音声合成

Speech Synthesis / ニック・キャンベル (Nick Campbell)

音声合成はメディア変換技術である．当初の開発目的は符号化などのために文テキストを音声に変換する朗読器として考えられていた．しかし現在では，単なる朗読器から「話す機械」へと大きく進化してきており，文テキストを音声に変換するという捉え方は適当ではなくなってきている．将来は人間に代わって情報を相手に伝える，対話のできる「コミュニケーション・ロボット」の役割が期待できる．

これに伴い，音声合成の技術的な関心も，音声の生成から文章の読み上げ特徴の生成へ，そしてヒューマン・インタフェースへの応用へと拡大してきた．音そのものの研究を第一世代の音声合成，韻律の研究とコーパスベースを第二世代とすると，第三世代ではニュアンスの研究が重要課題となる．

音声合成の研究はダドレイ（Dudley）のボコーダ［0］をはじめ，ファント（Fant［1］）やクラット（Klatt［2］）などに代表される音声のモデル化から始まった．これらの研究では，音声生成を音響的に捉えて，声道の特性を表わすモデルと音源の特性を表わすモデルの二つで行った．基本単位は音素で，パラメータ行列によってスペクトル特徴を規定し，これらを連続になるように補間した．規則による合成ゆえにパラメータ量は小さく抑えられており，補間処理は十分とはいえないものの合成音声の了解性は高く，人間の声を規則によって生成することが可能となった．

音声合成における次の顕著な展開はコーパスの利用である．オリーヴ（Olive［3］）らは，人間の発声した自然音声を，ダイフォン単位（先行音素と後行音素の半音素ペアで音素と音素の接続区間を含む）に切り出した．基本単位を大きくすることによって基本単位のもつ情報量が増大し，合成音声の品質を向上させることができた．しかし，韻律変換のための信号処理が必要であり，後に波形編集法（PSOLA）などが開発され，またLPC法，ケプストラム法などの符合化が利用され，合成音声の機械音化を免れなかった．

藤村（Fujimura［4］）はデミシラブル単位での音声合成を試み，匂坂（Sagisaka［5］）は可変長の単位による合成の手法を考案した．この時期，発話音声データベースの構築は世界的に進められ，また韻律に関しても特に基本周波数や音韻継続長などの研究が大きく進展した．信号処理による影響を取り除くためキャンベル（Campbell［6］）は基本単位を音素に戻し，韻律情報までを含んだ単位選択手法を提案した．

これらの試みにより，第二世代では情報量の拡大をキーとして音声合成の品質を大きく向上させてきた．しかし人間の声と聞き違えるほどには達せず，チューリング・テストにはとても合格しない．

多くの人は人間らしくない不自然な合成音声を嫌う．そのため駅の構内アナウンスに見るような録音接続方式の音声合成が一般的に使われている．

第三世代の音声合成は発話を3次元で定義する．第1次元は音素特徴を，第2次元は韻律特徴を，そして第3次元はテキスト内容に合う適当な声や発話スタイルを決定する．さらに親しみやすい人間らしい音声合成が必要とされる．

音声合成の処理はテキストから読み（発音と韻律）を予測し音の流れを決定する．テキスト入力のみでは必要な発声の情報をすべて予測することはできない．その予測には，文構造に関する情報や書き言葉を話し言葉に変換する技術が必要である．それに対して，機械との対話による「コミュニケーション・ロボット」，あるいは電子化情報コンテンツを提示する際の意味・概念構造からの音声合成であれば，表わしたい意味や強調も明確となり，聞きとりやすい，適切な音声の選択ができるはずである．

また，テキストへのXMLマークアップによって音色などを含む声のニュアンスを制御する情報を付加すれば，話者選択，発話スタイルや感情，声の響きなどのキメ細かい制御が可能となる．

人間の脳は，処理速度は遅いがメモリの容量は大きい．音声合成技術が進化するにつれて，パラメータが飛躍的に増加するとともにその結果として計算量は減少しており，音声合成の手法は人間の脳の仕組みに，より近くなってきたと言えよう．

以上の詳細については，COCOSDA音声合成委員会のホームページを参照されたい．

参考文献

[0] Dudley, H.: "The Carrier Natural of Speech", *Bell System Tech. J.*, Vol. 19, 495-515, 1940.

[1] Fant, G.: "Acoustic Theory of Speech Production", Mouton, The Hague, 1960.

[2] Klatt, D. H.: "Software for a cascade/parallel formant synthesizer", *J. Acoust. Soc. Amer.* 67, 971-995, 1976.

[3] Olive, J. P.: "Rule synthesis of speech from dyadic units", 568-570, *Proc. IEEE-ICASSP77*, 1977.

[4] Lovins, J. B., Macchi, M. J. & Fujimura, O.: "A demisyllable inventory for speech synthesis", In J. J. Wolff and D. H. Klatt (Eds.), ASA *50 Speech Communication Papers, 519-522, 1979.

[5] Sagisaka, Y.: "Speech Synthesis by Rule using an Optimal Selection of Non-Uniform Synthesis Units", 679-682, *Proc. ICASSP*, 1988.

[6] Campbell, W. N.: "Synthesis units for natural English speech", Transactions of the Institute of Electronics, *Information and Communication Engineers*, 55-62, Vol. SP 91-129, 1992.

[7] COCOSDA's speech synthesis working group home page: www.itl.atr.co.jp/cocosda/synthesis

音声認識……84
ヒューマン・インタフェース…90
チューリング・テスト……382