

# 音声合成のためのテンプレートを用いた録音文セット生成システムとラジオ番組「気象通報」への適用について

Sentence-Generating System for Speech Synthesis Using Templates and Application for the "Weather Report" Radio Program

正会員 世木 寛之<sup>†1,†3</sup>, 正会員 田高 礼子<sup>†1</sup>, 正会員 清山 信正<sup>†2</sup>,

正会員 都木 徹<sup>†1</sup>, 正会員 斎藤 英雄<sup>†3</sup>, 名誉会員 小澤 慎治<sup>†4</sup>

Hiroyuki Segi<sup>†1,†3</sup>, Reiko Takou<sup>†1</sup>, Nobumasa Seiyama<sup>†2</sup>, Tohru Takagi<sup>†1</sup>, Hideo Saito<sup>†3</sup> and Shinji Ozawa<sup>†4</sup>

**Abstract** The design method of a sentence set for a speech-synthesis database strongly influences the quality of the synthesized speech. To minimize the costs associated with making the speech recordings and constructing the speech database, the number of the sentence set should be limited. However, if a sentence set does not include sufficient data, the quality of the synthesized speech can be inadequate. In this paper, we propose a method for generating a sentence set from templates. When applied to the templates in the "Weather Report" radio program, the proposed method reduced the number of the sentence set to less than several percent of that required by a comparison method. In addition, the mean opinion score of speech samples synthesized using the proposed method was 4.32 on a five-point scale.

キーワード：音声合成，グラフ，気象通報，文セット，データベース，テンプレート

## 1. ま え が き

音声合成は20年ほど前から放送で利用されるようになった<sup>1)</sup>。読み手(アナウンサ等)がいなくても情報が伝えられるため、放送局にとって非常に利便性が高いからである。

これまでに放送局で使われてきた音声合成システムの用途としては、天気予報の自動送出<sup>1)~4)</sup>が多く、他にも交通情報<sup>5)</sup>、スポンサー名テロップの読み上げ<sup>6)</sup>などがある。これらの音声合成システムでは、録音した音声を実音部分で接続して再生する方式(録音編集方式<sup>7)</sup>)が採用されている。録音編集方式以外の音声合成方式<sup>8)~10)</sup>も存在するが、肉

声と同等の品質を実現できるという保証がないため、これまでは放送に利用するのは難しかった。

しかし、筆者らは、録音した音声を有音部分で接続して再生する方式(波形接続方式<sup>11)~12)</sup>)を用いた株価音声合成システムを試作し、2006年10月から2008年3月までの毎週金曜日に、地上デジタル音声放送実用化試験放送(通称デジタルラジオ)での運用を経て<sup>13)</sup>、2010年3月29日からラジオ第2放送での運用を行っている。このシステムを用いると、「3,465円」などの株価と「134円高」などの前日比の1億までの数値について、肉声とほぼ変わらない自然性を持つ合成音を作成することができる。

波形接続方式とは、あらかじめ準備しておいた音声データベースに含まれている基本音声合成単位を組合せて接続することによって、音声合成したいテキスト(これを、合成対象テキストと呼ぶ)の合成音を作成する方式である。音声データベースは、いろいろな文章を発声者に読み上げさせて録音し、各録音音声を基本音声合成単位の切れ目で分割することで作成する。

この波形接続方式をいろいろな番組制作に利用するためには、発声者にどのような文を読み上げさせるかという録音文セットを決定しなければならない。読み上げやデータベースの構築化の手間を考えると、発声者にあらかじめ読み上げさせる録音文の量は、少なければ少ないほど良い。一方で、充分な数のデータ量を揃えておかないと、自然性の高い合成音を作成できない。このため、合成対象テキス

2010年5月31日受付、2010年9月21日再受付、2010年10月25日採録

†1 NHK 放送技術研究所 (人間・情報科学)

(〒157-8510 東京都世田谷区砧1-10-11, TEL 03-5494-1125)

†2 財団法人NHKエンジニアリングサービス

(〒157-8510 東京都世田谷区砧1-10-11, TEL 03-5494-2400)

†3 慶應義塾大学 大学院理工学研究科 開放環境科学

(〒223-8522 神奈川県横浜市港北区日吉3-14-1, TEL 045-563-1141)

†4 愛知工科大学 大学院工学研究科 情報メディア

(〒443-0047 愛知県蒲郡市西迫町馬乗50-2, TEL 0533-68-1135)

†1 NHK Science and Technology Research Laboratory (Human & Information Science)

(1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510, Japan)

†2 NHK Engineering Services, Inc.

(1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510, Japan)

†3 School of Science for Open and Environmental Systems, Graduate School of Science and Technology, Keio University

(3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, 223-8522, Japan)

†4 School of Media and Information, Graduate School of Technology, Aichi University of Technology

(50-2 Umanori, Nishisako-cho, Gamagori, Aichi, 443-0047, Japan)

トに応じて、なるべく少ないデータ量で自然性の高い合成音を作成できる録音文セットを決める手法が重要となる。

筆者らの提案した文献13)の手法では、合成対象テキストをあらかじめリストアップしておき、合成するために必要な基本合成単位を含む録音文セットを、逐次貪欲アルゴリズムを用いて抽出する。

例えば、1億までの数値を音声合成したい場合には、1億までの数値をあらかじめリストアップし、その中から、1億までの数値を音声合成するのに必要な最少個の録音文(この場合は数値)を抽出する。この結果、約4,000個の録音文セットが抽出されれば、これらの録音文セットを読み上げて音声を録音し、音声データベースを構築することで、1億までの数値を音声合成することが可能になる。

上記システムでは、波形接続方式を株価の合成に適用していたため、合成対象は数値に限られており、リストアップされる文数も一定の範囲に抑えられていた。しかし、より一般的な用途に適用した場合、合成対象テキストによっては、リストアップする際に、膨大な数になる場合がある。例えば、NHKラジオ第2放送の「気象通報」という番組で用いられている文は、

「[X1]の、北緯[X2]度[X3]分、東経[X4]度[X5]分には、[X6]ヘクトパスカルの、台風第[X7]号が、[X8]へ毎時[X9]キロメートルで進んでいます」という形式で記述される。これを以下ではテンプレートと呼ぶことにする。「[X1]」は、場所の変数を表し、例えば、「日本海」・「渤海」・「黄海」など200種類の要素が入る。その他の変数にも対応した要素が入る。このようなテンプレートの場合、各変数に入る要素を組合せると、天文学的な数のテキストがリストアップされてしまう。その結果、録音文セットの生成に膨大な処理時間を要するため、テンプレートで記述されるテキストの合成音を作成するために、文献13)の手法を用いることは非現実的である。

この他にも、録音文セットを生成する手法としては、交換法を使用するもの<sup>14) 15)</sup>や、貪欲アルゴリズムを使用するもの<sup>16) ~18)</sup>が提案されているが、いずれの手法でも、合成対象テキストはあらかじめすべてリストアップする必要があるため、適用することができない。

そこで本研究では、変数や分岐で記述されたテンプレートに対して、合成対象テキストをすべてリストアップするのではなく、類似度の高いテンプレートを統合することでテンプレートの削減を行う。

さらに、この統合されたテンプレートから合成対象テキストを生成したとしても、依然として膨大な数のテキストがリストアップされてしまい、録音文の生成に膨大な時間がかかることが予想されるため、リストアップ作業をすることなく、テンプレートから直接録音文セットを生成する手法を提案する。

本研究では、この提案法の有効性を確認するため、テンプレートからランダムに録音文セットを生成した手法と比

較して、必要な文数が数%以下に削減されることを示す。

さらに、提案法で生成した録音文を読み上げて音声を収録し、合成音の作成を行い、自然性の評価を行うことで、提案法により生成された録音文が音声データベースを構築するために有効であることを検証する。

以下、2章で、全体のシステム構成について述べる。次に、3章で、提案する録音文セットの生成システムについて述べる。4章では、提案法と、テンプレートからランダムに録音文を生成した手法とで録音文セットの生成を行い、提案法を用いると同じカバレッジを得るために必要な文数が削減されることを示す。5章では、提案法を用いて作成した合成音の主観評価実験を行う。最後に6章でまとめを述べる。

## 2. 全体のシステム構成

全体のシステム構成を図1に示す。

図1では、上半分が音声合成を行う部分を示し、下半分が録音文セットを生成する部分を示す。まず、上半分の音声合成を行う部分では、音声合成システムは入力テキストを読み込み、音声データベースを用いて合成音を作成し出力する。一方、下半分では、録音文セット生成システムは入力テキストをすべて記述することのできるテンプレートを読み込み、音声合成の際に用いられる音声データベースを作るために必要な録音文セットを生成する。このテンプレートは、音声合成システムに入力が見込まれるすべてのテキストを記述できる必要がある。このため、テンプレートは複数あっても構わない。生成された録音文セットは事前にアナウンサーや声優などにより音声収録を行う。その後、音声合成システムへ音声データベースとして組み込みを行うことで、音声合成が可能になる。

録音文セット生成システムから生成される録音文セットの量が多いと、データベースの構築時や読み上げ時の作業が増えてしまう。このため、でき上がりの合成音の自然性

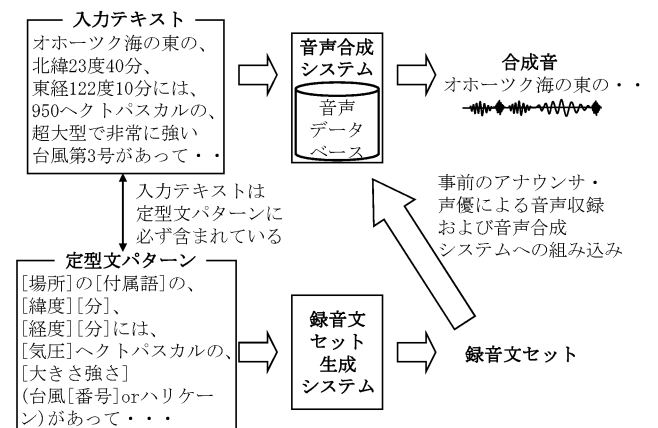


図1 全体のシステム構成  
Overview of the system.

を損なうことなく，録音文セットの大きさを極力削減することが本研究の課題である．次章では，テンプレートから録音文セットを効率的に生成するシステムを提案する．

### 3. 録音文セット生成システム

#### 3.1 全体の流れ

録音文セットを効率的に生成するため，以下の手順を行う．それぞれの詳細については次節以降で述べる．

- (1) 変数と分岐で記述された複数のテンプレートを読み込む．
- (2) 入力された複数のテンプレートの類似している部分を共有化するため，動的計画法<sup>19)</sup>を用いて各テンプレートを比較し統合する．
- (3) 統合したテンプレートにおいて，読点がなく連続して発声する箇所では，調音結合（例えば「あいう」と発声したときの「い」には，前に発声した「あ」の口の形と舌の位置の影響と，後に発声する「う」の口の形と舌の位置の影響が表れている現象）を考慮して音声波形の接続を行う必要がある．この場合，同じ「い」の表記でも，前に発声した音素と後ろで発声した音素が違えば別の音として区別されるため，収録するデータ量が増加してしまう．そこで，そのような箇所では，逐次貪欲アルゴリズム<sup>13)</sup>を適用することで，収録するデータ量の削減を行う．
- (4) 統合したテンプレートにおいて，生成する録音文が最少になる最適化問題を設定する．
- (5) 最適化問題をシンプレックス法<sup>20)</sup>で解く．
- (6) 解に基づき，統合したテンプレートにおける録音文の生成を行う．

#### 3.2 テンプレートの読み込み

テンプレートは，「[X1]」の形式で表記される変数と，「([X1] or [X2] or [X3] ···)」の形式で表記される分岐と，「<[X1]>」の形式で表記される省略で記述されるものとする．例えば，

「[X1] 時間後の [X2] 日 [X3] 時には，<半径> [X4] キロの円内に達し，(台風orハリケーン) に変わる見込みです．」

のように表すことができる．ここで，「(台風orハリケーン) に変わる見込みです」は，「台風に変わる見込みです」と「ハリケーンに変わる見込みです」のどちらも選択可能であることを意味する．また，「<半径> [X4] キロの円内に達し，」は，「半径 [X4] キロの円内に達し，」でも「[X4] キロの円内に達し，」でもどちらも選択可能であることを意味する．

それぞれの変数，例えば，例文の [X1] から [X4] に入る言葉は，あらかじめリストアップしておくことが必要で，これもテンプレートと同様，録音文セット生成システムへ入力する．

録音文セット生成システムは，このような形式で表記される複数のテンプレートを読み込み，グラフとして解釈す

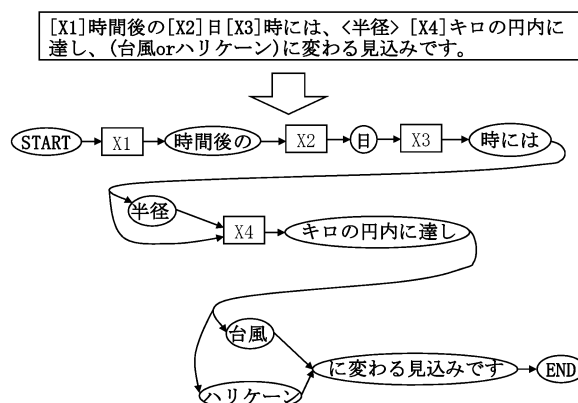


図2 テンプレートからグラフへの変換  
Transformation from templates to graphs.

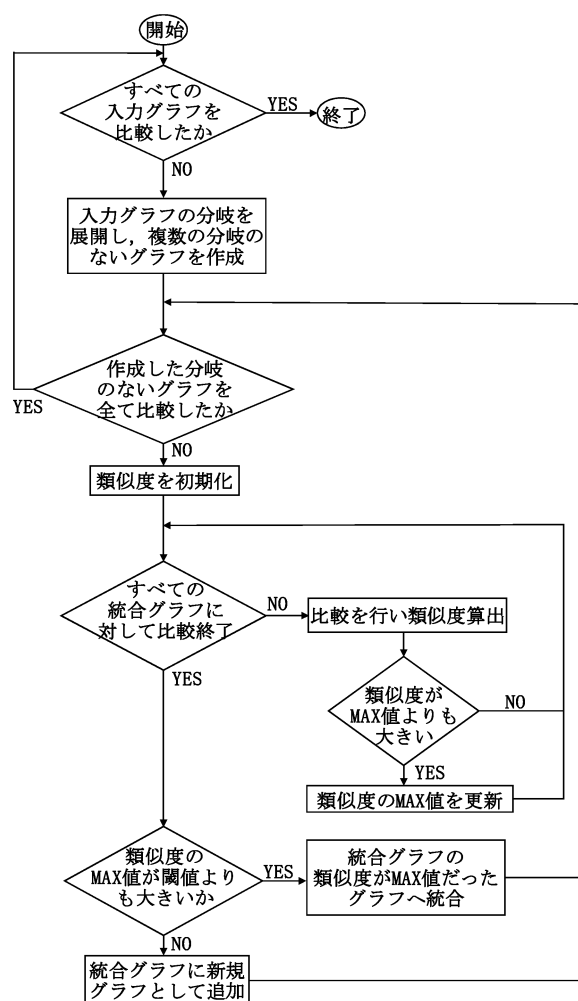


図3 グラフの比較・統合の流れ  
Flow chart for comparison and unification of graphs.

る．これを図2に示す．読み込む際には，文頭にスタートノードを，文末にエンドノードを付加する．

#### 3.3 動的計画法によるグラフの比較と統合

入力された複数のグラフの類似している部分を共有化するため，各グラフを比較し統合する．

グラフの比較の際には動的計画法<sup>19)</sup>を用いて，図3に示

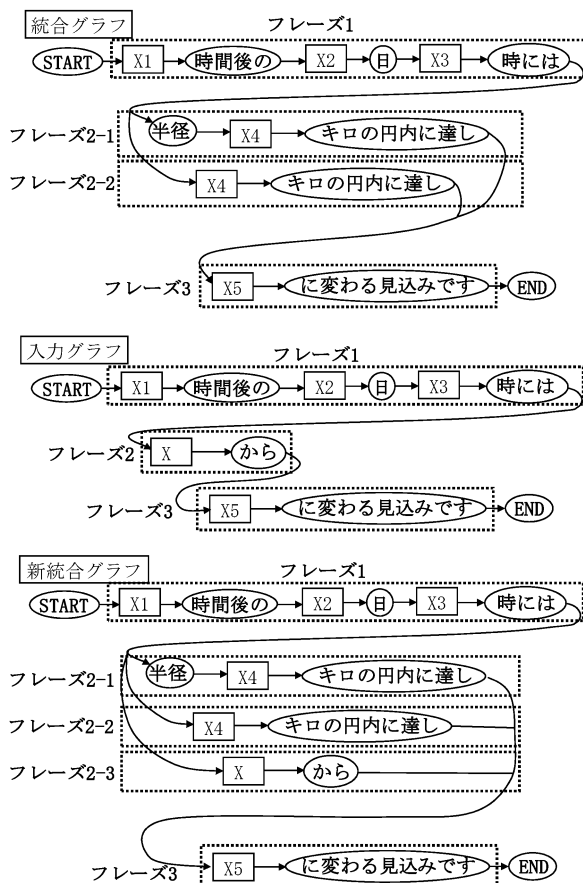


図4 統合グラフと入力グラフとの比較・統合  
Comparison and integration of unified graphs and input graphs.

す流れで処理を行う。

- (1) 比較の順番は、入力されたグラフの順番で行う。そのため、まず1番目と2番目のグラフの比較を行い、次にその統合結果と3番目のグラフとの比較を行う。4番目以降についても、同様に、それまでの統合結果と4番目以降のグラフとの比較を行う。
- (2) 入力されたグラフに分岐が含まれる場合には、分岐を文頭から文末まで展開したグラフを複数作成し、その結果得られた分岐のないグラフ一つずつと統合グラフとの比較を行う。これは、比較するグラフに両方とも分岐がある場合、動的計画法による比較ができないためである。
- (3) 比較の際には、動的計画法で計算した類似度が閾値よりも高い場合には、類似度が高いと見なして統合を行うが、類似度が閾値よりも低い場合には統合せず、新規グラフとして統合グラフに追加を行う。

動的計画法の比較の単位は、文頭・読点・句点で挟まれたフレーズとする。図4に、入力グラフ、統合グラフ、統合グラフに入力グラフを統合した新統合グラフを示す。点線で囲まれた単位がフレーズを表す。

また、動的計画法による統合グラフと入力グラフの類似度計算を図5に示す。図5では水平方向に進む場合には「脱落誤り」として0.7を累積スコアに加算し、垂直方向に進む

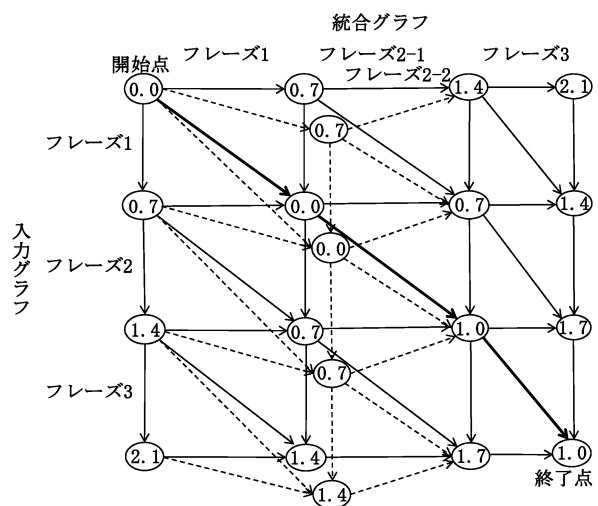


図5 統合グラフと入力グラフの類似度計算  
Calculation of similarity between unified graphs and input graphs.

場合には「挿入誤り」として0.7を累積スコアに加算する。斜め方向に進む場合には、対応するフレーズが同じ時には「正解」として累積スコアはそのまま、異なる時には「代替誤り」として1.0を累積スコアに加算する。図5の点線部分は、統合グラフのフレーズ2-2部分と入力グラフとの対応関係を表す。

図5では、左上が開始点でこの時の累積スコアは0.0である。各点では、点に集まるパスのうち、累積スコアが最も小さいパスを記録する。図5の右下にある終了点に到達するまで、各点でこの計算を繰り返し行う。終了点に到達したら、累積スコアが最小となるパスを、開始点まで逆に辿っていく。その結果、得られたパスにおける「脱落誤り」、「挿入誤り」、「代替誤り」、「正解」の組合せが、統合グラフと入力グラフの最適なマッチングを表すことになる。

図5では、終了点で累積スコアが最小になるパスは二つあるが、同じスコアの場合には、最初に計算したパスを優先することにするため、「フレーズ1：正解」、「フレーズ2-1：代替誤り」、「フレーズ3：正解」の組合せが累積スコア最小パスとなり、最適なマッチングを表すことになる。類似度 $P$ は、「正解」の数を $H$ 、「代替誤り」の数を $S$ 、「脱落誤り」の数を $D$ 、「挿入誤り」の数を $I$ として、次式で定義する。

$$P = \frac{H - I}{H + S + D} \quad (1)$$

このため、厳密にはタスクに依存する話だが、傾向としては、閾値を高い値に設定すると、類似度の高いグラフしか統合しないことになり、生成される文数は増加する。その結果、合成音の自然性は向上する可能性があるが、収録しなければならない録音文の数は増えてしまう。一方、閾値を低い値に設定すると、生成される文数は減る傾向にあるものの、合成音の自然性は低下する可能性が生じる。

したがって、閾値の設定は、タスクごとに録音文の数と合成音の自然性に依存することになるため、後述する実験では、予備実験の結果から、入力グラフと統合グラフの類

似度Pが閾値0.3以上のときに統合を行い、0.3未満の場合には入力グラフを別のグラフとして統合グラフに追加した。「閾値0.3以上」とは、「統合グラフと入力グラフで30%以上一致する部分がある」ことを意味することになる。

したがって、統合グラフは、複数のグラフを保持することになる。すでに統合グラフに複数のグラフが含まれる場合には、すべての統合グラフと比較を行い、動的計画法で計算した類似度が一番高いグラフに統合する。

統合する際には、動的計画法により求められた結果に基づき、「代替誤り」および「挿入誤り」と判定されたフレーズをグラフに付加する。また、「脱落誤り」と判定された際には、省略の分岐をグラフに付加する。図4では、入力グラフのフレーズ2が「代替誤り」と判定されたため、フレーズ2を統合グラフに追加する。

### 3.4 変数が連続している箇所に対する逐次貪欲アルゴリズムの適用

フレーズ内で変数が連続している場合、合成音の自然性を保つためには、すべての可能性のあるフレーズを録音することが望まれるが、その場合、録音文の量が非常に多くなる。例えば、図4のフレーズ1の「[X1] 時間後の [X2] 日 [X3] 時には、」の場合、[X1]に入る要素数が24、[X2]に入る要素数が31、[X3]に入る要素数が24であれば、 $24 \times 31 \times 24 = 17856$ のフレーズを収録する必要がある。

このため、提案法では、逐次貪欲アルゴリズム<sup>13)</sup>を用いることで、合成音を作成するのに必要となる最小限のフレーズを生成する。図4の「[X1] 時間後の [X2] 日 [X3] 時には、」というフレーズの場合であれば、93フレーズとなり、収録しなければならないフレーズ数を大幅に削減することができる。この逐次貪欲アルゴリズムを適用して得られた93フレーズを変数Y1と新たに置く。Y1のリストとして93フレーズが含まれることになる。この処理の結果を図6に示す。

### 3.5 統合したグラフにおける録音文セット最適化問題

統合したテンプレートにおいて、生成する録音文セットの大きさが最少になる最適化問題を設定する。生成する録音文セットは、各変数の要素をすべて含む必要があるため、

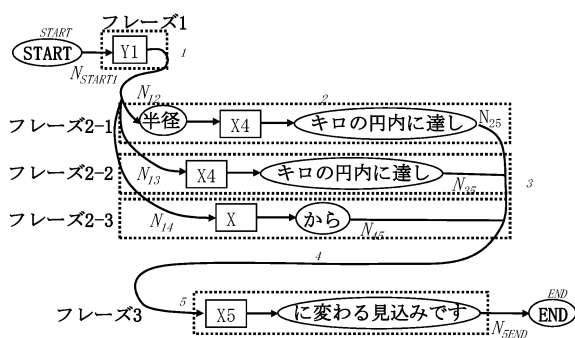


図6 逐次貪欲アルゴリズム適用後のグラフ  
Graphs after application of one-by-one greedy algorithm.

録音文セットがあるフレーズを含む回数は、そのフレーズに含まれる変数の要素数以上にならなくてはならない。例えば、変数Y1に含まれる要素数が93であれば、録音文セットがフレーズ1を含む回数 $F_1$ は、 $F_1 \geq 93$ となる。同様に、図6で変数X4、X5、X6の要素数がそれぞれ、50、2、10であれば、録音文セットがフレーズ2-1、2-2、2-3、3を含む回数 $F_2$ 、 $F_3$ 、 $F_4$ 、 $F_5$ は、それぞれ $F_2 \geq 50$ 、 $F_3 \geq 50$ 、 $F_4 \geq 10$ 、 $F_5 \geq 2$ となる。

収録時と合成時の音声データの文内出現位置を揃えたほうが、でき上がりの合成音の自然性は高くなるため、文頭にあるフレーズは文頭で発声し、文末にあるフレーズは文末で発声したほうがよい。このため、スタートノードから始まり、各ノードを通して、エンドノードに辿り着くような録音文のみ生成することにする。このような条件を考慮すると、あるフレーズを含む回数は、そのフレーズに入ってくるすべての経路を通る数の和になり、かつそのフレーズから出て行くすべての経路を通る数の和になる。図6の場合には、 $F_{START} = N_{START1}$ 、 $N_{START1} = F_1$ 、 $F_1 = N_{12} + N_{13} + N_{14}$ 、 $N_{12} = F_2$ 、 $N_{13} = F_3$ 、 $N_{14} = F_4$ 、 $F_2 = N_{25}$ 、 $F_3 = N_{35}$ 、 $F_4 = N_{45}$ 、 $N_{25} + N_{35} + N_{45} = F_5$ 、 $F_5 = N_{5END}$ 、 $N_{5END} = F_{END}$ の等式が成り立つ。ただし、 $F_{START}$ 、 $F_{END}$ は、録音文セットがスタートノードとエンドノードを含む回数であり、 $N_{ij}$ はフレーズ*i*からフレーズ*j*の経路を通る回数である。

以上より、録音文セットを生成する問題は、これまでに得られた不等式および等式を満たす中で、最少となる $F_{START}$ を得る最適化問題に帰着できる。この問題は、シンプレックス法<sup>20)</sup>を用いて解くことができ、その結果、各フレーズを通る回数を求めることができる。

### 3.6 録音文セットの生成

前節で、録音文の生成数と各フレーズおよび経路を通る回数決定された。録音文の生成を行うには、まず、スタートノードからフレーズ1への経路を考えて、フレーズ1における変数Y1の要素の1番目を用いる。次に、フレーズ1からフレーズ2-1、フレーズ2-2、フレーズ2-3への経路が考えられるが、まずフレーズ2-1への経路を考え、フレーズ2-1における変数X4の1番目の要素を用いる。さらに、フレーズ2-1からフレーズ3への経路を考え、フレーズ3における変数X5の要素の1番目を用いる。最後に、フレーズ3からエンドノードへの経路を考え、エンドノードで1文の生成が終了するので、スタートノードからエンドノードまでに用いるのが決定された要素をすべて接続すると1文が完成する。仮に、フレーズ1の1番目が「12時間後の20日午前3時には、」、フレーズ2-1の1番目が「半径10キロの円内に達し、」、フレーズ3の1番目が「台風になる見込みです。」であれば、「12時間後の20日午前3時には、半径10キロの円内に達し、台風になる見込みです。」という文が生成されることになる。

以降、先ほどと同じように、スタートノードからフレーズ1への経路を考えるが、すでにフレーズ1における変数Y1

の1番目の要素は使用済みなので、2番目の要素を用いる。もし、すべての要素が使用済みになった場合には、使用済みの要素を繰り返し用いることにする。次に、フレーズ1からフレーズ2-1、フレーズ2-2、フレーズ2-3への経路が考えられるが、シンプレックス法で求められた回数 $N_{12}$ だけフレーズ1からフレーズ2-1への経路を通っていないければ、フレーズ2-1における変数X4の2番目の要素を用いる。ここでもし、シンプレックス法で求められた回数 $N_{12}$ だけフレーズ1からフレーズ2-1への経路をすでに通っているのであれば、フレーズ2-2におけるX4の1番目の要素を用いるようにする。

このように、回数 $F_{START}$ だけ、録音文の生成を繰り返し行う。

## 4. 録音文セット生成実験

### 4.1 NHKラジオ第2放送の番組「気象通報」について

ラジオ第2放送の番組「気象通報」は、1928年11月5日から続く長寿番組である<sup>21)</sup>。毎日3回、1回につき20分間、全国の天気を伝えている。一般的な天気予報番組が各都市の天気予報を伝えるのに対し、「気象通報」は、山や海など各都市以外の場所において、自分で天気の予報を行わなければならない人に向けて、天気の概況を伝えることを目的としている。本番組を聴きながら天気図を描いている聴取者がいるため、アナウンサは天気図を描く際に必要な間を空けながら、正確に読んでいく必要があり、難度の高い業務である。このため、合成音による自動化が望まれている。

番組は、①「各地の天気」、②「船舶の報告」、③「漁業気象」、の3部構成になっていて、どの文も、3.2節で述べた変数と分岐と省略を用いたテンプレートで記述できる。例えば、③では、

「[場所]の、[緯度]、[経度]には、[気圧]ヘクトパスカルの、<[台風の大きさ・強さ]>台風第[数値]号があって、(ほとんど停滞していますor[方角]へ、毎時[速度]キロで進んでいます)。」

「なお、台風第[数値]号の中心は、[数値]パーセントの確率で、[数値]時間後の、[日付]、[時間]には、[場所]の、[緯度]、[経度]を中心とする[半径]キロの円内に達する見込みです。」

といったテンプレートが使われる。

### 4.2 録音文セット生成実験

提案法の有効性を示すために、気象通報の文章の一部を記述している9つのテンプレートを利用して、録音文セットの生成実験を行った。

従来法<sup>14)~18)</sup>はいずれの手法も、録音文セットの候補となるテキストをあらかじめすべてリストアップする必要がある、膨大な数になるため今回のタスクには適用することができない。したがって、ここではテンプレートからランダムに録音文セットを生成する手法と比較することにする。ランダムに録音文セットを生成する手法では、まずテ

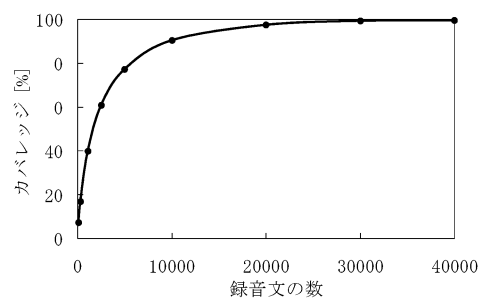


図7 ランダムに録音文セットを生成した際のカバレッジ  
Coverage of variable elements in the method of generating a sentence set to random.

ンプレートをランダムに選択し、次にテンプレートの分岐および省略の経路をランダムに選択し、さらに各変数に含まれる要素をランダムに選択する。この操作を文末まで繰り返し行い、一つの文を生成する。そして、この手順を生成する文の数だけ繰り返し行う。その結果、得られた録音文セットに含まれる各要素のカバレッジを計算した。これを図7に示す。録音文の数を増やすとカバレッジは上昇し、40,000文で99.7%に達する。文献14)と同様に、図7でも、カバレッジが大きくなるにつれて必要とされる文数が急速に大きくなるのがわかる。

一方、提案法により生成された録音文の数は1,085となり、このときのカバレッジは100%である。ランダム生成によるカバレッジが95%以上の場合と本手法を比較すると、録音文セット数が数%以下に削減できることとなる。したがって、削減効果は非常に大きいことがわかる。

## 5. 合成音主観評価実験

提案法により生成した録音文セットを用いて作成した合成音の自然性を評価するため、4.2節で得られた録音文セットの音声収録を行い、作成した合成音に自然音声を加えて、5段階の品質評価実験を行った。

音声データベースは、4.2節で得られた1,085文の録音文セットを読み上げることで構築した。読み上げたのは、気象通報の番組を担当している男性のアナウンサである。録音は防音室で行った。

評価用テキストは、録音文セットを生成する際に用いた9つのテンプレートで記述できる文章で、かつ録音文セットには含まれない63文を用いた。この評価用テキストを文献22)の手法で音声合成し、合成音の作成を行った。また、録音文セットを読み上げたアナウンサが、評価用テキストを読み上げるにより自然音声を収録した。評価には、この自然音声を加えた計126音声を使用した。

評価は、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。聴取レベルは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名である。各試行では、評価データをラン

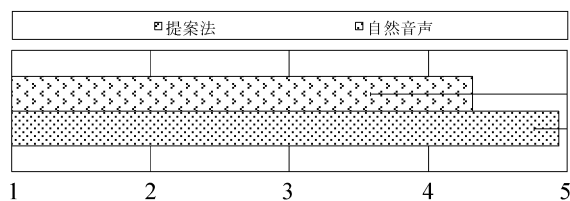


図8 提案法で作成した合成音の自然性

Mean Opinion Score (MOS) for speech samples synthesized by the proposed method.

ダムな順序で提示し、評定者には自然性の違いを、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)の5段階で評価するよう指示した。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<sup>23)</sup>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、本論文では合成音の自然性のレベルを具体的に知りたかったため、文献<sup>24)</sup>で行われているような5段階で評価することとした。評価に先立ち、音声波形データベース内の音声を3文章聞かせて、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。また、各音声の受聴は一回のみに限定した。なお、評価は、適度な時間間隔で休憩をはさみながら行った。

図8にMean Opinion Score (MOS)と標準偏差を示す。合成音のMOSは4.32となった。自然音声の評価値の4.94と比べると、自然性は低下するものの、比較的自然性の高い合成音が作成できているのがわかる。

本実験で得られた4.32は、文献<sup>13)</sup>で得られた4.90に比べて低い値となっている。低下の原因としては、文献<sup>13)</sup>では株価の合成音を評価したため、合成音の長さが平均2.4秒と短かったのに対し、本実験では合成音の長さが平均27秒と長くなったため、一つの合成音において、自然性の劣化が生じる可能性が増えたことが挙げられる。例えば、ほとんどの部分では自然性が高いものの、ある一箇所では自然性が劣化したために、合成音の評価が5から4になってしまうケースがしばしば見られた。また、合成音における自然性の劣化部分を調べると、誤ったアクセントになっていたり、音声の接続部分で接続前後の音が合っていないかたりしている。このため、録音文セットを生成した際に出力される統合グラフを利用し、統合グラフのノード間に接続部分を限定することができれば、正しいアクセントの実現が可能になり、最適な場所での音声の接続を実現できる可能性があるため、合成音の自然性をさらに向上できる可能性がある。

## 6. む す び

合成対象テキストが、変数・分岐・省略を用いたテンプレートで記述できる場合に、従来のように合成対象テキストをリストアップすることなく、類似度の高いテンプレートを統合することでテンプレートの削減を行い、テンプレ

ートから直接録音文セットを生成する手法を提案した。

テンプレートからランダムに録音文セットを生成した手法と比較した結果、提案法では、同じカバレッジを得るために必要な文数が数%以下まで削減できることを示した。

また、提案法で生成した録音文セットを読み上げて音声を収録し、合成音を作成して主観評価実験を行った結果、MOSで4.32という評価が得られた。放送利用を考えると、さらなる自然性の向上が望まれる。今後、録音文セットを生成した際に出力される統合グラフを利用した音声合成システムを検討するなどして、合成音の自然性の向上を図って行きたい。

本論文では、気象通報で用いられるテンプレートを利用して、提案法の有効性を示した。提案法は、気象通報に特化したものではなく、コールセンターの自動応答音声の作成や交通情報の自動作成など発声内容が限定されたタスクに応用できるはずである。他のタスクへ適用していくことで、提案法のさらなる有効性を検証していきたい。

## 〔文 献〕

- 1) 平岡, 内山: “「早朝・深夜の天気予報」の番組制作 天気予報アナウンスコメント自動送出”, テレビ誌, 41, 8, pp.742-743 (1987)
- 2) 北浜, 今川, 三浦: “自動音声付天気予報システム”, 第31回民放技術報告会予稿集, pp.94-95 (1994)
- 3) 澤口: “天気情報アナウンス送出装置”, 映像学技報, 21, 53, pp.25-30 (1997)
- 4) 足立, 渡辺: “自動天気音声アナウンス送出システム”, 第57回「明日の放送と技術フォーラム」講演予稿集, pp.27-28 (2004)
- 5) 河上, 下野: “CG, 音声合成を用いた全自動道路交通情報システム”, 第39回民放技術報告会予稿集, pp.148-149 (2002)
- 6) 大田, 岩下, 津浦, 近藤, 大須賀: “提供テロップ・アナウンスコメントの作成送出システムの概要”, テレビ学技報, 17, 23, pp.7-12 (1993)
- 7) 八木: “映像メディア技術”, オーム社, pp.201-211 (2007)
- 8) 板倉: “音声の線形予測分析”, 信学技報, SP98-55, pp.25-32 (1998)
- 9) 広瀬: “韻律と音声言語情報処理”, 丸善, pp.118-127 (2006)
- 10) E. Moulines and F. Charpentier: “Pitch Synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones”, Speech Communication, 9, pp.453-467 (1990)
- 11) A. J. Hunt and A.W. Black: “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database”, Proc. ICASSP, 1, pp.373-376 (1996)
- 12) キャンベル, ブラック: “CHATR:自然音声波形接続型任意音声合成システム”, 信学技報, SP96-7, pp.45-52 (1996)
- 13) 世木, 清山, 田高, 都木, 大出, 今井, 西脇, 小山: “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”, 映像学誌, 62, 1, pp.69-76 (2008)
- 14) 河井, 樋口, 山本: “基本周波数および音素持続時間を考慮した音声合成用波形素片データセットの作成”, 信学論, J82-D-II, 8, pp.1229-1238 (1999)
- 15) 磯, 渡辺, 桑原: “音声データベース用文セットの設計”, 音講論, 1, 2-2-19, pp.89-90 (1988)
- 16) J. V. Santen: “Diagnostic Perceptual Experiments for Text-To-Speech System Evaluation”, Proc. ICSLP, 1, pp.555-558 (1992)
- 17) 磯貝, 水野, 阿部: “形態素・音韻連鎖を考慮した音声合成コーパス用文セット生成の検討”, 音講論, 1, 1-6-17, pp.255-256 (2003)
- 18) C. Kuo and J. Huang: “Efficient and Scalable Methods for Text Script Generation in Corpus-Based TTS Design”, Proc. ICSLP, 1, pp.121-124 (2002)
- 19) 上坂, 尾関: “パターン認識と学習のアルゴリズム”, 文一総合出版, pp.91-102 (1990)

- 20)プレス：“ニューメリカルレシピ・イン・シー”，技術評論社，pp.317-329（1993）
- 21)NHK放送文化研究所：“20世紀放送史（資料編）”，日本放送出版協会，pp.126（2003）
- 22)世木，田高，清山，都木：“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”，情処学論，50，2，pp.575-586（2009）
- 23)音声入出力方式標準化委員会：“音声合成システムの性能評価方法ガイドライン”，電子情報技術産業協会，JEITA-IT-4001（2003）
- 24)H. Zen, T. Toda and K. Tokuda: "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006", IEICE Trans. Inf. & Syst., E91-D, 6, pp.1764-1773（2008）



**世木 寛之** 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．2009年，米カーネー  
メロン大学客員研究員．正会員．



**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  
にて音声合成などの研究に従事．正会員．



**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，同放送技術研究所勤務．  
2007年～2010年，（財）NHKエンジニアリングサービスに  
出向．話速変換，声質変換，音声合成などの研究に従事．  
1995年，日本音響学会佐藤論文賞受賞．正会員．



**都木 徹** 1981年，電気通信大学電気通信研究科  
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，字幕制作などの研究  
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．博士（工学）．正会員．



**斎藤 英雄** 1992年，慶大院理工学研究科電気工学  
専攻博士課程修了．同年，慶大助手．その後専任講師，  
助教授を経て，2006年，慶大理工情報工学科教授．この  
間，1997年～1999年まで，学術振興会海外特別研究員と  
して米国カーネギーメロン大学ロボティクス研究所に滞  
在し，主にVirtualized Realityの研究に従事．2000年～  
2003年，JSTさきがけ研究21「情報と知」領域研究員兼務．  
2006年より，JST CREST研究代表者．現在，主に，コン  
ピュータビジョンとそのVR技術への応用に関する研究等  
に従事．博士（工学）．正会員．



**小澤 慎治** 1967年，慶應義塾大学工学部電気工学  
科卒業．1974年，同博士課程修了．1970年，同電気工学  
科助手．同助教授，同理工学部情報工学科教授を経て，  
現在，愛知工科大学工学部情報メディア学科教授．画像  
と音声のデジタル情報処理に従事．道路画像の解析，  
スポーツ映像の解析に興味を持っている．電気学会ITS技  
術委員会委員長，2006年本学会会長などを歴任．工学博  
士．名誉会員．