社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IN2016-64 (2016-11)

アニメ動画における音声の周波数スペクトルを用いた声優認識

榮田　基希†　　服部　峻††

†,††室蘭工業大学 ウェブ知能時空間研究室　〒 050–8585 北海道室蘭市水元町 27–1
E-mail: †16043009@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

あらまし　アニメ動画から音声が流れたとき，それが誰であるか調べようとするならば，エンドロールを探すといった手間を掛ける必要がある．音声から声優認識することが出来るようになれば，手間が掛からずに声優名が分かるだけではなく，その声優の他作品の出演情報やブログ，関連動画や関連商品，これからのイベント情報など幅広く情報を取得することが出来る．これまでの研究では，Web から取得したアニメ動画のキャスト情報や音声の振幅に基づく類似度計算による絞り込みによって声優認識を試みたが，声優認識精度として良好な結果を得ることが出来なかった．そこで本稿では，声優認識するために音声の振幅ではなく周波数パワースペクトルを活用する．データベースに登録されている声優が個々に持つ特有の周波数パワースペクトルのパターンである「特有パワースペクトル」を自己相関分析によって予め特定しておき，実際に流れている再生中の動画の周波数パワースペクトルと，データベースに登録されている個々の声優が持つ特有パワースペクトルとを比較することによって声優認識を行う新しいシステムを提案する．
キーワード　声優認識，音声認識，特有パワースペクトル，自己相関

Voice Actor Recognition Using Frequency Spectrum in Anime Video

Motoki EIDA† and Shun HATTORI††

†,†† Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan
E-mail: †16043009@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

Abstract　When we hear someone's voice from an anime video, we need to carry extra burdens of searching the end roll of the anime video in order to know about whose voice it is. If a system can recognize a voice actor from his/her voice on behalf of us, not only we can know about the voice actor's name without carrying extra burdens, but also we can acquire widely information about him/her such as his/her appearance information, blogs, related videos, related goods, and event information in the future. Our previous research has been tackling a system of voice actor recognition with filtering by cast information extracted from the Web and similarity calculation based on voice amplitude, but the system could not give enough good performance as voice actor recognition accuracy. Therefore, this paper proposes a novel system of voice actor recognition that utilizes not voice amplitude but frequency power spectrum. Our proposed system identifies the "characteristic power spectrum" for each of individual voice actors who are registered in the database of the system by auto-correlation analysis in advance, and recognizes a voice actor from a voice in a playing anime video by comparing the voice's frequency power spectrum with each individual voice actor's characteristic power spectrum registered in the database.
Key words　Voice Actor Recognition, Speech Recognition, Characteristic Power Spectrum, Auto-Correlation

1. ま え が き

近年，日本には様々な娯楽メディアがあり，我々はそれらを普段の生活の中で目や耳にする機会が多くなっている．情報通信機器の普及により多くの人にとって，パソコンやモバイル端末などの機器で番組や動画の視聴，ゲームなどが今では手軽に行うことが出来る．このような娯楽メディアに触れる機会が多くなって来ると，どこかで聞いたことがある音声が流れて来ることがある．

その音声の発生源がアニメ動画の場合，誰の音声であるかを知る為には，エンディングのスタッフロールまで飛ばしたり，Web で作品のタイトル名やキャラクター名で検索したりするな

どの余計な労力を掛ける必要が出て来る．例えば，あるユーザが適当なアニメを視聴していた際，そのアニメの中に出て来たキャラクター A の音声がユーザの聞いたことのある音声であったとする．そこで，そのユーザがキャラクター A の声優について調べようとするならば，エンディングまで飛ばしたり，アニメタイトルやキャラクター名で Web 検索して，そのアニメの公式サイトやウィキペディアなどを探そうとするであろう．しかし，知りたいキャラクター A が作中の目立たない配役だった場合，Web で検索を掛けても中々出て来ないことも考えられる．また，主要なキャラクターではない場合，キャラクター名を記憶していない可能性もあり，エンディングのスタッフロールが流れても分からないだろう．その上，脇役であった場合，スタッフロールには男の子 B，男の子 C というようにキャラクター名を不明瞭に表記していることもあり，どの場面に出て来たキャラクターか分からないことも考えられる．

ここで，労力を掛けずに声優名を知るためには，アニメ視聴中に音声が流れたらリアルタイムに声優名を認識して自動的に画面に表示するシステムが必要になる．前回の研究で我々は，アニメ動画から流れる音声波形データとデータベースに予め登録してある各声優の音声波形データを使って類似度の計算を行い声優を判定する手法 [1] を提案した．しかし，音声を用いた声優認識の精度として良好な結果を得ることが出来なかった．精度が悪かった理由として主に考えられるのは，音声の振幅の波形データを単純に使用していた点と，データベースに登録する各声優の音声波形データを無作為に選出していた点である．

そこで本稿では，人の声が個々に持つ特有の周波数パワースペクトルと，実際に流れている動画の音声のパワースペクトルとを比較して声優を判定する手法を提案する．従来と同様に判定する際に，視聴中のアニメ動画のタイトルが特定されていることで，そのタイトルに基づいて Web 検索されたキャスト情報によって声優をキャスト陣のみに絞り込んでいる状態を想定する．また，声優データベースには，各声優が個々に持つ特有パワースペクトルのパターンを個々の声優の音声データから自己相関分析によって探索し，1 つずつ予め登録しておく．

最終目標としては，アニメ動画が流れている最中に声優を判定することだが，その際に複数の問題が生じる．例えば，BGM と声優のセリフの区別や，オープニングやエンディングの楽曲中における音楽と歌声の区別，及び，声優のセリフが 2 人以上重なる場合などがある．そこで本稿では，アニメ動画をフルに視聴している状況を想定するのではなく，アニメの 1 話に出て来るキャラクターが 1 人で話しているシーンの部分をキャラクター分だけ切り取り，それらの動画を声優認識対象の動画として評価実験を行う方法を採用する．本研究では，声優データベースに登録しておく時に使われる声優毎の音声データの種類や，特有パワースペクトル探索のための類似性の計算式の種類，また，実際に流れているアニメ動画の音声のパワースペクトルとデータベースに登録されている特有パワースペクトルとの類似性を比較する際の計算方法に依って声優認識精度が変わると考えられる為，様々な組み合わせについて比較実験を行う．

2. 関 連 研 究

話者認識する手法は多々ある中，アニメの音声から声優を認識する研究は中々見つからない．これまでの我々の研究 [1] では音声波形をそのまま用いて声優認識を試みてみたが，本研究では音声の主要な要素の一つである周波数に注視する．文献 [2] では，音声の有声音に限定して，基本周波数推定の方法として，音声の時間波形に対する周期性に着目した分析法やパワースペクトルの調波構造に着目した方法，及び，特徴量を用いた方法などが用いられている．パワースペクトルに着目した方法では，パワースペクトルの最も低いピークの周波数（f_0）を抽出することや f_0 の整数倍にピークを有する調波構造のピーク間隔を推定することでも基本周波数が推定できると記述されている．この従来研究においては，パワースペクトルのピーク間隔を推定することで基本周波数を推定する方法が記述されているが，本研究では基本周波数を特定するのではなく，周期毎（時間毎）にパワースペクトルを取得し，それらのパターンを観測することで声優が個々に持つ特有のパワースペクトルを特定する手法を用いる．関連研究で着目している基本周波数と本研究で着目する特有パワースペクトルの違いを図 1 に示す．

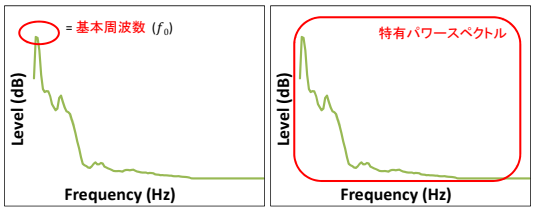

図 1　関連研究の基本周波数と本研究の特有パワースペクトル

3. 提案システム

アニメ動画に流れる音声から声優名を認識するため，声優に限定しない一般の話者認識に関する従来研究 [3,4] を参考にして，人それぞれには周波数毎に個人差があると考えた．そこで，人の声質にはバンド幅毎にそれぞれ違う強さ（dB）を持つと仮説を立て，声優が個々に持つ特有の周波数パワースペクトルのパターンである「特有パワースペクトル」を特定できれば声優認識が可能になると考えた．この仮説から，声優の人数分の特有パワースペクトルをアニメ動画やラジオ動画の音声データから分析し，声優毎に 1 つずつデータベースに予め登録しておく．本稿では，データベースに登録されている各声優の特有パワースペクトルと，アニメ動画から流れる音声データのパワースペクトルとの間の類似度計算を，データベースに登録されている声優の数だけ行うことによって，その音声の持ち主が誰であるかを判定するシステムを提案する．

では，本研究のシステムに関して詳しく説明を行う．本稿における声優認識システムを図 2 に示す．アニメ動画から流れる音声データを取得してスペクトル表示するのに，JavaFX に用意されているインターフェースである AudioSpectrumListener を用いる．本研究のシステムで用いた設定は，通知間隔が 0.1

秒，バンド数が 128，感度閾値が −60 デシベル（dB）である．感度閾値においては最低値を 0 にしたい為，各バンド毎の値に +60dB 加算して，正規化を行っている．処理 1 では動画や wav データを再生して音声スペクトルを表示させる．再生位置 t 秒（t∈{0.1，0.2，0.3，…}）において，新しく生成される音声スペクトルの各バンド幅の dB の数値（以下，$\boldsymbol{v_t}$）を 0.1 秒毎に取得する．次の処理 2 では，$\boldsymbol{v_t}$ と予めデータベースに用意されている各声優 i が個々に持つ特有パワースペクトルの各バンド幅の dB の数値（以下，$\boldsymbol{a_i}$）を使って，$\boldsymbol{v_t}$ が生成される度に類似度や相関を計算する．最後の処理 3 では，アニメ動画が終了した後，処理 2 で計算された 0.1 秒毎の類似度や相関が予め決められていた閾値を上回った回数が一番多い $\boldsymbol{a_i}$ を持つ声優 i をアニメ動画から流れる音声の持ち主の声優であると判定する．処理 3 の詳細は 5 章で説明する．

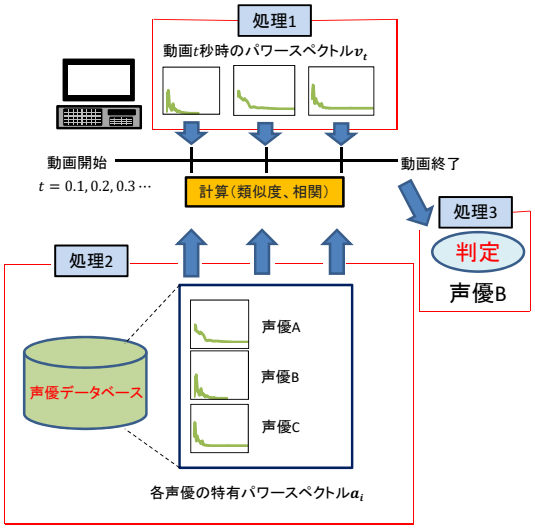

図 2　提案システム

4.　声優データベース

本章では声優データベースの詳細について説明する．データベースに入っている声優 1 人分の要素を図 3 に示す．声優 1 人につき 128 個のバンドぞれぞれが下限 0dB，上限 60dB の数値から成る特有パワースペクトルが登録されている．

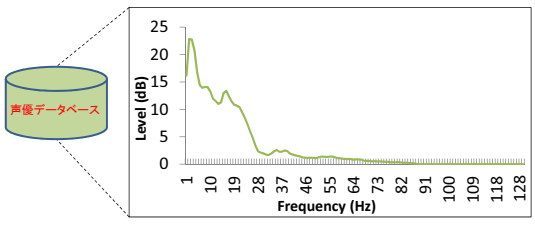

図 3　声優データベースの声優 1 人分の要素（特有パワースペクトル）

4.1　声優データベースの構築

データベースの中身には個々の声優の声質の特徴が表れている特有パワースペクトルを登録する必要がある．声優の特徴が表れている数値を見付ける方法として，ある声優の音声が流れている動画から取得できるパワースペクトルの数値から，頻出している類似パターンを探し出すことである．良く頻出しているパワースペクトルの類似パターンを見付け出す為に，自己相関に倣った手法を用いる．その際の類似性の計算式として，コサイン類似度，及び，相関係数の 2 種類を用いる．

声優の特有パワースペクトルを見付ける方法として，まず初めにデータベースに登録する声優の音声が流れている数多のアニメ動画，及び，ラジオ動画から音声スペクトルの各バンド幅分の dB の数値を取得する．目的の声優の音声が流れている部分だけの動画を切り取る際には以下のことに気を付ける．

- BGM が無い
- 効果音が無い
- 時間が短い音声は入れない（例：相槌など）
- こもっている音声を入れない（例：心の声など）
- 他の人の音声とかぶっていない

これらの注意点は，出来る限りノイズを含まない生の音声データのパワースペクトルの数値を取得する為である．次に，切り取った動画の音声データから音声スペクトルの数値を取得した後，各声優毎にそれらを 1 つにまとめる．最後に，声優毎に 1 つにまとめたパワースペクトルの集合に対して，繰り返し頻出しているパターンを解析して特有パワースペクトルを探し出す．

4.2　特有パワースペクトル探索の為の類似性の計算式

音声データから取得したパワースペクトルの時系列パターンの中で，自己と類似/相関するパターンが繰り返し頻出する特有パワースペクトルを探索する為の類似性の計算式について説明する．各声優毎に 1 つにまとめたパワースペクトルの時系列パターンを $\boldsymbol{P_t}$（t∈{0.1，0.2，…，T}）と表す．ある瞬間の t 秒におけるパワースペクトル $\boldsymbol{P_t}$ は，バンド数である 128 個の dB の数値 $P_{t,1}, P_{t,2}, \cdots, P_{t,128}$ から成り，それぞれを +60dB で正規化した値で書き直したものとする．$\boldsymbol{P_{t'}}$（t'∈{0.1，0.2，…，T}）は，$\boldsymbol{P_t}$ のコピーであり，全く同じパワースペクトルの時系列パターンを表している．本稿では，自己の時系列パターンとの類似性や相関を計算する関数 Auto$(\boldsymbol{P_t}, \boldsymbol{P_{t'}})$ として，以下のコサイン類似度 sim$(\boldsymbol{P_t}, \boldsymbol{P_{t'}})$ と相関係数 correlation$(\boldsymbol{P_t}, \boldsymbol{P_{t'}})$ のいずれかを用いる．

（1）　コサイン類似度

$$\boldsymbol{P_t} = (P_{t,1}, \cdots, P_{t,128}), \quad \boldsymbol{P_{t'}} = (P_{t',1}, \cdots, P_{t',128})$$

$$\mathrm{sim}(\boldsymbol{P_t}, \boldsymbol{P_{t'}}) = \frac{\sum_{j=1}^{128} P_{t,j} \cdot P_{t',j}}{\sqrt{\sum_{j=1}^{128} P_{t,j}} \sqrt{\sum_{j=1}^{128} P_{t',j}}}$$

（2）　相関係数

$$\boldsymbol{P_t} = (P_{t,1}, \cdots, P_{t,128}), \quad \boldsymbol{P_{t'}} = (P_{t',1}, \cdots, P_{t',128})$$

$$\mathrm{correlation}(\boldsymbol{P_t}, \boldsymbol{P_{t'}}) = \frac{\sum_{j=1}^{128} (P_{t,j} - \overline{P_t})(P_{t',j} - \overline{P_{t'}})}{\sqrt{\sum_{j=1}^{128} (P_{t,j} - \overline{P_t})^2} \sqrt{\sum_{j=1}^{128} (P_{t',j} - \overline{P_{t'}})^2}}$$

4.3 特有パワースペクトル探索アルゴリズム

本節では，各声優のセリフ毎に切り取った動画から取得したパワースペクトルの時系列パターンを 1 つにまとめたファイルから，声優の特有パワースペクトルを探索する手法について説明する．図 4 のように，0.1 秒毎のあるパターン $\boldsymbol{P_t}$ を 1 つずつベースとして，もう 1 つの比較対象の全パターン $\boldsymbol{P_{t'}}$ 各々との類似度や相関を計算する．計算した類似度や相関が予め定めた閾値を上回った場合，ベースにしているパターン $\boldsymbol{P_t}$ のカウント数を増やす．$\boldsymbol{P_t}$ と $\boldsymbol{P_{t'}}$ との全ての組み合わせの計算を終えた後，閾値を上回ったカウント数を一番多く獲得した $\boldsymbol{P_t}$ を，動画の音声データから得られた声優の特有パワースペクトルと特定する．しかし，閾値を上回ったカウント数が同数になることがある．その場合に，計算から得られた類似度や相関のうち閾値を上回った時だけ加算した平均で比較して，類似度や相関の平均が最も大きかったパターン $\boldsymbol{P_t}$ を動画の音声データから得られた声優の特有パワースペクトルと特定する．

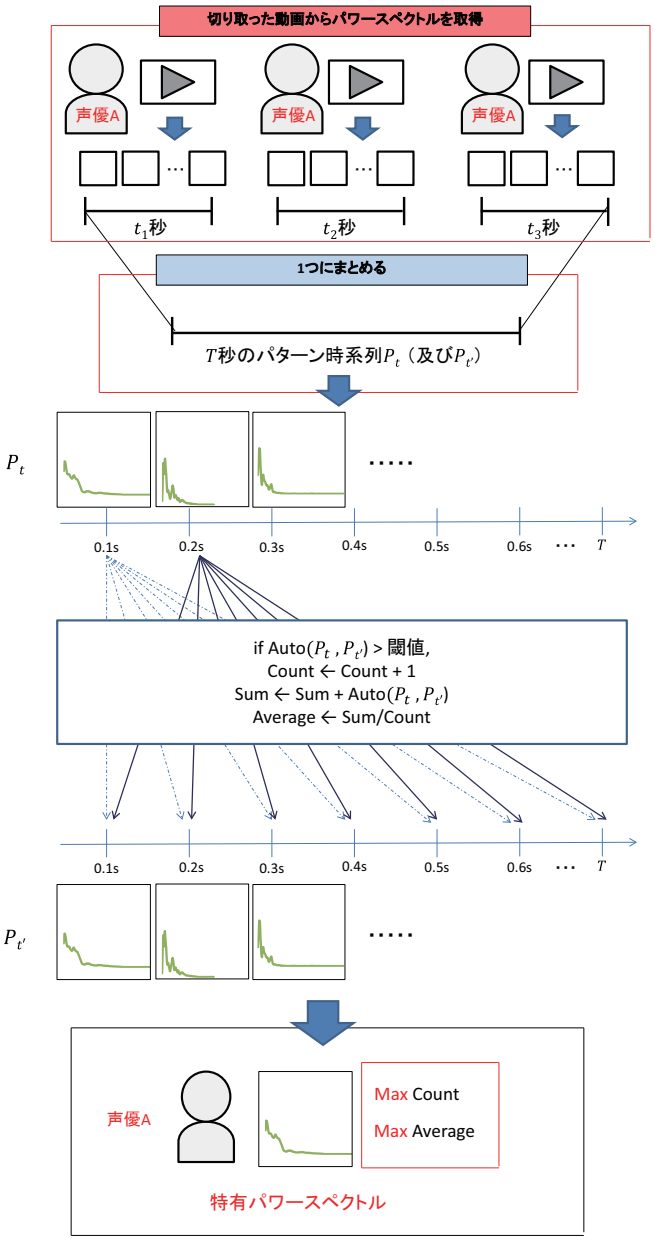

図 4　特有パワースペクトル探索アルゴリズム

5.　声優認識アルゴリズム

本章では，実際にアニメ動画から流れる音声から，そのアニメ動画内に出て来た声は誰の声優のものかを認識する手法を説明する．声優認識する為の提案システムの詳細は 3 章に載っている図 2 に示している．図 2 のようにアニメ動画が流れると 0.1 秒毎のその瞬間のパワースペクトルを取得した $\boldsymbol{v_t}$ と，声優データベースに予め用意されている各声優 i の特有パワースペクトル $\boldsymbol{a_i}$ との，それら 2 つの情報を用いて算出された類似度や相関を利用して，その動画内に流れた音声の持ち主の声優を判定する．但し，アニメ動画から音声が流れていない無音の時には $\boldsymbol{v_t}$ の数値は取得されない為，計算されない．$\boldsymbol{v_t}$ を取得して正規化を行ったバンド幅の dB の数値を $v_{t,1}, v_{t,2}, \cdots, v_{t,128}$ と置き直す．本稿では，声優認識する為に，互いのパワースペクトル $\boldsymbol{v_t}$ と $\boldsymbol{a_i}$ がどのくらい類似しているかを算出する計算式として，以下のコサイン類似度や相関係数のいずれかを用いる．

（1）　コサイン類似度

$$\boldsymbol{v_t} = (v_{t,1}, \cdots, v_{t,128}), \quad \boldsymbol{a_i} = (a_{i,1}, \cdots, a_{i,128})$$

$$\mathrm{sim}(\boldsymbol{v_t}, \boldsymbol{a_i}) = \frac{\sum_{j=1}^{128} v_{t,j} \cdot a_{i,j}}{\sqrt{\sum_{j=1}^{128} v_{t,j}} \sqrt{\sum_{j=1}^{128} a_{i,j}}}$$

（2）　相関係数

$$\boldsymbol{v_t} = (v_{t,1}, \cdots, v_{t,128}), \quad \boldsymbol{a_i} = (a_{i,1}, \cdots, a_{i,128})$$

$$\mathrm{correlation}(\boldsymbol{v_t}, \boldsymbol{a_i}) = \frac{\sum_{j=1}^{128} (v_{t,j} - \overline{v_t})(a_{i,j} - \overline{a_i})}{\sqrt{\sum_{j=1}^{128} (v_{t,j} - \overline{v_t})^2} \sqrt{\sum_{j=1}^{128} (a_{i,j} - \overline{a_i})^2}}$$

アニメ動画の再生が終了した後，声優を判定する処理に入る．再生中のアニメ動画から取得される 0.1 秒刻みのパワースペクトル $\boldsymbol{v_t}$（$t \in \{0.1, 0.2, 0.3, \cdots\}$）と各声優 i の特有パワースペクトル $\boldsymbol{a_i}$ との類似度や相関に対する閾値を予め定めておき，その定めた閾値を上回った時の $\boldsymbol{a_i}$ の声優 i のカウント数を+1 加算する．アニメ動画が終了した後，流れている間に加算された各声優 i のカウント数で比較して，閾値を上回ったカウント数を一番多く獲得した声優 i をそのアニメ動画から流れた音声の声優であると判定する．しかし，声優データベースに登録する特有パワースペクトルを特定する時と同様に，閾値を上回ったカウント数が同数になることがある．その場合には，特有パワースペクトルを特定する時と同様に，閾値を上回った時だけの類似度や相関の値をそれぞれ加算した平均で比較して，類似度や相関の平均が最も大きかった $\boldsymbol{a_i}$ を持つ声優 i を，そのアニメ動画から流れた音声の声優であると認識する．

6. 評価実験

アニメ動画 1 件を認識対象にして，本システムの声優認識精度に関して評価実験を行う．本評価実験では，Web からアニメ動画のキャスト情報を取得して，データベースの声優の人数を各話毎に絞り込めた状況であると仮定する．今回用いる評価実験用の 1 件のアニメ動画から得られるキャラクターと声優のペアは 6 組であるが，6 組中 2 組は次回予告の音声から登場する為，声優データベースには含むが評価実験の認識対象動画には含まない．以下に実験準備の条件を示す．

- 声優データベースの人数は 6 人に固定する
- 声優データベースを作成する時に用いた動画の種類には，認識対象動画と同一のアニメの各話の「アニメ声」と，声優の「素の声」に近いラジオからの 2 種類
- 認識対象動画は評価実験用の 1 件のアニメ動画からキャラクターの音声が流れている場面を切り取る
- 認識対象動画は BGM がある（物音，効果音も含む）場合と BGM が無い場合の 2 通り
- 認識対象動画の動画時間は各声優で統一していない
- 認識対象動画は 1 件に固定する為，各キャラクターの音声が流れている認識対象動画の数は統一していない

本システムの声優認識の精度を測る指標として，正答率と平均順位の以下の式を用いる．

$$\text{正答率} = \frac{\text{正解数}}{\text{認識対象動画の数（各声優毎）}}$$

$$\text{平均順位（6 人中）} = \frac{\text{データベース内の正解の順位の合計}}{\text{認識対象動画の数（各声優毎）}}$$

本システムは以下の項目に関して比較実験を行う．

（1） 声優データベースの作成に用いる類似性の計算式と，声優認識の際に用いる類似性の計算式
（2） 認識対象動画の BGM がある場合と無い場合
（3） データベースの作成に用いるアニメ動画の話数の違い
（4） 声優データベースの作成に用いる動画の種類

6.1 類似性の計算式の種類の比較

声優データベースを作成する時に用いる類似性の計算式には，4.2 節で論述されているコサイン類似度と相関係数の 2 種類を用いる．声優認識を行う時に用いる類似性の計算式は，5 章で論述されているコサイン類似度と相関係数の 2 種類を用いる．本実験では，この 2×2 通りの全 4 通りの組み合わせを用いた認識結果をそれぞれ比較して，本システムではどの計算式が最適なのかを比較していく．比較する際には以下の条件を固定する．

- BGM が無い場合の認識対象動画を用いる
- データベースに用いる動画はアニメ動画で話数は同じ

実験結果を表 1 に示す．表 1 から，声優データベース作成時に用いる計算式には相関係数が好ましいことが分かる．このような結果が得られたのは，類似性の計算式がコサイン類似度と相関係数との違いにより，各声優毎に選び出された特有パワースペクトルのパターンが異なる為であると考えられる．

6.2 認識対象動画の BGM がある場合と無い場合

認識対象動画には BGM がある場合と BGM が無い場合の 2 種類がある．この 2 種類に対してどちらの精度がより高いか比較実験を行う．比較する際には以下の条件を固定する．

- 声優データベース作成時に用いる計算式は相関係数，声優認識の時に用いる計算式はコサイン類似度を用いる

実験結果を表 2 に示す．表 2 によると，声優 C の結果に注目すると常に BGM がある場合の方が精度が高くなっている．声優 A や声優 B に注目すると，時々 BGM がある場合の方が精度が高くなっている箇所があるが，基本的には BGM が無い場合の方が精度は高い．声優 C の結果ではなぜ，BGM がある場合の方が高いか考察したところ，声優 C の認識対象動画を切り取った場面の BGM は全てあるワンシーンの同じ一定の音であった．つまり，声優 C の特有パワースペクトルとその部分の BGM の良く出て来るパワースペクトルが，類似していたのではないかと推察される．この結果から，本システムの声優認識は BGM に影響されている為，BGM に関する処理（例えば BGM 中のノイズ除去や BGM 自体の除去など）を行う必要があると考えられる．

6.3 声優データベース作成に用いる話数毎の違い

認識対象動画と同じタイトルのアニメ動画から声優データベースに登録する「アニメ声」を取得する．認識対象動画を 1 話に固定した時に，声優データベースに登録する「アニメ声」をアニメの初期の音声から最終話の音声に変化していくにつれて，声優認識精度がどう変化していくのか評価実験を行う．比較する際には以下の条件を固定する．

- 声優データベース作成時に用いる計算式は相関係数，声優認識の時に用いる計算式はコサイン類似度を用いる

実験結果を表 2，3 に示す．表 2，3 によると，初期の頃は認識精度が良くないが，中盤の辺りで精度が良くなり，終盤で悪化している．これは，初期の方では声優によるキャラクターの「アニメ声」が定着されていないからではないかと推測する．

表 1　類似性の計算式の種類に依る各声優の認識精度の比較

声優データベースのアニメの話数（2 話 (C 以外)，6 話 (C)）	認識対象動画	BGM 無し							
	データベースの計算式	コサイン類似度				相関係数			
	声優認識の計算式	コサイン類似度		相関係数		コサイン類似度		相関係数	
	評価尺度	正答率	平均順位	正答率	平均順位	正答率	平均順位	正答率	平均順位
認識対象動画（1 話の各声優のセリフ，BGM 無し）	声優 A (動画 10 件)	40%	1.9 位	40%	1.9 位	40%	1.8 位	40%	1.8 位
	声優 B (動画 9 件)	11%	2.8 位	11%	2.8 位	33%	1.8 位	33%	1.8 位
	声優 C (動画 6 件)	0%	3.0 位	0%	3.0 位	17%	2.3 位	17%	2.3 位
	声優 D (動画 2 件)	50%	1.5 位	50%	1.5 位	50%	1.5 位	50%	1.5 位

表 2　認識対象動画の BGM の有無と話数の異なる声優データベースに依る認識精度の比較

類似性の計算式	声優データベース：相関係数，声優認識：コサイン類似度							
データベース (アニメの話数)	2 話 (C 以外)，6 話 (C)				6 話 (D 以外)，7 話 (D)			
認識対象動画	BGM あり		BGM 無し		BGM あり		BGM 無し	
評価尺度	正答率	平均順位	正答率	平均順位	正答率	平均順位	正答率	平均順位
声優 A (動画 BGM あり 10 件，無し 10 件)	0%	3.4 位	40%	1.8 位	20%	2.2 位	70%	1.3 位
声優 B (動画 BGM あり 11 件，無し 9 件)	55%	1.5 位	33%	1.8 位	55%	1.5 位	56%	1.4 位
声優 C (動画 BGM あり 11 件，無し 6 件)	45%	1.9 位	17%	2.3 位	27%	2.5 位	0%	3.0 位
声優 D (動画 BGM あり 0 件，無し 2 件)	———		50%	1.5 位	———		0%	2.5 位

声優データベース：相関係数，声優認識：コサイン類似度												
···	9 話 (B，C)，6 話 (A)，8 話（D）				11 話 (C 以外)，12 話 (A)				13 話 (A 以外)，12 話 (A)			
	BGM あり		BGM 無し		BGM あり		BGM 無し		BGM あり		BGM 無し	
···	正答率	平均順位	正答率	平均順位	正答率	平均順位	正答率	平均順位	正答率	平均順位	正答率	平均順位
	10%	2.4 位	70%	1.3 位	0%	4.2 位	20%	2.5 位	0%	4.5 位	0%	3.3 位
···	90%	1.2 位	89%	1.1 位	82%	1.2 位	89%	1.1 位	27%	2.0 位	67%	1.4 位
	82%	1.2 位	67%	1.3 位	36%	2.9 位	0%	4.2 位	100%	1.0 位	67%	1.3 位
···	———		0%	3.0 位	———		50%	1.5 位	———		100%	1.0 位

表 3　声優データベースの種類の違いと話数の異なる声優データベースに依る認識精度の比較

類似性の計算式	声優データベース：コサイン類似度，声優認識：相関係数							
データベース (アニメの話数)	———		2 話 (C 以外)，6 話 (C)		6 話 (D 以外)，7 話 (D)		9 話 (B，C)，6 話 (A)，8 話（D）	
データベースの種類	ラジオ		アニメ		アニメ		アニメ	
評価尺度	正答率	平均順位	正答率	平均順位	正答率	平均順位	正答率	平均順位
声優 A (動画 BGM 無し 10 件)	40%	2.4 位	40%	1.9 位	90%	1.1 位	100%	1.0 位
声優 B (動画 BGM 無し 9 件)	22%	3.0 位	11%	2.8 位	56%	2.7 位	67%	1.6 位
声優 C (動画 BGM 無し 6 件)	17%	3.8 位	0%	3.0 位	0%	3.8 位	17%	2.3 位
声優 D (動画 BGM 無し 2 件)	0%	2.5 位	50%	1.5 位	50%	1.5 位	50%	1.5 位

声優データベース：コサイン類似度，声優認識：相関係数				
···	11 話 (C 以外)，12 話 (A)		13 話 (A 以外)，12 話 (A)	
	アニメ		アニメ	
···	正答率	平均順位	正答率	平均順位
	10%	2.4 位	10%	3.6 位
···	78%	1.2 位	56%	1.6 位
	82%	1.2 位	0%	3.3 位
···	0%	3.0 位	100%	1.0 位

6.4　声優データベースの作成に用いる動画の種類

声優データベースの作成に用いた動画には，認識対象動画と同じアニメの BGM の無い部分を切り取った箇所と，ラジオの動画の音声から取得した「素の声」の 2 種類がある．どちらの声優データベースを用いた方がより声優認識の精度が高いかを比較実験する．比較する際には以下の条件を固定する．

- BGM が無い場合の認識対象動画を用いる
- データベース作成時に用いる計算式はコサイン類似度，声優認識の時は相関係数を用いる

実験結果を表 3 に示す．表 3 によると，対象動画のアニメタイトルと同じ「アニメ声」の方が「素の声」より全体的に認識精度が高いのが分かる．しかし，部分的に「素の声」の方が認識精度が高いので，「アニメ声」と「素の声」を組み合わせた声優データベースを用意することで認識精度の改善が期待される．

7.　今後の研究課題

音声の周波数スペクトルに基づく声優認識の精度を改善する為には，以下のような点が今後の研究課題として挙げられる．

- BGM の対策（ノイズ除去）
- 声優データベースに登録する動画の種類の数
- バンド幅，及び，感度閾値の範囲を広くする
- 周波数のフィルタリング（低/高パスフィルタなど）
- 各話毎の特有パワースペクトルではなく，アニメの 1 話から最終話まで通した各声優の特有スペクトルを用いる
- 関連研究 [2] の特徴周波数と我々が提案した特有パワースペクトルとを組み合わせた手法
- 特有パワースペクトルが瞬間瞬間，局所的に固まって頻出しているのか，大局的に分散しているのかの違い

また，Web からキャスト情報だけでなく，登場人物の性別や年齢，性格なども取得して声優認識に活用できないか検討する．

文　献

[1] 榮田 基希, 服部　峻, “アニメ動画の音声とキャスト情報を用いた声優認識,” 電子通信学会, 情報ネットワーク研究会信学技法, Vol.115, No.405, pp.7–12 (2016).

[2] 森勢 将雅, “2 群 9 編 2 章 2-2 基本周波数推定（歌声に関する視点から）,” 電子情報通信学会「知識ベース」, pp.6–10 (2012).

[3] 早川 昭二, 板倉 文忠, “音声の高域に含まれる個人性情報を用いた話者認識,” 日本音響学会誌 51 巻 11 号, pp.861–868 (1995).

[4] 横山 雅夫, “音声に含まれる個人性情報,” 福島大学行政社会学会, 行政社会論集, 第 4 巻第 3 号, pp.96–113 (1992).