2L2-OS-09a-3

聴覚系モデルを用いた音のテクスチャ変換

Sound texture transfer using a model of the auditory system

上村 卓也[*1] 寺島 裕貴[*1] 古川 茂人[*1]
Takuya Koumura Hiroki Terashima Shigeto Furukawa

[*1] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

We propose a method for converting texture of a sound while retaining its content. The method is based on the previous studies on image texture conversion and the studies on sound texture synthesis. The proposed method uses a model of the auditory system which takes two sounds as inputs and generates another sound that has the content of one of the input sounds and the texture of the other input. Examining the amplitude envelopes of the auditory filter outputs indicated that generated sounds inherited the characteristics of the input texture sound. Not all features were identical to the input texture sound, however, suggesting that the effect of content sound was also substantial. We compared different weightings of texture and content, and confirmed that increasing the weight of texture resulted in the generated sound which was more similar to the input texture than input content.

1. はじめに

近年、機械学習の発展により、画像中の物体認識や音声認識などの課題において機械による自動認識がヒトのパフォーマンスに匹敵しつつある(Hinton et al., 2012; Krizhevsky, Sutskever, & Hinton, 2012; Piczak, 2015; Simonyan & Zisserman, 2014; Yoshioka et al., 2015)。画像や音から得られる情報は、物体のカテゴリ・位置・形状・発話内容などの情報だけではない。例えば画像からは物体表面の質感、音からは場の雰囲気などの情報も得ることができる(Gaver, 1993; Schwartz & Nishino, 2016)。画像や音のテクスチャも、このような情報のひとつである(Julesz, 1981)。テクスチャは、物体のカテゴリ・位置・形状などと異なり、刺激中のどこを切り出しても同様に知覚される。例えば、ある自然画像中の異なる部分を切り出したとき、そこには異なるカテゴリの物体が含まれることが多い(図 1)。一方、あるテクスチャ画像の異なる部分を切り出しても、そこに含まれるテクスチャは同じである。したがって、テクスチャの知覚は、物体認識とは異なるメカニズムによって実現されている可能性がある(McDermott, Schemitsch, & Simoncelli, 2013)。

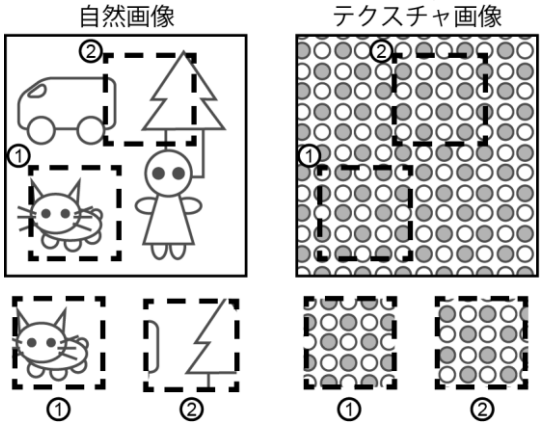

図 1：自然画像とテクスチャ画像の例。二箇所の破線の部分を切り出し、下に示した。

連絡先：上村卓也，NTT コミュニケーション科学基礎研究所，〒243-0198 神奈川県厚木市森の里若宮 3-1，046-240-3657，komura.takuya@lab.ntt.co.jp

過去の研究により、視覚系の機能を模擬したモデルを用いてテクスチャ画像を合成できることがわかっている(Portilla & Simoncelli, 2000)。この手法では、あるテクスチャ画像が、視覚系の機能を模擬した階層的なウェーブレットフィルタに入力される。そして、フィルタの出力画像から、様々な空間的な周辺統計量が計算され、それを元に同じ周辺統計量を持つ新たな画像が合成される。後に、深層畳み込みニューラルネットワークが視覚系の機能を模擬したフィルタバンクの役割を果たすことがわかり、階層的ウェーブレットフィルタの代わりに深層畳み込みニューラルネットワークを用いると、より複雑なテクスチャ画像を合成できることがわかった(Gatys, Ecker, & Bethge, 2015b)。

音についても同様に、聴覚系の機能を模擬したモデルを用いてテクスチャ音を合成できる(Mcdermott, Oxenham, & Simoncelli, 2009; McDermott & Simoncelli, 2011)。この手法では、あるテクスチャ音が、聴覚系の機能を模擬した階層的なフィルタバンクに入力される。そして、フィルタの出力波形から、様々な時間的な周辺統計量が計算され、それを元に同じ周辺統計量を持つ新たな音が合成される。これらの研究から、テクスチャは、刺激の周辺統計量（画像の場合は空間的な・音の場合は時間的な周辺統計量）に基づいて知覚されることが示唆される。

さらに、画像のテクスチャ合成を発展させ、ある画像のテクスチャを別のテクスチャに変換する手法が考案された(Gatys, Ecker, & Bethge, 2015a)。この手法では、ある画像から「内容」を表す特徴が抽出され、別の画像からテクスチャを表す特徴が抽出される。そして、内容の特徴とテクスチャの特徴を組み合わせ、一方の画像の内容ともう一方の画像のテクスチャを持った新たな画像が合成される。つまり、ある画像の内容を保持したままテクスチャのみを変換した、といえる。内容を表す特徴は、深層畳み込みニューラルネットワークの各層の変数値であり、空間的に非定常である。テクスチャを表す特徴は、テクスチャ合成と同様にニューラルネットワークの各層の空間周辺統計量である。よって、画像の内容は表象の空間パターンに基づいて知覚され、テクスチャは空間の位置によらない周辺統計量に基づいて知覚されることが示唆される。なお、この研究では絵画の画風変換に対象を絞っているため、画像のテクスチャをスタイルと呼んでいる。

我々は、これらの手法から着想を得、音の「内容」を維持したまま、テクスチャを変換することを試みた。これまでの研究から、

音のテクスチャは時間的な周辺統計量を元に合成できることがわかっている。音の「内容」の定義は現時点であいまいである。一方、画像については、ある画像の表象の空間パターンを内容とし、別の画像の空間的な周辺統計量を組み合わせ、画像のテクスチャを変換した例がある(Gatys et al., 2015a)。本研究では、このアナロジーとして、音の「内容」は波形のある種の時間パターン（2.4節参照）に、テクスチャは時間周辺統計量（2.3 節参照）によって表現されるという、暫定的な仮説を設定した。そのうえで、ある音の波形の時間パターンと、別の音の時間周辺統計量を組み合わせ、音のテクスチャ変換を試みた。

2. 方法

2.1 テクスチャ変換の枠組み

音のテクスチャ変換の枠組みを、図 2 に示した。主な構成要素は、聴覚系モデル、内容音、テクスチャ音である。まず、内容音とテクスチャ音から、それぞれ聴覚系モデルの出力波形が抽出され、内容音の出力波形からその時間パターンが、テクスチャ音からその時間周辺統計量が計算される。次に、合成対象の元となる新たな音から内容・テクスチャに対応する特徴が計算され、それぞれを内容音とテクスチャ音の値に近づけるように、合成対象の音を少しずつ変形していく。合成対象の元となる音としては、白色雑音を用いた。合成音の統計量が元の統計量に十分近づいたら、変換を停止する。合成音の変形には、次の誤差関数を最小化するような確率勾配法を用いた。

$$L_{total} = \alpha L_{content} + (1-\alpha) L_{texture}$$

$L_{content}$ は合成音と内容音の、内容を表す特徴（2.4 節）の平均二乗誤差、$L_{texture}$ は合成音とテクスチャ音の、テクスチャを表す特徴（2.3 節）の平均二乗誤差である。重み係数 $\alpha \in [0, 1]$によって、合成音を内容音とテクスチャ音のどちらにより近づけるかを調整した。

2.2 聴覚系モデル

聴覚系モデルは、バンドパスフィルタバンクと静的非線形変換によって構成した(McDermott & Simoncelli, 2011)。まず、入力音がバンドパスフィルタバンクにより、フィルタの通過帯域ごとに分けられる。次に、フィルタ通過後の波形からヒルベルト変換により振幅包絡波形が抽出され、非線形に圧縮される。最後に、周波数帯域ごとの振幅包絡線が、バンドパスフィルタバンクに入力される。はじめのバンドパスフィルタバンクと非線形圧縮は、蝸牛の機能を模擬したものである。振幅包絡線の抽出は、聴覚末梢神経系の性質を模擬したものである。二番目のバンドパスフィルタバンクは、聴覚系に存在するといわれている変調フィルタバンクを想定したものである。モデルの実装・パラメータは、過去の研究と同じものを用いた(McDermott & Simoncelli, 2011)。

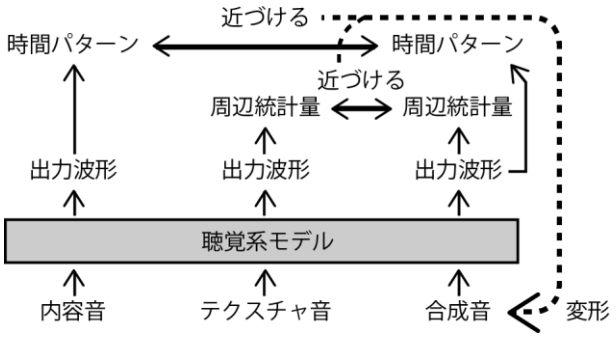

図 2：手法の枠組みの模式図。

2.3 テクスチャを表す特徴

音のテクスチャを表す特徴として、モデル通過後の波形のいくつかの特徴量について、周辺統計量を計算した。具体的には、過去の研究によりテクスチャ合成に有効であることがわかっている次の統計量を用いた(McDermott & Simoncelli, 2011)：聴覚フィルタ通過後の波形の分散、振幅包絡線の平均・分散・歪度・周波数チャネル間の相関係数、変調フィルタ通過後の波形のパワー。

2.4 内容を表す特徴

音の内容を表す特徴として、モデル通過後の波形の時間パターンを用いた。候補として、聴覚フィルタ通過後の波形・振幅包絡線の波形・変調フィルタ通過後の波形を検討したが、本研究では、計算時間短縮のため、振幅包絡線の波形のみを用いることとした。

2.5 データ

合成の元となるテクスチャ音として、過去のテクスチャ合成の研究で用いられた音を用いた(McDermott & Simoncelli, 2011)。内容音として、音声と音楽から切り出した音を用いた(Architectural Institute of Japan, 2004)。内容音は、残響が小さい環境で録音された音である。

3. 結果

3.1 テクスチャ変換後の音と振幅包絡線

音声と音楽を、鳥の鳴き声のテクスチャと水の泡のテクスチャに変換した（**図 3**）。内容音の重み α は 0.5 とした。内容音の振幅包絡線が、テクスチャ音の音圧の高い周波数帯域で強調され、音圧の低い帯域で抑制されているように見える。コントロールとして、内容音とテクスチャ音を単純に加算した音についても振幅包絡線を示した（**図 4**）。こちらは、内容音の振幅包絡線とテクスチャ音の振幅包絡線を足し合わせたもののように見える。よって、テクスチャ変換の結果、内容音の振幅包絡線にテクスチャ音の特徴が反映されたものが合成されたことがわかる。

3.2 テクスチャ変換後の周辺統計量

合成された音の周辺統計量が内容音とテクスチャ音のどちらに近いかを調べるために、合成音（内容：音声、テクスチャ：鳥の鳴き声）について、聴覚フィルタ通過後の波形の分散と振幅包絡線の平均を周波数チャネルごとに示した（**図 5**）。合成された音の聴覚フィルタ通過後の波形の分散が、全帯域で内容音（音声）ではなくテクスチャ音（鳥の鳴き声）のものと似ていることがわかる。一方、振幅包絡線の平均は、低周波帯域ではテクスチャ音と近く、高周波帯域では内容音と近かった。よって、合成音の全ての周辺統計量が全ての帯域においてテクスチャ音に近づいているわけではない。コントロールとして、音声と鳥の鳴き声の音を単純に加算した音についても同じ統計量を計算した。聴覚フィルタ通過後の波形の分散は、内容音とテクスチャ音のものの中間の値を取っていた。振幅包絡線の平均は、低周波帯域で内容音に近く、高周波帯域でテクスチャ音に近かった。

3.3 内容とテクスチャの重み

音を合成する際の、内容音とテクスチャ音の重み係数 α が合成結果に与える影響を調べた。音声のテクスチャを鳥の鳴き声に変換した合成音について、内容音の重みが 0.2 のものの振幅包絡線を示した（**図 6**）。重み係数が 0.5 のもの（**図 3**）と比較

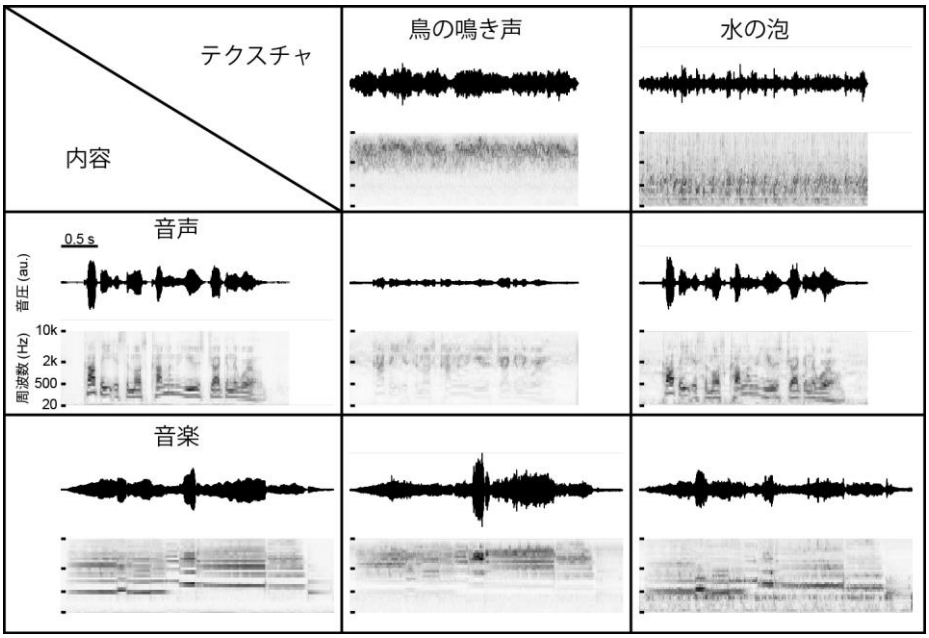

図 3：テクスチャ変換後の音と振幅包絡線。一番左の列に元の内容音、一番上の行に元のテクスチャ音を示し、それぞれが交差する枠にテクスチャ変換結果を示した。それぞれの音について、上側に音波形、下側に周波数チャネルごとの振幅包絡線を示した。

すると、音声の重みが大きい方が、振幅包絡線が音声に近くなっていることがわかる。

考察

3.4 音におけるテクスチャと内容の知覚

ある音の時間変化する特徴と別の音の時間的に定常な特徴を組み合わせ、一方の音の内容を持ち、他方の音のテクスチャを持つ音を合成できた。この結果は、音のテクスチャは時間周辺統計量に基づいて知覚されるという過去の研究と一貫する(McDermott et al., 2013; McDermott & Simoncelli, 2011)。本研究の結果から、音声や音楽などの内容とテクスチャが異なる統計量に基づいて知覚される可能性が新たに示唆された。また、本研究の結果は、音声認識には振幅包絡線の時間変化が重要であるとする過去の研究と一貫する(Houtgast & Steeneken, 1973; Pešán, Burget, Hermansky, & Vesely, 2015)。

3.5 内容とテクスチャのバランス

本研究では、合成音の周辺統計量が、必ずしもテクスチャ音のものと近くはなかった。内容を表す波形の時間パターンと、テクスチャを表す時間周辺統計量は、当然独立ではない。つまり、どちらかが変化すればもう片方も変化する。本研究では、内容とテクスチャの重みを、1つの係数によって全ての統計量について均一に定めていた。内容とテクスチャのバランスをどのように取ると効果的なテクスチャ変換を実現できるのかを検討することは、今後の課題である。音声のテクスチャを鳥の鳴き声に変換した例では、周波数帯域ごとに内容音・テクスチャ音の影響が異なっているように見えた。これは、音声が低周波帯域に高いパワーを持ち、鳥の鳴き声が高周波帯域に高いパワーを持つというように、元の音の周波数帯域がはっきりと分かれていたことが原因かもしれない。

3.6 声質変換との関係

音の質感変換に関する研究は、声質を対象としたものが多い(Tao, Kang, & Li, 2006; Toda, Black, & Tokuda, 2007)。声質変換では、音声の内容を保持したまま話者や情動を変換する。声質変換では、変換後の音も音声である。一方、本研究のテクスチャ変換による合成音は、音声の内容を持った音声ではない音である(例えば水の泡など)。この点で、本手法は声質変換とは異なる。しかし、話者・情動・残響などが振幅包絡線から計算される特徴量よって表現されるという研究もあり(Unoki, Furukawa, Sakata, & Akagi, 2003; Van Vuuren & Hermansky, 1998; Wu, Falk, & Chan, 2011)、これらの質感も聴覚モデルにより抽出された定常的な統計量の操作によって変換できるかもしれない。

音声のテクスチャを変換する手法として、ヴォコーダによる声質変換がある。ヴォコーダは音声のコーディングを目的としており、音楽などへ適用することは想定されていない。一方、我々のモデルは、聴覚において内容とテクスチャを分離することを目的としており、対象となる音は音声に限定されない。

謝辞

本研究は JSPS 科研費 JP15H05915(新学術領域研究、多元質感知)の助成を受けたものです。

参考文献

Architectural Institute of Japan. (2004). *Sound Material in Living Environment*. Gihodo Shuppan.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015a). A Neural Algorithm of Artistic Style. *arXiv Preprint*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015b). Texture Synthesis Using Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* (pp. 262–270).

Gaver, W. W. (1993). How Do We Hear in the World? Explorations in Ecological Acoustics. *Ecological Psychology*, *5*, 285–313.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., … Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *Signal Processing Magazine, IEEE*, *29*, 82–97.

Houtgast, T., & Steeneken, H. J. M. (1973). The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility. *Acta Acustica United with Acustica*, *28*, 66–73.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, *290*, 91–97.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
Mcdermott, J. H., Oxenham, A. J., & Simoncelli, E. P. (2009). Sound texture synthesis via filter statistics. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 297–300).
McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nat Neurosci*, *16*, 493–498.
McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*, 926–940.
Pešán, J., Burget, L., Hermansky, H., & Vesely, K. (2015). DNN derived filters for processing of modulation spectrum of speech. In *Sixteenth Annual Conference of the International Speech Communication Association* (pp. 1908–1911).
Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In *IEEE 25th International Workshop on Machine Learning for Signal Processing* (pp. 1–6).
Portilla, J., & Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, *40*, 49–70.
Schwartz, G., & Nishino, K. (2016). Perceptual Material Attributes Arise in Local Material Recognition. *arXiv Preprint*.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint*.
Tao, J., Kang, Y., & Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 1145–1154.
Toda, T., Black, A. W., & Tokuda, K. (2007). Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 2222–2235.
Unoki, M., Furukawa, M., Sakata, K., & Akagi, M. (2003). A method based on the MTF concept for dereverberating the power envelope from the reverberant signal. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.* (Vol. 1, pp. 888–891).
Van Vuuren, S., & Hermansky, H. (1998). On the importance of components of the modulation spectrum for speaker verification. In *ICSLP*.
Wu, S., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, *53*, 768–785.
Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., … Nakatani, T. (2015). The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 436–443).

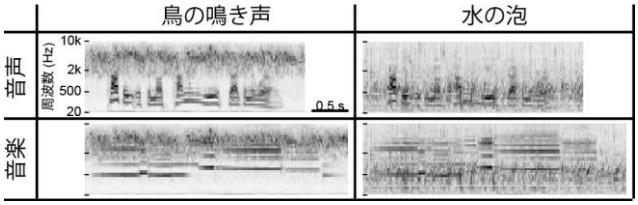

図 4：内容音とテクスチャ音を加算した音の振幅包絡線。
各行に内容の種類、各列にテクスチャの種類を示した。

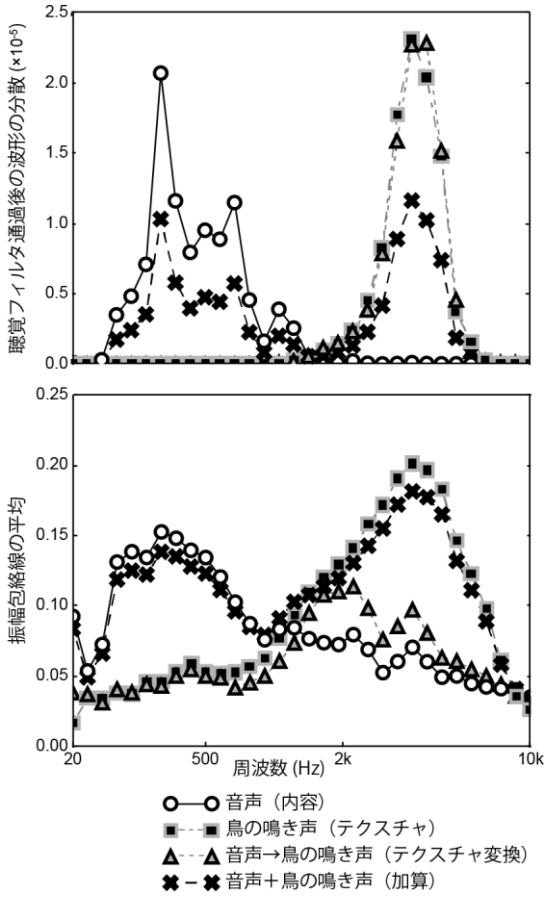

図 5：周辺統計量の例。横軸に周波数チャネル、縦軸に周波数チャネルごとの統計量を示した。

図 6：合成音の振幅包絡線。内容：音声、テクスチャ：鳥の鳴き声。重み係数を 0.2 とした。