口調・声質を学習する音声合成システム—TOS Drive TTS

文章を入力すれば，どのような文章でも音声に変換して出力する。それがテキスト音声合成技術です。

当社は，人間が発声した音声データを手本にして学習し，その人に似た口調，似た声色の音声を合成する技術を開発しました。TOtally Speaker Driven Text To Speech（TOS Drive TTS）と呼ばれるこの技術は，パソコン（PC）のテキスト読上げソフトウェアやカーナビゲーション（以下，カーナビと略記）の音声出力などに応用され，製品化されています（**図1**）。

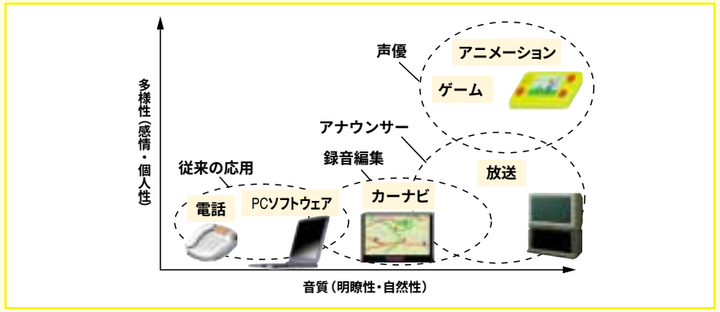

図1．音声合成の応用分野　従来の文章読上げソフトウェアに加えて，カーナビで音声合成の採用が増えています。音質の向上に伴って，その応用分野は放送にも広がり，更にエンターテインメント分野にも広がっていきます。

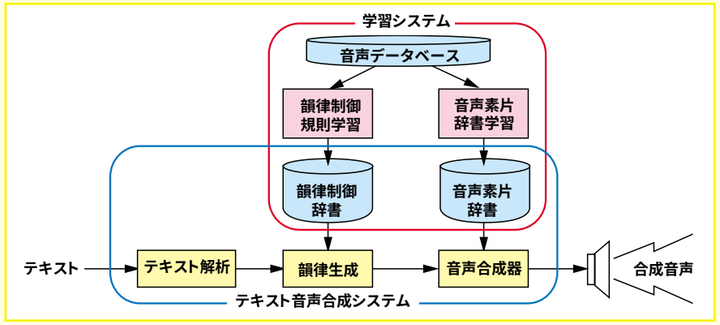

図2．TOS Drive TTS　ナレーターが発声した大量の音声データベースを基に，韻律制御辞書と音声素片辞書を自動学習し，ナレーターの口調・声質に似た，自然で高品質な音声を合成します。

■音声合成器

「音声合成なんて簡単」と思われる方もいるかもしれません。「ひらがなを一文字ずつ録音して，並べ替えて順番に再生すればいい」と。確かに，音声合成器は，音声の短い区間のデータ（音声素片）をつなぎ合わせて音声を作り出しています。ただし，つなげるだけではなく，録音した声の高さを，作りたい声の高さに変える処理を行います。自然な抑揚の音声を作るためには，任意の高さの声を合成することが必要なのです。この“高さを変える”処理によって音質が劣化し，不明瞭（ふめいりょう）な，機械的な，鼻に掛かった声になってしまうということが長年の問題でした。

そこで，“声の高さを変えない”という方式が登場しました。つまり，あらかじめあらゆる高さの声を録音しておくのです。当然，システムは巨大になり，数百Mバイトの記憶容量が必要です。これに対して，当社は，“高さを変えても劣化しない音声素片”を作り出す技術を開発しました。これは，様々な高さで発声されたナレーターの肉声と，それらと同じ高さになるように声の高さを変更して作られた合成音を比較し，その差が最小となるような音声素片を作る技術です。この，音声素片の閉ループ学習という技術によって，数百Kバイト程度のコンパクトな音声素片で，ナレーターの声色に近い高音質な合成音声が実現できました（**図2**，**図3**）。

■韻律制御

人間に近い合成音声を実現するうえでもう一つ大切なのが，声の高さの変化パターン（ピッチパターン）の制御です。従来は，「単語の品詞が何で，アクセント位置がどこで，文字数がいくつのときは，声の高さがどう変化する」というような規則を研究者が作っていました。時間と労力を費やして詳細な規則を作り上げた結果，確かに正しいアクセント・イントネーションの合成音ができるようになりましたが，どうしても人間がしゃべっているようには聞こえず，単調で機械的な印象がなくなりませんでした。

当社は，肉声を手本として学習することでこの問題も解決しました。ナレーターの発声した大量の音声データのピッチパターンから，文節単位の典型的なパターン（代表パターン）をあらかじめ数十個抽出しておき，これらの中から文節ごとに選択されたパターンをつなぎ合わせることで，文のピッチパターンを生成します。

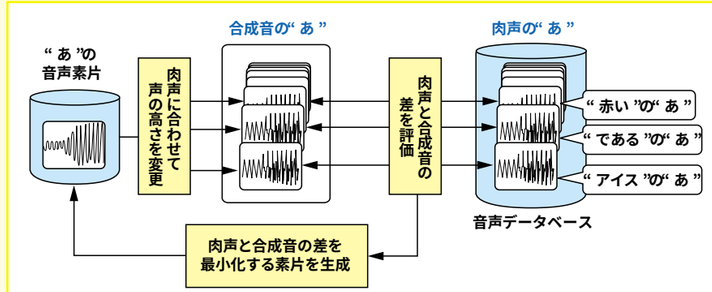

図3．音声素片の閉ループ学習　様々な声の高さの肉声と，それらと同じ高さになるように合成された音声との差を評価し，その差を最小化することにより，声の高さを変えても劣化しない音声素片を生成します。

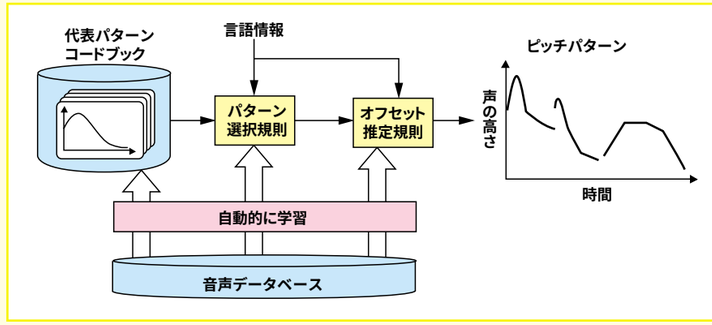

図4．ピッチパターン制御モデル　代表パターンコードブック，パターン選択規則，オフセット推定規則を音声データベースから自動学習することで，自然な抑揚を実現することができます。

代表パターンはもちろん，各文節の言語情報に基づいて適した代表パターンを選択するための規則も，学習によって自動的に作成できます。ナレーターの口調に似た，自然な抑揚の合成音声がこの技術によって実現できました（**図4**）。音声素片とピッチパターン制御規則を肉声から学習するTOS Drive TTSは，言語に依存しない技術なので，日本語以外の言語にも応用することが可能です。既に，アメリカ英語に応用して，その有効性が確かめられており，イギリス英語，ドイツ語，フランス語などの欧州言語についても開発を進めています。

■音声合成の応用分野

当社の音声合成技術は，PC“Dyna-Book”にプリインストールされている“東芝音声システム”やパッケージソフトウェアの“LaLaVoice$_{TM}$ 2001”などのテキスト読上げソフトウェアとして製品化されています。そのほかにも，ベルの代わりに“○○さん”と，だれからの電話か声で教えてくれる電話機に使われています。また，最近は，カーナビの音声出力のためのミドルウェアとして，当社の音声合成が広く使われるようになってきました。しゃべることが決まっているはずのカーナビですが，膨大な数の地名や施設名称を録音して蓄積するのは，DVDでもできません。音声合成なら，テキストさえ入れておけばOKなので，記憶容量の点でも，語彙（ごい）のメンテナンスの点でもつごうが良いのです。今後は，インターネットに接続して，電子メールやWebを読み上げるカーナビが増えると予想され，音声合成は必須の技術となるでしょう。また，更に音質が向上すれば，ニュースや天気予報などの放送に使うことも，可能になります。

現在のところ，合成された音声は，文章を朗読するような調子の音声ですが，声質や口調を様々に変化させたり，感情のこもった音声を合成したりする技術が開発されれば，音声合成の応用は，エンターテインメントの分野にも広がっていくことでしょう。将来は，“映像はコンピューターグラフィックス，音声は合成音声”の映画を作ることも夢ではありません。

研究開発センター
マルチメディアラボラトリー研究主務
籠嶋　岳彦