日本人が聞き取りやすい音響特性を持つ英語音声補正システム

水谷淳　山田貴弘　市村哲
東京工科大学　コンピューターサイエンス学部

1. はじめに

現在，英語を映画やビジネス等，様々な場面で日本でも耳にすることが多く，英語コンテンツの入手もインターネットの普及により容易になった．しかし，一般的に日本人にとってアメリカ人等のネイティブ英語を聞き取り，理解することは困難と言われている．この困難な理由として，日本語と英語の音響的特徴の違いが挙げられる．
そこで本研究では，日本語と英語の音声の特徴の違いの中でも，1 単語中の音節数の違いに着目した．日本人が聞き取りやすい音響特性を持つ英語音声補正システムを提案する．

2. ネイティブ英語聞き取りの問題

日本語と英語の音響的特徴の違いの一つとして，1 単語中の音節数の違いがある．図 1 で示すように，英語は日本語に比べて 1 単語に対する音の区切れる数が少なくなる．これにより，聴きなれている日本語に比べて速く聞こえ，聞き取りを困難にさせる要因になっていると考えられる．

□ 単語：Subject
■ 日本語　sa - bu - je - e - cu - to (6 音節)
■ 英語　sub - ject (2 音節)

図 1.　音節数の違い

また，アクセント等のリズムの違い[1]が聞き取りを困難にしていると考えられる．

3. 提案

本稿では，1 単語中の音節数の違いを考慮した補正方法を提案する．音の移り変わる部分のみ伸長を行い，擬似的に音節数を増やし，聞き取りやすくする．具体的には周波数解析より音の区切れ位置の特定を行い，さらに区切れ位置付近で周期性を持つ波形の抽出を行った．その波形を連続して音声波形中に一定回数挿入を行った．

English voice correction system for Japanese speakers.
Atsushi Mizutani , Takahiro Yamada , Satoshi Itimura
School of Computer Science, Tokyo University of Technology

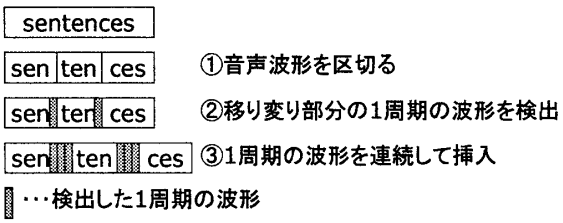

図 2.　音節数の違いを考慮した補正方法

4. 実装

4.1. システム概要

本システムの処理フローを図 3 に示す．まず，音声波形データを離散フーリエ変換し，対数スペクトルに変換後，その結果を逆フーリエ変換することでケプストラム[2]を求めた．このケプストラムからスペクトル包絡を求めた．その際のフーリエ変換条件は，フレーム長を 1024 サンプル(23ms)，フレーム間隔を 512 サンプル(11.5ms)，窓関数はハニング窓とした．

区切れ位置の特定方法として，各フレームのスペクトル包絡の 1～13000Hz までの各周波数スペクトル値の合計を求め，各フレームで隣接する 2 フレームと合計値を比較した．隣接する 2 フレームより値が低くなった場合は付近の音声より音が小さくなり，音が移り変わり部分であると考え，音が区切れた部分と判断した．

音節数の違いを考慮した補正として，区切れ位置付近の周期性を持つ波形の特定を行い，連続して挿入を行った．まず，入力した音声波形の区切れ位置を基点として AMDF 法によって周期性をもつ波形 *Tp* を特定し，*Tp* を連続して挿入した．

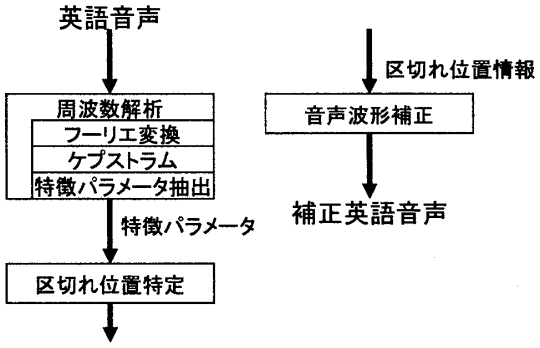

図 3.　システムの流れ

4.2. ノイズ除去

波形 *Tp* と挿入された元の波形のつながる部分の前後の差が大きくなる場合，ノイズが生じるため，差を極力少なくする補正を加えた．

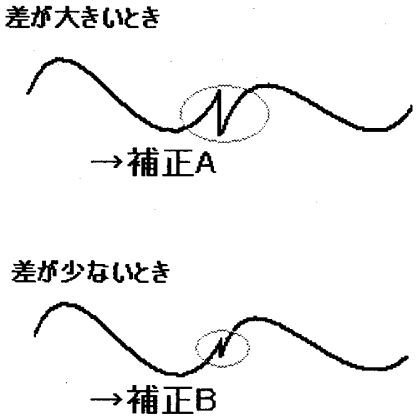

図 4. ノイズ処理条件

図から差が大きいときは補正 A，少ないときは補正 B とする．

・補正 A

波形 *Tp* と挿入された元の波形のつながる部分の値を比べ，徐々に値を近づけていく処理を行った．あらかじめ 15 サンプル用意しる．差を 15 で分割し，その値を 15 サンプルに入れ，波形 *Tp* の前後に付け加えた．

・補正 B

波形 *Tp* のつながるサンプルから，元の波形から後 1 サンプルずつ，波形 *Tp* から前 1 サンプルずつ値を比べていく．差の値が最も低くなるサンプルを基準に，それまで比べられたサンプルを削除した．

4.3. 挿入回数の可変補正

区切り位置が，ちょうど母音であるとき違和感があることを発見した．これは母音部分に補正がかかると，間延びして聞こえるように感じるためであると考えた．そのため，区切り位置が母音である場合，補正の挿入回数を少なくするように改良した．母音部分 1 フレームのスペクトラムを図 5 に示す．

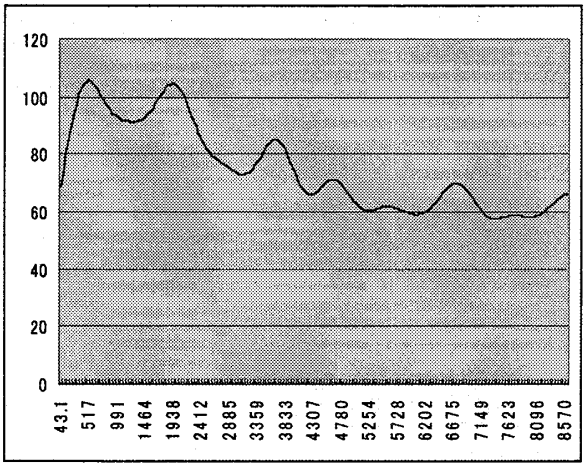

図 5. 母音部分のフレームのスペクトラム値

この図のように，第 1 フォルマント，第 2 フォルマント，場合によっては第 3 フォルマントが大きく間隔を空けずに平均 85 以上の値を示していることが確認できた．また，大きな 1 つ目の山があった先には，10 以上の差をつけた谷ができていることが見て取れた．

この結果から，対数スペクトル値を出す処理時に，今回発見した条件に満たすフレームを記録しておき，区切り位置処理時にフレームが一致するかを調べた．この条件を付け加え，条件一致した場合には通常 0.06s の最大挿入時間を 0.04s となるようにした．

5. まとめ

今回，実装したシステムの評価から元音声よりも補正音声のほうが聞き取りやすいという結果が得られた．また，ノイズ削除，挿入回数の可変補正がない音声と補正のある音声を比べた結果，ノイズに関しては大きく削減し，可変補正も効果が得られた．

6. 参考文献

[1] 清水克正：英語音声学 理論と学習，勁草書房，1995

[2] 鹿野，伊藤，河原，武田，山本：音声認識システム，オーム社，1995