4D－7

波形の時間的変化に基づいた日本語音声合成

鈴木 将貴　中西 正和

慶應義塾大学大学院 理工学研究科 計算機科学専攻

1. はじめに

テキストからの音声合成における韻律情報決定の第一段階と音声波形生成の第二段階について、これらを独立したものとして扱い、音声波形生成のみを行うシステムをパーソナルコンピュータ上で実装した。韻律情報についてはモデルを簡素化し、情報を直接与えることで前段階の処理を省いている。合成単位としては日本語の母音と子音とからなる33種類の音素を単位とする。音声合成における波形生成の研究は、周波数領域におけるものが中心であったが、本研究では波の形そのものに注目し、その形の時間の経過に伴う変化を疑似的に生成することにより音声を合成する。アルゴリズムを単純化することで計算量を押え、特別なハードウェアを必要とせずに音声波形を生成している。

2. 韻律情報

音声波形の周波数成分の変化は図1に示す$F0$モデルとして捕らえられることが知られている[1]。

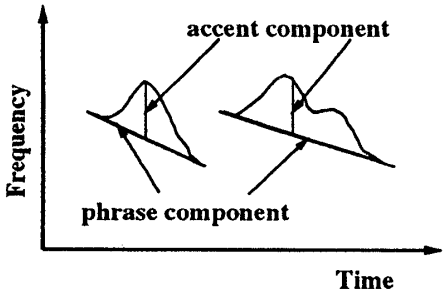

図1: $F0$モデル

音素ごとにアクセントの有無を判定し、アクセント成分を一定値とする各音素に固有の周波数を求める。ただし、各音素が出力されている間、ずっと一定の値を保つことにすると、アクセントが付いている音素とアクセントが付いていない音素との間の周波数の変化が急激であり、その差が目立ってしまう。実際の発話では周波数はある程度滑らかに変化していると考えられるので、このような場合、音素から音素へのわたりの部分で滑らかに接続するために修正を加える。

3. 母音の合成

母音の音声波形は子音のものに比べて、はるかに単純な形をしており、細かい変化はあるものの、基本的には山と谷の連続として捕らえることが可能である。それぞれの山や谷がどのような形をしているか、まず実際の波形からその特徴を抽出する。ここで、特徴としているのは山や谷の音素の中でのタイミング、またその高さや深さ、斜面がどのような傾きをもっているかなどであり、それぞれ数値化し、パラメータを保存しておく。

母音にはその母音に固有の周波数が与えられているので、その時間的な長さに応じて、保存しておいたパラメータをもとに音声波形を伸縮させる。

4. 子音の合成

子音の音声波形は母音に比べて複雑であり、モデル化するのは難しい。そこで、子音を2つの部分に分ける。前半部は子音の最初の部分であり、時間的に継続しない部分である。また、後半部は子音の後部であり、時間的に継続可能な部分である。波形を生成する際には前半部は1回のみ出力し、後半部は子音の長さに応じて繰り返し出力する。

Japanese Speech Synthesis Based on Change of Waveform in Time Course

Masataka SUZUKI
Masakazu NAKANISHI
Department of Computer Science, Graduate School of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223, Japan

5. 音素間のわたりの部分の合成

合成単位を音素としているので、音素間のわたりの部分の波形を生成するためには、隣接した2つの音素から中間的な波形を合成しなければならない。音素を大きく母音と子音の2種類に分けると、その接続の仕方は4通り考えられる。そのそれぞれに対して、以下のように接続方法を考える。

5.1 子音から母音への接続

子音から母音への接続はさらに2通りの場合が考えられる。すなわち、子音と母音を切り離して考えられる場合と、子音から母音へと徐々に変化していく、または子音と母音が重なりあって切り離すことが不可能な場合である。前者の場合は単純に接続する。音量等の調節は行うが、子音の波形を出力し、その後に母音の波形を出力する。後者の場合は子音のモデル化が困難なため、波の形の変化を合成によって行うのは難しい。そこで、単純な波形の重ね合わせを行う。子音から徐々に母音の割合が大きくなるように、音量を調節することにより、急激な変化を和らげることができる。また、接続の形態が前者か後者かは前に来る子音の種類によると考えられ、母音には無関係に子音だけを見て判別する。

5.2 母音から母音への接続

英語では二重母音として扱うべきものであるが、日本語を扱う本研究ではバラバラな音素の連続として捕らえる。母音から母音への移り変わりにおいては、波の形が前の母音とも後の母音とも異なる、両者の中間の形を経て変化することによって、接続の際の急激な変化をなくし、滑らかな接続が行われると考えることができる。母音の出力の際にその波形をパラメータとして合成するが、中間の波形を生成するためにもそのパラメータを用いる。合成はパラメータを前の母音のものから後の母音のものへと徐々に変化させ、わたりの部分の各時点における波の形を決定する。

5.3 子音への接続

子音への接続は基本的に何も行わない。すなわち、新たな音節が始まるときには、前の音節とのつながりは弱いと考え、音量の調整あるいはポーズによるタイミングの調整などは行うが、波の形そのものを変えるようなことは行わない。

6. 結果

生成された音声は自然とは言い難いのが現状である。とくに音素間のわたりの部分では波形がどのように変化していくのかのモデルを単純化しており、変化の仕方を実際に発話された音声波形の変化の様子に即したものにするためには、音声波形のさらなる分析が必要と思われる。

7. まとめ

周波数領域ではなく、時間領域を中心として処理を行い、音声波形の生成を行った。音素間のわたりの部分においては、隣接した2つの音素からその中間の波形を生成した。まだ自然さを欠いているのが現状であり、より実際の音声波形に近づけるような、わたりの部分の波形変化のモデルを確立する必要がある。

参考文献

[1] K. Hakoda and H. Sato. Prosodic Rules in Connected Speech Synthesis. *Trans. IECEJ*, J63-D(9):pages 715–722, 1980.

[2] K. Takahashi, K. Iwata, Y. Mitome, and K. Nagano. Japanese Text-to-Speech Conversion Software for Personal Computers. In *Proc. ICSLP 94*, pages 1743–1746, 1994.

[3] K. Momosaki et al. A Japanese TTS (Text-to-Speech Synthesis) Software for Personal Computers. In *Proc. the Autumn Meeting of the Acoustical Society of Japan*, pages 327–328, 1994.