考慮した指標が得られ，これを考慮したバイナリマスキングを行うことで分離精度の向上がみられた．

キーワード：時間周波数マスキング, 瞬時周波数, convex clustering, グラフ隣接行列, 重み付き無向グラフ
Keywords: time-frequency masking, instantaneous frequency, convex clustering, graph adjacency matrix, weighted undirected graph.

1. ま え が き

ステレオ音源分離は分離に用いることができる位相の情報が少ないため，レベル比を分離情報として用いる手法が検討されてきたが [1]，位相自体にも音源情報が含まれており，これを考慮することで分離精度の向上が期待できる．本研究では位相情報を用いた非混合度評価指標を提案し，これを分離に用いるために convex clustering を導入した．評価指標に加え，音の時間連続性を考慮し，作成したステレオ音楽信号の分離を行った．

2. 瞬時周波数に基づく非混合度評価指標

ステレオ音楽信号は時間周波数領域で各チャンネルごとにモノラル音源に実数係数をかけることで作成される．この場合，チャンネル間で振幅には差が生まれるが，位相は変化しないので，各チャンネルの位相スペクトログラムは同じはずである．しかし，位相スペクトログラムは音源が混ざると必ずしもチャンネル間で値が等しくなるとは限らない．従って，時間周波数ビンのチャンネル間の位相差が 0 に近いかどうかで非混合度の評価ができると考えられる．

評価の単純な手法としてチャンネル間の位相を比較することを考える．各チャンネルのスペクトログラムの位相を $\varphi_\alpha[t,f] = \mathrm{Arg}(X_\alpha[t,f])$ とする．ただし Arg は複素偏角，X_α ($\alpha \in \{1,2\}$) は各チャンネルのスペクトログラム (1 が左，2 が右チャンネル) を表す．位相差が 0 である時間周波数ビンの評価を最大値 1 とするために，例として評価式を

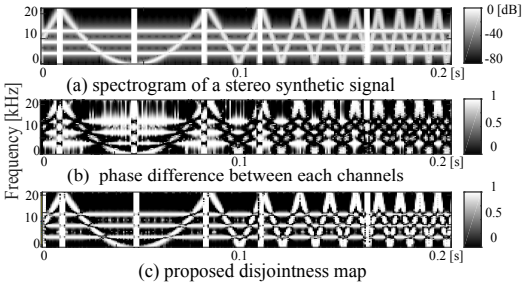
$$\exp(-\zeta|\varphi_1 - \varphi_2|) \qquad (1)$$

とする．ただし，ζ は任意値である．(1) は音源の構造的な情報を考慮していないため，図–1 (b) に示すように音源が混ざった時間周波数ビンが密集すると，音源が一つで構成されている時間周波数ビンの評価も低下してしまい，評価の都合がよくない．

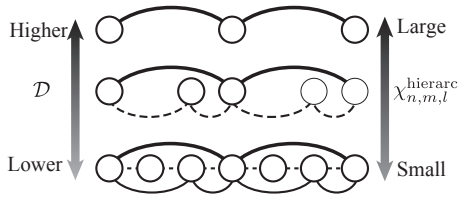
そこで，瞬時周波数 [2] の比較に基づく評価指標を提案する．瞬時周波数は，位相の時間微分 $\varphi'_\alpha = \partial\varphi_\alpha/\partial t$ で定義される値であり，位相の時間変化を考慮するので，構造情報を含む評価指標が作成できる．瞬時周波数は位相同様に音源が混ざることに敏感な性質をもつので，時間周波数ビンが一つの音源から構成されているならば φ'_1 と φ'_2 は一致するはずである．ただし，その絶対値は中心周波数から離れるほど大きくなる．また，瞬時周波数の値が小さい時間周波数ビンは中心周波数付近の時間周波数ビンであると考えられるので，チャンネル間のずれを許容することが求められる．従って，瞬時周波数の値，およびチャンネル間のずれの許容範囲を調整できる評価式を用いることが望ましい．本論文では，各チャンネルの瞬時周波数の差に基づく非混合度評価指標を楕円の方程式を用いて

$$\mathcal{D} = \exp(-\{a(\varphi_1'^2 + \varphi_2'^2) - 2b\varphi'_1\varphi'_2\}) \qquad (2)$$

と表す．ただし，$a = (r_1^2 + r_2^2)/(2r_1^2r_2^2)$，$b = (r_1^2 - r_2^2)/(2r_1^2r_2^2)$ とし，r_1，r_2 はそれぞれ楕円の長辺と短辺を表す．(2) の 2 次式は $\varphi'_1 = \varphi'_2$ を長辺に取る楕円の方程式であるので，その長辺は瞬時周波数の値に対応し，短辺はチャンネル間のずれに対応する．$\mathcal{D} \in [0,1]$ は φ'_α が 0 に近いほど 1 となり，値の大きさとチャンネル間の差が大きくなるほど 0 に近づく．$\mathcal{D}$ による評価指標を図–1 (c) に示す．図–1 (c) は図–1 (b) と比べて評価の低い時間周波数ビンが密集している中でも単一音源で構成され



図–1 位相差に基づく非混合度評価指標と瞬時周波数に基づく提案評価指標の比較



図–2 評価指標に基づく path graph リンク

る時間周波数ビンの評価が高い指標であることがわかる

3. バイナリマスキングへの導入

評価指標 $\mathcal{D}$ を考慮したバイナリマスキングを行うために，convex clustering を用いてバイナリマスクを作成した [3]．convex clustering は導入する情報に合わせて点同士をつなぐ重み付きグラフリンクを作成することで自由度の高いクラスタリングを行うことができる．本研究では $\mathcal{D}$ を導入することでグラフリンクに優先度を与えることを考える．隣接点同士をつなぐ path graph リンクは一つの音源で構成される時間周波数ビンのレベル比は各音源ごとに似た値をとるという考えに基づくものであり [3]，これを階層的に与えることで $\mathcal{D}$ が高い点同士のリンク強度を高めることができる．階層 path graph リンクを**図–2** に示す．ただし，線の太さはリンク強度を表す．$\mathcal{D}$ の値に閾値を設け，作成した各階層ごとに path graph リンク [3] の重みを調整するパラメータを掛けることでリンク強度を調整できる．クラスタリングに導入する階層 path graph リンクは全階層の和によって与えられる．

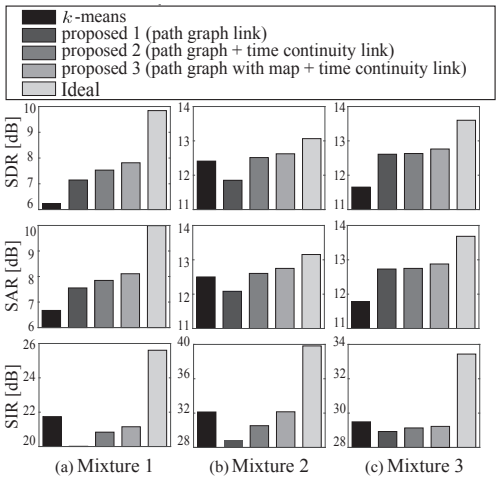
4. 数値シミュレーション

表–1 に各ステレオ音楽信号の構成音源，混合比を示す．ただし，構成音源は BASS-dB データベース (注1) よりダウンロードした 3 つのトラック (Mix 1 : Anabelle Lee, Mix 2 : Wreck，Mix 3 : Life as a distributed infobeing) から 3 つのモノラル音源を選択した．ステレオ音楽信号に対し k-means と convex clustering 3 手法 (「path graph」,「path graph+時間連続」,「時間連続+非混合度評価指標」) の分離結果を比較した．ただし path graph，時間連続リンクは [3] を用いた．分離結果は SDR，SAR，SIR で評価した．

(注1)：http://bass-db.gforge.inria.fr/BASS-dB/

表–1 実験に用いた音源と混合比．

Mixture 1	Electric guitar 1	Acoustic guitar	Electric guitar 2
Mixture 2	Bass	Drums	Electric guitar
Mixture 3	Keyboards	Drums	Acoustic Guitar
Left	0.25	0.5	0.75
Right	0.75	0.5	0.25



図–3 実験結果 (a) Mixture 1 (b) Mixture 2 (c) Mixture 3

図–3 に実験結果を示す．ただし，理想マスクは測定データから作成した．時間連続リンクを用いることで SDR，SAR は従来の k-means クラスタリングよりも高い数値を得ることができたが，SIR には低下がみられた．特に図–3(b) の SIR には差が大きくみられることから，ドラムが混ざる音源には時間連続を考慮するだけでは先験情報が不十分であると考えられる．また，マップに基づくリンクを与えることで，時間連続性のみ考慮した結果よりも全評価に改善がみられた．

5. むすび

本研究では瞬時周波数に基づく各時間周波数ビンの非混合度評価指標を提案した．また提案の有効性を確かめるために，音源分離手法として convex clustering を導入した．瞬時周波数を用いることで構造情報を考慮した評価指標が作成でき，これを分離に用いることで精度に改善がみられた．今後は評価指標の安定性の追求及び応用先について検討する．

参考文献

[1] S. Araki, H. Sawada, R. Mukai and S. Makino. “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” Signal Process., vol. 87, no. 8, pp. 1833–1847, 2007.

[2] S. A. Fulop and K. Fitz. “Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications,” J. Acoust. Soc. Am., vol. 119, no. 1, pp. 360–371, 2006.

[3] 蛭間涼，矢田部浩平，及川靖広，“凸最適化を用いた重み付き方位クラスタリング”，音講論集， pp. 473–474, Sep. 2017.