CE-002

統計的歌声合成の音素タイミングモデル構築に向けた歌唱データベースの統計解析

Statistical analysis of the singing database for phonetic timing model in the statistical singing synthesis

森勢 将雅†‡
Masanori Morise†‡

1. はじめに

歌声合成とは歌詞と譜面を入力すると歌声波形を生成する技術の総称であり，代表的な歌声合成ソフトウェアである VOCALOID [1]の発売後は，多くのクリエイタが歌声合成ソフトウェアを活用したコンテンツを公開している．歌声合成は，テキストから音声を生成するテキスト音声合成と類似した技術で構成されており，初期のアルゴリズムには Unit selection [2]が用いられてきた．テキスト音声合成に関しては，その後，HMM (hidden Markov model)による統計的パラメトリック音声合成[3]が提案され，Unit selection よりも相対的に少ない学習データから，読み上げ内容を理解できる程度に自然な音声が生成可能となった．2013 年には，DNN (Deep neural network)を用いたテキスト音声合成 [4]が発表され，2017 年に提案された Tacotron [5]をはじめとする End-to-End 方式により，人間の音声と等価な品質での音声合成が実現された．

歌声合成に関してもテキスト音声合成と同様に進化を続けており，2016 年には DNN を用いた方法が提案されている[6]．高い品質での歌声生成を実現する統計的手法（以下では統計的歌声合成とする）では，Sinsy [7]や Neural parametric singing synthesizer [8]などが提案されている．統計的歌声合成の品質は，現状で十分に高い一方多くのデータ量を必要とし計算コストが高い．そのため，現在の検討は，より効率的に学習を進めるため CNN（convolutional neural network）による歌声合成[9]や，計算量の削減[10]など複数の方向性に発展しつつある．計算量の削減は，FastSpeech [11]や Human-in-the-loop 型音声合成[12]のように，合成に関し人間の操作が入ることを想定した研究に貢献する．これは，人間の操作が入るため波形の合成までにかかるステップが多く，各 DNN の出力にかかる時間が長い場合は作業のストレスに繋がるためである．

筆者らは，統計的歌声合成の次のステップとして，文献[11, 12]と同様に，ユーザが高さなどの情報を操作することが可能な統計的歌声合成技術に向けた検討を進めている．本論文では，この技術への要求事項である，効率的なモデル学習の基盤構築を狙う．効率的な学習のためには，CNN の利用のようにニューラルネットワークの構造を工夫するアプローチだけではなく，入力する学習データを吟味し，精度向上に有効な特徴量を選定するアプローチも存在する．本論文では，推定に適した特徴量の策定を目指し統計的な性質の解析を実施する．統計的手法は，1 名の話者や歌手により構築されたデータベースを用いるため，特定のデータベースに絞り統計的な性質について調査した．

現状の統計的歌声合成では，歌詞と譜面の情報から波形に至るまでの処理を，複数の DNN（処理の一部には DNN を用いないモデルも含む場合もある）を組み合わせることで実現している．本論文では，その中でも歌詞と譜面に関する情報のみから生成が必要な音素タイミングモデルに着目する．このモデルの目的は，歌詞と譜面の情報に対し音素境界の時間的なズレを推定することである．

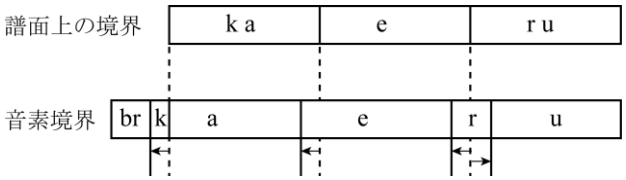

図 1：音素タイミングのズレ（/br/はブレスを示す）．以下ではタイムラグと表記する．

2. 関連研究と比較した本研究の位置付け

本研究の最終的な目標は，ユーザが高さ等のパラメータを操作可能な統計的歌声合成技術の開発であり，実現の基盤となる効率的な学習モデルの構築が本論文の狙いである．本章では，既存の技術やフレームワークについて説明し，本論文で実施する統計解析の位置付けを明確にする．

2.1 音素タイミングモデルの定義

音素タイミングモデルは，図 1 に示す譜面上の境界と実際の歌声の音素境界のズレ（タイムラグ）を推定するためのモデルである．歌詞と譜面に関する情報のみが与えられているため，入力する特徴量は，それらのみから選定しなければならない．具体的には，第一に，音素や譜面上の音高と音符長，それらの時間的な遷移が候補となる．加えて，音符の直前にブレス（図 1 の/br/に相当する）や促音などの特殊な成分が含まれることや，音符にスラーやスタッカートが含まれかを示す情報も利用可能である．利用可能の特徴量は言語のシラブルにも影響するため，本論文では日本語に限定して議論を進める．

統計的歌声合成では，歌詞と譜面から波形に至るまで複数の推定モデルを組み合わせるため，音素タイミングモデルについても複数のアプローチが存在する．例えば Sinsy では，DNN を用いず 1 次元ガウシアンを用いた方法を採用している[13]が，Neural parametric singing synthesizer では，類似した部分を Note timing model とし，DNN を用いて推定している[14]．タイムラグの推定後には，各音素の持続時間を推定するモデル，音素と持続時間から音高や音色に関する情報を推定するモデルなど，複数のモデルを経て最終的な波形を出力する．本論文では，図 1 に示す音素タイ

† 明治大学，Meiji University
‡ JST さきがけ，JST PRESTO

ミングモデルに限定し，タイムラグと対応付けが容易な特徴量の策定を目指す．

2.2 解析対象とする歌唱データベース

テキスト音声合成や統計的歌声合成では，特定の話者・歌手を対象に多数の音声・歌声を収録する必要がある．この際，音素の境界に相当する音素ラベルが付与されていること，および統計処理に対し十分な量を確保していることが重要である．音声の読み上げでは，NIT ATR503 M001 [15]や JSUT コーパス [16]などが，条件を満たすデータベースとして公開されている．音素ラベルは提供されていないが多人数の音声を対象にしたデータベースでは，LibriTTS [17]などが公開されている．LibriTTS は音素ラベル情報を提供していないが，1,000 人以上の音声を収録した大規模な音声データベースである．

歌声合成に関してもいくつかのデータベースが公開されており，音素ラベルまで含むものでは NUS-48E [18]，NIT SONG070 F001 [19]，JSUT-song [20]が公開されている．音素ラベルは手動で設定する必要があり，熟練度や好みにより傾向がばらつくことが想定される．本論文では，筆者らが構築し，1 名が全ての楽曲に対し統一した基準でラベル付けを実施した東北きりたん歌唱データベース [21]を利用する．本データベースは，プロ声優である茜屋日海夏氏が所属するアイドルグループ i☆Ris がリリースした楽曲 50 曲分，音声区間約 57 分から構成される．歌詞と譜面に相当する情報は MusicXML で公開されており，統計的歌声合成に利用可能である．現在本データベースを用いた統計的歌声合成システムとして NEUTRINO [22]がリリースされており，ニコニコ動画においてもすでに 3,000 件以上の動画が公開されている．このことからも，十分に高い品質の歌声を生成できるポテンシャルがあるデータベースであり，解析対象としても適切と言える．ただし，既存の楽曲を歌う都合上，音声読み上げで利用される音素バランス文のような音素バランスの調整は行えていない．音素の出現頻度を解析すると，音素の出現頻度は大きく偏っていることが確認されている[23].

2.3 解析に向けた前処理

本解析を行う前に，いくつか外れ値に相当する成分を除去する前処理を実施している．まず，MusicXML と音素ラベルとの間には，全体的な時間ズレとなるオフセットが含まれていた．この影響を除去するため，筆者が 1 ms 単位で時間をシフトしながら楽曲を聴取し，違和感の無い範囲となるように補正用のオフセットを与えた．また，今回収録した楽曲では，部分的に英語読みの歌詞があることなどにより，1 つの音符に 2 モーラ以上割り当てられることもある．このような音符では，前半と後半のモーラのタイムラグに大きな差が生じることから，該当する音符は解析に含まないようにした．

本論文では，以上の前処理を経て残った音素を対象に，統計的な解析を実施することとした．本解析では，推定対象となるタイムラグの分布に対する解析と，歌詞と譜面から得られる特徴量の情報とタイムラグ情報との相関関係を明らかにする解析を目標としている．これらの解析結果から，音素タイミングモデルを構築するために重要な特徴量を策定する．

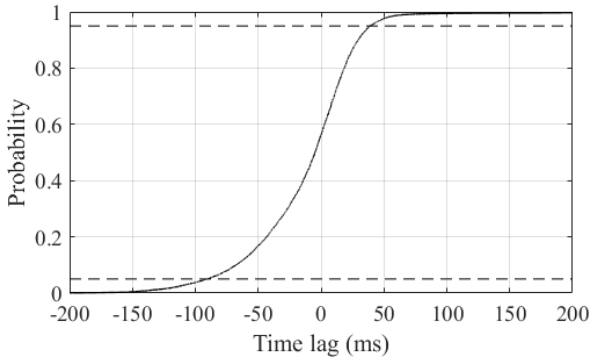

図 2：全音素を対象としたタイムラグの累積分布関数．破線は 5%と 95%を示すためにプロットしており，対応する時刻は，それぞれ-91 ms と 39 ms である．

3. 統計解析

本解析では，全音素を対象としたタイムラグの分布を出発点に，音素の種類に対する分布の差を解析する．DNN の学習時に分布の平均と標準偏差の正規化を実施するため，音素の種類により分布が異なる場合は，音素を適切に区分することで精度の向上につながる．その後，譜面から得られるいくつかの情報を対象に，タイムラグの分布との相関を分析する．相関の無い特徴量を省くことで，学習効率の改善が期待される．

3.1 全音素を対象としたタイムラグの分布

はじめに，ブレスや無音などの特殊なものを除く音素のタイムラグを対象として，累積分布関数を算出した．図 2 が結果であり，この図は合計 18,028 の音素から算出されたものである．図中の破線は 5%, 95%の範囲を示すためにプロットしている．それぞれの境界となる時刻は-91 ms と 39 ms であることから，実際の音符よりもやや手前の時刻で発声している傾向が確認できる．

以下の解析を実施する前に，本分布からいくつかの外れ値を除去することとした．具体的には，歌手が譜面とは明らかに異なるタイミングで発声している音符が観測され，特に，タイムラグの絶対値が 200 ms 以上となるものが 0.35%程度存在していた．これらの音素の前後を含む音源を聴取したところ，これらは歌手により意図された逸脱表現であると解釈できた．本解析では，意図せず表現されるタイムラグの統計解析を目的としているため，対象から除外することとした．以下の分析では，これらの音素を省き，タイムラグの絶対値が 200 ms 以内のもの 17,965 音素を対象とする．

3.2 タイムラグの分布に基づく音素の分類

タイムラグの分布は，音素毎に異なることが想定される．例えば，母音に限定しても，音符上子音が手前に来る母音と，単独で発声された母音（以下では単母音とする）であれば，発声タイミングに差が生じるだろう．子音についても，摩擦音や破裂音では差が生じると考えられる．既存のモデルでは，音素の分類を特徴量として用いており，例えば有声音／無声音の括りが用いられている．ここでは，全音素についてタイムラグの中央値を算出し，中央値に基づいて音素を分類する．分析に用いた歌唱データベースでは

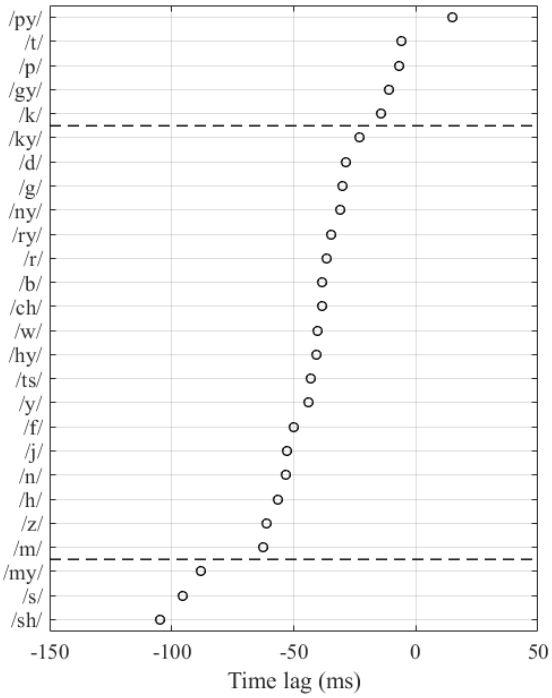

図 3：各音素について算出したタイムラグの中央値.

27 の子音（「ん」に相当する/N/を除く）が用いられているが，/v/については解析対象に含む音素の中での出現回数が 0 のため，結果には含んでいない.

全音素と分布の中央値をプロットしたものが図 3 である．上からタイムラグの値が大きい音素となるようにソートしており，破線に基づいて音素を 3 種類に区分している．一番上の区分は，/t/, /p/, /k/などが存在することから，破裂音に相当する音素の区分となる．破裂音は持続時間も短く，音符のタイミング付近で母音を発声する都合上，タイムラグが他の子音よりも 0 に近い傾向といえる．一番下の区分は，/sh/は/s/があることからも摩擦音に対応しており，摩擦音は音符のタイミングよりも 100 ms 程度早く発声していることが確認できる.

音素区分の境界については，3 音素区分それぞれについてタイムラグの分布を求め，分布から求められる標準偏差の和が最小となるように設定した．例えば，境界に存在する/ky/について，/k/と同一のカテゴリにした場合としない場合とで分布を算出して両分布の標準偏差を求め，標準偏差の和がより小さい区分にした結果が図 3 の境界である.

以下の分析では，図 3 に基づいて算出した 3 種類の子音区分に加え，母音についても表 1 に示すように，単母音と同一音符で子音＋母音で構成される母音とを分けることにした．これらの各区分についてタイムラグの分布を算出し，分布の平均値が異なり全体から求めた標準偏差を減少させることができれば，学習モデルの推定精度を向上できることにつながる．なお，母音については，音符に単独で存在させられるという理由から，「ん」に相当する/N/は母音に分類することにした.

3.3 分類された音素とタイムラグの分布

音素の区分による分布の傾向を計測するため，表 1 に示す 5 種類の音素区分それぞれについて，タイムラグの分布

表 1：5 種類の音素区分と該当する音素．/N/は，音符に単独で含められるため母音に分類している.

音素区分	該当する音素
単母音	/a/，/i/，/u/，/e/，/o/，/N/
母音	同音符内の先頭が子音時の/N/を除く 5 母音
子音 1	/py/，/t/，/p/，/gy/，/k/ （主に破裂音）
子音 2	/ky/，/d/，/g/，/ny/，/ry/，/r/，/b/，/ch/，/w/，/hy/，/ts/，/y/，/f/，/j/，/n/，/h/，/z/，/m/
子音 3	/my/，/s/，/sh/　（主に摩擦音）

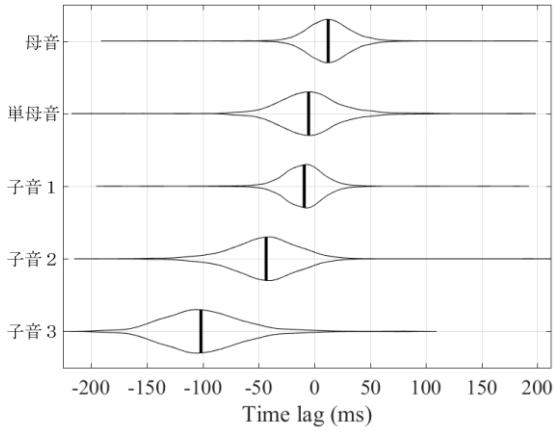

図 4：各音素区分について求めたタイムラグ分布のバイオリンプロット.

を算出した．バイオリンプロットとして表示した算出結果を図 4 に示す．各プロットの縦線は，中央値に対応する．最も遅れているのは子音の後に続く母音であり，単母音と子音 1（破裂音）が概ね近い中央値となる．その他の子音については，音符本来の時刻よりも早く発声していることが確認できる．子音 3（摩擦音）については，他の子音と比較して音符よりも早く発声しているといえる．音素の種類に応じて分布の中央値などが異なることからも，学習データとして用いる際の正規化では，音素の区分ごとに正規化したほうが高い性能が得られることを示唆する.

従来のモデルでは，有声音・無声音などの区分やシラブルに関する情報を用いている．例えば文献[14]では，Note Timing のモデルにおいて音素クラスの one-hot ベクトルを，Phoneme Duration のモデルでは，音節核など音節に関する区分の情報を用いている．様々な言語に対応することを想定すると，音節に関する細かな情報の区分は有効に働く可能性がある．一方，日本語に限定すればモーラ単位となるため，母音と単母音の区別，および子音に関する出力すべき情報に基づく区分でも有効であると考えられる.

音素区分については，主に破裂音，摩擦音とその他子音に相当する 3 区分としているが，この区分数の妥当性については，音素タイミングモデルの構築時に検討する必要がある．特に今回の検証では歌手が 1 名であるため，表 1 の分類は，歌手に対する依存性も存在すると考えられる．その場合，本解析と同様に音素毎のタイムラグについて分析することで，適切な区分が可能であるといえる.

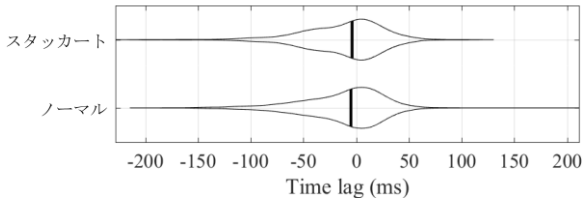

図 5：スタッカートを含む音符と含まない音符のタイムラグ分布の差

3.4 スタッカートの有無の影響

同じ音符に対する歌唱でも，技巧を付与するため表現を変化させることがある．歌唱技巧の代表的なものとして，ビブラートやポルタメントが存在する．これらは音符に対して明確に記述されるわけではなく，歌手個々人の判断により必要に応じて表現される．譜面上で陽に与えられる表現としては，スラーやスタッカートが該当する．今回用いた歌唱データベースにスラーは含まれていないが，スタッカートは全体の約 2.5%にあたる 435 音素分含まれている．スタッカート歌唱とスタッカートを含まない歌唱（ノーマル）との分布を比較することで，スタッカートの有無が分布に与える差を検証する．

図 5 が，スタッカートの有無について算出したバイオリンプロットである．どちらの図も音素の区分をしていないため，分布の形状は図 2 の累積分布関数に近い．両分布の中央値の差は 1 ms であり，分布の標準偏差の差は 2 ms であった．図 4 で示した差と比較すると差は小さく，スタッカートの有無が音素のタイミングに与える影響は 実質的にないと考えられる．

3.5 音高や音符長がタイムラグに与える影響

各音符には，音高と音符長が設定されている．音符長については，楽曲のテンポから単位を秒にした長さに変換可能である．ここでは，音高については音楽で用いられる単位として cent を利用し，音符長については時間(s)を単位として相関分析を実施する．高さの単位である cent は，半音の差が 100 cent となるものであり，ここでは C4 が 4,800 となるように以下の式により変換した．

$$c = 1200\ \log_2\left(\frac{f}{f_c}\right) + 4800,$$

ここで，cが変換後の音高(cent)，fが変換前の音高(Hz)であり，f_cが C4 に相当する周波数(Hz)である．これら音符の情報とタイムラグとの相関を確認することで，特徴量としての有効性について議論することが期待できる．音符長については，同じ音符でも楽曲のテンポにより実際の長さが変化するため，楽曲のテンポを用いて単位を時間に変換することとした．

各音素区分について算出した相関係数をまとめたものが表 2 である．相関係数を観測すると，子音２の音符長との相関が 0.1 以上である一方，それ以外では 0.1 未満であり，全ての条件で相関は低いことが確認された．既存のモデルには，音高や音符長を特徴量として用いることもあるが，これらの情報は相関が低く，音素タイミングの推定精度向上には寄与しない可能性を示唆する．

表 2：各音素区分に対する音高・音符長とタイムラグとの相関分析結果

音素区分	音高（cent）	音符長(s)
単母音	0.07	0.08
母音	−0.06	0.08
子音１	−0.05	0.05
子音２	−0.03	−0.11
子音３	0.01	0.01

表 3：4 つの音素区分に対する 1 つ前の音符との音高・音符長の差とタイムラグとの相関分析結果．ここでの母音は 1 つ前が同じ音符の子音であり，音高・音符長差が必ず 0 となるため解析対象から外した．

音素区分	音高差（cent）	音符長差(s)
単母音	−0.09	0.18
子音１	−0.02	0.04
子音２	−0.06	0.15
子音３	−0.12	0.16

表 4：各音素区分に対する 1 つ先の音符との音高・音符長の差とタイムラグとの相関分析結果．子音の次には同じ音符内で必ず母音が到来するため，解析対象から外した．

音素区分	音高差（cent）	音符長差(s)
単母音	0.01	−0.06
母音	0.04	−0.04

3.6 前後の音高と音符長がタイムラグに与える影響

当該音符に対する情報だけではなく，前後の音符に対する音高・音符長も入力特徴量として利用されることがあるため，3.5 節と同様に相関分析により検証する．１つ前の音素で解析する場合，同じ音符の場合は音高・音符長のどちらも同一の値になるため，前後の音素が異なる音符に含まれるという条件に限定して解析対象となる音声を選定した．前後の音符が休符の場合も解析対象から外している．１つ前については単母音と子音１，２，３の 4 種類について，１つ先については単母音と母音の 2 種類が解析対象となる．HMM を用いたモデル構築では，前後の音素間の音高・音符長の差を用いていたが，近年の DNN ベースのモデル構築では，複数の音符の情報をそのまま入力する場合もあり，これは，前後の情報に対し任意の重み付けが可能であることを意味する．一方，重みの最適化を含む解析では実験条件の組み合わせ数が増えるため，本解析では単純な音符間の差を算出することにした．音高については cent 単位で差分を求めることにした．

それぞれの解析結果を表 3, 4 に示す．表 3 からも，主に破裂音で構成される子音１については，音高・音符長のどちらも相関が 0.05 未満と低い傾向が確認できる．摩擦音に相当する子音３では，どちらの相関も絶対値が 0.1 以上であった．これは高い相関ではないものの，他の音素区分よりは相対的に強い相関である．全音素区分について音高差と音符長差を比較すると，音符長差のほうが相対的に高い相関係数であるといえる．今回の結果は単純な差分で算出しているが，適切な重み付けがなされている場合，さらに

強い相関になる可能性が残されている．音符間の差ではなく，DNN によるモデルで両方の音符長を用いることで適切な重みを学習することは，音符単独の音高・音符長よりも相対的に有効な特徴量となる可能性がある．

1 つ先の音符との差分では，どちらの相関も絶対値が 0.06 以下であった．これは，1 つ前の音素よりも低い値であり，こちらの特徴量はタイムラグに与える影響は小さい可能性を示している．以上をまとめると，1 つ前の音素との音符長差については他の情報よりも相対的に強い相関であり，DNN によるモデルに現音符と 1 つ前の音符の音符長を用いることは，精度向上に寄与する可能性がある．

3.7 前後の音素がタイムラグに与える影響

最後に，モデル構築にあたり，特殊な音素として促音やブレス（/br/），休符（無音）などが前後に含まれる場合，それが目的となる音素のタイムラグに影響するかどうかを解析する．1 音符につき 1 モーラという条件で解析しているため，上述の特殊な音素が直前に含まれる音素は，単母音と子音に限定される．直後が特殊音素にある場合は，単母音と母音に限定される．促音，ブレス，無音の存在がタイムラグに与える影響を，その他の音素と比較することが解析の目的である．

解析結果を図 6, 7 に示す．図 6 については，無音からブレスを挟まず発声された音素が 8 つしか存在しなかったため，今回の解析対象に無音を含めていない．図 6 からは，1 つ前の音素がその他の場合（今回では，単母音か母音のみ）と比較して，促音とブレスと間で約 13 ms，ブレスとその他音素の場合で約 18 ms の差が観測されている．これらの結果は，1 つ前の音素区分の情報は推定結果の精度向上に寄与する可能性を示唆する．

図 7 の結果からは，1 つ先に特殊音素が存在する場合でもしない場合でも，中央値の差は小さいことが観測できる．促音と無音の分布が歪んでいる原因は，解析対象となる音素数にある．今回の解析では，促音は 54 音素，無音は 182 音素であった． 特に，ブレスの場合とその他の場合との分布の差はほぼ等しく，1 つ先の音素は現在の音素のタイムラグに与える影響は，少なくとも 1 つ前の音素よりも少ない結果であった．

4. 考察

本解析では，いくつかの視点から統計的な性質について解析した．ここでは，それらの結果から，音素タイミングモデル構築に向けて議論する．

4.1 音素タイミングモデルに適した音素区分

本解析で示したように，音素を適切に区分することで，各分布の標準偏差を抑制することが可能である．図 2 に示した全音素から求めた分布の標準偏差は 45.5 ms であるが，5 つの音素区分に分けることで，母音，単母音，子音 1，子音 2，子音 3 の順に 21.2, 30.6, 21.6, 32.8, 36.8 ms に標準偏差を抑制している．

今回の解析対象となるデータベースでは，特定音素の出現回数が極端に少ないため，複数音素をまとめて区分している．現状の音素タイミングモデルにおいても，音素系列に関する one-hot ベクトルと有声音・無声音などの音素区分の one-hot ベクトルを併用している．全ての音素について十分な量が確保できる場合は全音素について分布の正規化を実施することが有効であると考えらえるが，限られた音素数で十分な精度を達成することを目指す場合，今回のように標準偏差を目安にして音素を区分することは有効であると考えられる．

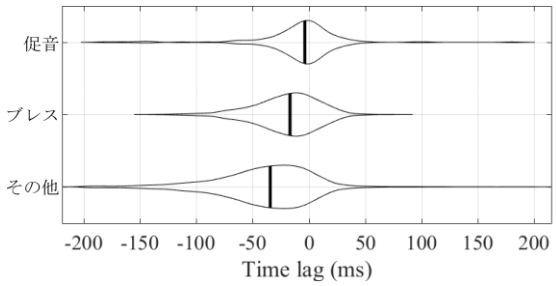

図 6：目的音素の 1 つ前に特殊音素があることがタイムラグの分布に与える変化．無音は，該当するパターンが 8 個しか観測できなかったため，含めていない．

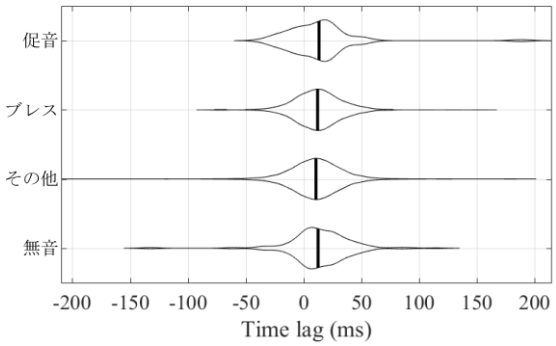

図 7：目的音素の 1 つ先に特殊音素があることがタイムラグの分布に与える変化

4.2 音素タイミングモデル構築に有効な特徴量の策定

音素タイミングモデルに用いる特徴量の詳細は，いくつかの文献で定義されている．ただし，シラブルなど言語固有の問題があるため，英語歌唱[24]や韓国語[25]，中国語[26]など言語ごとに合成法が検討されている．本論文では，日本語の歌声合成を前提に議論する．

例えば Neural parametric singing synthesizer の音素タイミングモデル[14]では，音素ではなく音素クラスの情報を one-hot ベクトルして与えており，対象時刻と 1 つ前の音符の音符長を与えている．Phoneme duration モデルでは，1 つ先の音素の情報も加えている．論文中では，音素クラスを具体的に定義していないが，これは言語のシラブルなどにより結果が変わるためであると考えられる．本論文では，日本語に限定し，タイムラグの標準偏差を抑制する区分を提案した．

今回の解析では，音高や音符長とタイムラグとの相関は，子音 2 の音符長との相関以外で 0.1 未満と低いことが確認された．1 つ前の音符との音符長の差については，主に破裂音に対応する音素カテゴリ以外において，0.15 以上と弱いものの，他の条件よりは相対的に強い相関が観測された．1 つ前の音符との音高差，および 1 つ先の音符との音高差，音符長差については子音 3 の音高差を除き 0.1 未満の相関

であった．一般的に相関係数の絶対値が 0.2 未満であればほぼ相関が無いといえるが，本解析では相関を求めるためのデータ量が 1,000 以上である場合も多く相関係数は低くなりやすい．そのため，手前の音符からの差については，各音符に適切な重みを与えることでより精度向上に寄与する特徴量になる可能性がある．

前後に特殊な音素が存在する場合の解析結果から，1 つ前の特殊音素は，種類によりタイムラグの中央値に影響することが示された．1 つ先の特殊音素は，タイムラグへの影響は実質的に無いといえる．以上のことから，音素に対する one-hot ベクトルと音素区分に対する one-hot ベクトル，および 1 つ前の音符の情報は有効であると考えられる．音符に与えられるスタッカートは，分布には影響しないため不要であると考えられる．

4.3 本統計解析の普遍性について

今回の考察は，1 名の女性声優から構築された歌唱データベースに限定したものであり，普遍性が高いとは言い難い．歌手の差だけではなく，ラベル付けの精度や担当者の癖による影響についても議論が必要である．本解析の最終的な価値は，本論文で示した特徴量を用いて構築した音素タイミングモデルが既存のモデルよりも高い精度を達成できることで示されるが，それは今後の検討課題とする．

構築したモデルの推定精度は，単純な特徴量の選定だけではなく，ハイパーパラメータのチューニング結果にも影響される．モデル構築は，本解析で策定した特徴量を利用するだけではなく，これらパラメータの最適化まで含めて実施する必要がある．

5. おわりに

本論文では，統計的歌声合成を実現するために必要な DNN モデルのうち，歌詞と譜面に対する音素タイミングを出力するモデルの特徴量策定に向けた統計解析を実施した．女性歌手 1 名にターゲットを絞っているが，解析の結果，音素の種類（母音，単母音，主に破裂音，摩擦音，その他に区分される 3 種類の子音カテゴリ）によりタイムラグの差が顕著にあることと，単母音と子音の場合において，1 つ前の音が促音やブレスの場合にタイムラグの分布に変化が生じることが示された．一方，音高や音符長，および前後の音符の変化量に対する相関は，1 つ前の音符長との差以外について，概ね 0.1 以下と低いことが確認された．

今後の課題は，本論文で得られた知見を組み込んだ音素タイミングモデルの構築である．従来のモデルで用いている特徴量を対象とし，それらと本研究で得られた特徴量のみで推定した場合との性能比較が必要となる．音素タイミングモデルの構築に加え，最終的な目的である波形まで出力するニューラルネットワークまで構築し，データベースの収録対象となる歌手と等価な歌い方まで再現可能な歌声合成システムの開発を目指す．

謝辞

本研究は，JST さきがけ JPMJPR18J8 の支援を受けた．

参考文献

[1] H. Kenmochi and H. Ohshita, “VOCALOID － Commercial singing synthesizer based on sample concatenation,” in Proc. INTERSPEECH 2007, pp. 4009-4010 (2007).
[2] A. Hunt and AW. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in Proc. ICASSP’96, pp. 373-376 (1996).
[3] H. Zen, K. Tokuda, AW, Black, “Statistical parametric speech synthesis,” Speech Communication, Vol. 51, No. 11, pp. 1038-1064 (2009).
[4] H. Zen, A. Senior, M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc. ICASSP2013, pp. 7962-7966 (2013).
[5] Y. Wang, RJ. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R.A. Saurous, “Tacotron: Towards End-to-End speech synthesis,” arXiv:1703.10135 (2017).
[6] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, “Singing voice synthesis based on deep neural networks,” in Proc. INTERSPEECH 2016, pp. 2478-2482 (2016).
[7] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, K. Tokuda, “Recent development of the HMM-based singing voice synthesis system – Sinsy,” in Proc. SSW7, pp. 211-216 (2010).
[8] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer,” in Proc. INTERSPEECH 2017, pp. 4001-4005 (2017).
[9] 中村和寛, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一, “歌声合成における CNN に基づく音声パラメータ生成手法の検討,” 音講論(春), pp. 1033-1034 (2019).
[10] 中村和寛, 高木信二, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一, “CNN に基づく歌声合成における計算量削減の検討,” 音講論(秋), pp. 939-940 (2019).
[11] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, T. Liu, “FastSpeech: Fast, robust and controllable text to speech,” in Proc. NeurIPS 2019, pp. 3171-3180 (2019).
[12] D. Kondo and M. Morise, “Human-in-the-loop speech-design system and its evaluation,” in Proc. APSIPA ASC 2019, pp. 608-612 (2019).
[13] K. Saino, H. Zen, Y. Nankaku, A. Lee, K. Tokuda, “An HMM-based singing voice synthesis system,” in Proc. INTERSPEECH 2006, pp. 2274-2277 (2006).
[14] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” Applied science, Vol. 7, No. 12, pp. 23-page (2017).
[15] HTS Working Group, “The NITech Japanese speech database NIT ATR503 M001,” [Online] http://hts.sp.nitech.ac.jp/ (2020/06/10).
[16] R. Sonobe, S. Takamichi, H. Saruwatari, “JSUT corpus: free large-scale Japanese speech courpus for end-to-end speech synthesis,” arXiv, vol. 1711.00354 (2017).
[17] H. Zen, V. Dang, R. Clark, Y. Zhang, R.J. Weiss, Y. Jia, Z. Chen, Y. Wo, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in Proc. INTERSPEECH 2019, pp. 1526-1530 (2019).
[18] Z. Duan, H. Fang, B. Li, K. Sim, Y. Wang, “The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech,” in Proc. APSIPA ASC 2013, pp. 1-9 (2013).
[19] HTS Working Group, “The NITech Japanese speech database NIT SONG070 F001,” [Online] http://hts.sp.nitech.ac.jp/ (2020/06/10).
[20] S. Takamichi, N. Tanji, H. Saruwatari, “JSUT collection ver. 1,” https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song (2020/06/10).
[21] 小川樹，森勢将雅，“アニメソングの統計的歌声合成に向けた歌唱データベースの構築，”日本音響学会 2019 年秋季研究発表会, pp. 1091-1092 (2019).
[22] https://n3utrino.work/ (2020/06/10)
[23] 小川樹，森勢将雅，“東北きりたん歌唱データベースを対象とした歌声の統計的解析，”日本音響学会 2020 年春季研究発表会, pp. 1121-1122 (2020).
[24] K. Nakamura, K. Oura, Y. Nankaku, K. Tokuda, “HMM-based singing voice synthesis and its application to Japanese and English,” in Proc. ICASSP 2014, pp. 265-269 (2014).
[25] J. Lee, H. Choi, C. Jeon, J. Koo, K. Lee, “Adversarially trained End-to-End Korean singing voice synthesis system,” in Proc. INTERSPEECH 2019, pp. 2588-2592 (2019).
[26] X. Li and Z. Wang, “A HMM-based mandarin Chinese singing voice synthesis system,” IEEE/CAA Journal of Automatica Sinica, Vol. 3, No. 2, pp. 192-202 (2016).