話者のイントネーションを模倣する
インタラクティブ声質変換システムの構築

足立 吉広† 森島 繁生††

†成蹊大学大学院 〒180-0001 東京都武蔵野市吉祥寺北町3-3-1

Tel:0422-37-3742 Fax:0422-37-3726

††早稲田大学理工学部 〒169-8555 東京都新宿区大久保3-4-1

Tel:03-5286-3510 Fax:03-5286-3510

E-mail:dm033203@cc.seikei.ac.jp, shigeo@waseda.jp

あらまし

音声への感情付加や発話強調、方言の付加などをインタラクティブに実現するシステムを提案する。これは予め蓄えられた自然音声もしくは合成音声に対して、マイクから入力された話者の音声を実時間分析して得られた韻律パラメータにしたがって韻律制御し、話者情報や言語情報はオリジナルのままイントネーションのみを実時間制御するシステムである。この制御法としてイントネーションの要素である話速、基本周波数、音量を、手本となる参照音声から自動抽出して、この値のみを変換することにより実現する。特に本研究では、話速制御で用いる補間法についてさまざまな手法を検討し、その合成音声品質について評価を行った。

Interactive Speech Conversion System
Tracing Speaker Intonation Automatically

† Yoshihiro ADACHI, †† Shigeo MORISHIMA

† Graduate School of Engineering, SEIKEI University

3-3-1, Kichijoji-Kitamachi Musashino, Tokyo, 180-0001, JAPAN,

Tel:0422-37-3726 Fax:0422-37-3742

†† Science & Engineering, Waseda University. 3-4-1 Okubo Shinjuku-ku Tokyo, JAPAN 169-8555

Tel:03-5286-3510 Fax:03-5286-3510

E-mail:dm033203@cc.seikei.ac.jp, shigeo@waseda.jp

Abstract

In this paper, an interactive speech conversion system which can modify emotional factor in sppech, emphasize utterance and append a resional dialect is presented. Prosody of stored sample voice is converted to a new prosody given by realtime analysis of microphone captured reference voice. Only intonation is controlled by converting utterance speed, pitch and power of speech by keeping speaker characteristic and linguistic information as original. Especially, an interpolation method of utterance speed is originally proposed and a quality of synthesized speech is evaluated subjectively.

1．はじめに

映画などに登場する人物の印象は、映像から受け取る表情の印象のみならず、同時に提示される音声の印象によっても大きく左右される。この顔の印象と声の印象が異なる場合に違和感を生じることもある。筆者のグループでは音声翻訳と唇動画像合成によって、映画の吹き替えなどを自動化するマルチモーダル翻訳システムを提案した[1]がその音声合成部分では、感情の制御は行われておらず無感情の音声合成に留まっている。

本稿では、このように予め蓄えれた音声や感情表現を伴わない合成音声に対して、マイクから入力された音声（参照音声）を実時間で韻律情報分析した結果に基づき、イントネーションをインタラクティブに付加する声質変換システムについて述べる。

３次元ＣＧによって故人となった名優をディジタルアクターとして再登場させたり、登場人物の声を音声翻訳技術によって自動吹き替えを行うシステムが現実のものとなりつつある。しかし、故人の音声は再収録することができないため、過去の作品等の音声データベースからコーパスを生成し、必要な台詞の音声を再合成しなくてはならない。また素片合成による音声合成システムは個人の特徴は保存されるが、感情をこめた音声を合成するためには、データベースのバリエーションを豊かにせねばならず、そのような収録が不可能な場合もある。そこで、話者情報や台詞はそのままに、イントネーションだけを声質変換できるシステムが存在すれば、リファレンスの音声を模倣する形で、声優の声のイントネーションにしたがって、故人の合成音声を感情を込めたものに変換することが可能となる。

まず、過去の音声データベースや音声合成によって必要な俳優の台詞の音声が存在するとする。これが処理対象となる入力音声である。次に、マイクに向かって「こういう音声を作りたい」という見本となる音声を発声する。これが参照音声である。参照音声は、発声終了後速やかに、発話速度、ピッチ、パワーの分析が行われて韻律情報が自動的に抽出される。入力音声も同様の処理によって、予め韻律情報分析が行われ､参照音声の韻律と同一になるように、発声の都度インタラクティブに変換が施され、暫く後に合成音声が出力される｡この処理を繰り返して、所望の感情のこもった音声ないしは方言を含む音声を作成することができるシステムを構築した。

音声の感情を規則化したり、制御する研究[2-11]はこれまで数多く発表されてきた。北原らは感情情報に関する特徴量として基本周波数、振幅、発話時間を扱ったが、合成音声は感情を上手く表現できていなかった[4]。そこで杉本は感情制御の物理量変換ルールとしてパワーと基本周波数・ホルマントのシフト・高域スペクトルの増加を結びつけた[5]。岩見は混合正規分布(Gaussian Mixture Model: GMM)に基づく声質変換法を用いた感情音声合成手法を提案している[6]。これは感情音声と平静音声から声質変換規則を学習し変換規則を任意の読み上げ口調の発話に適用し感情音声を生成するものである。一方飯田らは感情音声のコーパスを作成し、同じ感情分類から選出した素片を波形接続することにより合成音声を生成、評価した[7]。こうした感情制御の応用として阿部らは、音声ガイダンスの作成等の音声メッセージの作成時に感情を込めて台詞を読ませるツールを開発している[8]。

本システムのように、韻律情報変換を行って声質変換を実現するシステムにおいては、参照音声といかに同じイントネーションが生成できるかという点は勿論のことであるが、変換の前後において音質の劣化を最小限に食い止めることが課題となる。従来から著者らのグループではイントネーション変換システムの構築を行ってきた。しかしシステムの一部である話速変換で用いているスペクトログラムを時間軸上で伸縮させる為の補間法については根拠なく用いていた[3]。そこで本稿では補間法の違いによる性能比較の為に３つの手法を比較検討した。

まず第２章では、システムの全体概要を述べる。３章では韻律情報の抽出手順について述べる。第４章では韻律情報の変換方法、第５章では声質変換音声の評価結果を示す。最後に第６章で結論を述べる。

2．システム概要

システムの全体像を図１に示す。入力音声はイントネーション変換の処理対象となる音声で、たとえば故人の声のデータベースから作成されたものであり、予め存在するものとする。参照音声はイントネーションの見本となるものであり、その都度、マイクからユーザが入力することができる。この参照音声は、発話が終了するごとに発話速度情報、ピッチ情報、パワー情報が順に自動的に抽出される。これらの韻律パラメータを用いて、もとの入力音声が

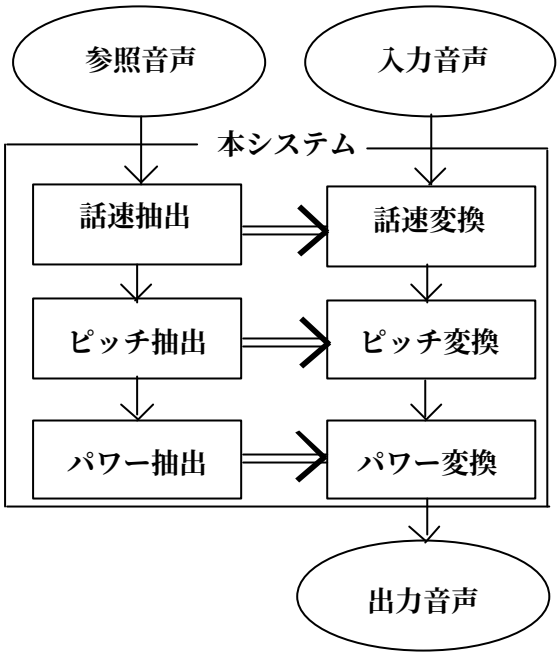


図1　システムの全体像

処理されて、参照音声を模倣したイントネーションが付加される。イントネーション変換プロセスは、まず発話速度を変換して参照音声と時間同期させた上で、ピッチ情報、パワー情報の変換が行われ、合成音声が出力される。この際、入力音声と参照音声の話者は当然異なってもかまわないが、台詞の内容は一致していることが条件となる。この参照音声として、声優が発声する感情のこもった音声を入力すれば、感情情報も思うように付加した台詞音声を合成することが可能となる。処理時間は、参照音声入力（サンプリング周波数16[kHz]，ビットレート16[bit]，音声長2.56[sec]）から、イントネーションを変換付加した台詞音声の合成まで、Pentium4(2.0MHz)搭載のＰＣで処理時間はおよそ7分34秒となる。この処理を繰り返すことによって、所望の感情のこもった音声や方言を付加された音声を合成することができる。

3．韻律情報抽出

3.1．話速抽出

参照音声や入力音声から話速を抽出するにはまず、音声を音素ごとに区切るセグメンテーションと呼ばれる処理を行う。入力された音声を自動セグメンテーションするために、ケンブリッジ大が開発したHidden Markov Model ToolKit(HTK)[12][21]に収録されている音声認識ツール群を用いた。本セグメンテーションシステム（以下、HTKセグメンテーション）では入力音声を音響特徴量に変換し、音響特徴パラメータとして12次元のメルケプストラム(MFCC)+12次元のデルタケプストラム+デルタパワーの計25次元を特徴ベクトルとして用いている。そして音響特徴量、音素ネットファイル、音響モデル（男性話者用、女性話者用、話者非依存）、音素辞書、音素ラベルを用いて音素セグメンテーションを行う。またHidden Markov Model(HMM)は、IPA(Imformation-technology Promotion Agency)が無料配布している音響モデルを使用している。HTKセグメンテーションにより音素の境界を取得することが出来れば（図2)、各音素の継続長を抽出し、話速パラメータとして得ることが出来る。

3.2．ピッチ抽出

ピッチは音声の基本周波数の遷移であり、この変化が発話の抑揚となる。また発話の全体的なピッチの違いは男女の声の高さを区別したり、子供と大人といった年齢の違いにも現れる。

ピッチの抽出法については、高精度な基本周波数抽出法として知られる、河原らが発表した周波数から瞬時周波数への写像の不動点を用いた音源情報の抽出法[13]を用いた。この手法はフィルタの中心周波数からフィルタの出力の瞬時周波数への写像の不動点を用いて、音声等の周期的信号の基本周波数と雑音成分を推定するものである。この手法を用いてピッチを抽出した例を図3に示す。

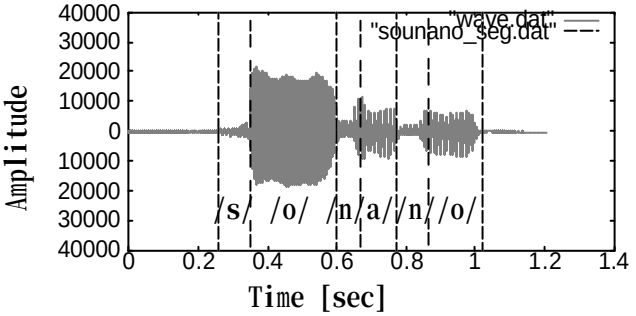


図2　セグメンテーションによる話速抽出
発話内容：「そうなの」

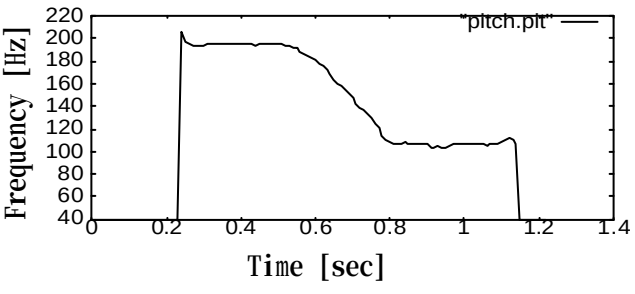
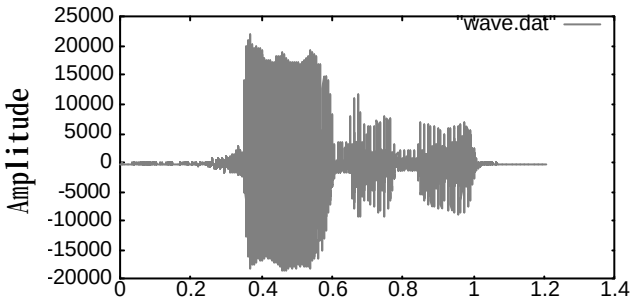
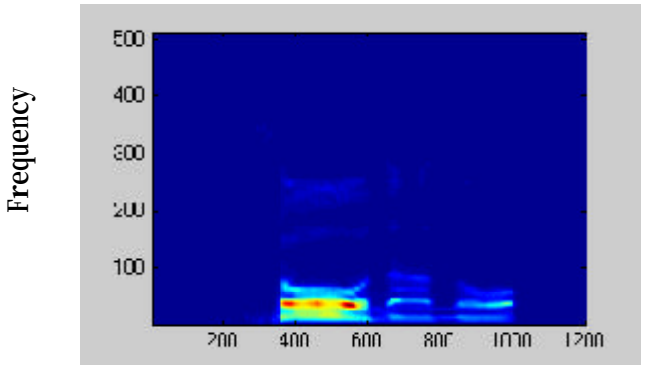


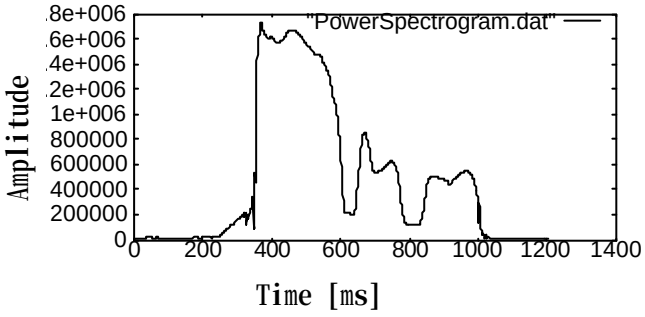
図3　ピッチ抽出 発話内容：「そうなの」



a) 音声



b) スペクトログラムに変換



c) パワー成分抽出

図4　パワー抽出　発話内容：「そうなの」

3.3. パワー抽出

音声の音量遷移を表すパワーは発話においてアクセントの役割を果たす。パワー抽出の流れを図4に示す。まずは音声（図4-a)をSTRAIGHT-coreを用いてスペクトログラムに変換する(図4-b)。STRAIGHT-coreというのは高品質音声分析変換合成法STRAIGHT (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram)[14]で用いられているスペクトログラムを抽出するアルゴリズムである。このスペクトログラムから周波数方向の振幅値の合計の遷移を図4-cのように求め、これをパワーとした。

4. 韻律情報変換

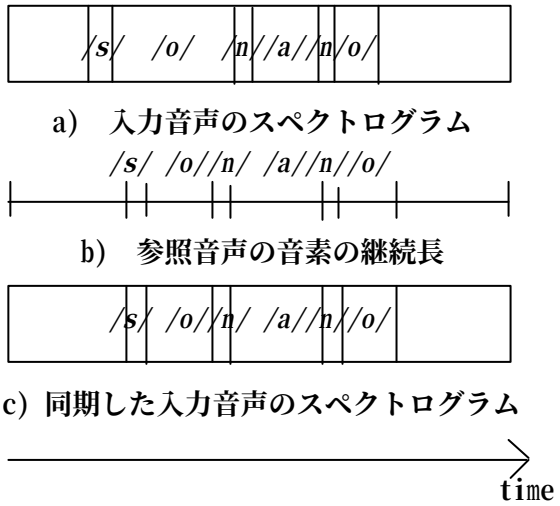
韻律情報変換は音声を一度パラメータ化し、所望の値に変換させ、その後音声に変換し出力する。これは音声をピッチとスペクトログラムに分解するVOCODERの考えに基づいており、音声波形に対して制御するよりも出力音声の音質劣化を防ぐことが出来るためである。音声のパラメータへの変換、またパラメータから音声への変換はVOCODERの原理に基づいているSTRAIGHTを応用した。話速、ピッチ、パワーといった各韻律情報の変換方法について本章で述べる。

本稿では話速変換のアルゴリズムに重点を置いているが、音声の話速を変換する技術は様々なものが発表されている。芹沢はウェーブレット変換を用いた手法[15]を提案している。池沢らは無音区間を伸縮する技術に加え話速を非線形に伸縮することによって聞きやすい、ゆっくりした音声に変換する研究を行っている[16]。

4.1. 話速変換

図1のシステム全体像で示した通り、入力音声に対して最初に行う韻律情報変換は話速の変換である。これはまず入力音声を参照音声に時間軸上で同期させることにより両者の対応をとりやすくするためである。

話速変換ではまず入力音声に対しHTKセグメンテーションを行い、音素ごとの継続長を抽出する。次にSTRAIGHT-coreにより入力音声から得られたスペクトログラムを、時間軸上で音素単位に分割する(図5-a)。そしてその音素単位のスペクトログラムを、参照音声から得られている対応する音素の継続長（図5-b）にあうように時間軸方向に伸縮する。分割された入力音声の全ての音素の継続長が、参照音

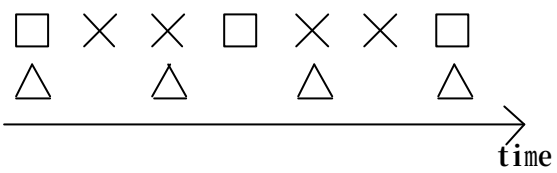


a)　入力音声のスペクトログラム

b)　参照音声の音素の継続長

c) 同期した入力音声のスペクトログラム

図5　スペクトログラムの伸縮概念図



□：元のデータ
×：補間により求めたデータ
△：新たなデータ

図 6　線形補間法

声から得られた継続長と同じになるように伸縮を行った後、再び一つなぎにすることにより、参照音声に同期したスペクトログラムを得ることが出来る(図 5-c)[17]。

ここでスペクトログラムの伸縮には 3 種類の補間法を試みた。線形補間法、Catmull-Rom スプライン補間法、3 次スプライン補間法である。これらは全て 1 次元の配列を伸縮するものである。

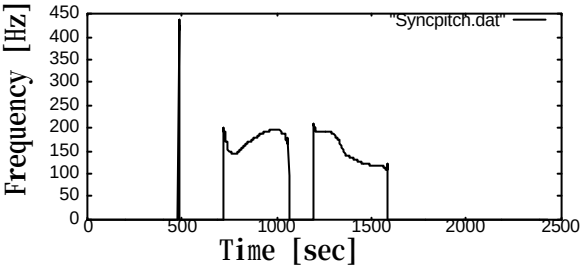
線形補間法とは標本値間を一度線形に補間しその補間された配列から等間隔に所望の長さとなるよう標本値をサンプリングする手法である。つまり、長さ l 個の配列を l' に伸縮する場合、まず元の配列の標本値間に $l'-2$ 個づつ、線形補間によって補間値を求める。こうして出来た長さ $(l-1)\times(l'-2)+1$ の配列から、$(l-1)$ 個ごとに l' 個サンプリングすることによって長さ l' 個の配列を生成する。こうすることにより、元の配列の最初と最後を採用し、その間を偏りなく補間することが出来る。この具体例を図 6 に示す。これは長さ 3 の配列を 4 に変化させたい場合である。図 6 の□で示したものが元のデータであり、その間に挟まれている×印は元のデータから補間により得られた補間値である。この□と×で構成される長さ 7 の配列から△で示される位置の値をサンプリングし、長さ 4 の新たな配列を生成する。

Catmull-Rom スプライン補間法とは標本値間を Catmull-Rom スプラインを用いて補間し、補間された配列から所望の長さとなるように標本値を等間隔にサンプリングする手法である。つまり線形補間法で標本値間を線形補間していた部分を Catmull-Rom スプラインで補間したものである。Catmull-Rom スプラインは 4 つの制御点、P_{i-1} P_i、P_{i+1}、P_{i+2} を用いて計算され、P_i と P_{i+1} の両点通り、その間を滑らかにつなぐ特徴がある。P_i と P_{i+1} の間の補間値 $x_i(t)$ を求める式を(1)に示す。

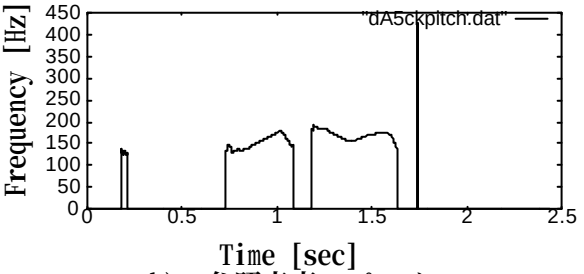
$$x_i(t)=\frac{1}{2}\begin{bmatrix}t^3 & t^2 & t & 1\end{bmatrix}\begin{bmatrix}-1 & 3 & -3 & 1\\ 2 & -5 & 4 & -1\\ -1 & 0 & 1 & 0\\ 0 & 2 & 0 & 0\end{bmatrix}\begin{bmatrix}P_{i-1}\\ P_i\\ P_{i+1}\\ P_{i+2}\end{bmatrix} \quad \cdots(1)$$

ここで t は媒介変数であり、$t=0$ の時 $x_i(0)=P_i$ となり、$t=1$ のとき $x_i(1)=P_{i+1}$ となる。

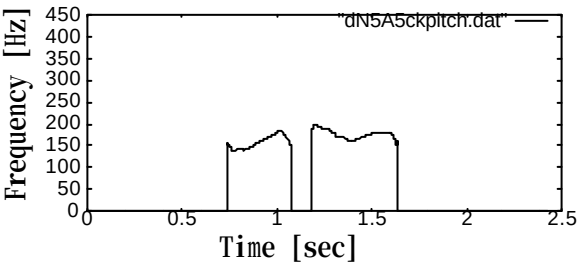
3 次スプライン補間法は W.H.Press 等が書いた「Numerical Recipes in C」に載っている spline，splint 関数を用いた。これは $x_1<x_2<\cdots<x_N$ を満たす点と関数値 $y_i=f(x_i)$ を入れた配列 $x[1..n]$，$x[1..n]$ が与えられ、点 1，n での補間関数の 1 階導関数の値 $yp1$, ypn が与えられたとき、各点 x_i での補間関数の 2 階導関数の値を入れた配列



a)　入力音声のピッチ(話速変換後)



b)　参照音声のピッチ



c)　出力音声のピッチ

図 7　ピッチ変換　（発話内容：どうしても）

$y_2[1..n]$を返す。この$y_2[1..n]$と小さい順に並んだ標本点xa_iと関数値を入れた配列$x_a[1..n], y_a[1..n]$が与えられ、xの値が与えられたとき、3次スプラインによる補間値を返すものである[18]。

4.2. ピッチ変換

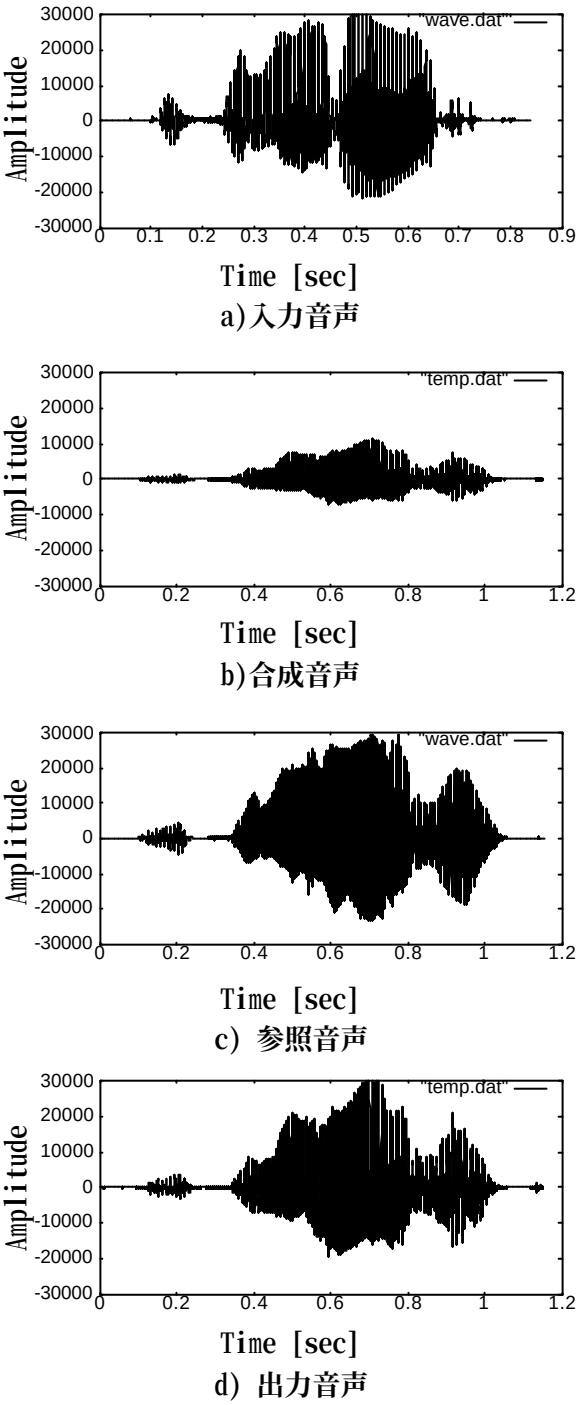
図7にピッチ変換過程を示す。図7-aは入力音声のピッチ、図7-bは参照音声のピッチであり、図7-cは出力音声のピッチである。入力音声のピッチ変換は、話速変換の後に行われる為、既に参照音声と入力音声の時間軸上での同期がとれており、音節を気にすることなく入力音声のピッチを参照音声から抽出されてたピッチと同じにすることによって行う。ただし、男性の声を女性の声のイントネーションに模倣するといった、音声全体のピッチの変化が著しい場合、音質や話者性が損なわれる可能性がある。そこで、入力音声のピッチの平均は韻律情報変換後も同値となるようにした。つまりピッチ変換前にピッチが抽出された区間のみから平均値を算出し、ピッチ変換が行われた後も平均値は同じになるよう両音声のピッチの平均値から補正率を割り出し調整した。

4.3. パワー変換

パワー変換は、入力音声に対してパワー抽出を行い、参照音声のパワーと単位時間ごとに比較し補正率を割り出す。この変換も話速変換の後に行われている為、両者のパワーは時間軸で同期している。時刻tにおける参照音声のパワーを$P_{Model}(t)$、入力音声のパワーを$P_{Input}(t)$としたとき(2)式で表される比率$r(t)$を、スペクトログラムの時刻tにおける周波数方向の振幅値にかける。

$$r(t) = \frac{P_{Model}(t)}{P_{Input}(t)} \qquad \cdots(2)$$

このようにスペクトログラムを変化させた場合、図8-aのような入力波形は図8-bに示すような出力波形となる。これは図8-cに示す参照音声の波形と振幅のレンジが異なるものの同様な形をしているので、参照音声、合成音声から振幅のピーク値を検出し、その比を割り出し、波形レベルで各サンプル値に直接かけあわせることによって図8-dに示すような波形となり、その結果パワーの遷移、振幅値のピーク共に揃う。



a)入力音声
b)合成音声
c) 参照音声
d) 出力音声

図8　パワー変換
発話内容：「いつもどおり」

5. 声質変換音声の評価実験

5.1. 評価実験方法

評価実験は客観評価、主観評価を行った。評価内容は、客観評価では線形補間法、Catmull-Rom スプライン補間法、3次スプライン補間法の3通りを用い

て声質変換した3種類の音声からピッチ、パワーを抽出し、参照音声のそれらと同じものになっているか相関を取る[19][20]。主観評価は声質変換音声と参照音声を聞き比べた時の印象に違いがあるかどうかを主観で答えるものである。

評価に用いた音声は、20代前半の男性5人が無感情に発話した「いつもどおり」、「そういうこと」、「どうしても」という発話内容の音声と、特徴的なイントネーションで発話した同一内容の音声で、これらを16[kHz]、16[bit]、モノラルで録音したものである。

5.2. 評価実験結果

最初に客観評価の結果を示す。スペクトログラムの伸縮法を線形補間法、Catmull-Romスプライン補間法、3次スプライン補間法としたときの3種類の出力音声と、手本とした参照音声とのピッチ、パワーの相関の結果を図9、図10に示す。相関値が1に近づくほど、参照音声のパラメータに近いことを示し、1と成るときは完全に一致することを示す。

図9、図10の結果からCatmull-Rom補間法が最も参照音声に近づける補間法と言えるであろう。

次に主観評価の結果を表1に示す。評価実験の被験者は20代前半、10名である。まず補間法の異なる

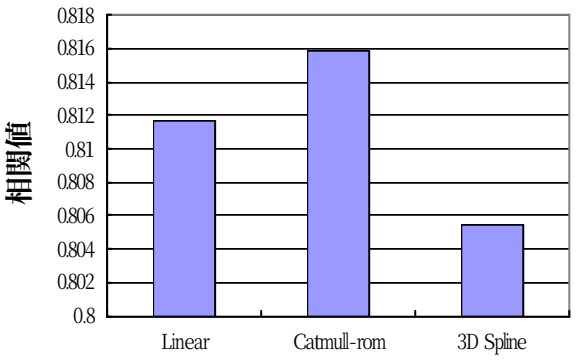


図9　ピッチの相関

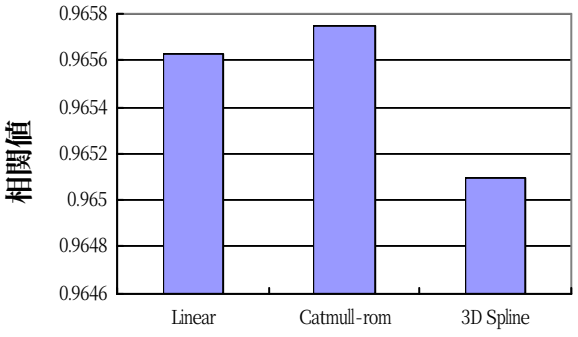


図10　パワーの相関

表1　主観評価実験結果

補間法の違いによる差が知覚された割合	4%
参照音声と韻律変換音声の印象が異なる割合	51%

3種類の声質変換音声間で、その違いが認識されるかを調査したところ、4%の確率で違いが知覚された。これは補間法にほぼ差がないことになるが、違いが知覚された場合、どの補間法が最も音質が良いかを調査したところ、全150音声の回答のうち3次スプライン補間法が最もよいとの回答が3音声、次に線形補間法が2音声、そして最後にCatmull-Romスプライン補間法が1音声であった。これは誤差の程度に収まる数値と考えられるので、補間法による違いはないと考えられる。また参照音声と声質変換音声の受ける印象が同じかどうかの調査では、51%の確率で異なることがわかった。話速、ピッチ、パワーといったパラメータは客観評価の結果から模倣していることがわかるが、それだけでな足りないパラメータがあるため、聞いたときの印象がことなるということを示唆している。

6. まとめ

故人となった名優をディジタルアクタとして再登場させたり、登場人物の声を音声翻訳技術によって自動吹き替えするといったことの実現を考えた場合、故人の音声は再収録することが出来ない為、過去の音声データベースからコーパスを作成し必要な台詞の音声を再合成する必要がある。この時同じ発話内容であってもイントネーションの違いにより不自然性が生じてしまう為、コーパスに多くのバリエーションを蓄えなくては対応しきれない。またコーパスに新たなバリエーションに追加したくても、再収録が不可能な為そのようなことは出来ない。そこで話者情報や台詞はそのままでイントネーションだけを変換する技術が実現すれば、リファレンスの音声を模倣する形で、声優の音声のイントネーションに従って故人の合成音声を感情を込めた形に変換することが出来る。

これまで著者等のグループはイントネーションの自動変換の実現に向けて研究を行ってきた。本稿ではイントネーション変換のアルゴリズムに組み込まれている話速変換で用いる補間法について3つの手

法を用いて比較検討を行った。従来はスペクトログラムの補間法にCatmull-Romスプライン補間法を用いていたが、そこに根拠はなかった。そこで、補間法について検証するため、線形補間法と3次スプライン補間法を用いて3種類の韻律情報変換音声を生成し、客観、主観の評価実験を行った。客観評価実験では、3種類の韻律情報変換音声と参照音声との韻律情報の相関をみることで、正確に韻律情報を模倣できているのかを調査した。その結果はわずかではあるがCatmull-Romスプライン補間法を用いたとき、最も似たパラメータを持つ音声が生成されることがわかった。また主観評価実験では3種類の韻律情報変換音声の中で最も音質が良いと感じるのはどれかを答えてもらうことで音質劣化を防げている手法を調べた。その結果96%の確率で音質に差が感じられないとの結果であった。また参照音声と韻律情報変換音声の印象を比べたところ、51%の回答が印象は異なるという結果を得た。これは本稿で扱ったパラメータ以外に印象を構成するものが存在することを示唆している。

本システムにより自然音声のイントネーション変換を行うことができるようになったが、参照音声と声質変換後の音声を聞いて受ける印象は異なることが主観評価実験結果からわかったので、両者に含まれる印象を構成するパラメータを明らかにし､制御、評価することが今後の展開として考えられる。韻律情報を変換するように印象も変換することが出来れば、音声のイントネーション変換から感情変換に拡張することができると考えられる。

参考文献

[1]緒方信、中村哲、森島繁生「ビデオ翻訳システム－自動翻訳合成音声とのモデルベースリップシンクの実現－」インタラクション2001 pp.203-210, 2001

[2]足立吉広、前島謙宣、四倉達夫、森島繁生「音声のパラメータ変換によるイントネーション変換システムの構築」情報通信学会　2002.3.19

[3]中野篤志、足立吉広、森島繁生「音声の韻律情報の変換によるイントネーション変換システム」情報通信学会　2003.3.26

[4]北原義典、東倉洋一「音声の韻律情報と感情表現」信学技法 ,SP88-158, Mar,1989

[5]杉本隆「音声中の感情表現に関連する物理量とその制御に関する研究」北陸先端科学技術大学情報科学研究科平成11年度修士論文 2000.3

[6]岩見洋平「声質変換法を用いた感情音声合成手法」NAIST-IS-MT0151013、2003-3

[7]Iida,A., Higuchi, F., Campbell, N., Yasumura, M., “A corpus-based speech synthesis system with emotion,”Speech Communication. Vol. 40/1-2 pp. 161 - 187.

[8]阿部匡伸 , 水野秀之 , 中嶌信弥「様々な音声表現を実現できる音声作成ツール Speed97」研究報告「音声言語情報処理」アブストラクト No.017 - 012

[9]門谷信愛希 , 阿曽弘具 , 鈴木基之 , 牧野　正三「音声に含まれる感情の判別に関する検討」研究報告「音声言語情報処理」アブストラクト No.034 - 008

[10]小池和仁 , 斎藤博昭 , 中西正和「感情音声の合成」研究報告 「音声言語情報処理」 アブストラクト No.024 - 012

[11]須田俊之「音響声道模型を用いた感情音声の生成に関する研究」http://hafu0103.me.kagu.sut.ac.jp/haralab/pdf/03tyuukan/suda.pdf

[12]“HTK Book”http:htk.eng.cam.ac.uk

[13]河原英紀、Parham Zolfaghari、Alain de Cheveigne、Roy.D.Patterson「周波数から瞬時周波数への写像の不動点を用いた音源情報の抽出について」信学技報　音声研究会　1999.7.8

[14]河原英紀「聴覚の情景分析が生み出した高品質VOCODER:STRAIGHT」　日本音響学会誌 54 巻 7 号 (1998), pp.521-526

[15]芹沢, 菊池, 中静, 佐々木, 渡辺「 ウェーブレット変換を用いた話速変換」電子情報通信学会技術研究報告, DSP94-42, pp. 145-152, June 1994.

[16]池沢龍 , 中村章 ,清山信正 , 都木徹 , 宮坂栄一「話速変換に伴う時間伸張を吸収するための一方法」研究報告 「ヒューマンインタフェース」 アブストラクト No.044 - 007

[17]松井九美 , 河原英紀 , 「STRAIGHT による Morphing 音声の聴覚的印象について」 日本音響学会春季大会講論, I, pp.523-524, Tokyo, Mar 2003.

[18]W.H.Press, B.P.Flannery, S.A.Teukolsky, W.T.Vetterling「Numerical Recipes in C 日本語版」株式会社技術評論社、平成5年6月25日

[19] 石村貞夫 「すぐわかる多変量解析」 東京図書株式会社　1992.10.26

[20]http://psy.isc.chubu.ac.jp/~mizunor/teaching/freshman_b/excel_function/equation_txt2.html

[21]http://htk.eng.cam.ac.uk/docs/history.shtml