# 誰でも色々な声が出せる声質変換システム UNVOCS の構造

2022 年 3 月

この文書では、筆者が開発を進めている声質変換システム「UNVOCS」の仕組み等について解説する。

**開発経緯**

現在一般に利用できる声質変換システムは、主に以下の 2 つのいずれかのアルゴリズムで声質変換を行う。第 1 の方法は変換元音声から音声認識によって音素・音高などの要素を抽出し、それらをもとに変換先話者の声を合成するものであり、第 2 の方法は変換元話者と変換先話者の音声間の変換係数を導いておき、それによって変換元音声を直接変換するものである。第 1 の方法はゆかりねっと[1][2]や Seiren Voice[3]などが採用しており、第 2 の方法は Voidol[4]などが採用している。

いずれの手法でも、変換先話者の種類を増やすのは容易ではない。変換先話者の十分な音声データが必要となるからである。第 2 の手法では、それに加えて変換元話者の音声データも必要となる(Voidol では変換係数を変換元話者に応じて作成するのではなく、予め用意しておくことによってこの点を回避しているが、その代償として変換の品質は低下している)。そして十分な音声データが必要となる原因の核心は、変換過程において機械学習を用いるところにある。

したがって、**機械学習を用いずに声質変換を行えるなら、必要な音声データの量を削減することができるであろう。**本システムでは、変換元話者と変換先話者の間の変換係数を機械学習を用いずに推定する(上述の 2 つの手法のうち第 2 の方法を採用したことになる)。これにより、**変換元の話者と変換先の話者の母音の音声のみを基準とし、両話者間の声質変換を実現する**ことが可能となった。

**変換方法**

人間の声は、スペクトル包絡と基本周波数によって特長づけられる。スペクトル包絡は声質に、基本周波数は声の高さに対応する。よってスペクトル包絡の変換によって声質の変換が可能である。

最初に述べたように、今回開発した変換アルゴリズムは母音の音声を基準としているが、より正確に表現するならば母音のスペクトル包絡の構造に基づいている。**図 1** は AHS 社の音声合成システム、VOICEROID2 結月ゆかり[5]により作成した日本語の母音のスペクトル包絡である(スペクトル包絡の計算には WORLD[6](D4C edition [7])を使用した。以降単にスペクトル包絡と述べたときには WORLD により求めたスペクトル包絡のことを指す)。**図 2** に示したように、スペクトル包絡は各母音のピーク位置に基づいて 9 つの領域に分割することができる。すなわち、

・第 1 の領域では全母音に
・第 2 の領域では「え」、「お」に
・第 3・第 4 の領域では「あ」に
・第 5 の領域では「え」に
・第 6 の領域では「い」に
・第 7 の領域ではいずれかの母音に

それぞれピークが見られる。

残る高音域については、いずれかの母音に比較的大きなピークがみられる部分を第 8 の領域に、目立ったピークが見られない部分を第 9 の領域に分ける。当然ながら、各領域の幅は話者によって異なっている。変換過程においては、これらの領域が本質的な役割を果たす。

さて、変換元話者の「あ」の音声のスペクトル包絡、および変換元音声のある時刻 t におけるスペクトル包絡に対して、上記の領域を設定し、各領域においてスペクトル包絡の平均を求める。変換元話者の「あ」の音声のスペクトル包絡、および変換元音声の時刻 t におけるスペクトル包絡の第 i 番目の領域( i=1～9 )の平均をそれぞれ a( i )、St( i )とおこう。

その上で、変換先話者の「あ」の音声のスペクトル包絡の第 i 番目の領域に St( i ) / a( i )を掛けたものを、変換後音声の時刻 t におけるスペクトル包絡とする。

このようにして求めたスペクトル包絡と、元の音声の基本周波数をもとに、WORLD で音声を再合成し、目的の音声を得る。なお、基本周波数については定数を加えて高い声にしたり、定数を掛けて抑揚の強い声にしたりと

いった変換が可能である。

以上が変換過程の定量的表現であるが、この過程の意味するところを定性的に述べよう。

同じ音を発したとしても、話者が異なればスペクトル包絡の形状は異なる。しかし、**異なる音のスペクトル包絡間の関係性はどの話者であっても大きくは違わない**であろう。すなわち、**変換元話者の基準となる音のスペクトル包絡と、変換したい音声のある時刻のスペクトル包絡との関係は、変換先話者の基準となる音のスペクトル包絡と、変換後の音声の同じ時刻のスペクトル包絡との関係に等しい**と考えられる。このことを用いると、変換後の音声のスペクトル包絡を全ての時刻について計算できる。

ただし、**変換元話者と変換先話者の基準となるスペクトル包絡の形状の違いを計算に反映させる必要がある。**これがスペクトル包絡を 9 領域に分け、各領域について変換を行う理由である。

基準となる音声として「あ」を用いたのは、低周波数の領域の音の成分が多く、他の母音と比べて変換後の音声の品質が良好だったためである。

変換した音声のサンプルについては、既に公開している動画[8][9]を参照されたい。

**課題**

一番の問題は、変換後の音声にノイズがのることである。これについては、変換の際に概ね 5000Hz 以上の高音域を減衰させることで一応の解決が図れているが、より根本的な原因を究明して対策を講じたいところである。

**応用の可能性**

声質変換ソフトとしてだけではなく、音声合成ソフトと組み合わせることで、実質的に任意の声が出せる音声合成ソフトとして運用することも可能となる。

また変換過程からも分かるように、変換において必要なパラメータは、話者のスペクトル包絡の領域の分割と「あ」のスペクトル包絡の形状である。したがってこれらを人工的に作成すれば、いわゆる「無生物音源」、すなわち録音された人間の声に基づかない音声を合成することも可能である。

**本システムの改良にご協力ください**

筆者は音声合成に関しては素人であるため、特に音声合成に詳しい方々からのご意見をお待ちしている。

**外部リンク**


[1]https://www.nicovideo.jp/watch/sm28022694

[2]http://www.okayulu.moe/

[3]https://seiren-voice.dmv.nico/

[4]https://crimsontech.jp/apps/voidol/

[5]https://www.ah-soft.com/voiceroid/yukari/

[6]M. Morise, F. Yokomori, and K. Ozawa: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016.
https://www.jstage.jst.go.jp/article/transinf/E99.D/7/E99.D_2015EDP7457/_article

[7]M. Morise: D4C, a band-aperiodicity estimator for high-quality speech synthesis, Speech Communication, vol. 84, pp. 57-65, Nov. 2016.
http://www.sciencedirect.com/science/article/pii/S0167639316300413

[8]https://www.nicovideo.jp/watch/sm39377001

[9]https://www.nicovideo.jp/watch/sm39592238

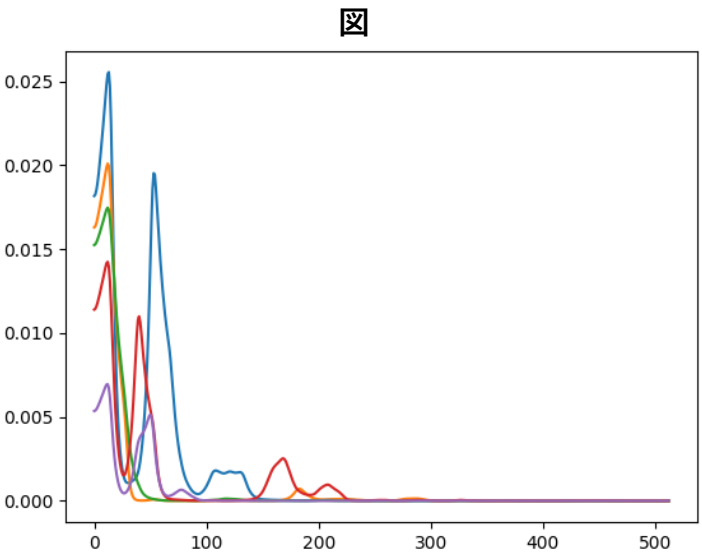


**図 1**

青線、橙線、緑線、赤線、紫線の順に、あ、い、う、え、おのスペクトル包絡。サンプリング周波数は 16000Hz である。

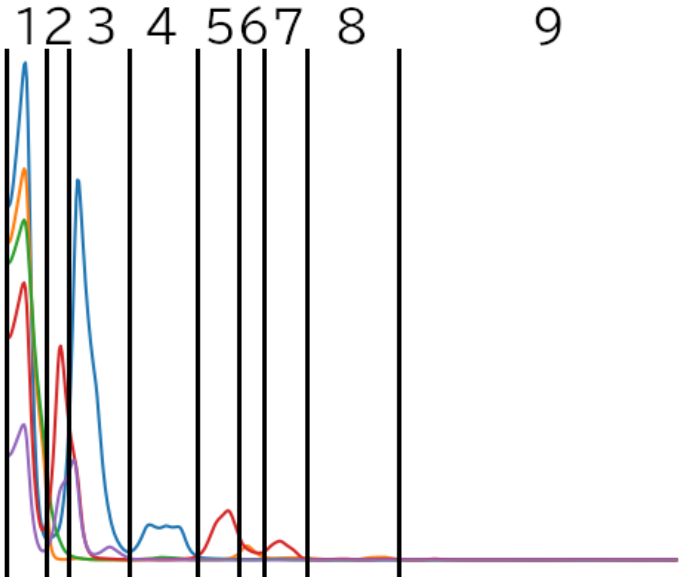


**図 2**

**図 1** のスペクトル包絡を本文の通り分割したもの。