人の声と背景音のボリュームバランスを調整する音源分離技術

Audio Source Separation Technology to Control Volume Balance between Voices and Background Sounds

広畑　誠　　小野　利幸　　西山　正志
■HIROHATA Makoto　■ONO Toshiyuki　■NISHIYAMA Masashi

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，かつ手軽に映像コンテンツを楽しめるようになった。しかし，人の声が音楽などの背景音に埋もれて聴き取りにくかったり，逆に歓声など背景音が小さくて臨場感が楽しめなかったりと，快適に視聴できない場合があった。

東芝は，入力信号音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開発した。これによって，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高める，歌の練習をするといった，様々な視聴ニーズに対応できる新しい高音質化機能を実現した。

The wide dissemination of audiovisual (AV) products has provided users with easy and diversified styles of viewing and listening to video contents. However, it is not always possible to view video contents comfortably because of an imbalance in the volumes of voices and background sounds.

Toshiba has developed an audio source separation technology to extract voice and background sound source signals from audio signals. This new technology realizes a more enjoyable viewing experience by allowing users to adjust background sounds and hear voices more easily, thus providing highly realistic sensations when watching programs such as sports matches and enhancing the experience of karaoke while watching music programs.

1　まえがき

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ている。しかし，人の声が音楽などの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て臨場感が楽しめない場合など，快適に視聴ができない場合がある。そのため，人の声や背景音の強調など，音源ごとにボリュームバランスを調整する高音質化機能への期待が高まっている。

人の声を強調できるイコライザ機能や，背景音を抽出して強調できるボーカル抑制機能[1]など，音源ごとにボリュームバランスを調整することを目的とした高音質化機能は，これまでにも開発されてきた。しかし，従来のイコライザ機能は，事前知識に基づいて制御するため，シーンによって人の声が十分に強調できない場合があった。また，従来のボーカル抑制機能は，ステレオ信号のセンタ成分（同一位相で同一振幅の成分）中に含まれる人の声を抑制するため，モノラル信号やセンタ成分中に含まれる背景音の強調には対応できなかった。このように，従来技術では，シーンや音声の収録方式など，入力信号音の特性によって性能が大きく劣化するという課題があった。

そこで東芝は，新たに入力信号音から背景音の情報を逐次学習し更新しながら，人の声と背景音の音源信号を推定する，音源分離技術[2]を開発した（**図1**）。開発技術は計算量が少なく，デジタル映像機器全般に汎用的に適用できる。また，人の声と背景音を個別に逐次学習するため，アマチュアが作成したコンテンツなど，多様なコンテンツに対しても効果を発揮できる。ここでは，開発した音源分離技術の概要と，その効果を確認した実験結果について述べる。

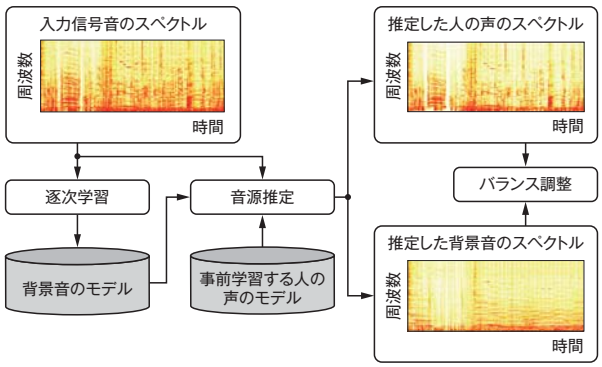

図1．音源分離技術の概要 ― 背景音の情報を逐次学習しながら，人の声と背景音をそれぞれ推定する。
Overview of audio source separation technology

2　音源分離技術

入力信号音の変化に対応しながら，人の声と背景音の音源信号を推定するには，背景音の成分を入力信号音から逐次学習する必要がある。背景音の成分の学習に関する従来技術として，学習区間の平均スペクトルで背景音の成分を近似する方式がある[3]。この技術は，高速で高精度な処理ができるもの

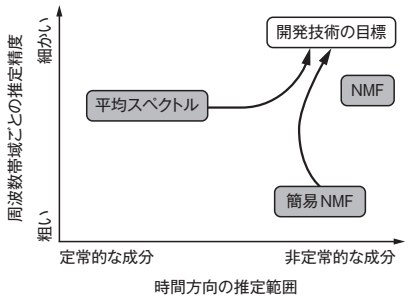

図2．開発技術と従来技術の関係 — 開発技術は，簡易NMFにおける推定精度の低化を平均スペクトルで補い，更に相乗効果によって精度の向上を目指した。

Relationship between conventional and newly developed technologies

の，定常的な成分の推定に限定され，時間的に変動する非定常的な成分を推定することはできない。一方，非定常的な成分を推定する方法として，非負値行列因子分解（NMF）[4]に基づく方式が提案されている。しかし，推定精度を高めるために，スペクトルを細かく分析し計算の繰返し回数を増やすと，リアルタイム処理が困難になる。

スペクトルの分析単位を粗くし，繰返し計算も大幅に削減した簡易NMFを適用すれば，NMFにおける計算量を削減できるが，周波数帯域ごとの推定精度が低下してしまう。そこで，周波数帯域ごとの推定精度が高く，かつ少ない計算量で非定常的な成分を求めることができる平均スペクトルに着目し，簡易NMFを平均スペクトルによって補間する音源分離方式を開発した（**図2**）。以下に音源分離を行ううえでポイントとなる，背景音の逐次学習，及び人の声と背景音の音源推定について述べる。

2.1　背景音の逐次学習

背景音の成分の学習は，時間的に不変の定常的な成分の学習と，時間的に変動する非定常的な成分の学習に分けて行う。学習は，左チャンネル（L）信号と右チャンネル（R）信号の和信号（モノラル信号なら入力信号音）から定期的に取得した，一定時間分のスペクトルを用いて行う。

まず，背景音の定常的な成分として，取得した学習スペクトルの平均スペクトルを求める。ただし，学習スペクトルには人の声の成分も含まれている可能性があるため，あらかじめ定めた帯域重み付けにより学習スペクトルの背景音の成分を強調しておく。

次に，背景音の非定常的な成分を学習するためNMFを用いる。NMFを用いれば，スペクトルを構成する基本要素である基底を学習できる[5]。このとき，計算の繰返し回数を増やすことで，精度よく推定できるが，計算量の削減を優先し，簡易NMFを適用する。一方，人の声の特性は，背景音に比べて大きく変化しないとの仮定から，大量のデータを用いて事前に学習しておく。

2.2　人の声の音源推定

人の声と背景音の音源信号の推定は入力信号音のスペクトルに対して行うが，人の声は中央に定位しやすいとの仮定から，和信号のスペクトルから推定する。人の声のスペクトルは，周波数帯域ごとに重みを制御したフィルタと和信号スペクトルの積で求める（**図3**）。

人の声又は背景音を高精度に推定するには，推定に用いるフィルタの生成が重要である。推定フィルタは，時間的に変動する音源の非定常的な成分を推定した後，過去の推定値と時間的に不変の定常的な成分の値をそれぞれ適用し，その結果を重畳することで生成する（**図4**）。

まず，非定常的成分の推定ステップでは，逐次学習した背景音の基底と，事前に学習した人の声の基底とを用いて，背景音と人の声のスペクトルを推定する。

ここで，背景音の基底は人の声の基底に比べて少ないデータで学習しているため，背景音の成分は人の声の成分に比べて，時間的にも周波数的にも十分な推定ができていない。そこで，過去の推定値を適用するステップにおいて，現在の推定値と過去の推定値の最大値を求め，背景音の時間的なふらつきを抑えるように，劣化成分を補う。更に背景音の周波数的な劣化成分を補うため，定常的な成分を適用するステップにおいて，現在の推定値と背景音の定常的な成分として細かい分析単位で求めた平均スペクトルの最大値を求め，周波数的な解像度を高める。

最後に混合処理では，二つの適用ステップを経て得られた成分を統合し，人の声の成分と背景音の成分の比に基づいて

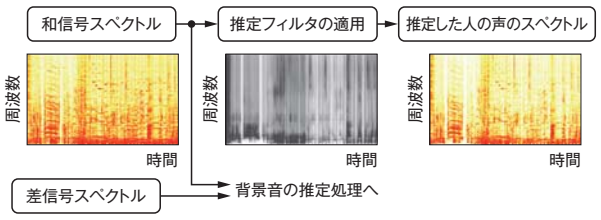

図3．人の声の推定プロセス — 入力信号音のうち，和信号のスペクトルに対して，帯域ごとに制御できる人の声の推定フィルタを適用する。

Process of voice spectra estimation

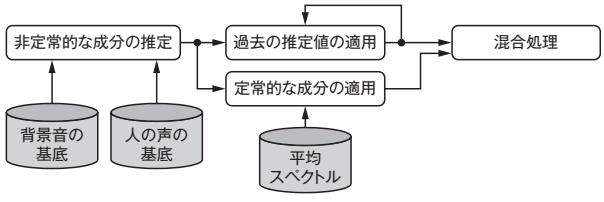

図4．音源推定フィルタの生成処理の流れ — 非定常的な成分を推定した後，過去の推定値と定常的な成分を考慮して推定フィルタを生成する。

Flow of audio source estimation filter generation

人の声のスペクトルの推定フィルタを生成する。同様に背景音のスペクトルの推定フィルタも生成するが，差信号の活用を考えて，推定結果は仮の背景音としている。

2.3　背景音の音源推定

人の声の音源推定では使用しなかった差信号は，中央に定位した人の声の成分が除かれているため，背景音を表すことが多く，従来のボーカル抑制技術では主な出力成分とされてきた。しかし，モノラル信号の場合，和信号に含まれる背景音の成分は全く推定できず，音声フォーマットの変更で生じたノイズなど不要な成分を出力してしまうという問題があった。

また一般的には，和信号と差信号のゲイン差が大きいときは，差信号に背景音が含まれないと判定し，例えば和信号から求めた仮の背景音が利用できる。しかし，和信号に人の声が含まれるシーンでもゲイン差が大きくなるため，シーンによって正しく判定できないという問題があった。

そこで，和信号から求めた仮の背景音を判定に加えることで，人の声が含まれるかどうかにかかわらずゲイン差を一定に保つことに成功し，シーンによらない正確な判定を実現した。

3　実験による効果の確認

開発技術と従来技術をそれぞれ用いて，人の声に背景音を重畳した混合音に対して，人の声及び背景音の推定結果を比較した。

3.1　推定した人の声の比較

まず，人の声の音源推定において，従来技術には，人の声の強調を目的とした一般的なイコライザによる方式と，簡易化なしのNMFによる方式を用いた。背景音として音楽を重畳した際の推定結果を**図5**に示す。

それぞれの画像は，背景音を重畳した後の入力信号音のスペクトル，背景音を重畳する前の人の声のスペクトル，イコライザを用いた従来技術，NMFを用いた従来技術，及び簡易NMFを用いた開発技術の推定結果に関するスペクトルを示す。図5中の破線で囲まれた周波数の高い帯域において，開発技術は，従来技術と同じように背景音の成分を抑制しながら，人の声の成分も損なわずに推定できていることがわかる。また，実線で囲まれた低い周波数帯域においては，開発技術によって，背景音の成分をより抑制できていることが確認できる。

3.2　推定した背景音の比較

次に，背景音の推定結果を**図6**に示す。従来技術には，差信号を主な出力成分とする一般的なボーカル抑制方式を用いた。それぞれの画像は，背景音を重畳した入力信号音のスペクトル，重畳した背景音のスペクトル，従来のボーカル抑制方式の推定スペクトル，及び和信号から推定した仮の背景音の成分を用いた開発技術の推定スペクトルである。

推定結果の比較は，(1)L信号とR信号が異なる，差信号に背景音が含まれる場合と，(2)L信号とR信号が同じ又はモノラル信号が入力となる，差信号に背景音が含まれない場合の二つの条件下で行った。

差信号に背景音が含まれる(1)の場合では，開発技術と従

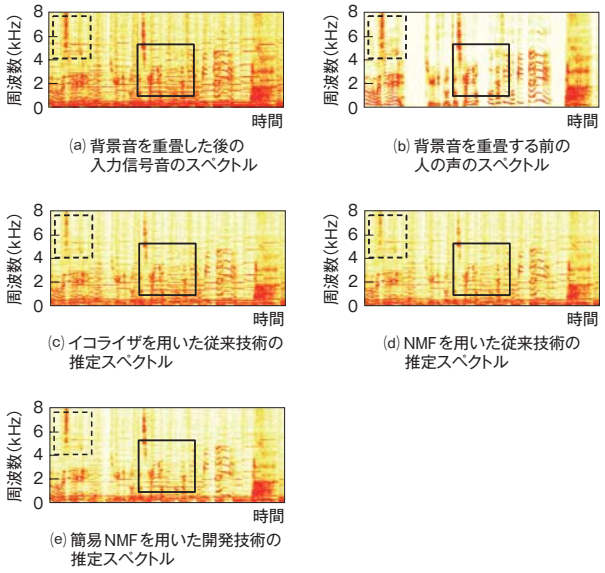

図5．推定した人の声のスペクトル比較 ― 従来技術に比べ開発技術では，より高い精度で元の人の声を推定できている。

Comparison of estimated voice spectra using conventional and newly developed technologies

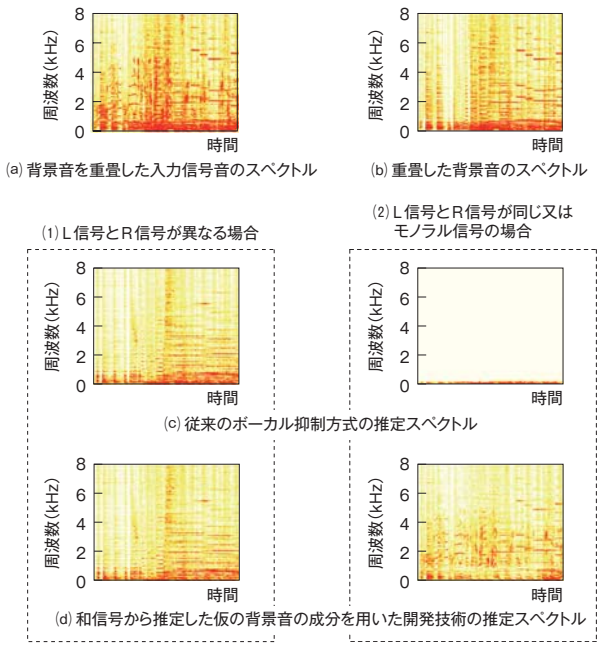

図6．推定した背景音のスペクトル比較 ― 従来技術に比べ開発技術では，入力信号音の特性によらず安定して背景音の推定ができている。

Comparison of estimated background sound spectra using conventional and newly developed technologies

来技術との差は小さいものの，差信号に背景音が含まれない⑵の場合では，従来技術がほとんど推定できないのに対し，開発技術は⑴の場合に近い推定結果が得られていることがわかる。

3.3　バランス調整機能の評価

開発技術を用いると，人の声と背景音のボリュームバランスを段階的にコントロールできる機能を実現できる（**図7**）。推定した背景音を抑制した後に推定した人の声と混合すれば，背景音の抑制を弱めにすると人の声は少し強調され（人の声の強調 弱モード），背景音の抑制を強めにすると人の声は強く強調される（人の声の強調 強モード）。同様に，人の声を抑制することで背景音を強調することもできる。図7のようなユーザーインタフェースを提示すれば，ユーザーは好みに応じて人の声又は背景音を強調するモードを選択できる。

開発技術をレグザタブレットAT500に実装したところ，リアルタイム動作を行うには十分な計算量に抑えることができた。また，バランス調整機能を評価するため，様々なジャンルから計14種類のコンテンツを無作為に選び，12名の被験者に対して，主観評価実験とアンケート調査を実施した。主観評価実験では，全てのコンテンツにおいて用意した計四つの強調モードの効果を確認できた。またアンケート調査では，各強調モードの利用シーンについて意見を収集した。主な意見を**表1**に示す。

背景音と人の声の強調の弱モードではともに，臨場感を味わいたい，せりふを聴き取りたいといった汎用的な利用を意識した意見を収集することができた。一方，強モードでは，実況者の声を小さくしたい，歌の練習をしたい，発音をチェックしたい，歌詞を覚えたいといったように，特殊な用途での利用を意識した意見が得られた。開発技術を活用して音源のボリュームバランスを段階的にコントロールできる機能は，様々な価値の提供につながることを確認できた。

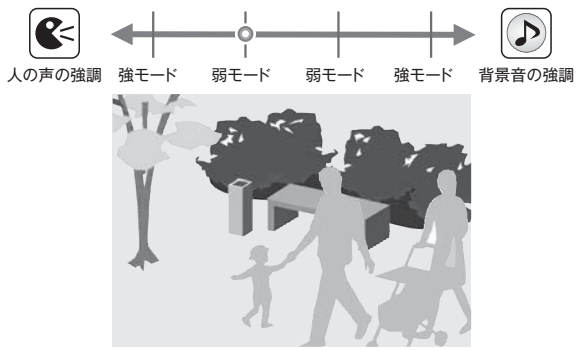

図7．アプリケーション画面のイメージ ― 人の声と背景音の音源バランスを段階的に調整して，映像を視聴できる。
Example of display of prototype application for tablets

表1．各強調モードの利用シーンに関するアンケート結果
Results of questionnaire on scenes of usage of each emphasis mode

強調モード	利用したいシーン
人の声の強調　強モード	映画（洋画）番組で発音をチェックしたいとき 音楽番組で歌詞を覚えたいとき
人の声の強調　弱モード	映画，ドラマ，アニメ番組でせりふを聴き取りたいとき
背景音の強調　弱モード	スポーツ番組や音楽番組で臨場感を味わいたいとき
背景音の強調　強モード	スポーツ番組で実況者の声を静かにしたいとき 音楽番組で歌の練習をしたいとき

4　あとがき

映像コンテンツの快適な視聴を目指して，当社が開発した音源分離技術について述べた。この技術によって，人の声と背景音のボリュームバランスを調整し，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高める，歌の練習をするといったユーザーが求める視聴を実現することができる。また，開発技術はデジタル映像機器全般に適用可能である。

今後も様々な視聴環境において，高音質化に対する要求がますます高まると考えられる。当社はこれからも，汎用的に適用できる高音質化技術を開発していく。

文　献

⑴　日本電信電話．ステレオ音響信号処理方法及び装置並びにステレオ音響信号処理プログラムを記録した記録媒体．特許第3670562号．2005-04-22.
⑵　広畑　誠 他．“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2013年春季研究発表会講演論文集．東京，2013-03，日本音響学会．論文番号3-10-18.
⑶　Boll, S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics, Speech and Signal Processing. **27**, 2, 1979, p.113－120.
⑷　Lee, D.D. et al. Algorithms for Non-negative Matrix Factorization. Proc. NIPS. **13**, 2000, p.556－562.
⑸　Schmidt, M. N. et al. "Wind Noise Reduction using Non-Negative Sparse Coding". 2007 IEEE Workshop on Machine Learning for Signal Processing. Thessaloniki, Greece, 2007-08, p.431－436.

広畑　　誠　HIROHATA Makoto
研究開発センター 知識メディアラボラトリー研究主務。
音響信号処理の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.

小野　利幸　ONO Toshiyuki
研究開発センター マルチメディアラボラトリー研究主務。
画像処理及び音響信号処理の研究・開発に従事。電子情報通信学会会員。
Multimedia Lab.

西山　正志　NISHIYAMA Masashi, Ph.D.
研究開発センター インタラクティブメディアラボラトリー研究主務，博士（学際情報学）。画像認識及び信号処理の研究・開発に従事。電子情報通信学会会員。
Interactive Media Lab.