外国語の音声合成について *

◎ニック キャンベル
(ATR 音声翻訳通信研究所)

要旨

本稿では、第一言語の音声データベースから外国語の音声を合成する手法について提案する。音響的、韻律的に高品質な発話を生成するため、本方式では目的言語の話者データベースから作成した中間情報を用いて2段階で音声を合成する。実音声ケプストラム・ターゲットを得るため、ネイティブ話者の発話をモデルとして利用する。

1 はじめに

現在、ATR 音声翻訳通信研究所では音声言語を自動翻訳するための基盤技術が研究されており、それには通信やマルチメディア・アプリケーション等の利用が考えられる。このメディア変換技術は、音声の認識結果を、言語翻訳に渡し、音声合成を用いて声に変換するという、言語間の音声情報の伝送である。

音声翻訳における出力音声は、これまで出力言語（ターゲット言語）の母語話者の合成音声で出力するのが一般であった。そのため、ことばの意味は伝わるものの、入力話者の声と出力話者の声が異なり、音質的にも、また話者の特徴的にも違和感が持たれた。

そこでこの問題を解決するため、入出力を同一話者の声とする手法を考案した。出力ターゲットとなる第二言語（以下 L2、外国語）を中間情報として利用し、本人の声の音質と特徴をそのままに、本人の声で、L2 のネイティブ（母語話者）が話すように音声出力。ネイティブ話者の音声特徴を合成のための単位選択ソースとする物理的方法を用いて、音声を合成する。

2 コーパス・ベース音声合成

コーパスを用いた音声合成は現在さまざまな言語で行なわれているが、その利用は単一言語枠内に限られていた。多言語音声合成システム CHATR は、2段階処理（BTTS 法と呼ぶ）による多言語化を可能にした。これまでの多話者多言語に加え、同一話者による多言語発話を実現した。[1]。

CHATR は、単位波形接続のソースに大規模な自然発話コーパスを利用する。これらの音声単位は、ターゲットとなる韻律に近いが、接続した単位を滑らかにするためにはターゲットコストとジョイントコストの2つの値の最小値が選ばれる。単一言語における音声合成の場合、音声データベースから最適な音声単位を選択するため、合成する発話内容の音響及び韻律ターゲットを予測する必要がある。外国語を合成する場合にも、単位選択のターゲットは予測に依存していた。一般に韻律は予測できるが、細かい発音の違いなどまで区別することができなかった。

BTTS では、一旦、ターゲット言語の音声データベースを用いて音声波形を生成し、その物理的特徴を単位選択のベースとして利用する。単位選択のターゲットに L2 話者（ネイティブ）の音声波形を用いることにより、単に L2 の音響および韻律特徴を予測する場合に比べて、より希望に近い音響及び韻律特徴を実現できる。

3 話者間のマッピング

2つの言語が混在するテキスト、例えば、本稿の次の段落を音声合成で変換すれば、以下に含まれる英単語は、次の3種類の出力が考えられる。方法1：英語話者の声、方法2：同じ日本人の声（カタカナ英語）、方法3：同じ日本人の声（ネイティブに近い英語）。

同一話者による多言語音声合成の方法2では、発話にふさわしい音素を決定するため、音素をそれぞれの言語で対応させたマッピングテーブルを使用する。例えば、英語の cap と cup の母音、ramp と lamp の第一子音は異なるが、いずれの発音も日本語では区別されていない音である。

日本語話者のデータから上記の単語の合成音声を生成する場合、母音はいずれもア行の

* "Foreign-language Speech Synthesis", by Nick Campbell (ATR Interpreting Telecommunications Research Labs.)

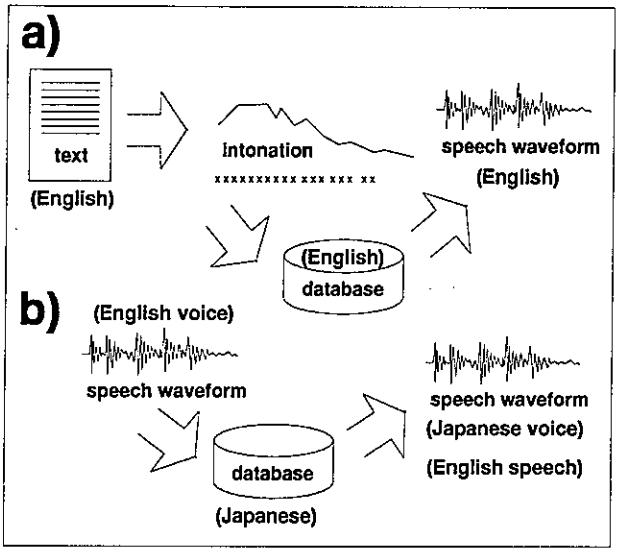

Figure 1: 2段階の音声合成

/a/、子音はラ行にマッピングされ、lamp, lump, ramp, rump の発音の区別がなくなり、英語学習初心の話し手が話すような「カタカナ英語」の発音になる。

このマッピングにおいても、日本語話者から英語にかなり近い音声を得ることができる。音素ラベル（音韻記号）のみで選択すれば、細かい区別ができないまでも、実母音空間では変差があるため、適切なターゲットさえあれば、実音声に近い音素の選択が可能になる。例えば、日本語の自然発話コーパスから /a/ や /l/ に対応する英語の近似値も選ぶことができる。この場合は日本語の音素と英語の音素のケプストン距離が有効な手段である。

4 韻律特徴

言語の違いは単に音響的なものだけではなく、韻律的な違いもある。例えば、日本語では音韻時間長の変化は少ないが、英語ではストレスによる音韻時間長の変化は大きい。

ターゲット言語で既に学習されている規則を用いて韻律の変差を予測することはできるが、異なる言語間では韻律変化によってスペクトルに影響がでる。例えば、英語における強調発話での母音のスペクトル傾斜は通常発話のものと異なる。同じ /a/ でもストレスの有無により音韻継続時間長やスペクトル包絡も異なる場合が多い [2]。同じ長さの日本語からの音を取ったとしても、それが必ずしも適切な音とは限らない。しかし、L2 話者（ネイティブ）による音声をスペクトル・ターゲットとして用いることにより、日本語データベースから最も最適な音声サンプルを選択することが可能となる。

5 ケプストラム・ターゲットによる単位選択

同じ音素グループ内での音響的特徴を区別するスペクトル・ターゲットを得るために、ターゲット言語で入力話者の音質に近いデータベースを選び、その音声データを用いてCHATRでターゲット発話を合成する。

その音声波形、つまりケプストラム情報から、より正確なスペクトル及び韻律のターゲットが定義できる。また、音声波形とそれぞれの単位候補のケプストラム距離をフレーム単位で比較することにより、L1 日本語データベースから最適な音素単位の選択が可能である。（図1）

「BTTS」は、'Bilingual TTS' ではなく、物理的ターゲット単位選択の略称である。

6 むすび

本手法は2段階の音声合成と、算出の複雑な距離関数があるため、比較的時間を要し、リアル・タイムでの実現には至っていない。しかし言語間で声の特徴が似通ったデータベースがあれば、ケプストラム・ターゲットによる単位選択により、同一話者での多言語音声合成が生成できる。

謝辞

本研究用に音声コーパスを提供して頂いた香港大学 Chorkin Chan 教授ならびにドイツキール大学 Klaus Kohler 教授に感謝の意を表したい。また中国語の音声合成に携わった Ming Yue Xie-Zhang さんにも感謝します。

References

[1] CHATR 音声合成ホームページ: http://www.itl.atr.co.jp/chatr （ATR ITL）.

[2] Stress, Loudness, and Spectral Tilt 音響学会, 3-4-3, pp 279-280 1995