個人適応音声情報配信システムにおける音声情報生成方法

坂元 盛浩　　齊藤 ゆみ
オムロン株式会社　コントロール研究所

1.はじめに

人は，自動車の運転などの作業を行いながら，音声により言語的な情報を入手することができる．このような「ながら」の状態での情報入手手段の代表例がラジオであるが，不特定多数のリスナーを対象としているため，提供される情報が個人にマッチしていない場面が多々ある．このような背景から，興味や嗜好により提供する情報を個人適応する音声情報配信システムの検討を行っている．このシステムでは，提供する情報として，ニュースなどのフロー系情報を想定している．

本システムでは，個人適応した情報を音声で提供するため，大量の音声データが必要になる．このデータを録音して用意するには，非常にコストがかかる．そこで音声合成によりテキスト情報から音声情報を生成することが必要になる．

本稿ではまず，人の肉声でニュース文を読み上げることができる波形接続方式の音声合成エンジン CrsyTalk™ の技術的特徴について述べる．そして，上記情報配信システムを前提に，合成音声の品質向上を目的とした言い換えによる音声情報生成方法について詳述する．

2.音声合成エンジン CrysTalk™

筆者らは，アナウンサーの肉声に近いニュース読み上げを実現することを目標に，音声合成エンジン CrysTalk™ の開発を行っている．CrysTalk は，波形接続方式[1]の音声合成エンジンであり，音素環境を考慮した可変長音素列を選択単位とする合成手法を用いている．CrysTalk の構成を**図 1**に示す．CrysTalk は，漢字かな混じり文に読みとアクセントを付与する「言語処理」，全体の抑揚や“ま”の目標値を決定する「韻律予測」，音声部品データベースから最適な部品を選択する「単位選択」，部品

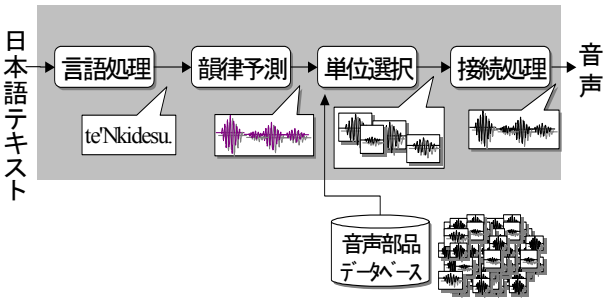

図 1　CrysTalk の構成

を１つの波形に接続する「接続処理」から構成される．音声部品データベースは，特徴データを付与した録音音声データを細かく断片化した音声部品から構成される．

従来の可変長音素列を選択単位とする手法[2]では，各選択単位の音素長にかかわらず，候補データの最小数は一定である．その結果，長い音素列を選択単位とするためには，大規模な音声部品データベースが必要となった．

CrysTalk の技術的な特徴の１つは，長い音素長の選択単位に対してはデータ数が少なくても候補とすることにある．これにより，比較的小規模な音声部品データベースでも長い音素長の音素列を選択単位とすることができる．

一方で，データ数が少ない選択単位では，韻律目標の F_0 値から大きく離れる候補が選択される確率が高まる．そこで，韻律目標の F_0 値を，データ数が少なく，かつ音素長が長い音素列に合せて修正する仕組みを導入している．修正の前後で韻律目標の形状は，前後の音素間で相対的に維持されるため，修正によるアクセント異常等の悪影響は発生しにくい．その結果，自然な韻律の音声を合成することができる．

3.言い換えによる音声情報生成方法

波形接続方式で合成される音声の品質は，音声部品データベースへの依存度が高い．音声部品データベースは，通常は人が原稿を読み上げた録音音声を部品として分割して作成するため，原稿に出現する語彙に対しては音声品質が高く，原稿に出現しない語彙に対しては品質が低くな

Voice Generating Method of Personalized Voice Information Providing System
Morihiro Sakamoto, Yumi Saitoh
Control Technology Laboratory, Omron Corporation

る傾向がでる．

一方，合成するテキストは，ニュースなどのフロー系情報を対象としているため，情報の内容が同一であれば入力テキストと一字一句違わずに合成する必要はないと考える．

そこで，音声部品データベースの内容を鑑みてテキスト情報を適切に言い換えることで，音声化したときの品質を向上する方法を考案した．

言い換えにより情報の内容に差異が起こらないことが必須であるため，通称や略称による言い換えと，表記ゆれによる言い換えの２つのみを行うこととした．以下，例を示す．

(1) 通称と略称(通称略称辞書として整備)
「首相」「内閣総理大臣」「総理大臣」など
(2) 外来語の表記ゆれ
「インターフェース」「インタフェース」
「インタフェイス」など

次に，言い換えの処理手順について説明する．予め，音声部品データベースの元となる原稿の中に上記の通称略称辞書の単語が含まれる場合は，辞書の該当単語をマーキングしておく．同様に，原稿に含まれる表記ゆれが起こりえるカタカナ言葉は，表記ゆれ単語リストに登録しておく．

入力されたテキストは形態素解析されて単語が抽出される．本稿では，単独の名詞または複合名詞を対象とする．これらの単語に対して，以下のルールで言い換えの処理を行う．

A) 抽出した単語 X が通称略称辞書に含まれ，かつマーキングされていないなら，マーキングされている同義の単語 Y に置き換える．

B) 抽出した単語 X がカタカナである場合，表記ゆれ単語リストに含まれるかをチェックする．含まれない場合は，単語 X の表記ゆれの単語 Z を順次生成し，表記ゆれ単語リストに含まれるかをチェックする．単語 Z がリストに含まれる場合は，その単語 Z で置き換える．

４．考察

言い換えによる音声品質の向上を以下の手順で検証した．

[手順 1] 単語 A と単語 B が前述した言い換えの関係にあり，単語 A が音声部品データベースに含まれ，単語 B が含まれないような単語セット(言い換え単語セット)を用意する．

[手順 2] 単語 A と「です。」を組み合わせた文章 A と，単語 B と「です。」を組み合わせた文章 B を用意する．

[手順 3] 文章 A，文章 B に対して，それぞれ音声合成処理を実施し，内部で計算している合成コストを得る．(合成コストは，選択した音声部品に関する前後音素環境に関するコストと韻律目標との差に関するコスト，音声部品間の接続コスト，の３つのコストに一定の重みをつけた総和として算出する)

上記手順で１０個の言い換え単語セットに関して，算出したコスト比を**表 1**に示す．コスト比は（言い換え後のコスト)/(言い換え前のコスト)である．また，表中の矢印は，左側の単語を右側の単語に置き換えることを示す．10 件中 9 件の言い換え単語セットでコストが低くなっていることがわかる．

表 1　言い換えによる効果(コスト比)

言い換え単語セット	コスト比
米国→アメリカ	0.751
コンピュータ→コンピューター	1.010
首相→総理大臣	0.082
レポート→報告書	0.763
サラリーマン→会社員	0.384
日本銀行→日銀	0.411
農水省→農林水産省	0.970
自由民主党→自民党	0.278
短期大学→短大	0.745
エーリアン→エイリアン	0.814

５．おわりに

波形接続方式音声合成エンジン CrsyTalk™ の概要とその技術的特徴を述べるとともに，合成音声の品質向上が期待できる言い換えによる音声情報生成方法について述べた．

今回述べた韻律目標の修正と，言い換えによる音声情報生成について，それぞれ評価実験を通じて効果を確認していく予定である．

参考文献

[1] Campbell, W.N. and Black, A.W.:CHATR:自然音声波形接続型任意音声合成システム,信学報,SP96-7,pp45-52(1996)

[2] 世木寛之,都木徹:可変長の音素環境依存音素列を単位とする波形接続型音声合成,信学技法,SP2003-83,pp.55-60(2003)