1T-1

Text to Speech と同期して動作する唇モデル

酒巻　亮†　　杉本　富利‡
東洋大学大学院工学研究科情報工学専攻†
東洋大学工学部情報工学科‡

1．はじめに

現在、日常生活のさまざまな場面において音声合成システムが用いられ、多くのサービスが提供されている。しかし、音声は聴覚にのみ働きかける情報であるので、音声が聞き取り難い環境においては情報を伝えることが困難な場合がある。聴覚だけでなく他の感覚器においても音声情報を伝達することのできるインタフェースが実現すれば、音声情報の補助的な情報伝達手段としても有効であると考えられる。

そこで本研究では、発話内容を唇の動きから理解する読唇術[1]に着目し、人間の発声時における唇形状の変化特性に基づいて、テキスト音声合成と同期して唇の形状を変化させる唇モデルの作成を試みた。

音素や音節単位に対応する画像を表示させてアニメーションを実現させる類似研究[2]もあるが、本研究では音声合成に用いられる音素コンテキストという概念を導入し，この音素コンテキストのパターンの組み合わせによって唇のアニメーションを実現する。

2．日本語の音素

発生音の最小単位は「子音」と「母音」の音素であり、日本語では 1 個または 2 個の音素の組み合わせで音節が形成される。さらに、音節を連ねて一つの単音が表現できる[3]。このことから、単語は一連の音素のつながりであり、このつながりの一部を取り出したものが音素コンテキストである。

2.1　音素コンテキスト

単語発声時に、それぞれの音素がその前後の音素からどの程度の影響を受け、また与えているのかについて考えていく必要があるが、本研究では、ある音素は後方の音素からは何の影響も受けなく、前方の音素からのみ影響されると仮定する。そこで前方の音節について見ていくと、音節であることから当然に母音が含まれている。母音の唇の形の変化は、発声時の唇の形の変化に最も大きな影響を与えている。そして、子音は各子音行の特徴を付加するために存在し唇の形の変化にはあまり大きな影響は与えないと考えられる。よって唇の変化を母音から母音への形の変化を基本とし、その変化過程の中間に挟まれている子音の影響によって起こる一連の変化として捉えるとよい。本研究では母音＋子音＋母音の流れを一つのセットとし、これを音素コンテキストとする。ここで同じ音素コンテキストの中に母音が二つあるので、前者を「第一母音」後者を「第二母音」とする。第一母音は音素全体を表すわけではなく、母音発声完了時の最後の唇の形のみを表すものとする。また、発話開始時は第一母音が存在しないので、子音＋母音の組み合わせを音素コンテキストとする。よって、本研究では図 1 で示す二種類の音素コンテキストのそれぞれの特徴について解析する。

2.2　音素コンテキストの種類

第一母音と第二母音の組み合わせは 25 種類あり、これを大分類とする。第一母音と第二母音の間には十数種類の子音が挟まれており、ひとつの大分類の中で間に挟まれる子音によって唇の形状の変化が異なるので、それぞれの変化過程毎にその変化を再現するための唇の特徴点のトレースデータを作成する。子音＋母音の音素コンテキストは 117 種類、母音＋子音＋母音の音素コンテキストは 682 種類存在する。そして、この音素コンテキストを組み合わせることによって、ほとんどの日本語の単語は表現可能である。

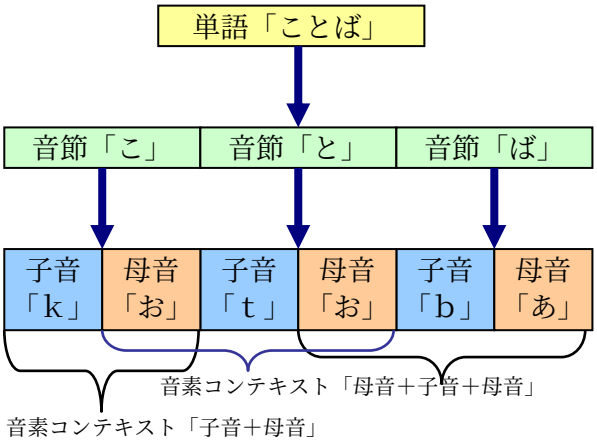

図 1　音素コンテキスト

A Lip Model Synchronizing with Text to Speech
†Ryo SAKAMAKI: Dept. of Information & Computer Sciences, Graduate School of Engineering, Toyo University
‡Futoshi SUGIMOTO: Dept. of Information & Computer Sciences, Faculty of Engineering, Toyo University

3．発声時の唇の変化の特徴抽出

唇モデルを作成するための前提として唇の形状の変化特性を抽出しておかなければならない。そこで、単語発声時の唇の変化の映像と音声から変化特性を抽出する。

3.1　単語発声の録画・録音

録画にはハイスピードカメラを使用した。録画環境は 240fps、サイズ 256×256（ピクセル）である。録画の際、被験者の唇の縁に特徴抽出がしやすいように 8 つの点を打っておく。

3.2　音声解析

録画した映像を音素ごとに区切るために、図 2 に示すように映像と同時に録音した音声を解析し、そのスペクトログラムを手掛かりとして唇の映像を音素毎に区切る。ここでは音声解析ソフトを使い、フォルマント[4]の音素による特徴の違いに着目して時間軸を各音素ごとに区切り、その時間幅を唇の映像の時間軸に当てはめる。

ここで第一母音の終わりから第二母音の終わりまでの間が一つの音素コンテキストであり、その先頭の部分は前の音素コンテキストの第二母音の終端であり、その時の唇の形が次の第一母音の形となる。そこから子音を挟み第二母音に入り一つの音素コンテキストが終わる。この一連の流れに対応した映像の変化特性を抽出していく。

3.3　変化特性の抽出

音声解析をもとに一つの音素コンテキスト部分を抜き出した映像をイメージトラッカーというソフトを使い、被験者の唇に打った点の 2 次元平面状での動きを時間軸に沿ってトレースしたデータを得ることができる。

4．唇モデルの実現

唇モデルは 2 次元で描き、スプライン曲線を使い前述したトレースデータを音素コンテキスト毎に読み込み、それらをつなぎ合わせて図 3 に示すような 2 次元で描かれる唇モデルのアニメーションを作成する。

5．まとめと今後の課題

現在、基本システムは完成しほとんどの日本語の単語の表現は可能にすることができた。しかし、それぞれの音素コンテキストの時間幅を一定に設定しているので、口の動作に違和感を感じる部分もあるので改善が必要である。そして、モデルに陰影を加えるなどしてよりリアリティのあるモデルとし、より情報伝達能力の高い唇モデルへと改良する。

参考文献

[1] 佐藤則之“聴力障害者へのミューラー・ウォール読唇法”，ろう教育研究年報（金沢大学教育学部）10，pp.21-42，1962.3
[2] 小林隆夫，益子貴史，徳田恵一“音声およびテキストからの音声同期唇アニメーションの自動作成”
[3] 千駄ヶ谷日本語教育研究所，『日本語教育講座 II 音声，語彙・意味』
[4] レイ・D・ケント/チャールズ・リード著，荒井隆行，菅原勉　監訳，『音声の音響分析』，海文堂.

図３　唇のアニメーション（例：ひだり）

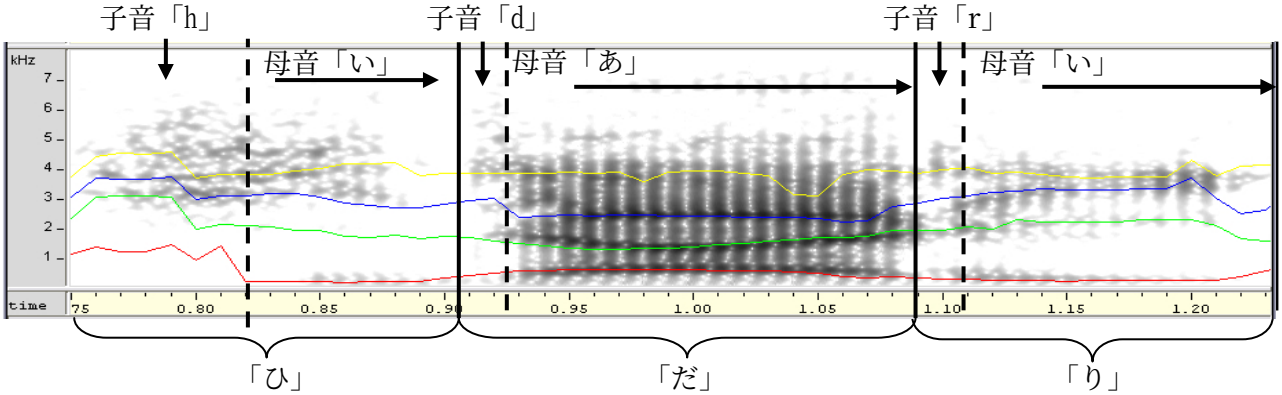

図２　音声解析（例：ひだり）