

声と背景音のボリュームバランス調整に向けた音源分離

西山 正志^{†1} 広畑 誠^{†1} 小野 利幸^{†1}

本稿では、音の入力信号から背景音情報をリアルタイムに学習することで、人の声と背景音の音源信号を推定する、遅延なし音源分離手法を紹介する。本手法をデジタル映像機器に適用することで、映像コンテンツを視聴している際に、人の声を聴き取りやすくする、背景音を静かにする、スポーツの臨場感を感じやすくする、歌の練習に用いるといった新しい機能をユーザに提供することを狙う。実際に、遅延なし動作で端末を選ばないため、処理が軽く、効果をその場で体感しながら視聴できる効果を確認した。また、声と背景音を個別に学習するため、多様なコンテンツで声と背景音を精度よく推定できる効果を確認した。

1. はじめに

様々なデジタル映像機器の普及に伴い、映像コンテンツの視聴方法が多様化し、かつ手軽に行えるようになった。しかし、人の声が音楽など背景音に埋もれて聴き取りにくい場合や、歓声など背景音が小さく臨場感を楽しめない場合など、必ずしも快適な視聴ができるとは限らなかった。そのため、人の声や背景音の強調など、音源毎のボリューム調整を目的とした高音質化機能への期待が高まっている。しかし、従来の高音質化機能では、映像シーンや音声の収録方式など、入力信号の特性が変わると、音源信号の推定が不十分で、音質が劣化するという問題があった。

そこで、本稿では、音の入力信号から背景音情報を学習、更新しながら人の声と背景音の音源信号を推定する、遅延の生じない音源分離手法を紹介する。図1に開発した手法の概要を示す。この手法はデジタル映像機器全般で汎用的に活用することが考えられ、以下にその可能性の具体例を列挙する。

本手法は処理が軽く、端末に依存しないため、効果をその場で体感しながら視聴できるといったことが考えられる。また、声と背景音を個別に逐次学習するため、シーンや収録方式によらず多様なコンテンツへ適用することが考えられる。更に、音楽、歓声、雑音など様々な背景音へ柔軟に対応できるため、ホームビデオのようなアマチュアが作成したコンテンツに対しても有効であると考えられる。すなわち、多様なコンテンツに対して効果を汎用的に発揮できる可能性がある。

また、本手法をデジタル映像機器に適用することで、映像コンテンツを視聴している際に、人の声を聴き取りやすくできる、背景音を静かにできる、スポーツの臨場感を高められる、歌の練習ができるといった新しい機能をユーザに提供できるポテンシャルを秘めている。

提案手法の詳細^[6]は発表済みであり、本稿では、その応用の可能性と技術の概要を紹介する。

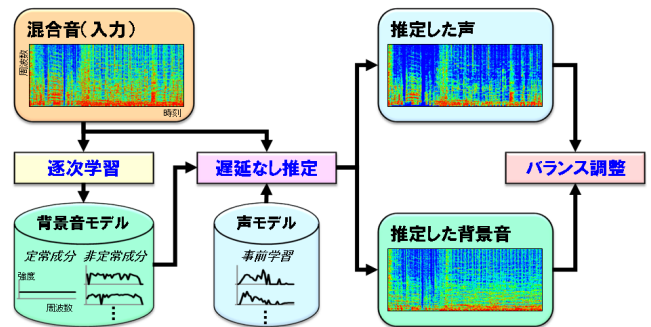


Fig. 1 Overview of proposed method

図1 提案手法の概要

2. 問題設定

人の声を強調できるイコライザ機能や、背景音を抽出・強調できるボーカル抑制機能^[1]など、音源毎のボリューム調整を目的とした高音質化機能は従来から広く開発されている。しかし、従来のイコライザ機能は、予め定めた声が含まれやすい帯域は強調し、声が含まれにくい帯域は抑制する制御を行うため、シーンによって声が十分に強調できない場合や、背景音が抑制できない場合がある。また、従来のボーカル抑制機能は、L Rチャンネルのセンター成分（同一位相、同一振幅）に記録された人の声を抑制するため、センター成分に記録された背景音も抑制してしまう難点がある。このように、従来手法では、シーンや音声の収録方式など、入力信号の特性が変わると、音源信号の推定が不十分で効果が得られない問題があった。

以下では、上記の問題点を解決するために、入力信号から背景音情報を更新しながら人の声と背景音の音源信号を逐次推定する遅延なし音源分離手法を紹介する。本手法は、イコライザ機能やボーカル抑制機能に比べ、音源信号の推定が高精度に行える効果がある。

3. 音源分離技術

本手法は、背景音成分を一定時間毎に入力信号から学習、更新することで、入力信号の特性変化に対応する。背景音

^{†1}(株)東芝 研究開発センター
TOSHIBA Corporation
Corporate Research and Development Center

成分の学習において、学習区間の平均スペクトルで成分を近似する方法^[2]は、高速な処理が可能であるが、時間毎に変動する非定常成分が推定できない。一方、NMF^[3]に基づく非定常成分を推定する方法^[4]では、スペクトルを用いたベクトル演算が必要なため、スペクトルが高次元になるにつれ、少ない計算コストで遅延なしの処理を実現するのが困難になる問題があった。

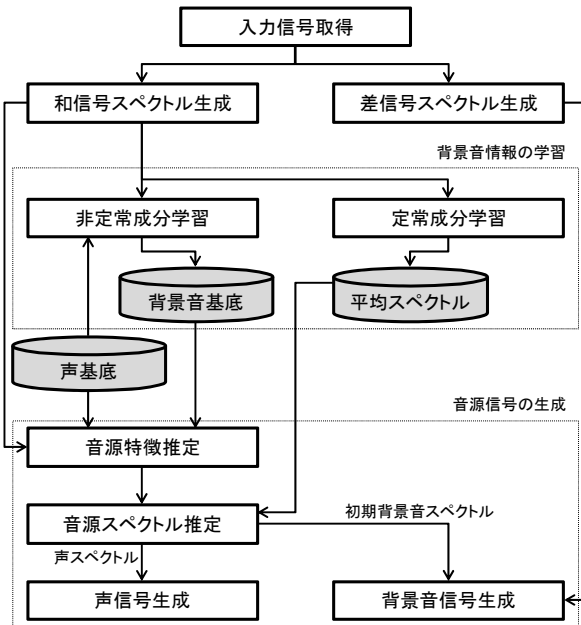


Fig. 2 Flow diagram of proposed method

図 2 音源分離手法の流れ

そこで本手法は、背景音情報の学習について NMF のベクトル演算を低次元化して行い、低次元化による推定劣化を補うため、平均スペクトル（定常成分）も活用して音源信号の生成を行う。

本稿で提案する音源分離手法の流れを図 2 に示す。2ch 以下の信号を対象とし、まず、L 信号と R 信号の和信号及び差信号から FFT によりスペクトル $X(\omega, t)$, $S(\omega, t)$ を生成する。なお、入力信号が 1ch の場合は入力信号を和信号とし、差信号は 0 とする。以下、処理の詳細を説明する。

3.1 背景音情報の学習

背景音情報は、L 信号と R 信号の両方の成分を含む和信号から学習する。

定常成分学習

まず、背景音の定常的な成分を抽出するため、複数のスペクトル $\{Y(\omega, t)\}_{t=k, \dots, k+K-1}$ の平均値 $A(\omega)$ を求める。 $Y(\omega, t)$ は背景音学習用に声成分を抑制したスペクトルであり、入力スペクトル $X(\omega, t)$ に対する予め定めた帯域毎の重み付けで求める。なお、開始時刻 k 、スペクトル数 K は実験的に設定した。

非定常成分学習

背景音の非定常成分は、NMF を用いた学習により求めた複数の基底と、時刻毎に推定した異なる係数（アクティビティ）との演算により表現することが可能である。NMF における計算コストを削減するため、スペクトル $Y(\omega, t)$ をメル周波数軸上での 32 次元の特徴ベクトル $Y_{mel}(t) = [Y_{mel}(0, t), \dots, Y_{mel}(31, t)]^T$ に変換する。特徴変換後、複数の特徴 $Y_{mel} = [Y_{mel}(k), \dots, Y_{mel}(k+K-1)]$ と事前に作成した声基底行列 $V = [V_0, V_1, \dots, V_{J-1}]$ を用いて、式 (1) を近似する背景音基底 $B = [B_0, B_1, \dots, B_{J-1}]$ を学習する。なお、 V_j, B_j は 32 次元のベクトル、 U_V, U_B は基底に対応する係数行列で、基底数 J は声及び背景音共に 64 とした。

$$Y_{mel} = [V \ B] \begin{bmatrix} U_V \\ U_B \end{bmatrix} \quad (1)$$

3.2 音源信号の生成

入力信号に対し、遅延なしで音源分離を行うために、背景音情報の更新処理とは異なり、スペクトル単位で音源信号を生成する。

音源特徴推定

まず NMF を用いて音源毎の特徴を推定する。入力スペクトル $X(\omega, t)$ をメル周波数軸上での特徴ベクトル $X_{mel}(t)$ に変換した後、声基底 V と背景音基底 B を用いて式 (2) を近似する係数ベクトル $U(t) = [U_V(t), U_B(t)]^T$ を推定する。基底と係数の対応関係から声特徴 $V_{mel}(t)$ と背景音特徴 $B_{mel}(t)$ を求める。

$$X_{mel}(t) = [V \ B] \begin{bmatrix} U_V(t) \\ U_B(t) \end{bmatrix} \quad (2)$$

$$V_{mel}(t) = V \cdot U_V(t) \quad (3)$$

$$B_{mel}(t) = B \cdot U_B(t) \quad (4)$$

音源スペクトル推定

音源スペクトルの推定では、NMF を用い背景音の非定常成分を考慮して求めた推定値（音源特徴推定の結果）と、背景音の定常成分（平均スペクトル）を考慮して求めた推定値をアルファブレンドする。

まず、推定した音源特徴 $V_{mel}(t)$ と $B_{mel}(t)$ を周波数軸上に逆変換し、非定常成分を考慮した声成分 $V_1(\omega, t)$ 、背景音成分 $B_1(\omega, t)$ を求める。ここで、NMF において、背景音基底は声基底に比べ少ないデータで学習しているため、背景音成分は十分な推定ができていないと考えられる。そこで、劣化成分の補間処理として、背景音成分 $B_1(\omega, t)$ には過去の推定値との最大値 $B'_1(\omega, t)$ を求める。

$$B'_1(\omega, t) = \max_{k \leq t' \leq t} B_1(\omega, t') \quad (5)$$

次に、背景音の定常成分（平均スペクトル）を考慮した声成分 $V_2(\omega, t)$ 及び背景音成分 $B_2(\omega, t)$ を求める。定常成分に時刻毎の変動成分を加えるため、非定常成分を考慮した背景音成分 $B_1(\omega, t)$ と平均スペクトルの最大値で背景音成分 $B_2(\omega, t)$ を設定する。

$$B_2(\omega, t) = \max(B_1(\omega, t), A(\omega)) \quad (6)$$

一方、声成分 $V_2(\omega, t)$ は入力スペクトルと背景音成分 $B_2(\omega, t)$ の差分とする。

$$V_2(\omega, t) = X(\omega, t) - B_2(\omega, t) \quad (7)$$

背景音の非定常成分、定常成分を考慮して求めた音源成分の推定値をアルファブレンドし、声と背景音成分の比から声スペクトル及び背景音スペクトル $V(\omega, t)$, $B(\omega, t)$ を求める。アルファブレンドするために用いた変数 α は予備実験の結果、0.8 に設定した。

$$V_3(\omega, t) = (1 - \alpha)V_1(\omega, t) + \alpha V_2(\omega, t) \quad (8)$$

$$B_3(\omega, t) = (1 - \alpha)B'_1(\omega, t) + \alpha B_2(\omega, t) \quad (9)$$

$$V(\omega, t) = \frac{V_3(\omega, t)}{V_3(\omega, t) + B_3(\omega, t)} \cdot X(\omega, t) \quad (10)$$

$$B(\omega, t) = \frac{B_3(\omega, t)}{V_3(\omega, t) + B_3(\omega, t)} \cdot X(\omega, t) \quad (11)$$

声信号生成

声信号は入力した和信号に含まれるとの仮定を置き、声スペクトル $V(\omega, t)$ に IFFT を行うことで求める。

背景音信号生成

背景音信号生成にあたり、音源スペクトル推定で求めた背景音スペクトル $B(\omega, t)$ は初期背景音スペクトルとして扱い、入力した差信号にて表現される成分を加える。しかし実際には、差信号が 0 となる（L 信号と R 信号が同じ）またはノイズが含まれるコンテンツも存在するため、差信号スペクトル $S(\omega, t)$ の大きさに応じて最終的な背景音スペクトル $B'(\omega, t)$ を求め、IFFT を行う。なお、 ε は閾値、 $\beta(\omega)$ は帯域毎の重み係数である。

$$B'(\omega, t) = \begin{cases} \beta(\omega) \cdot B(\omega, t) + S(\omega, t) & |S(\omega, t)| \geq \varepsilon \\ B(\omega, t) & |S(\omega, t)| < \varepsilon \end{cases} \quad (12)$$

4. 評価実験

4.1 処理コスト測定

本手法はスペクトル単位の逐次処理を行っており、遅延のない音源分離処理であるかは、実時間での動作が可能か否かで確認できる。サンプリングレート 48kHz、ビットサイズ 16bits の 2ch 信号を対象として、1 秒間の入力信号に対する動的命令回数を求めた結果、Intel® Core™ i5 の CPU で測定した命令回数は 326 M cycles であった。なお、スペクトルの次元数は 1024 とした。表 1 に示したように、背景音情報の学習と音源信号の生成は、処理過程にて入力スペクトルを 1/32 のサイズのメル周波数軸に変換したことで、FFT 及び IFFT に比べ少ない命令回数に抑えることができた。次に、モバイル機器での処理時間を NVIDIA® Tegra® 3 1.3GHz CPU で計測した結果、入力 1 秒あたり 603ms となり、実時間動作を確認できた。

表 1 計算コスト

Table 1 Calculation cost

処理項目	命令回数	占有率
FFT 及び IFFT	223 M cycles	68%
背景音情報の学習	17 M cycles	5%
音源信号の生成	86 M cycles	26%
合計	326 M cycles	100%

4.2 評価尺度

評価では、混合前の声及び背景音と、声と背景音の混合音から分離した信号の近さを測定した。混合前の信号が再現できれば、より効果的な音源調整が可能だと考えられる。目的音に対し生成した信号の近さを測る尺度として、ITU-T 勧告 P.862 の PESQ^[5] を用いた。この尺度は背景音の評価向けには設計されていないが、予備実験の結果、主観評価結果との相関が確認できたため用いることとした。

4.3 比較手法

声の評価では、混合音である入力信号（Input）、提案手法で生成した声信号（Proposed）、従来手法として声の強調を目的としたイコライザの出力（Conventional）を比較した。なお、従来手法における信号の帯域制御は、一般的に用いられるパラメータ値を参考に実装した。

背景音の評価では、入力信号、提案手法で生成した背景音信号（Proposed）、従来手法としてボーカル抑制の出力（Conventional）を比較した。なお、従来手法では差信号と中域成分を除いた和信号を加算する一般的な手法を実装した。

4.4 評価データ

人の声には男性 3 名、女性 2 名が発声している計 20 秒のデータを用いた。一方、背景音は、音楽、雑音、歓声といった背景音の種類毎に評価を設定した。音楽背景音 (Music) には、ドラマ、音楽番組、ニュースなどのコンテンツに流れる BGM を想定した、音楽 2 種類を用いた。雑音背景音 (Noise) には、収録環境を想定し、展示会場と街中の音を用いた。歓声背景音 (Sports) には、野球場での音 (屋外競技) とバレーボールが行われている体育館での音 (屋内競技) を用いた。また、声と背景音の混合率 (SNR) には 0dB, 5dB, 10dB, 15dB の 4 種類を用いた。

4.5 声の評価

図 3 に、評価に用いたスペクトルの例として、混合前の声信号、音楽を SNR 0dB で混合した入力信号、従来手法及び提案手法の出力を 8kHz (中域成分) まで示している。提案手法では、特に 2~8kHz でより声信号に近い信号を生成していることが分かる。

図 4 は、声の評価結果であり、背景音の種類毎に PESQ の平均値を求めた。結果より、提案手法は従来手法であるイコライザに比べ、混合前に近い声信号を得ることが確認できた。

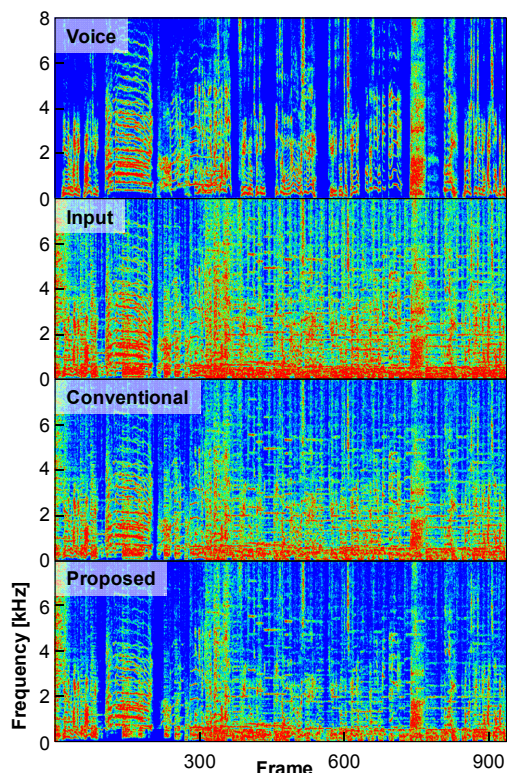


Fig. 3 Samples for voice evaluation

図 3 声の評価に用いたスペクトルの例

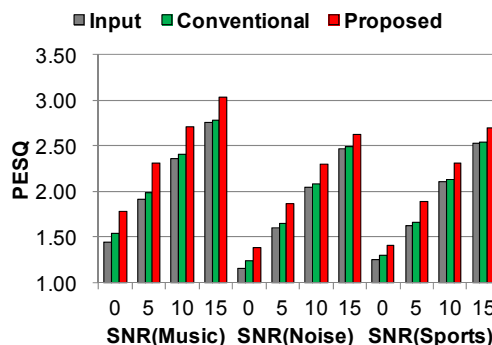


Fig. 4 Evaluation results of voice

図 4 声の評価結果

4.6 背景音の評価

図 5 には、評価に用いたスペクトルの例として、SNR 0dB で混合する前の音楽信号、従来手法及び提案手法の出力を示している。なお、従来手法及び提案手法の 2 つ目の出力は、差信号が 0 となる (L 信号と R 信号が同じ) 場合である。差信号が 0 の場合、従来手法であるボーカル抑制では、背景音のほとんどの成分が抑制された。一方、提案手法では、差信号が 0 でない場合に比べ推定精度の劣化は認められるものの、元の信号に近い出力が得られた。

図 6 は背景音の評価結果である。SNR が小さく背景音の占める割合が大きくなるにつれて、従来手法よりも提案手法の方が効果的であることを確認した。これは提案手法が和信号から抽出した背景音成分を活用したことで、従来手法では表現しにくい中域成分も出力できたためだと考えられる。

また、図 7 はモノラル録音のように入力信号の差信号が 0 である場合の評価結果である。提案手法では、差信号が 0 でない場合と比べ性能劣化があるものの、入力信号を処理しない場合に比べ、効果を維持することができた。

最後に、提案手法を用いることで、声と背景音のボリュームバランス調整に向けた音源分離の機能を実現したサンプルアプリケーションを図 8 に示す。ユーザは声と背景音のボリュームバーをそれぞれ自由に調整することで、好みに応じた音源毎のバランスを制御することができる。

5. おわりに

本稿では、音の入力信号から人の声と背景音の音源信号を推定する音源分離技術を紹介した。本技術は、映像コンテンツを視聴する際に、人の声を聴き取りやすくできる、背景音を静かにできる、スポーツの臨場感を高められる、歌の練習ができるといったユーザメリットを提供することを狙っている。また、本技術はデジタル映像機器全般に汎用的に活用できると考えている。

この手法は、背景音の定常・非定常成分を学習、更新しながら、声及び背景音の音源信号を遅延なしで推定する。

客観評価にて、音源信号の推定が、イコライザ機能やボーカル抑制機能に比べて高精度に行えることを確認した。

今後の課題として、デジタル映像機器に組み込んだ形でのアプリケーションの実現、非定常成分の推定に NMF 以外の方法を用いた場合や MOS 値による主観評価について検討する必要がある。

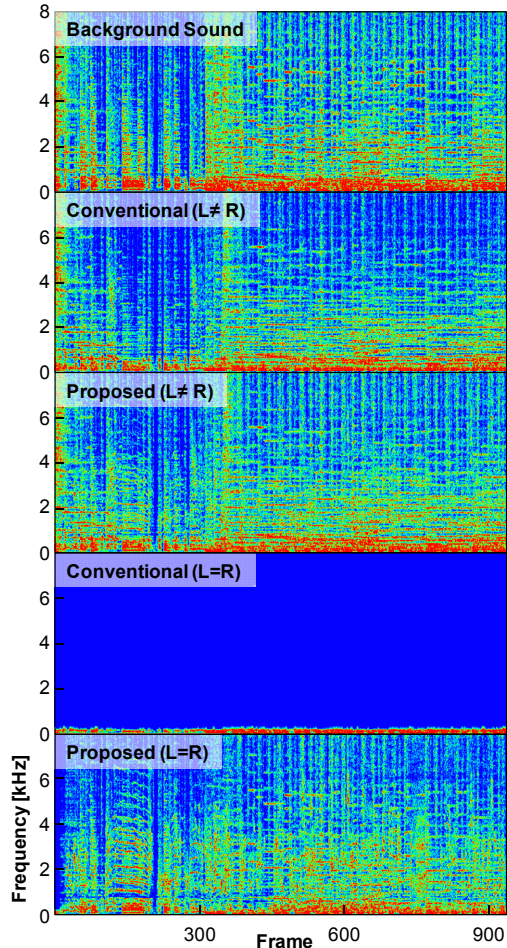


Fig. 5 Samples for background sound evaluation
図 5 背景音の評価に用いたスペクトルの例

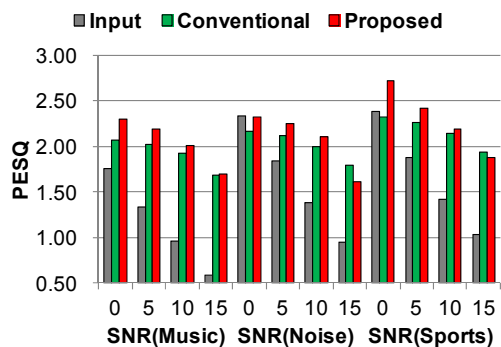


Fig. 6 Evaluation results of background sound for stereophonic audio
図 6 背景音の評価結果

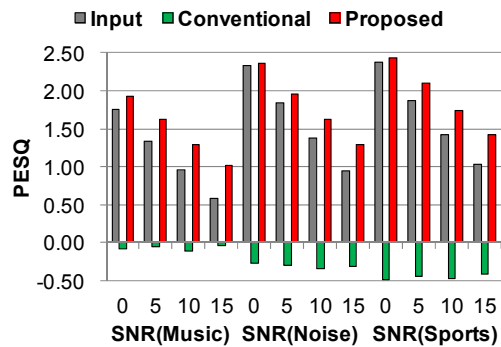


Fig. 7 Evaluation results of background sound for monophonic audio

図 7 入力信号の差信号が 0 である場合の評価結果

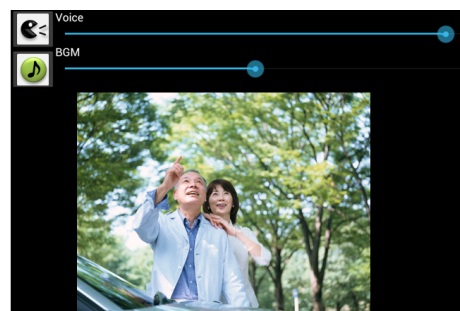


Fig. 8 Sample application

図 8 サンプルアプリケーション

参考文献

- [1] Stereo Signal Processing Apparatus, U.S. Patent No. 6,690,799, 2004, H. Iwase and T. Kubota.
- [2] S.F. Boll, Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Trans. ASSP, vol. 27, pp.113-120, 1979.
- [3] D.D. Lee and H.S. Seung, Algorithms for nonnegative matrix factorization, Proc. NIPS, vol. 13, pp. 556-562, 2000.
- [4] M. N. Schmidt, J. Larsen and F. Hsiao, Wind Noise Reduction using Non-negative Sparse Coding, Proc. MLSP, 2007.
- [5] PESQ, ITU-T Recommendation P.862, 2001.
- [6] 広畑誠, 小野利幸, 西山 正志, 多様な映像コンテンツに対応した遅延なし音源分離技術, 日本音響学会春季研究発表会, 2013.