擬音語アニメーションによる動画音響の可視化手法

王 方舟　　柏野 邦夫　　永野 秀尚　　五十嵐 健夫*

概要.
動画に含まれる音の情報は，動画の視聴体験を構成する重要な要素である．一方で，実際に視聴者が動画を視聴する際は，状況によっては必ずしも音を伴って視聴できない場合が存在する．従来，このような状況下で音の情報を視覚的に補う手段としては字幕が用いられてきたが，そのほとんどが人の発した声を文字に書き起こしたものであり，声以外の一般的な音の情報に関する表現力は非常に乏しいのが現状である．本稿では，動画中に生起する音の種類を自動で判別し，擬音語（オノマトペ）を用いて可視化する手法を提案する．生成された擬音語は音の変化に合わせてアニメーションされる．これにより，動画中に含まれる一般的な音の種類およびそのダイナミクスを自然な形で可視化し，視聴者に伝えることが可能になる．

1　はじめに

多くの動画コンテンツにおいて，音は重要な役割を担う．人の話すセリフやナレーションなどの音声情報はそれ自体が不可欠な情報であるし，それ以外の音であっても，例えばカーチェイスのシーンでのエンジンの唸り，サッカーの試合のシーンでの人々の歓声など，様々な場面において音は視聴者の感情に影響を及ぼし，映像と合わせて一つの視聴体験を構築する．一方で，視聴者が動画を視聴する状況は様々であり，必ずしも常に音が聞き取れるとは限らない．そのような状況下では，動画本来の視聴体験が大きく損なわれることがある．

従来，このような状況下で音の情報を伝達する手段としては，字幕が使われてきた．一般的に使われている字幕は，静的なテキストをシーンの進行に合わせて画面下に表示するものであるが，1) 人手による事前の準備が必要であり，2) 人の声以外の効果音や環境音に対する表現力が低く，3) 音の生起のタイミングや変化を表現できない，などの問題点がある．特に，効果音や環境音については，単に「ガラスの割れる音」など音の状況を説明する文字列を配置するものが多く，音が本来持つ視聴体験を高める働きは失われてしまっている．

この問題を解決するため，本研究では，動画中に生起する音の種類を認識して擬音語に変換することによって可視化し，さらに音の変化に合わせて適切にアニメーションする手法を提案する．提案手法は入力された動画から音の情報を自動で認識するために人手の介在が不要である．さらに音を自然な視覚表現である擬音語に変換した上で，音の変化に合わせて適切にアニメーションすることにより，音のない状況下でも豊かな視聴体験を提供することを可能にする．

図 1. 擬音語アニメーションによる動画音響の可視化例

2　関連研究

音の種類を識別するための研究は従来より多くなされてきた [2] これに対し，石原ら [1] は擬音語の表現力に着目し，環境音を直接擬音語に変換する手法を提案している．山本ら [3] は，環境音を直接擬音語に変換し，音の大きさや質に応じた適切なフォントを用いて可視化する手法を提案している．一方で山本らの提案手法は極めて限定された実験環境を対象としており，単発の短音のみを対象とし，一度生成された擬音語表示は音の変化に対して静的であるなど，実在の動画に対して適用し得るものではない．

3　システム設計

本研究で実装したシステムは，自動車のレース動画を対象とし，動画中のエンジン音，およびドリフトやブレーキの際に発生するスキール音の 2 種類の音を自動認識して擬音語アニメーションを付与するものである．生成する擬音語の種類およびアニメー

* Houshu Oh and Takeo Igarashi, 東京大学大学院 情報理工学系研究科 コンピュータ科学専攻, Kunio Kashino and Hidehisa Nagano, NTT コミュニケーション科学基礎研究所.

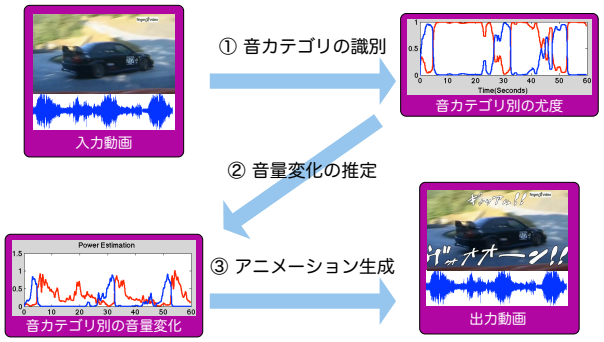

図 2. アルゴリズムの概要

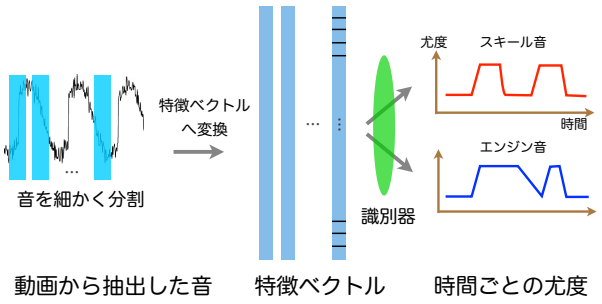

図 3. 音カテゴリの識別方法

ションのデザインは，考えうる最もシンプルな形として，1) 基本的な音の種類 (エンジン音またはスキール音) に応じてあらかじめ用意された擬音語画像を，2) 映像の固定位置に表示し，3) 音の大きさに合わせた拡大および縮小アニメーションを行うものとする．

4 アルゴリズム

提案するアルゴリズムは，音付きの動画を入力とし，擬音語アニメーションが付与された動画を出力するものである．アルゴリズムは大きく分けて，1) 音カテゴリの識別，2) 音量変化の推定，3) 擬音語アニメーションの生成，の 3 つのステップをたどる (図 2)．

アルゴリズムはまず，入力された動画の音の種類を識別する．この際に用いる識別器は，ラベル付きの音データを用いて SVM で機械学習を行うことで事前に構築しておく．構築する識別器は，0.75 秒の音の断片から生成された特徴ベクトルを入力とし，その音の断片がどの音カテゴリに属するかの尤度を計算するものである．入力された動画の音データを時間方向に細かく分割し，各断片を特徴ベクトルに変換して，構築した識別器に入力として与えることにより，動画の時系列に沿って音のカテゴリ別の尤度が計算される (図 3)．

アルゴリズムは次に，音のカテゴリ別に音量変化を推定する．推定は，前のステップで計算された音カテゴリ別の尤度に，元の音の音量を掛け合わせることによって行われる．

アルゴリズムは最後に，推定されたカテゴリ別の音量変化をもとに擬音語アニメーションを生成する．アルゴリズムは入力動画を 1 フレームずつ見ていき，それぞれのフレームに対応する音のカテゴリ別音量を前ステップの結果から得る．次に，音量が一定の値以上を超えた音のカテゴリについて，擬音語画像を生成する．現実装では，システムが予め音の種類の数だけ準備した擬音語画像をそのまま用いる．最後に，生成した擬音語を対応する音カテゴリの音量の大きさに合わせて拡大・縮小し，映像に埋め込む．

5 議論と今後の課題

環境音から擬音語を生成する研究は以前よりなされてきたが，その多くは入力となる音を日本語に直接変換することを目的としている [1]．このような手法は擬音語表現が豊富な日本語に対しては有効であるが，一方で他の言語に適用しにくいというスケーラビリティの問題や，音の混合に弱いという問題を抱えている．本稿で提案した手法は音を言葉に直接変換するのではなく，音の「種類」を認識し，対応する擬音語に変換するという手法を採用している．これにより，言語的な表現力を多少犠牲にしつつも，多言語拡張へのスケーラビリティと音の混合に対する一定のロバスト性を確保している．

今後の課題としては，音源物体の位置に合わせた擬音語の適切な配置がある．これは，一般物体認識の手法や Saliency 解析などを用いて行うことが考えられる．音の変化をより自然に捉えたアニメーションの生成も課題である．例えば，音量の起伏を複数の立ち上がりと減衰に分割することで，同一カテゴリの音についても複数の音がかぶさりあって生起している様子を捉えることが可能になると考えられる．さらにはアニメーションの要素として，拡大・縮小だけではなく，ゆるやかな減衰部分についてはゆっくりフェードアウトさせるなどといったより自然なアニメーションも実現できるようになると考えられる．

参考文献

[1] K. Ishihara, T. Nakatani, T. Ogata, and H. G. Okuno. Automatic Sound-Imitation Word Recognition from Environmental Sounds focusing on Ambiguity Problem in Determining Phonemes. *PRICAI 2004: Trends in Artificial Intelligence. Lecture Notes in Artificial Intelligence*, 3157:909–918, 2004.

[2] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

[3] 山本貴史, 松原正樹, 斎藤博昭. 擬音語と書体表現を用いた環境音の可視化. 芸術科学会論文誌, 11(1):1–11, 2012.