STRAIGHTに基づく柔軟な音声合成技術の開発

河原 英紀† 大西 壮登† 森勢 将雅† 高橋 徹† 西村 竜一†
坂野 秀樹‡ 入野 俊夫†
†和歌山大学システム工学部 ‡名城大学理工学部

1 はじめに

音声を用いたインタフェースが急速に普及している。インタフェースへの音声の応用は、システムを利用する際のユーザの負担の軽減に大きく貢献する。音声には文字情報だけではなく、音声でしか表すことのできない、個人性や感情などに代表される多様な非言語情報が含まれている。このような、非言語情報の存在が、ユーザ負担のない自然な音声対話の実現に重要な意味を持つ。本プロジェクトは、そのような自然な音声対話を実現するために、多様で自然な応答音声の合成技術を開発することを目的として進められた。

この目的を達成するための基盤技術として、聴覚における情報表現に基づく音声合成方式STRAIGHT[1, 2]を採用し、非言語情報の自由な制御手段として、このSTRAIGHTに基づくモーフィング[3]を採用して開発が進められた。本報告では、この「多様な音声合成プログラム」プロジェクトにおける研究開発の経緯と成果について紹介する。

2 開発目標

開発目標の一つとして、例えば応答音声に多様な声質や話し方の音声を作り込むような応用に必要な、非リアルタイムであるが高い品質を実現する一連のプログラムを設定した。もう一つの目標として、品質には限度があるものの、リアルタイムで動作する、声質や話し方の変換を実現する一連のプログラムを設定した。STRAIGHTおよびそれを用いたモーフィングは、行列をベースとした科学技術計算環境であるMatlab上で実装されているため、応用を容易にするために、開発目標のプログラムはCを用いて実装することとした。

3 基盤技術

STRAIGHTの基本構造は、1939年に提案されたChannel VOCODER[4]を踏襲している。STRAIGHTでは、入力された音声波形から、独立に操作できる音源情報(基本周波数および非周期性指標)とフィルタ情報(スペクトル包絡)が抽出され、実数値として表現されるそれらのパラメタのみから音声が再合成される。従来、このような分析合成に基づく音声合成技術は、柔軟な音声の加工が可能であるものの、高い品質の音声の合成は困難であると考えられてきた。一方、膨大な量の収録音声データベースから、条件に適合する波形素片を検索し編集する波形接続型音声合成(concatenative synthesis)は、容量制限の厳しい組込システムへの応用が困難であるという問題をかかえていた。これらの問題を解消する手段として、分析合成型でありながら高い品質の音声の合成が可能であるSTRAIGHTが基盤技術として採用された。

様々な非言語情報の属性のそれぞれと物理パラメタとの相関を調べることを通じて、非言語情報の操作システムを開発しようとする従来のアプローチとは異なり、本プロジェクトでは、事例に基づく操作システム[3]の開発を進めた。これは、前述のようにSTRAIGHTでは3種類のパラメタのみから高度に自然な音声が再合成できることによる。必要とする属性を有する実例さえ用意できれば、加工対象の音声のパラメタとそれらの実例のパラメタとの補間により、自由に目的とする属性を操作することができる。このようなアプローチを採ることにより「感情の推定や記述」という、人間にとってさえ困難な未解決の問題を直接扱わずに開発を進めることが可能となった。

4 開発の経緯

プロジェクトの推進にあたり、開発に直結する、(1) Matlabコードの体系化、(2)多様な発声を含む音声データベースの構築、(3)Cによるオフラインプログラムとオンラインプログラムの実装を進めるとともに、開発したプログラムの技術移転を側面から促進するために、(4)応用システムの開発例の蓄積、(5)国内外への普及活動を進めた。プロジェクト前半においては、音声データベースの構築と、Matlab版での分析とモーフィング事例の蓄積に基づくアルゴリズムの改良とコードの体系化を行い、プロジェクト後半において、それらに基

Development of versatile speech synthesis technology based on STRAIGHT
†Hideki KAWAHARA, Masato ONISHI, Masanori MORISE, Toru TAKAHASHI, Ryuichi NISIMURA, Toshio IRINO ‡Hideki BANNO
†Faculty of Systems Engineering, Wakayama University
‡Faculty of Science and Engineering, Meijo University

づくCでの実装を重点的に進めた。また、並行して、応用システムへのSTRAIGHTアルゴリズムの提供と、様々な機会を捉えた宣伝普及活動を進めた。

5 開発システム

Cによる実装では、オンラインプログラムの開発を先行させた。一括処理に基づいていたMatlabによる実装を根本的に見直し、幾つかの品質向上策を省略することにより、実時間動作する音声変換システムとして実現した[5]。オフラインプログラムの実装は、この開発経験に基づきAPIの体系化および分析パラメタの最適化を並行して進めた。この過程で、高品質領域におけるパラメタの冗長性と品質のトレードオフの定量的評価が進められた。

6 応用システムと普及活動

これらの、基盤となる開発システムを用いることで、様々な応用に適した音声合成／変換システムを構築することができる。

開発システムならびに、応用システムの普及には、基盤技術であるSTRAIGHTおよびモーフィングの知名度を向上させることが必要である。学会を通じた継続的な活動により、STRAIGHTは、音声知覚研究のデファクトスタンダードとして内外の研究機関で数多く使用されるに至っている[2]。STRAIGHTのスペクトル分析は、HMMに基づくTTSシステムに応用され[6]、別稿で詳しく説明される国際的な普及活動の一環であるBlizzard Challange 2005, 2006において、優れた成績を上げる原動力となった。また、STRAIGHTに基づく統計的音声モーフィングアルゴリズム[7]は、当初計画を超える成果である無音声通信での非可聴つぶやき声（NAM）の変換においても基幹技術として応用されている。一般の社会への広報も、日本科学未来館の企画展への感情モーフィング（図1）の出展を行った他、オンラインプログラムのCによる実装に基づき、webを介して音声変換を体験することのできるページを用意して進めている[8]。

7 まとめ

本プロジェクトでは、STRAIGHTとモーフィング技術を基盤として、自然な音声対話を実現するための多様で自然な応答音声の合成技術を開発してきた。こうして開発された、本プロジェクトによる成果である「多様な音声合成プログラム」および関連する技術資料は、

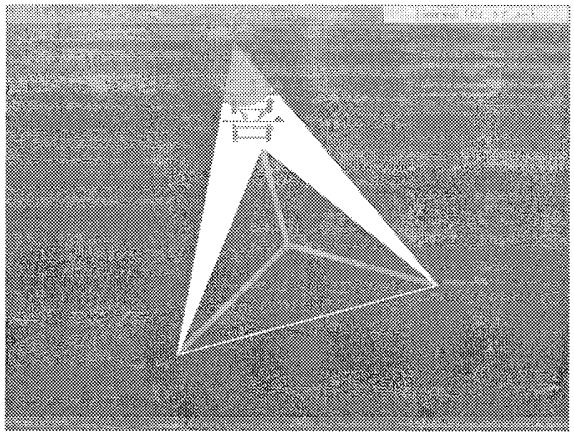

図1: User interface for morphing demonstration (courtesy of the Miraikan, designed by Takashi Yamaguchi)

ホームページからダウンロードすることが可能となっている。

参考文献

[1] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.

[2] 河原英紀. Vocoderのもう一つの可能性を探る–音声分析変換合成システムSTRAIGHTの背景と展開–. 日本音響学会誌, Vol. 63, No. 8, pp. 442–449, 2007.

[3] H. Kawahara and H. Matsui. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Proc. ICASSP 2003*, Vol. I, pp. 256–259, Hong Kong, 2003.

[4] H. Dudley. Remaking speech. *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177, 1939.

[5] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara. mplementatioin of realtime straight speech manipulation system: Report on its first implementation. *Acoustical Science and Technology*, Vol. 28, No. 3, pp. 140–146, 2007.

[6] Heiga Zen, Tomoki Toda, Masaru Nakamura, , and Keiichi Tokuda. Details of nitech hmm-based speech synthesis system for the blizzard challenge 2005. *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 1, pp. 325–333, 2007.

[7] T. Toda, A. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *Proc. ICASSP 2005*, Vol. I, pp. 9–12, 2005.

[8] 西村竜一, 三宅純平, 河原英紀, 入野俊夫. 音声入力・認識機能を有するwebシステムw3voiceの開発と運用. 情報処理学会研究報告, 第2007-SLP-68巻, 3 2007.