社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DPS－132 (28)
2007－GN－ 65 (28)
2007－EIP－ 37 (28)
2007／9／21

異文化コミュニケーションのための日本人に聞き取りやすい英語音声の研究

山田 貴弘† 水谷 淳‡ 市村 哲‡
†東京工科大学 バイオ・情報メディア研究科 ‡東京工科大学 コンピュータサイエンス学部

あらまし 本論文では、異文化コミュニケーションのために日本語と英語の音声の特徴の違い(音節数、リズム)に着目し、英語音声を日本人に聞き取りやすく補正する方法について提案する。また、音声の補正を行う上で音声を区切る必要がある。そこで、音が移り変わる区切れ位置の特定方法についても提案する。これら音声補正方法、区切れ位置特定方法を用いて、自動的に音声波形の補正を行うシステムを実装し、評価を行った。

Creating English voice intelligible for the Japanese for the cross cultural communication

Takahiro YAMADA† Atsushi MIZUTANI‡ Satoshi ICHIMURA‡

† Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology
‡ School of Computer Science, Tokyo University of Technology

Abstract In this research, we propose the method that creates English voice intelligible for Japanese people. For this purpose, we focused the difference (in the number of syllables, and a rhythm) of Japanese and English sound. We also propose a method of finding the segmentation position where the sound changes. The system was constructed, and evaluated.

1. はじめに

現在、英語人口はおよそ 10 億人であり、国際語としての性格がとても強い。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。

聞き取り能力が十分でない状態で異文化コミュニケーションを行おうとする場合、英日音声翻訳などの支援ツールが助けとなる。国際電気通信基礎技術研究所(ATR)では多言語音声翻訳システム ATR-MATRIX[1]等の研究開発が行われ、その後も多言語音声翻訳システムの研究開発が行われている。しかし、翻訳システムを使用する場合、英日方向の話し言葉翻訳精度は、相手に自分の言ったことが最低限伝わるレベルで約 55%程度[2]と必ずしも高いとはいえない。

また、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、1 単語における音の区切り数やアクセント方法、リズムのとり方等、複数の音響的特徴の違いが存在する。これら違いを克服して、日本人が英語聞き取りを上達するためには、長時間の聞き取り訓練や繰り返しの訓練が必要になってしまう。

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とリズムの違い[3]に着目し、周波数解析により英語音声の分割を行う。分割した音声に対して日本語と英語の特徴を考慮した補正を行うことで日本人にとって聞き取りやすい英語への補正法を提案する。

2. 日本人の英語聞き取りにおける問題

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。音節とは1個の母音を音節主音とし、その母音単独、あるいはその母音の前後に1個または複数個の子音を伴って構成する音声である。発話される音声の区切りの数は音節によって区切られるといわれている。

日本語ではほとんどの音節が(子音+母音)によって構成されているが、英語では(子音+母音+子音)、(子音+母音+子音+子音)等、一つの音節を構成する際の音の数が増えてしまう。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、言葉を発する際のリズムも音響的特徴の違いがある。日本語はモーラ型リズムであるのに対し、英語は強勢型リズムである。モーラは「拍」とも呼ばれ、母音または母音+子音から成り立つ。図2で示すようにモーラ型リズムでは、このモーラ一つ一つがそれぞれ等しい長さで発音されるリズムである。これに対し、強勢型リズムは文章中に強勢(アクセント)が現れる間隔が等しく発音されるリズムである。このリズムの違いが聞き取りを困難にしていると考えられる。

■日本語
こんにちは/ko - n - ni - chi - wa/ (6モーラ)
■英語
John saw a　black bird　yesterday

図2　リズムの違い

3. 提案

本研究では、1単語中の音節数およびリズムの取り方の特徴の違いに着目した、英語音声の補正方法について提案をする。

英語音声に対して周波数解析を行い、音が移り変わる区切れ位置の特定を行なう。そして、区切れ位置によって分割を行った音声波形に対して日本語と英語の特徴の違いを考慮した補正を行う。これにより、日本人にとって聞取りやすい英語音声を作成する。

3.1. 音節数の違いを考慮した音声補正方法

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(図3)。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う。その波形を連続して音声波形中に一定回数挿入を行う。音の移り変わる部分のみ伸長を行うため、分割された各音声を崩さずに音声をゆっくりにし、聞き取りやすくする。

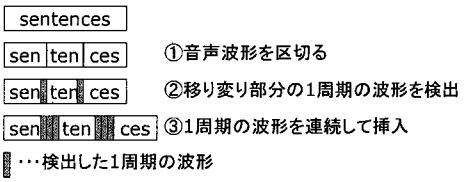

図3　音節数の違いを考慮した補正方法

3.2. リズムの違いを考慮した音声補正方法

リズムの違いを考慮した補正方法を提案する(図4)。日本語のリズムは各モーラの時間間隔が等しいとされている事から、音の区切れ位置の特定後、分割を行った各音声波形を本論文では音声セグメントと定義する。そして、最も長いとされる音声セグメントに近い長さになるよう他のセグメントのタイムストレッチを行う。タイムストレッチとは、音声波形の音の高さを変えずに時間を伸縮する事が手法であり、これを用いて音声を伸長すると聞き取りやすくなると言われている。

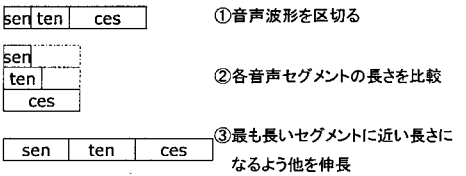

図4　リズムの違いを考慮した補正方法

3.3. 音の区切れ位置の特定方法の提案と検証

音の区切れ位置の特定方法のとして、フォルマント周波数と最大スペクトルの周波数による音の区切れ位置特定方法を提案および検証を行った。

この方法では、表1の音声波形データに対し、表2の条件でフーリエ変換を行い、各フレームの第1～3フォルマント周波数および、最大スペクトル成分を持つ周波数をスペクトル包絡から求める。フォ

ルマント周波数とは、母音の周波数スペクトルに現れる複数のピークであり、低周波数から順に第 1、第 2･･･と数字が割り当てられ、母音の識別を行う上で重要な要素といわれている。隣接するフレーム同士を以下の条件で比較を行い、I～III のいずれも満たさない場合を区切れ位置と判断し、特定を行った。

I. 最大周波数成分が 1000Hz 以上のフレームが連続している場合、子音部分とする。
II. 第 1 フォルマント周波数のスペクトルが一定値以下のフレームが連続する場合、無音や極端に音が小さい部分とする、
III. 隣り合うフレームの第 1~3 フォルマント周波数のいずれかが等しいフレームが連続する場合、母音部分とする。

表 1　音声波形データ

ファイル形式	wav
サンプリング周波数	44100Hz
チャンネル数	モノラル
ビット数	16bit
収録内容	英語音声のみ

表 2　フーリエ変換条件

フレーム長	1024 サンプル (23ms)
フレーム間隔	512 サンプル　(11.5ms)
窓関数	ハニング窓

提案した方法により、音の区切れ位置が特定できるか実験を行った。その結果、条件 I により子音に関しては区別することができたが、1000Hz に最大周波数成分を持つ子音部分を区切ることができない事が判明した。また、条件 III では、各フレームの周波数解析結果の誤差により、不必要な部分で音が区切られてしまい、誤差を許すと区切るべき部分で区切れないという結果が得られた。

以上から、適切に区切れ位置の特定ができない事が判明した。

4. 区切れ位置特定方法の改善

新たに周波数スペクトル合計値による音の区切れ位置特定の方法を提案する。そして、効果について検証を行った。

4.1. 各周波数スペクトル合計値による特定方法

この方法では、4.1 と同様の条件でフーリエ変換を行い、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する。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する。

合計 n 個の各フレームのスペクトル合計値を $S_{1\cdots n}$ とし、不必要なピーク部分の検出を避けるため、この各フレームのスペクトル合計値を以下の式により平均化を行う(図 5)。

$$SA_n=(S_{n-1}+S_n+S_{n+1})/3$$

平均化後、$SA_{n-1}>SA_n<SA_{n+1}$ となるようなフレームの特定を行い、音が区切れる部分と判断する。

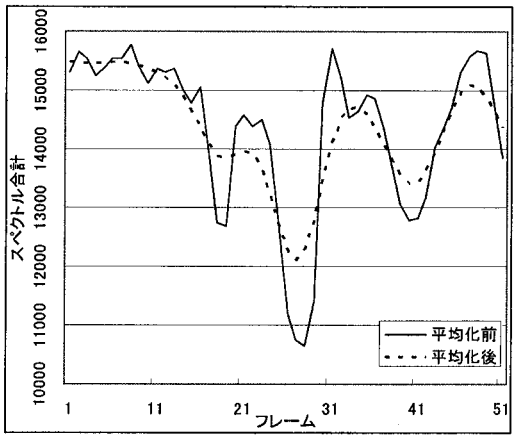

図 5　各フレームのスペクトラム合計値

4.2. 各周波数スペクトル合計値による特定結果

3.3.での検証に用いた音声波形データを用いて区切り位置を特定した結果を図 6 に示す。図 6 中 POS1 が 3.3.での結果であり、POS2 が 4.1.の方法での結果である。

子音部分、無音声部分のみ区間を区切ることは出来なかった。しかし、日本語の特徴として、子音+母音で区切られる場合が主なため、子音部分のみを区切れないことの影響は大きくないと考えられる。また、POS1 に比べ不用意に区切れる位置が少ない事から、本研究では 4.1.の方法で音の区切れ位置を特定することに決定した。

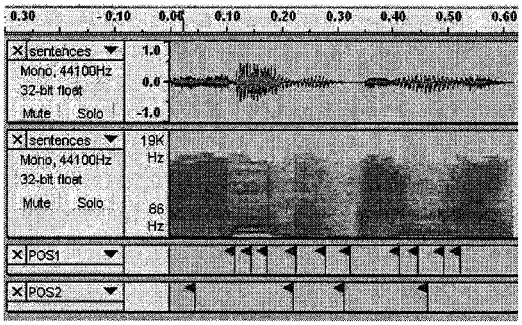

図 6　区切り位置特定結果

5. システム実装

本システムの処理フローを図 7 に示す。本システムは大きく分けて、(1)フーリエ変換による周波数解析パート、(2)解析結果を用いた区切れ位置特定パート、(3)音声波形補正パートの 3 つに分けられる。次項では、各パートでの処理の詳細を述べる。

実装画面を図 8 に示す。「開く」ボタンから音声ファイルを入力後、周波数解析が行われ各フレーム番号、時間、周波数スペクトル合計値がフレームデータに表示される。次に、「解析」ボタンを押すことで、音の区切り位置特定を自動的に行い、フレーム番号、区切り位置の時間が区切り位置特定結果に表示される。そして、「補正」ボタンを押すことでそれぞれの方法で補正を行った音声ファイルが出力される。

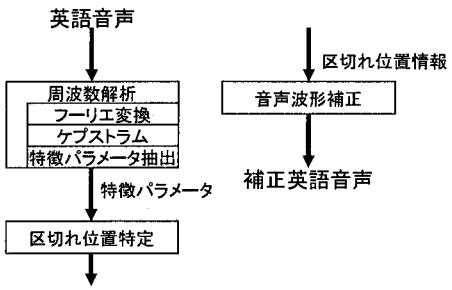

図 7　処理フロー

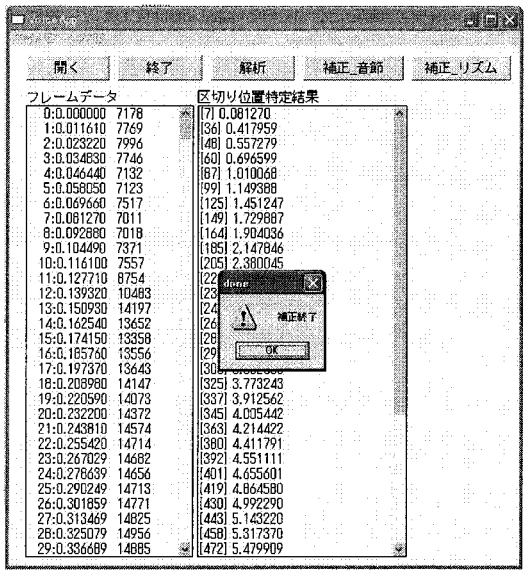

図 8　システム画面

5.1. 周波数解析

表 1 の音声波形データに対し、表 2 の条件で音声波形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[4]を求める。このケプストラムから 60 次以上の高ケフレンシー部分の除去し、再度フーリエ変換を行うことで各フレームのスペクトル包絡を求める。

5.2. 音の区切れ位置特定

5.1 にて得られた各フレームの特徴パラメータを用いて区切れ位置の特定を行う。特定方法として、4.2 の結果から各周波数スペクトル合計値による方法を用いた。

5.3. 音声波形補正

特定した区切れ位置情報を元に、音声波形の伸長を行う。両特徴の違いとも、AMDF(Average Magnitude Difference Function)を用いて周期性を持つ波形の抽出を行う。AMDF は自己相関関数の乗算部分を減算に置き換えることで、演算処理を通常の自己相関に比べ高速に行える手法である。また、リズムの違いを考慮した補正では PICOLA(Pointer Interval Controlled OverLap and Add)アルゴリズム[5]を用いてタイムストレッチを行う。PICOLA アルゴリズムは、音声波形中に同じような波形が繰り返し現れる概周期性を利用し、同じような波形を伸

長率に応じた間隔で挿入することにより時間の伸長を行うアルゴリズムである。

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、連続して挿入を行う。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* を連続して挿入する。この際に、波形 *Tp* を挿入する回数を以下の式から求める。

挿入回数＝最大波形挿入時間／波形 *Tp* の時間

リズムの違いを考慮した補正方法として、区切り位置に基づいて分割された音声セグメントの長さを比較し、各音声のセグメントの長さに応じたタイムストレッチを行う。各音声セグメントの中で最も時間が長い音声波形を探し、その音声セグメント長と他の各音声セグメント長との比率 *r* を求める。その比率 *r* の値により、PICOLA アルゴリズムで用いられる波形を伸長する割合を求める。

r ＝最長音声セグメント長／各音声セグメント長

6. 評価実験

補正効果の検証として被験者 10 名に対し、単語の聞き取りやすさに関しての評価実験を行った。評価は 1:聞き取りにくい、2:少し聞き取りにくい、3:どちらともいえない、4:少し聞取りやすい、5:聞取りやすい、の 5 段階とした。

6.1. 同一音声による聞き取り実験

実験 1 として、同一の音声を異なる補正条件で補正を行い、聞き比べを行った。音声を表 3 の条件で補正を行い、5 つの音声補正データを作成した。元音声と補正音声 a～e を聞き比べ、下記項目についての 5 段階評価および、a～e の音声で最も聞取りやすい音声ファイルの選択を行った。また、比較対象として、タイムストレッチによって、音声波形全体を伸長した音声ファイルを用意した。

表 3　実験 1 音声補正条件

条件	補正方法	補正後の時間長(sec)
元音声	—	(3.55)
条件 1	音節を考慮した補正 (最大挿入時間 0.6sec)	4.43
条件 2	リズムを考慮した補正 (r>1.5:伸長率 1.3 , r<=1.5:伸長率=1.1)	4.4
条件 3	リズムを考慮した補正 (r>1.5:伸長率=1.4 , r<=1.5:伸長率=1.1)	4.86
条件 4	音声波形全体をタイムストレッチ 伸長率 1.20	4.48
条件 5	音声波形全体をタイムストレッチ 伸長率 1.40	4.97

結果を図 9 に示す。条件 2 が 4.33 と最も聞取りやすいという評価が得られ、本研究で提案したリズムを考慮した手法が波形全体をタイムストレッチする手法より有効的である結果が得られた。また、音声の時間が他に比べて長い条件 3,5 と 4 にそれ程差が無いことから、一定以上音声を長くしても、聞き取りやすさは向上しないのではないかと考えられる。

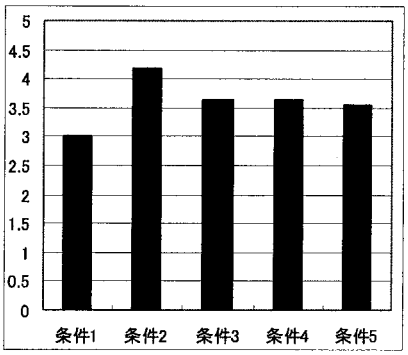

図 9　同一音声による聞き取り実験結果

6.2. 複数音声による聞き取り実験

実験 2 として、全て異なる内容の音声を試聴し、その音声が聞き取りやすいか検証を行った。評価用音声としてリスニング問題(音声 a)、TOEIC のリスニング問題サンプル(音声 b)、アメリカ報道メディアの podcast ニュース(音声 c)の 3 種類を用意し、各音声に対して表 4 の補正を行った。また、各条件でも、音声内容は異なっている。

表 4　実験 2 音声補正条件

条件	補正方法
条件 1	補正なし
条件 2	リズムを考慮した補正 (r>1.5:rate=1.3 , r<=1.5:rate=1.1)
条件 3	音声波形全体をタイムストレッチ 伸長率 1.2
条件 4	音節を考慮した補正 (最大挿入時間 0.6sec)

実験結果を図 10 に示す。全ての条件ではないが、補正なしである条件 1 に比べ聞き取りやすくなる結果が得られた。特に、音節数を考慮した補正である条件 4 が音声 a で特に効果が顕著に見られた。音声 c では補正なしの条件 1 に比べると聞き取りやすくはなったが、音声 a に比べ効果が見られず、評価が低くなるという結果が得られた。この原因として、音声 c は他 2 音声に比べて時間当たりの単語数が多く発話スピードが速い事が考えられる。音節数を考慮した補正の場合、音が移り変わる区切り部分の波形伸長は行っているが、各音声セグメントにタイムストレッチを行っていない。このため、一つ一つの音声セグメントが速いまま聞こえてしまい、効果が少なかったのだと考えられる。

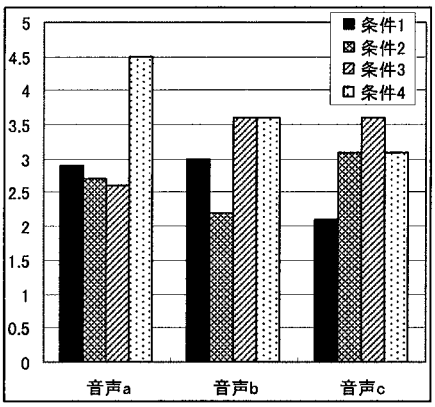

図 10　複数音声による聞き取り実験結果

7. まとめ

日本語と英語の発話の特徴の違いによる、英語音声の補正手法および、英語音声の音の区切れ位置の特定手法を提案し、その有効性を検証した。その結果、ある程度の聞き取りやすさの向上はみられたが、音声波形全体をタイムストレッチする方法と同等か若干下回る結果になった。これは、被験者の英語の聞き取り能力を考慮に入れていなかった為、一部分や、音声セグメント単位で補正をする事で音声に不自然性が現れ、逆に聞き取りにくくなった恐れがある。また、6.2 の実験結果について、文章内容が評価に影響を与えてしまう可能性が大きいと考えられるため、文章内容を踏まえた上でさらに評価を行う必要があると考えられる。

今後の課題として、手法の再検討および今回提案した 2 手法組み合わせた補正手法の提案および評価を行っていきたい。また、時間軸の伸長による補正以外にも、日本語の高低アクセントと英語の強勢アクセントに代表されるような、音声の振幅や周波数成分に関わる特徴の違いを取入れた補正を検討していきたい。

参考文献

[1] 菅谷史昭，竹澤寿幸，隅田英一郎，匂坂芳典，山本誠一：音声翻訳システム：ATR−MATRIX の開発と評価，情報処理学会論文誌，Vol. 43，No.7，pp.2230-2241(2002.7)

[2] 古瀬蔵，美馬秀樹，山本和英，Michael Paul，飯田仁：多言語話し言葉翻訳に関する変換主導翻訳システムの評価，言語処理学会第 3 回年次大会，pp.39-42 (1997.3)

[3] 清水克正：英語音声学 理論と学習，勁草書房，1995

[4] 鹿野，伊藤，河原，武田，山本：音声認識システム，オーム社，1995

[5]PICOLA and TDHS：
http://keizai.yokkaichi-u.ac.jp/%7Eikeda/research/picola.html