情報処理学会論文誌 Vol. 50 No. 2 575–586 (Feb. 2009)

ニュース番組の収録音声を利用した波形接続型音声合成システム

世木寛之[†1] 田高礼子[†1]
清山信正[†2] 都木徹[†1]

大規模な音声データベースから音声データを選択して接続する波形接続型音声合成が提案されている．この音声合成方式で利用される大規模音声データベースは，音韻バランスなどを考慮して選定された文章を，音声合成に適した話速やスタイルで読み上げることで作成されることが多い．一方，放送局では過去に放送された番組が大量に保存されているため，これらを音声データベースとして利用することが考えられる．本研究では，ニュース番組の収録音声を，波形接続型音声合成システムの音声データベースとして利用することを試みた．高い頻度で音声データベースに存在する音素列を，前後の音素環境を考慮して抽出した“音素環境依存音素列”を探索単位として合成音を作成し，5 段階のオピニオン評価実験を行った結果，MOS は 4.01 となり，「不自然な部分はあるが気にならない」という自然性を持つ合成音が得られた．特に，全体の 39.8%が 5 の「自然である」と評価され，自然音声と変わらない品質の合成音がかなりの頻度で作成されていることが分かった．次に，目標スコアを用いた場合と，用いない場合の合成音とを比較したところ，MOS の差は 0.18 となり，音声データベースの発話内容と合成する文が類似している場合には，必ずしも韻律予測せず目標スコアを考慮しなくても，自然性の高い合成音を作成できる可能性が示された．

Concatenative Speech Synthesis System Using Recordings of Japanese Broadcast News Programs as a Speech Database

Hiroyuki Segi,[†1] Reiko Tako,[†1] Nobumasa Seiyama[†2] and Tohru Takagi[†1]

Proposals have been made to implement a system that generates synthesized speech by concatenating segments of speech stored in large databases. While these databases are often created by recording sentences with a specific phonetic balance, read at a rate and in a style that are optimal for speech synthesis, this paper explores an alternative method of database creation, one that utilizes broadcast materials archived in networks. In our study, we used samples of recorded speech from news programs to create a speech database. An assessment of speech generated by the speech synthesis method using “context dependent phoneme sequences” as search units yielded the mean opinion score (MOS) of 4.01 in a one-to-five-scale rating. Overall, the samples were considered “somewhat unnatural but not bothersome.” In particular, 39.8% of the entire samples scored 5.0, demonstrating their highly natural-sounding quality. In addition, we compared the evaluation on “synthesized speech with target scores” and that on “synthesized speech without target scores.” The difference of MOS was 0.18. This result confirmed that prosody prediction or target scores are not necessarily required to create synthesized speech of natural-sounding quality when the content of input sentences is similar to the content of sentences stored in the database.

1. はじめに

日本語のテキスト音声合成（TTS: Text To Speech）システムはさまざまな方式が研究されているが，放送に利用するためにはいっそうの高品質化が期待されている．ニュース原稿や台本から自動的に合成音が生成できれば，ラジオの自動音声放送が可能になる[1)]．また，CG 画像と組み合わせればテレビ番組の自動作成[2)]も可能になる．さらに，受信機や番組制作装置において音声合成が可能となれば，目の不自由な方々や自動車の運転者でも，データ放送の文字情報を音声で聞き取ることができる[3)]．

近年，このような目的に利用可能な高品質な合成音を作成する手法として，隣り合う音声素片の基本周波数やケプストラムとの類似度（接続スコア）と，目標とする基本周波数や音素長などの韻律予測値との類似度（目標スコア）の和が高い音声素片を，大規模な音声データベースから選択して接続する波形接続型音声合成が提案されている[4)–8)]．この種の音声合成方式で利用される大規模音声データベースは，音韻バランスなどを考慮して読み上げ文章を選定し[9),10)]，スタジオで収録作業を行うことにより作成されることが多かった．また，読み上げる速度や，読み方などを指定して，後で合成しやすいように調整して収録することも行われていた[11)]．

†1 日本放送協会放送技術研究所
Science and Technical Research Laboratories, Japan Broadcasting Corporation
†2 財団法人 NHK エンジニアリングサービス
NHK Engineering Services, Inc.

これまでにも，100 時間規模の音声データベースを使用した研究は行われているが[12]，基本的に前述の収録方式に基づいたものである．一方，放送局では過去に放送された番組が大量に保存されているため，これらの収録音声を音声データベースとして利用することが考えられる．多種多様な番組があるため，発話内容，声質，感情など幅広い音声を合成できる可能性がある．本研究では，まずその中からニュース番組の収録音声を選び，波形接続型音声合成システムの音声データベースとして利用することを試みた．

ニュース番組の収録音声を音声データベースとして利用することを考えると，大量に存在することから，なるべく効率的に音声データベースの探索を行うことが望ましい．ニュース文を合成する場合，「でした」，「お伝えしました」など固定されたいい回しはそのまま使える可能性が高い．これまで使われてきた音節[4]，音素[5),6)]，半音素[7),8)] を音声素片の探索単位として用いると，「でした」の合成音を作るにも，音素の場合には「d」，「e」，「sh」，「i」，「t」，「a」の音声素片の組合せを考える必要があり，探索時間を余計に費やしてしまう．そこで，本研究では，「でした」のように高い頻度で音声データベースに存在する音素列を，前後の音素環境を考慮したうえであらかじめ抽出しておき，ひとかたまりの探索単位として利用する音声素片選択を行う．この探索単位は，あくまでも音声データベースで頻度の高い音素の並びであるため，単語のように言語上意味がある単位とは限らない．この探索単位を以下では，“音素環境依存音素列” と呼ぶことにする．

前述のように，波形接続型音声合成では，大規模な音声データベースから接続スコアと目標スコアの和が高くなる音声素片の組合せを選択するが，高品質な合成音を目的とすると，目標スコアの信頼性が問題になる．目標スコアとして，予測された基本周波数や音素長の値と音声素片の実際の値との差分をとることが多いが，予測された基本周波数や音素長に予測誤りや不自然さが含まれると，目標スコアの信頼性は低下する．しかし，現時点では，個人性を正確に反映した基本周波数や音素長の予測は，困難な課題である．

一方，ニュース番組の収録音声を音声データベースとして利用しニュース文を合成する場合には，音声データベースの発話内容と合成する文が類似しているため，目標スコアを考慮しなくてよい可能性がある．これまでの研究で，目標スコアを使わずに韻律コンテキストを用いる手法が提案されている[13]．しかし，日本語の場合，番組の収録音声を聞き取り文字に起こすのは比較的容易であるが，アクセントを聞き取り付与するのは標準アクセントを正確に判別できる人でないと難しい．そこで，本研究では，韻律予測を行わないだけでなく，韻律パラメータを予測するために用いる韻律コンテキスト情報さえも用いなくても，限定されたドメイン内では高品質な音声の合成が可能であることを示す．具体的には，音声データベースには含まれないニュース文を入力として，アナウンサの実際の発話である自然音声から抽出した基本周波数と音素長を使った予測誤りを含まない目標スコアを用いた場合と，目標スコアを用いない場合の合成音との比較を行い，目標スコアを考慮しなくても高品質な合成音を作成できる可能性を示す．

以下，2 章で，音声合成システムの概要について述べる，次に，3 章で，音声合成システムの各処理の詳細を述べる．4 章では，ニュース文を入力とした場合の合成音の自然性に対する評価試験と合成処理時間の検討を行う．5 章では，目標スコアがない場合でも，合成音の自然性が低下しないことを示すため，目標スコアを考慮した場合としない場合とで合成音を作成し，評価試験を行う．最後に 6 章でまとめを行う．

2. 音声合成システムの概要

音声合成システムの構成を**図 1** に示す．漢字仮名交じり文が音声合成システムに入力されると，テキスト解析部で形態素解析を行い音素に変換する．この際，音声データベースにアクセント情報がない場合には，アクセント情報は付与しない．

次に，韻律予測を行うが，5 章で検討するように，合成する文が，音声データベースの発話内容と類似していることがあらかじめ分かっている場合には，韻律予測をしない構成も設定できる．

続いて，この音素を，音声データベースの中で頻度が高い音素環境依存音素列のリストを

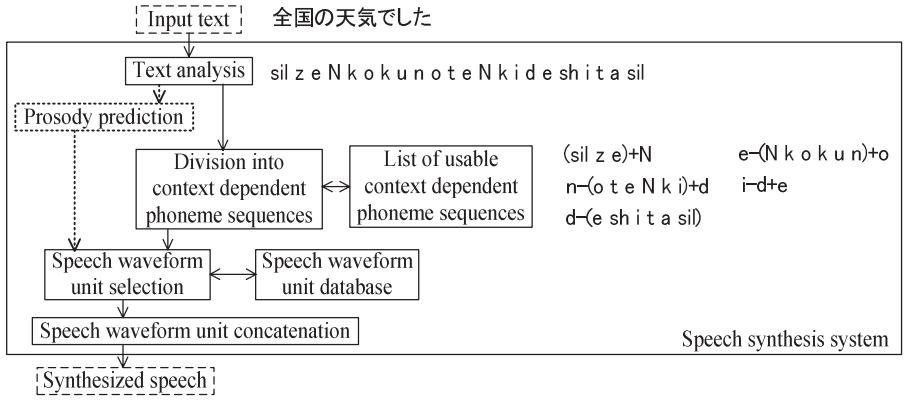

図 1　音声合成システムの構成
Fig. 1　Speech synthesis system.

用いて，個数が最少になりそれぞれの長さがなるべく均等になるように分割する．たとえば，「全国の天気でした (sil z e N k o k u n o t e N k i d e sh i t a sil)」は，「(sil z e)+N」，「e-(N k o k u n)+o」，「n-(o t e N k i)+d」，「i-d+e」，「d-(e sh i t a sil)」に分割される．ここで，「e-(N k o k u n)+o」は，前の音素が「e」で後ろの音素が「o」であるような音素列「(N o k u n)」の意味で，「sil」は文頭文末の無音である．分割の個数を最少にするよう分割する理由は，音声素片の探索回数を減らすためである．大規模音声データベースを利用した波形接続型音声合成の処理にかかる時間のほとんどは，音声素片の探索に費やされるため，探索回数を削減すると，音声合成の処理時間を大幅に短縮できる．また，音素列の長さがなるべく均等になる分割を選択する理由は，短い音声素片が連続して接続することで音質が劣化するのを防ぐためである．この分割方法の詳細については，3.3 節で述べる．また，音素環境依存音素列のリストの作成方法については 3.2 節で述べる．

この音素環境依存音素列への分割時に，前後 1 つずつの音素を考慮した音素（トライフォン）すら，音素環境依存音素列リストに存在していない場合には，音声素片の探索単位としてクラスタリングされたトライフォンで代用する．トライフォンのクラスタリングには Tree-based clustering[14] を使用する．

次に，分割された音素環境依存音素列に属する音声素片の中で，接続スコアと目標スコアの和が最大となる組合せの探索を行う．この探索方法の詳細は 3.4 節で述べる．

最後に，選択された音声素片の組合せについて接続を行い，合成音として出力する．

3. 音声合成システムの各処理について

3.1 合成用素片データベースの作成

音声データベースは，1996 年 6 月 3 日から 2001 年 6 月 22 日までに放送された NHK のニュース番組の中で同じ女性アナウンサが発声し，背景音が少ない 27,788 文と，同アナウンサが読み上げた音素バランス文 100 文の計 86.0 時間分を使用した．21,622 単語，総トライフォン数は 384 万，異なりトライフォン数は 8,771 である．ミキシング調整後の音声を 16 kHz，16 bit で収録したため，伝送による雑音などの問題はないが，VTR 再生，対談などの相手話者の声，効果音などがミキシングされた状態になっている．その中から比較的ノイズの少なく，アナウンサが単独で発声している部分を人手で抜き出して使用した．

音声データベースは文単位に分割され，それぞれの音声ファイルの発声内容は音素で書き起こされている．もともと音声認識を目的として作成された音声データベースであるため[15]，発声内容以外の情報は付与されていない．したがって，アクセントの情報も付与されていない．

音声データベースは，HMM を用いてアラインメントを行い[16)–18)]，各音素の発声時刻およびポーズの位置を自動的に決定する．また，基本周波数の抽出も自動的に行う[19]．アラインメント結果および抽出された基本周波数の結果には誤りを含む可能性もあるが，音声データベースの規模が大きいことから，人手による補正は行わなかった．

3.2 可変長の音素環境依存音素列リストの作成

音声データベースから前後の音素環境を考慮した可変長音素列リストを作成する方法について述べる．

合成の際，探索対象となる同じ音素列を構成する音声素片のデータ数（候補数）が多い方が，その部分で音質の良いつながりとなる最適な音声素片が見つかる可能性が高くなる．そのため，音声合成を行う際には，ある閾値以上のデータ数を持つ音素環境依存音素列のみ使用する．今回は予備実験の結果から，閾値を 100 個とした．

音声データベースにおける音素環境依存音素列は下記の手順で求める．

① まず，音声データベースの各文について，音素列に分解してトライフォンを生成し，それぞれのトライフォンに関してデータ数を計算する．たとえば，「sil ny u: s u d e s u sil」という文があった場合には，「(sil)+ny」，「sil-(ny)+u:」，「ny-(u:)+s」，「u:-(s)+u」，「s-(u)+d」，「u-(d)+e」，「d-(e)+s」，「e-(s)+u」，「s-(u)+sil」，「u-(sil)」の各データのカウント値を 1 つ増やす．音声データベースのすべての文でこの計算を行い，最後に閾値以上のカウント値を持つトライフォンを出力する．

② 次に，音声データベースの各文について，カウント値が閾値を超えたトライフォンに関して，音素の長さを伸ばし，データのカウント値を計算する．たとえば，閾値を超えたトライフォンが「u:-(s)+u」，「s-(u)+d」，「d-(e)+s」で，「sil ny u: s u d e s u sil」という文があった場合には，「u:-(s u)+d」，「s-(u d)+e」，「d-(e s)+u」の各データのカウント値を 1 つ増やす．先ほどと同様，音声データベースのすべての文でこの計算を行い，最後に閾値以上のカウント値を持つ音素環境依存音素列を出力する．

③ さらに，データ数が閾値を超えた音素環境依存音素列に関して，音素の長さを伸ばし，データ数を計算する．たとえば，閾値を超えた音素環境依存音素列が「u:-(s u)+d」のみで，「sil ny u: s u d e s u sil」という文があった場合には，「u:-(s u d)+e」の各データ数のカウント値を 1 つ増やす．先ほどと同様，音声データベースのすべての文でこの計算を行い，最後に閾値以上のカウント値を持つ音素環境依存音素列を出力する．

④ この操作を，閾値以上のカウント値を持つ音素環境依存音素列が存在する限り繰り返

し行い，対象としているすべての音素環境依存音素列の持つデータのカウント値が閾値を下回った時点で終了する．

⑤ 最後に，上記の計算を行っている途中に出力された音素環境依存音素列をマージし，可変長の音素環境依存音素列リストを作成する．

前述した 27,888 文の音声データベースに対して本手法を適用した結果，38,447 個の音素環境依存音素列が生成された．この音素環境依存音素列の最大の長さとなるものは 55 音素で，「(sil i Q p o: t o: ky o: k a b u sh i k i sh i j o: sh u y o: m e: g a r a n o h e: k i N k a b u k a n o o w a r i n e w a)+sp」となり，音声データベースに存在する決まったいい回しをまとめて探索することが可能で，その内部の音素の接続は考慮する必要がない分，探索時間を削減することができる．

3.3　入力テキストに対する音素環境依存音素列への分割方法

テキスト解析部から出力された 1 つの文に対する音素を入力として，個数が最少になる音素環境依存音素列に分割する．1 番目から r 番目までの音素を音素環境依存音素列に分割する最少分割点数 $M(r)$ は，

$$M(r) = \min_{0 \le k < r} [(M(k) + 1) \times \delta(e_{k+1,r})] \tag{1}$$

となる．ただし，$M(0) = -1$，$\delta(e_{k+1,r})$ は，$k+1$ 番目から r 番目までの音素からなる音素環境依存音素列 $e_{k+1,r}$ が音素環境依存音素列リストに含まれている場合は 1，そうでない場合は正の大きい値をとる関数とする．

したがって，入力音素の数が n であれば，Viterbi アルゴリズムを用いて，先頭の音素から順に $M(r)$ を求めていき，最終的に $M(n)$ を求めることができる．この最少分割点数を実現する分割方法で，入力音素を分割する．もし $M(n) = 0$ であれば，分割がなく，入力音素は音素環境依存音素列リストに含まれていたということである．また，$M(n) = n - 1$ であれば，入力音素はすべて音素 1 つずつに分割されたということである．図 1 の「(sil z e)+N」，「e-(N k o k u n)+o」，「n-(o t e N k i)+d」，「i-d+e」，「d-(e sh i t a sil)」の場合は，$M(22) = 4$ となった例である．

さらに，分割点数が最少になる分割方法が複数ある場合には，以下の式が最大になる分割方法を採用する．

$$E(r) = \prod_{j=1}^{M(r)+1} n_j \tag{2}$$

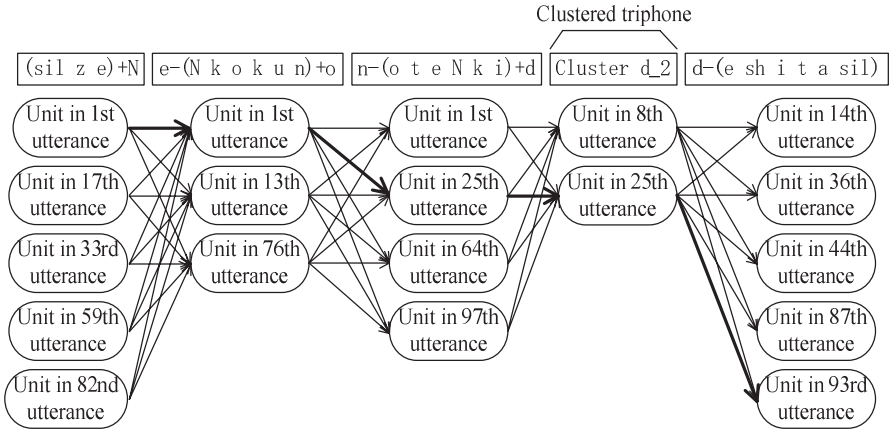

図 2　音素環境依存音素列とクラスタリングされたトライフォンによるビタービ・サーチ

Fig. 2　Viterbi search using context dependent phoneme sequences and clustered triphones as search units.

ただし，n_j は，j 番目の音素環境依存音素列の音素数である．

したがって，各音素環境依存音素列の長さがなるべく均等であるものが採用されやすい仕組みになっている．図 1 の場合は，$E(22) = 3 \times 6 \times 6 \times 1 \times 6 = 648$ となる．

3.4　音声素片の探索

入力文が音素環境依存音素列に分割された後，分割された音素環境依存音素列に属する音声素片の中で，接続スコアと目標スコアの和が最大となる組合せの探索を行う．探索では枝刈りは行わず全探索する．これを**図 2** に示す．図 2 で，素片探索回数は，「(sil z e)+N」と「e-(N k o k u n)+o」の間，「e-(N k o k u n)+o」と「n-(o t e N k i)+d」の間，「n-(o t e N k i)+d」と「i-d+e」の間，「i-d+e」と「d-(e sh i t a sil)」の間の計 4 回になる．また，仮説数は，「(sil z e)+N」で 5 個，「e-(N k o k u n)+o」で 3 個，「n-(o t e N k i)+d」で 4 個，「i-d+e」で 2 個，「d-(e sh i t a sil)」で 5 個となる．

前後の隣り合う音声素片 A と音声素片 B の接続スコア $S_C(A, B)$ は，

$$S_C(A,B) = -\left[a + (p_A^E - p_B^I)^2 \times w_1 + \sum_{i=1}^{d} \frac{(c_{iA}^E - c_{iB}^I)^2}{\left[\sigma_i^{T_A^E} + \sigma_i^{T_B^I} + (\mu_i^{T_A^E} - \mu_i^{T_B^I})^2 \right]} \times w_2 \right] \times \delta_{AB} \tag{3}$$

表 1　音響分析パラメータ
Table 1　Conditions and parameters of acoustic analysis.

Sampling frequency	16 kHz
Analysis window	Hamming
Window size	25 ms
Frame period	10 ms
Pre-emphasis coefficient	0.97
Filterbank channels	24
Lifter	22

と定義する．ただし，p_A^E，p_B^I は，A の終わり，および B の始めの基本周波数（無声フレームの基本周波数は，前後の有声フレームの基本周波数を平均した値を用いた），c_{iA}^E，c_{iB}^I は，A の終わり，および B の始めの i 次元目の特徴量，$\sigma_i^{T_A^E}$，$\sigma_i^{T_B^I}$ は，A の終わりのトライフォンが含まれるクラスタ T_A^E，および B の始めのトライフォンが含まれるクラスタ T_B^I の i 次元目の特徴量の分散値，$\mu_i^{T_A^E}$，$\mu_i^{T_B^I}$ は，A の終わりのトライフォンが含まれるクラスタ T_A^E，および B の始めのトライフォンが含まれるクラスタ T_B^I の i 次元目の特徴量の平均値，d は特徴量の総次元数，a は正の定数，w_1，w_2 は正の重み，δ_{AB} は A と B が不連続である場合を 1，連続している場合を 0 とする．なお，本研究では，予備実験の結果から，a は 10.0，w_1 は 0.006，w_2 は 0.25 とした．

第 1 項はデータベース上の連続した部分が選択されやすくするための項，第 2 項は，接続部分の基本周波数の違いを考慮する項，第 3 項は，特徴量の平均値と分散値を使い，各次元および各クラスタリングされたトライフォン別に正規化した距離で特徴量の違いを考慮するための項である．分母の正規化係数は，付録 A のような手順で導出した．特徴量としては，12 次元 MFCC（Mel-Frequency Cepstrum Coefficients）と対数パワー，およびそれぞれの 1 次，2 次の回帰係数の全 39 個を使用する．これらを計算するための，使用する音響分析パラメータは**表 1** のとおりである．特にベクトル量子化などの処理は行っていない．

μ_i^T および σ_i^T は，音声データベースのアラインメント結果を利用して次式から求める．

$$\mu_i^T = \frac{\sum_{f \in F^T} c_i^f}{\sum_{f \in F^T} 1} \tag{4}$$

$$\sigma_i^T = \frac{\sum_{f \in F^T} \left(c_i^f - \mu_i^T\right)^2}{\sum_{f \in F^T} 1} \tag{5}$$

ただし，c_i^f は f 番目のフレームの i 次元目の MFCC の特徴量，F^T はトライフォン T にアラインメントされたフレームの集合とする．

さらに，入力された音素環境依存音素列 Y，音声データベース中の音素環境依存音素列 D の目標スコア $S_T(Y, D)$ を，以下のように定義する．

$$S_T(Y,D) = -\left(\sum_{j=1}^{n_Y} (p_Y^j - p_D^j)^2 \times \theta(T_Y^j)\right) \times w_3 - \frac{|l_Y - l_D|}{l_Y} \times w_4 \tag{6}$$

ただし，p_Y^j，p_D^j は，Y の語頭から j 番目のトライフォンにおける予測された基本周波数および D の語頭から j 番目のトライフォンの音素内平均基本周波数，n_Y は Y の音素数，$\theta(T_Y^j)$ は，Y の語頭から j 番目の音素 T_Y^j が母音もしくは半母音である場合 1 を返しそれ以外では 0 を返す関数，l_Y，l_D は Y の予測音素列長および D の音素列長，w_3，w_4 は正の重みである．なお，本研究では，予備実験の結果から，w_3 は 0.0625，w_4 は 4.0 とした．

第 1 項は，予測された基本周波数と音声データベースにおける音声素片の基本周波数との違いを考慮し，第 2 項は，予測された音素列長と音声データベースにおける音声素片の音素列長との違いを考慮するための項である．なお，今回行った基本周波数の抽出方法[19]では，誤抽出の際に，本来の値の倍の基本周波数を抽出する傾向があったため，式 (6) における差分の値が非常に大きくなるという問題があった．そこで，基本周波数の誤抽出が生じにくい母音および半母音に対してのみ基本周波数の差分を考慮した．

最後に，文末で接続スコアと目標スコアの和が最大となる音声素片の組合せとして，図 2 の太線で接続する音声素片が選択されたとすると，合成音 1 文について，音声データベース中の異なる発話の音声素片を接続している，もしくは同じ発話中でも隣り合わない音声素片を接続している回数は，「e-(N k o k u n)+o」と「n-(o t e N k i)+d」の間，「i-d+e」と「d-(e sh i t a sil)」の間の 2 回となる．

文献 20) では，ラティスを用いることで探索単位を決めることなく音声素片の組合せを計算する手法が提案されていて，合成単位は可変長の音素列である．これまでの波形接続型音声合成[4)–8)]や本論文でも同様に，合成単位は可変長音素列や可変長半音素列である．し

かし，大規模音声データベースが利用されるようになり，探索単位と合成単位はほとんどの場合一致しなくなった．たとえば，文献 6) の場合，探索単位は音素であるが，合成単位は音素ではなく可変長音素列である．本論文の場合には，この探索単位が音素環境を考慮した音素列になっている点が従来法と異なる．図 2 の場合には，「(sil z e)+N」，「e-(N k o k u n)+o」，「n-(o t e N k i)+d」，「i-d+e」，「d-(e sh i t a sil)」が探索単位となり，「sil z e N k o k u n」，「o t e N k i d」，「e sh i t a sil」が合成単位になる．

4. 合成音の自然性に関する評価と合成処理時間

4.1 実 験 条 件

大量のニュース番組の収録音声を利用した波形接続型音声合成の有効性について調べるため，音声データベースの大きさを全データベースの 8 分の 1（10.9 時間），4 分の 1（21.6 時間），2 分の 1（43.2 時間），1（86.0 時間）の 4 種類用意し，それらを用いて合成音を作成して 5 段階の絶対判断（MOS: Mean Opinion Score）による品質評価実験を行った．

評価用テキストは，2001 年 6 月 28 日から 6 月 29 日までに放送された NHK のニュース番組の中で，音声データベースと同じアナウンサの発声を書き起こした 40 文（1,444 単語，5,927 音素）を使用した．その結果，試料音声は，4 種類の音声データベースを用いて本手法により合成音を作成した 160 の合成音と，書き起こしの基となった 40 の自然音声の計 200 である．なお，本実験では韻律予測の誤りを除くため，目標スコアの韻律予測値は，自然音声から抽出した基本周波数と音素長を使用した．

実験は，許容騒音レベルが NC-15 である防音室内でスピーカを用いて行った．レベルは MCL（Most Comfortable Level）で，スピーカは DIATONE の DS-A3 を使用した．評定者は，音声の評価実験の経験のない 20 代の男性 5 名，女性 5 名の計 10 名である．

各試行では，評価データをランダムな順序で提示し，評定者は自然性の良し悪しを評価した．自然性の評価では，合成音の品質評価に対するガイドライン[21] のように 7 段階の両極尺度で評価する手法もある．しかし，本研究では合成音の自然性が具体的にどのくらいのレベルか知りたかったため，文献 22) で行われているように，**表 2** の 5 段階で評価することとした．

評価に先立ち，評定者に対して音声データベース内の音声を 3 文聞かせて，この程度の自然性の場合には，評価 5 の「自然である」と見なすようインストラクションを与えた．1 文の長さが，平均 148.2 音素，約 10 秒と長いことから，受聴は 1 回のみに限定した．なお実験は，約 10 分間評価を行い 10 分間休憩することを 1 時間 40 分繰り返したところで，1 時間の休憩を入れることで，評定者の集中力が持続するようにした．

表 2　5 段階評価試験の評価基準
Table 2　Perceptual scale for five-point MOS test.

5.	自然である
4.	不自然な部分はあるが気にならない
3.	少し気になる
2.	気になる
1.	非常に気になる

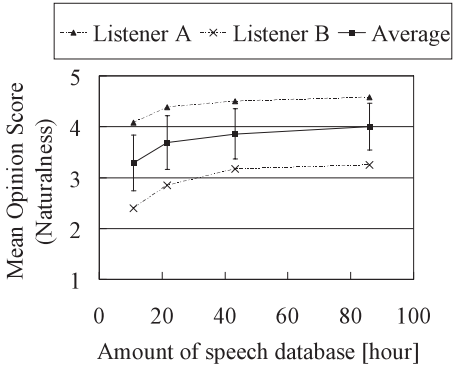

図 3　音声データベースの大きさによる自然性の変化
Fig. 3　MOS for various speech databases.

4.2 主観評価実験結果

［音声データベースの大きさによる自然性の変化］

音声データベースの大きさによる MOS の変化を**図 3** に示す．評価の結果を評定者ごとにみるために，図 3 には，話者ごとの MOS の標準偏差と，例として最も評価の高かった評定者 A の結果と，最も低かった評定者 B の結果を示している．

自然音声の評価は，全体の平均で 4.99 となった．合成音の評価で最も良かったのは，データをすべて使った 86.0 時間の音声データベースを使用したときで 4.01 となり，「不自然な部分はあるが気にならない」の自然性を持つ合成音が得られた．

図 3 から，音声データベースの大きさが大きくなると自然性の評価が向上していくこと

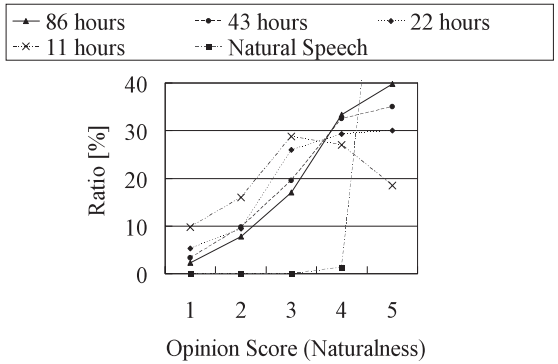

図 4　音声データベースの大きさによる自然性の分布
Fig. 4　Histogram for various speech databases.

が分かる．しかし，音声データベースの大きさが 43 時間を超えたところでほぼ飽和していて，これ以上増やしても自然性の改善効果は小さい．43 時間を超えたところで MOS が飽和する傾向は，評定者 A や B だけでなく，図示していない各評定者に関しても見られた．

1 文あたりの音声データベース中の異なる発話の音声素片を接続している，もしくは同じ発話中でも隣り合わない音声素片を接続している回数は，10.9 時間で 46.4，21.6 時間で 41.2，43.2 時間で 37.2，86.0 時間で 33.3 となり，音声データベースが大きくなると減少する．大きい音声データベースを使うと，合成する文に対して，連続して一致する音声素片の割合が増えることが分かる．

また，音声素片の平均音素数 $\overline{n}$ は，1 文あたりの音声データベース中の異なる発話の音声素片を接続している，もしくは同じ発話中でも隣り合わない音声素片を接続している回数を b，1 文に含まれる総音素数を N とすると，

$$\overline{n} = \frac{N}{b+1} \tag{7}$$

の関係がある．このため，音声素片の平均音素数は，10.9 時間で 3.13，21.6 時間で 3.51，43.2 時間で 3.88，86.0 時間で 4.32 となる．

［音声データベースの大きさによる評価値の割合の変化］

同じ実験で選択された各評価値の割合を**図 4** に示す．

自然音声が 5 の「自然である」と評価されたのは全体の 98.8%であるため，図示していない．図 4 から，86.0 時間の音声データベースを使用したときの合成音では，全体の 39.8%が 5 の「自然である」と評価されていて，本手法により自然音声と変わらない品質の合成音がかなりの頻度で作成されていることが分かる．

また図 4 から，音声データベースの大きさが小さくなると，5 の「自然である」および 4 の「不自然な部分はあるが気にならない」と評価された音声が減っていき，逆に 3 の「少し気になる」，2 の「気になる」，1 の「非常に気になる」と評価された音声が増える．86.0 時間の音声データベースによる合成音で 2 および 1 の評価を受けた合計は全体の 10.0%で，これらの合成音の生成される主な原因は，アラインメントの不整合や基本周波数の誤抽出であった．

表 3　合成処理時間
Table 3　Runtime of various speech databases.

音声データベースの大きさ (時間)	10.9	21.6	43.2	86.0
合成処理時間 (×実時間)	2.26	2.40	2.27	2.53
CPU TIME (×実時間)	0.27	0.21	0.23	0.29
音素環境依存音素列数	2791	6980	17448	38447
最長音素環境依存音素列の音素数	13	20	40	55
平均素片探索回数	103.1	85.6	70.7	57.4
平均仮説数	333.3	395.8	468.7	566.1

4.3　合成処理時間の検討

音声データベースの大きさによる音声合成処理時間の変化を**表 3** に示す．実行環境は，CPU XEON 3.20 GHz × 2（ただし，プログラムはシングルスレッド処理），メモリ 2 GB，OS Red Hat Linux Advanced Server 2.1（Kernel 2.4.9-e.24smp）である．

また，各音声データベースにおける音素数を横軸にとった音素環境依存音素列数を**図 5** に示す．音声データベースの大きさが大きくなると，音素環境依存音素列の音素数が増える

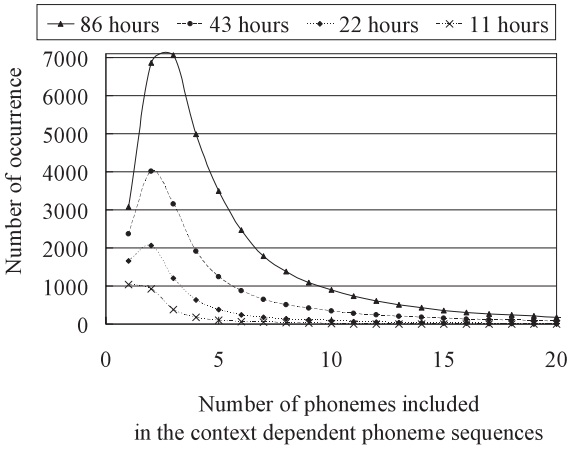

図 5　音素数に対する音素環境依存音素列数

Fig. 5　Histogram for the number of phonemes included in the context dependent phoneme sequences.

だけでなく，同じ音素数における音素環境依存音素列の種類も増えていることが分かる．その結果，1 文あたりの平均素片探索回数は減少する．また，音声データベースの大きさが 4 倍になると，探索単位が変わらなければ，単純に考えて，1 回の探索あたりの平均仮説数は 4 倍になる．しかし，本手法では，探索単位が長くなるため，平均仮説数は若干増加するだけにとどまる．そのため，音声データベースを大きくしても合成処理時間はそれほど増加しない．また，CPU TIME が実際の合成処理時間よりも少ないのは，処理時間のほとんどが合成素片データベースの音声特徴量を記述したファイルの読み込みに費やされているためである．

5. 目標スコアの有無による自然性の違い

5.1 実 験 条 件

音声データベースの発話内容と合成する文とが類似している場合には，目標スコアを考慮しなくても，自然性の高い合成音を作成できる可能性があるため，3.2 節で述べた 86.0 時間の音声データベースを利用して目標スコアの有無で合成音を作成し，5 段階 MOS 評価実験を行った．

評価用テキストは，2001 年 6 月 25 日から 6 月 29 日までに放送された NHK のニュース番組の中で，音声データベースと同じアナウンサの発声を書き起こした 100 文（3,370 単語，13,777 音素）を使用した．試料音声は，目標スコアを考慮した場合と，まったく考慮せず接続スコアのみ用いた場合とで作成した 200 の合成音と，自然音声の計 300 である．目標スコアを考慮する場合には韻律予測の誤りを除くため，目標スコアの韻律予測値は，自然音声から抽出した基本周波数と音素長を使用した．

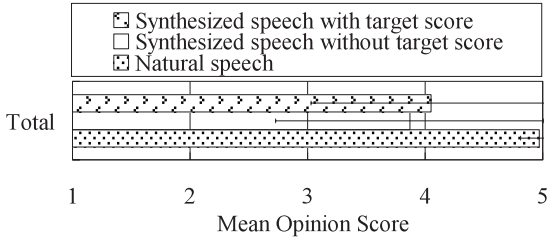

図 6　目標スコアの有無による自然性の変化

Fig. 6　MOS for synthesized speech with or without target score.

なお，その他の実験条件は 4.1 節と同様である．

5.2 実 験 結 果

図 6 に MOS と標準偏差を示す．目標スコアを考慮した場合の合成音の MOS は 4.05，考慮しない場合の合成音は 3.87 と評価され，MOS の差は 0.18 となった．自然音声の評価は 4.97 となった．したがって，目標スコアを考慮しない場合でも，音声データベースの発話内容と合成する文が似ている場合には，自然性の高い合成音が作成できることが分かる．

目標スコアを考慮した場合の合成音は，自然音声から抽出した基本周波数と音素長を使用しているため，一般的な韻律予測を行う場合には，本実験で得られた結果よりも自然性が低下することが予想される．これまでの研究でも，自然性の低下を示す結果が報告されているが[22),23)]，韻律予測を行う場合の自然性の低下の度合は，音声データベースが異なるため一概には比較できない．

5.3 アクセントの高低情報の再現性

目標スコアを考慮したとしても，最終的な選択結果に寄与しなくては意味がない．そこで，目標スコアを考慮せずに選択した音声素片系列に対して，事後的に式 (6) を用いて目標スコアを計算し，目標スコアを考慮した音声素片系列に対する目標スコアと比較した．目標スコアを考慮せずに選択した音声素片系列に対して，事後的に計算した目標スコアを 1 とすると，目標スコアを考慮した音声素片系列に対する目標スコアは 0.04 になった．式 (6)

は，目標とする基本周波数と音素列長に近いほど小さなスコアになるため，目標スコアを考慮して選択した音声素片系列では，目標スコアは十分考慮されている．

目標スコアを考慮しない場合でも，韻律が再現されたことについてさらに検討するため，アクセントの高低情報の再現性について調べた．まず，自然音声を人手で聞くことにより，評価用テキストのアクセントの高低情報を付与した．次に合成音のアクセントの高低情報を人手で聞くことにより付与した．最後に 2 つのアクセントの高低情報の比較を行い，アクセントの高低が一致した割合を求め，正解率とした．その結果，目標スコアを考慮した場合で 97.3%，目標スコアを考慮しない場合で 90.3%となった．目標スコアを考慮しなくても，かなりの割合でアクセントの高低が一致している音声素片を選択できていることが分かる．

しかし，一方で，目標スコアを考慮した場合としない場合とでアクセントの正解率が異なるにもかかわらず，MOS がそれほど低下しないことが疑問として残る．この理由としては，2 つのことが考えられる．

理由の 1 つは，目標スコアを考慮しない場合，アクセントの正解率が低下しても，接続される音声の前後の連続性が増すことで，その低下分を補ったということが考えられる．1 文あたりの音声データベース中の異なる発話の音声素片を接続している，もしくは同じ発話中でも隣り合わない音声素片を接続している回数を調べたところ，目標スコアを考慮した場合で 31.5，考慮しない場合で 19.4 となり，目標スコアを考慮しないと連続している音声からデータが選択される割合が高くなる．また，接続スコアは，目標スコアを考慮した場合を 1 とすると，考慮しない場合で 0.54 となり，目標スコアを考慮しない場合には，合成される音声の接続前後の基本周波数および特徴量の差は小さくなる．

理由のもう 1 つは，評定者によっては正しいアクセントを判別できない，もしくは間違ったアクセントでも気にならないことが考えられる．今回はアクセントの評価ではないため，合成音の評定者をアクセントの評価の際によく行われる「両親と本人が東京生まれの東京育ち」には限定していない．これまでの研究で，生粋の大阪人と方言環境の異なる大阪在住者に対してアクセントの識別能力の比較が行われた[24]．分析の結果，生粋の大阪人に比べると，母親と本人とが異なるアクセント体系を持つ地域に育った場合の方が，アクセントの識別ができ難いことが報告されている．したがって，母親と本人とが異なるアクセント体系を持つ地域に育った評定者の場合，共通語のアクセントとして間違っている合成音を聞いても，違和感を持たない可能性がある．

上述した部分で，目標スコアを考慮せずに作成した合成音で，アクセントの高低情報が約 9 割再現されていることと，連続している音声からデータが選択される割合が高くなることを示した．しかし，これらは音声素片を選択した結果であり，合成音の自然性が高いことに対する選択のメカニズムそのものは明らかになっていない．このため，以下では，その要因に関する若干の考察を試みる．

音声データベースの発話内容と合成する文の類似度が高い場合，文頭と文末の探索時に使われる音素環境依存音素列は，文頭と文末で発声されている音素環境依存音素列が用いられる．本実験で使用した評価文を調べた結果，文頭の探索時には，98%の場合，文頭で発声された音素環境依存音素列が用いられ，残りの 2%の場合，クラスタリングされたトライフォンが用いられている．つまり，98%の場合には，この後の処理でどのような音声素片が選ばれたとしてもすべて文頭の音声素片である．残りの 2%の場合には，クラスタリングされたトライフォンを探索単位としていることから，音素環境の異なる音声素片も含まれるため，文頭の音声素片が選ばれる保証はない．しかし，その 2%の場合にも結果的には文頭の音声素片が選択されている．

一方，文末の探索時に文末で発声された音素環境依存音素列が用いられたのは 100%である．つまり，文末の場合には，この後の処理でどのような音声素片が選ばれたとしても，それらはすべて文末の音声素片になる．したがって，本実験で使用した評価文章を音声合成したほとんどの場合，探索単位が決定した段階で，文頭と文末の基本周波数および特徴量は，本人が実際に文頭と文末で発声したことのある値になる．しかも，文末の音素環境依存音素列の長さは平均 9.3 音素と非常に長いため，本人が発声した値にかなり近いことが期待できる．

さらに，このような性質を持つ文頭と文末の間の音声素片は，本手法では接続スコアを考慮して探索を行うため，文頭の隣には，文頭の基本周波数・特徴量とそれほど大きく変わらない値を持つ音声素片が選択される．また，同様に，文末の隣には，文末の基本周波数・特徴量とそれほど大きく変わらない値を持つ音声素片が選択される．これを繰り返していくと，文頭から文末にかけて，それほど違和感を与えない音声素片が順次選択されると考えられる．

また，文中にポーズがある場合，ポーズ前後の探索時にポーズ前後で発声された音素環境依存音素列が用いられた割合は，ポーズ前で 97%，ポーズ後で 90%である．したがって，ポーズ前後においても，本実験で使用した評価文章を音声合成した多くの場合，探索単位が決定した段階で，ポーズ前後の基本周波数および特徴量は，本人が実際にポーズ前後で発声したことのある値になる．

定性的には上記のように考えられるが，アクセントの高低情報が再現される詳細なメカニ

ズムについては，今後明らかにする必要がある．

5.4 考　　察

今回の実験条件では，目標スコアを考慮しなくても合成音の自然性はそれほど低下しないという結果が得られたが，音声データベースの発話内容と合成する文が類似していない場合には，合成音の自然性が低下することが予想される．厳密な評価を行ってはいないが，小説，口語調の文章，見出し記事のような体言止めの文章などニュース番組の中には存在しないいい回しを含む文を合成した場合，合成音の自然性はニュース文を合成するときに比べて明らかに劣化する．また，同じニュースでも，ニュースで頻繁に取り上げられる政治や経済に関するトピックは，音声データベース中の頻度が高いため合成音の自然性は高いが，遺跡の発掘やスポーツなど，音声データベースに存在しないトピックでは，自然性が低下する．さらに，政治や経済のトピックでも，閣僚の名前や新規の固有名詞など音声データベースの収録時期には含まれない語句を含む場合には，必要となる音素の並びが異なることで，自然性が低下することがある．

現在の我々が試作した音声合成システムでは，目標スコアの有無は手動で切り替える必要があるが，将来的には，音声データベースの発話内容と合成する文の類似度が高いときには目標スコアを使わず，類似度が低いときには目標スコアを利用するといったシステム構成が望ましい．そのためには，今後，音声データベースの発話内容と合成する文の類似度が高いということが，具体的にどのような条件になるのか明らかにする必要がある．

また，今回使用した音声データベースは，1996 年から 2001 年までの番組音声を収録しているため，音声の収録時期に幅がある．このため，文献 25) で議論されているような声質時期差が問題になることが予想されるが，今のところ音声の収録年度による目立った違いは見つかっていない．それよりも，鼻声になっているなどの発声者の体調の問題や，番組の収録音声を使うことでの固有の問題が生じている．番組の収録音声を使うことでの固有の問題とは，収録音声のほとんどが単純な原稿の読み上げである中に，記者との掛け合いの音声や，明るいニュースの原稿の読み上げで気持ちがこもっている音声などが含まれていて，これらの音声が合成音の中に使われると，その部分だけ会話調に聞こえたり，明るい印象の読み上げに聞こえたりすることである．これらの問題を解決するためには，基準に該当しない音声を，音声データベースの中から削除すればよいが，標準的な読み上げ音声という基準を明確にするのは難しい．これを実現するには，発声スタイルを表す物理量を明らかにすることが今後の課題である．

6. おわりに

ニュース番組の収録音声を波形接続型音声合成システムの音声データベースとして利用することを試みた結果，5 段階の MOS 評価による合成音の平均評価値は 4.01 となり，「不自然な部分はあるが気にならない」という自然性を持つ合成音が得られた．特に，全体の 39.8%が 5 の「自然である」と評価され，自然音声と変わらない品質の合成音がかなりの頻度で作成されていることが分かった．また，音声データベースが大きくなると自然性の評価も向上するが，43 時間を超えたところでほぼ飽和することが分かった．さらに，合成処理時間について検討したところ，探索単位として音素環境依存音素列を用いた本手法では音声データベースを大きくしても合成処理時間はそれほど増加しなかった．

次に，目標スコアを用いた場合と，用いない場合の合成音とを比較することにより，目標スコアを用いた場合と用いない場合の合成音の MOS の差は 0.18 となり，音声データベースの発話内容と合成する文が類似している場合には，必ずしも韻律予測せず目標スコアを考慮しなくても，自然性の高い合成音を作成できる可能性が示された．

今後は，音声データベースの発話内容と合成する文の類似度が高いということが，具体的にどのような条件になるのか検討するとともに，他の話者のニュース番組の収録音声の利用や，ニュースだけでなくドラマやバラエティー番組などの収録音声を利用することで，番組の収録音声を利用した音声合成の有効性についてさらに検証を進めていきたい．

謝辞　ニュース番組の音声データベースの構築を行った音声認識グループの皆様に深く感謝いたします．

参 考 文 献

1) 世木寛之，清山信正，田高礼子ほか：高品質な株価音声合成装置の開発とデジタルラジオ放送での試験運用，映像情報メディア学会誌，Vol.62, No.1, pp.69–76 (2008).
2) 道家　守，林　正樹，牧野英二：TVML を用いた番組情報からのニュース番組自動生成，映像情報メディア学会誌，Vol.54, No.7, pp.1097–1103 (2003).
3) Matsumura, K., Kai, K., Hamada, H. and Yagi, N.: Transforming Data Broadcast Contents to Fit Different User Interfaces Generating a Readout Service for Mobile DTV Receiver, *Proc. 7th International Conference on Human Computer Interaction with Mobile Devices and Services* (*MobileHCI'05*), pp.323–324 (2005).
4) 村上仁一，水澤紀子，東田正信：音節波形接続による単語音声合成，電子情報通信学会論文誌（D-II），Vol.J85-D-II, No.7, pp.1157–1165 (2002).
5) Black, A. and Campbell, N.: Optimizing Selection of Units from Speech Database

for Concatenative Synthesis, *Proc. EUROSPEECH*, Vol.1, pp.581–584 (1995).
6) Hunt, A. and Black, A.: Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database, *Proc. ICASSP*, Vol.1, pp.373–376 (1996).
7) Conkie, A.: A Robust Unit Selection System for Speech Synthesis, *Proc. 137th meet. ASA/Forum Acusticum* (1999).
8) 戸田智基，河井　恒，津崎　実，鹿野清宏：素片接続型日本語テキスト音声合成における音素単位とダイフォン単位に基づく素片選択，電子情報通信学会論文誌（D-II），Vol.J85-D-II, No.12, pp.1760–1770 (2002).
9) 磯　健一，渡辺隆夫，桑原尚夫：音声データベース用文セットの設計，日本音響学会春期研究発表会，2-2-19, pp.89–90 (1988).
10) 河井　恒，樋口宣男，山本誠一：基本周波数及び音素持続時間長を考慮した音声合成用波形素片データセットの作成，電子情報通信学会論文誌（D-II），Vol.J82-D-II, No.8, pp.1229–1238 (1999).
11) 河井　恒，山本誠一：波形素片接続型音声合成システムのための波形素片データベースの作成，日本音響学会秋季研究発表会，3-5-5, pp.325–326 (1994).
12) 河井　恒，戸田智基，山岸順一ほか：大規模コーパスを用いた音声合成システム XIMERA，電子情報通信学会論文誌（D-II），Vol.J89-D-II, No.12, pp.2688–2698 (2006).
13) Chu, M., Peng, H., Yang, H. and Chang, E.: Selecting Non-Uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer, *Proc. ICASSP*, Vol.1, pp.785–788 (2001).
14) Young, S., Odell, J. and Woodland, P.: Tree-based State Tying for High Accuracy Acoustics Modeling, *Proc. ARPA Human Language Technology Workshop*, Vol.1, pp.307–312 (1994).
15) 安藤彰男，今井　亨，小林彰男ほか：音声認識を利用した放送用ニュース字幕制作システム，電子情報通信学会論文誌（D-II），Vol.J84-D-II, No.6, pp.877–887 (2001).
16) Young, S., Evermann, G., Mark, G., et al.: *HTK Book* (*for HTK Version 3.3*) (2005).
17) Kawai, H. and Toda, T.: An Evaluation of Automatic Phone Segmentation for Concatenative Speech Synthesis, *Proc. ICASSP*, Vol.1, pp.677–680 (2004).
18) 米澤朋子，水野秀之，阿部匡伸：HMM 音素モデルによる自動ラベリングのロバスト性の検討，電子情報通信学会技術研究報告，SP2002-74, pp.17–22 (2002).
19) 小林　載，島村徹也：対数スペクトルにクリッピングと帯域制限を用いる基本周波数抽出法，電子情報通信学会論文誌（A），Vol.J82-A, No.7, pp.1115–1122 (1999).
20) 匂坂芳典：種々の音韻連接単位を用いた日本語音声合成，電子情報通信学会技術研究報告，SP87-136, pp.47–52 (1988).
21) 電子協（編）：音声合成システム性能評価方法のガイドライン，日本電子工業振興協会 (2000).
22) 濱上知樹，古村光夫：深い意味や構造を意識せず抑揚を抑えて発声された音声の F0 パターンの分析と合成，電子情報通信学会論文誌（D-II），Vol.J81-D-II, No.6, pp.1047–1057 (1998).
23) 額賀信尾，永松健司，安藤ハルほか：韻律コーパス利用合成方式における文節選択基準と自然性の関係，日本音響学会春季研究発表会，2-P-19, pp.305–306 (1999).
24) 杉藤美代子：「花」と「鼻」，pp.119–120, 和泉書院 (1998).
25) Kawai, H. and Tsuzaki, M.: A Study on Time-Dependent Voice Quality Variation in a Large-Scale Single Speaker Speech Corpus Used for Speech Synthesis, *Proc. IEEE 2002 Workshop on Speech Synthesis*, pp.15–18 (2002).

付　　録

付録 A　接続スコア

まず，式 (3) の第 3 項の分子について，トライフォン T および T' にアラインメントされたフレームの集合 F^T および $F^{T'}$ に関する平均を計算する．

$$
\begin{aligned}
&\frac{\sum_{f\in F^T}\sum_{f'\in F^{T'}}\left(c_i^f-c_i^{f'}\right)^2}{\sum_{f\in F^T}\sum_{f'\in F^{T'}}1}\\
&=\frac{\sum_{f\in F^T}\sum_{f'\in F^{T'}}\left\{c_i^f-\mu_i^T-\left(c_i^{f'}-\mu_i^{T'}\right)+\mu_i^T-\mu_i^{T'}\right\}^2}{\sum_{f\in F^T}\sum_{f'\in F^{T'}}1}\\
&=\frac{\sum_{f\in F^T}\left(c_i^f-\mu_i^T\right)^2}{\sum_{f\in F^T}1}+\frac{\sum_{f'\in F^{T'}}\left(c_i^{f'}-\mu_i^{T'}\right)^2}{\sum_{f'\in F^{T'}}1}+\left(\mu_i^T-\mu_i^{T'}\right)^2\\
&\quad-2\frac{\sum_{f\in F^T}\sum_{f'\in F^{T'}}\left(c_i^f-\mu_i^T\right)\left(c_i^{f'}-\mu_i^{T'}\right)}{\sum_{f\in F^T}\sum_{f'\in F^{T'}}1}+2\left(\mu_i^T-\mu_i^{T'}\right)\frac{\sum_{f\in F^T}\left(c_i^f-\mu_i^T\right)}{\sum_{f\in F^T}1}
\end{aligned}
$$

$$
\begin{aligned}
&- 2\left(\mu_i^T - \mu_i^{T'}\right) \frac{\sum_{f' \in F^{T'}} \left(c_i^{f'} - \mu_i^{T'}\right)}{\sum_{f' \in F^{T'}} 1} \\
&= \sigma_i^T + \sigma_i^{T'} + (\mu_i^T - \mu_i^{T'})^2 \qquad (8)
\end{aligned}
$$

2 行目から 3 行目は，2 乗の中を展開した．また，3 行目から 4 行目は，式 (4) および式 (5) により 4 項目から 5 項目までが 0 になることを利用した．

(平成 20 年 5 月 30 日受付)
(平成 20 年 11 月 5 日採録)

世木　寛之（学生会員）

1994 年慶應義塾大学理工学部物理学科卒業．1996 年同大学大学院修士課程修了．同年 NHK 入局．山口放送局を経て，1998 年より放送技術研究所に勤務．音声合成，音声認識の研究に従事．2002 年電子情報通信学会論文賞受賞．2007 年慶應義塾大学大学院後期博士課程（社会人）入学．日本音響学会，映像情報メディア学会，電子情報通信学会各会員．

田高　礼子

2000 年東北大学工学部電気工学科卒業．2002 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．同年 NHK 入局．放送技術研究所にて音声合成等の研究に従事．

清山　信正

1989 年早稲田大学大学院修士課程修了．同年 NHK 入局．同年放送技術研究所勤務．2007 年より NHK エンジニアリングサービス勤務．話速変換，音声合成等の実用化研究に従事．1995 年日本音響学会佐藤論文賞受賞．

都木　徹

1981 年電気通信大学大学院電気通信研究科修士課程修了．同年 NHK 入局．長野放送局を経て，1984 年より放送技術研究所に勤務．音声知覚，声質変換方式，話速変換方式，音声合成，字幕制作等の研究に従事．2003 年映像情報メディア学会技術振興賞，2005 年第 13 回日本音響学会技術開発賞，2008 年第 54 回大河内記念技術賞受賞．日本音響学会，映像情報メディア学会，電子情報通信学会会員．博士（工学）．