2008--GN-67 (9)
2008/3/21

時間伸長処理と振幅平準化を加えた日本人に聞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 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。筆者らは、これまでに音節数、リズムの違いに着目をした時間伸長処理による補正方法の提案を行ったが、評価の結果、補正効果が安定していないことがわかった。そこで本論文では、音節数とリズムの違いに着目した補正方法の再検討および、アクセントの違いによる振幅平準化を提案する。再度評価を行った結果、リズムの違いによる補正に十分な聞き取りやすさ向上の効果が得られた。

Research of English voice with intelligible sound feature for the Japanese based on time stretch and balancing voice wave form

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
We previously proposed the method that created 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. The system was constructed, and evaluated. However, the result of the evaluation was insufficient. The reason was produced noise etc. In this research, we propose an improved the correction method. Also we applied the correction based on balancing voice wave form, and we evaluated it again. As a result, English voice modeled on a Japanese rhythm was the best.

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[3]。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。

筆者らはこれまでに日本人にとって聞き取りやすい英語音声を作成することを目的とし、日本語と英語の特徴の違いを考慮した補正方法を提案してきた。特徴の違いとして 1 単語中の音節数の違い、リズムの違いに着目をした音声補正システムを実装し、評価を行った結果、一定の効果は見られたが安定した結果が得られないことがわかった。

そこで本研究では、これまでに得られた問題点を踏まえた2補正方法の改良方法および、新たにアクセント方法の違いに着目をした補正方法を提案する。また、それぞれの補正方法を組み合わせ、再度評価を行った。

2. 日本人の英語聞き取りにおける問題

これまでに着目した日本語と英語の音響的特徴の違いの一つ目として、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

二つ目に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

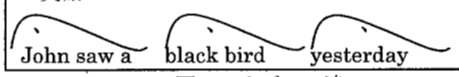

図2　リズムの違い

また、今回新たに着目した特徴の違いとして、アクセント方法の違いがある。図3、4は日本語と英語の音声波形である。日本語の場合、高低アクセントのため強勢部分が存在せず、各単語の音声の振幅の大きさはあまり差がない。しかし、英語音声はアクセント方法が強勢アクセントであるため、日本語に比べ音声の振幅の差が大きい部分が存在する。また、「popular」のように1単語内でもアクセント位置で振幅の大きさが違うことがわかる。

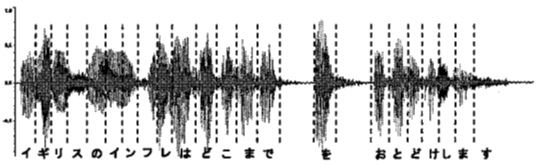

図3　日本語の音声波形

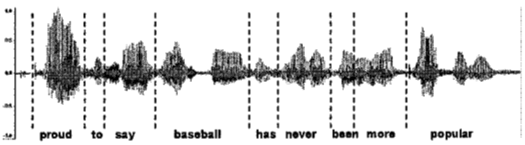

図4　英語の音声波形

3. 音節数の違いを考慮した補正方法とその改良

3.1. 提案補正方法

1 単語中の音節数の違いを考慮した英語音声の補正方法として、一部分のみを時間伸長する方法を提案する。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時間を行う。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。

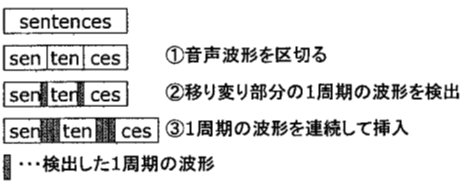

図5　音節数の違いを考慮した補正方法

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、任意で与える波形挿入時間(0.06sec)と波形

Tp の時間に応じて挿入する回数を以下の式から求める。

挿入回数＝波形挿入時間／波形 *Tp* の時間

3.2. 問題点と改良

これまでの評価結果より、ノイズが気になるという意見があった。ノイズが発生した理由として、連続して挿入する波形の両端のデータ値に差があるため、2パターンのノイズが作られていることがわかった(図 6)。そこで、各パターンに応じたノイズ低減処理を加えた。

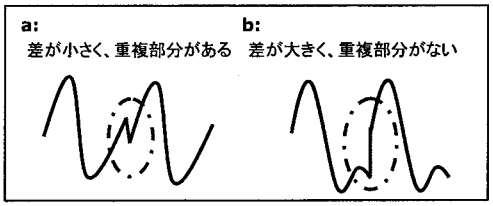

図 6　補正によるノイズの生成パターン

・パターン a

補正 a では挿入した波形と挿入された元の波形のつながる部分の前後の値を比べ、差が少ない部分を見つけ、その部分から重複する部分を間引く。挿入する波形のつながる波形部分から値を比べる際、元の波形から後 1 サンプルずつ、挿入波形から前 1 サンプルずつ比べていく。差の値が最も低くなる場所を基準に、それまで比べられた範囲を削除する。

・パターン b

パターン b では挿入した波形と挿入された元の波形のつながる部分の値を比べ、差が大きいために余計な波形部分を間引く事が困難である。そこで、両端の値を徐々に近づけていく処理を行うようにした。あらかじめ 15 サンプル分の追加エリアを用意し、差を 15 で分割し、その値を追加エリアに入れ、挿入波形の前後に付け加える。

また、音の区切れ位置が母音である場合に違和感があるという意見があった。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られる。そのため、区切れ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ル包絡を図 7 に示す。縦軸はデシベル[dB]、横軸は周波数[Hz]である。

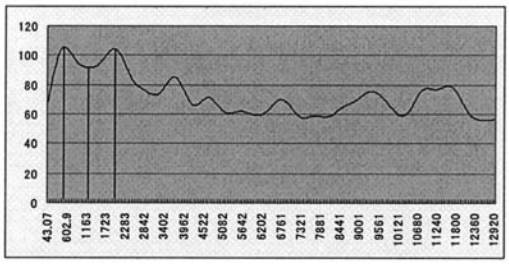

図 7　母音部分フレームのスペクトル包絡

母音部分では、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 フォルマントが連続するフレーム間で平均 85dB 以上を示していることが確認できた。また、第 1 フォルマント周波数と第 2 フォルマント周波数の間には、第 1 フォルマント周波数に対して 10dB 以上低くなる谷部分ができていることが見て取れた。そこで、区切れ位置の周波数スペクトルが以下の条件の場合、波形挿入時間を 0.06sec から 0.04sec に変更を行った。

・215～860,1500～2150Hz に 85db 以上の最大値を持つ
・2 つの最大値の間で 10dB 以上値が下がる

4.　リズムの違いを考慮した音声補正方法

4.1. 提案補正方法

リズムの違いを考慮した補正方法として、時間伸長する割合を可変にし、各音の長さを出来るだけ近づける方法を提案する。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、最長音声セグメントと各音声セグメントの比率に応じた伸長率により、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。

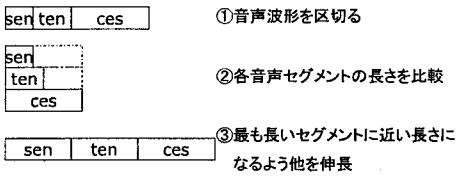

図 8　リズムの違いを考慮した補正方法

区切れ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 の値が 1.5 以上の場合と 1.5 以下の場合の 2 パターンの伸長率で、PICOLA アルゴリズム[4]を用いて伸長を行う。

r =最長音声セグメント長／各音声セグメント長

4.2. 問題点と改良

これまでは最長音声セグメントとの比率に応じた 2 パターンの伸長率でタイムストレッチを行っていた。これは、比率 r をそのまま伸長率にすると、比率が大きい場合、補正することで極端に間延びした音声になるのを防ぐためである。しかし、図 9 のように、最長音声セグメントが極端に長い場合、殆どの比率が 1.5 以上になってしまい、波形全体を一定の伸長率で伸長するのと変わらなくなってしまう。また、比率 r をそのまま伸長率にした場合、非常に間延びした音声が生成されてしまう。

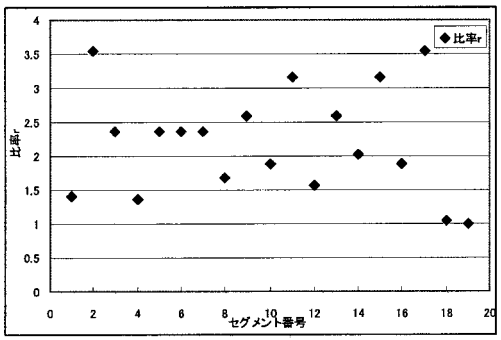

図 9　各音声セグメントの比率 r

そこで、平均長による伸長率の算出に変更した。まず、各音声セグメントの平均長を求める。その平均長以上の音声セグメントで更に平均長を求め、基準長とする。図 10 は図 9 と同様の音声の各音声セグメントの長さと、平均長、基準長を示したものである。極端に長くなっている最長音声セグメントに比べ、基準長は比較的長い音声セグメントの中でばらつきが少ない長さとなっていることがわかる。この 2 つの長さと比較し、基準長以上の場合は伸長率を 1.1、平均長以上の場合は伸長率を 1.2、平均長以下の場合は伸長率=基準長／各セグメント長とした。

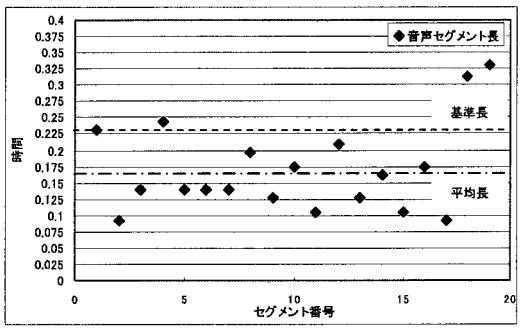

図 10　各音声セグメントの時間長

5. アクセントの違いによる振幅平準化

英語は日本語に比べて各音声の振幅の大きさに差がでている事に着目をし、振幅の小さい部分を増幅し、日本語のように各音声の大きさの差を減らす振幅平準化処理を加えた。

振幅平準化の処理は以下のようになる。まず、各音声セグメントの中で AMDF 法により自己相関を行い、音声波形を 1 周期毎に区切り、それぞれ 1 周期の波形の中で振幅の最大値を求める。音声セグメントの中で、各 1 周期の振幅の最大値を比較し、音声セグメント全体の振幅の最大値を求める。そして、各 1 周期の最大振幅値が一定値以上の場合のみ、音声セグメント全体の最大振幅値との比率に応じて波形を増幅する。また、前後どちらかの 1 周期が増幅しない場合は極端な波形の変化を防ぐため、比率を半分の値で増幅する。

6. 補正の組み合わせ

3 つの補正方法を組み合わせた音声の作成を行った。時間伸長による補正(1 単語の音節数の違い、リズムの違い)と振幅平準化の組み合わせでは、振幅平準化による各周波数スペクトルの変化により、音の区切れ位置の変化を避けるため、時間伸長による補正を行った後に振幅平準化を行った。

時間伸長による補正同士の組み合わせの場合では、リズムの違いを考慮した補正を行った後、1 単語の音節数の違いによる補正を行った。これは、音節数の違いによる補正では音の区切れ位置部分の波形を直接伸長することから、この処理後に音の区切れ位置が変化する可能性を防ぐためである。

7. システム概要

本システムの処理フローを図 11 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。

周波数解析パートでは、表 1 の音声波形データに対し、表 2 の条件で音声波形の離散フーリエ変換を行う。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[5]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。この結果を元に、音の区切れ位置の特定を行う。特定方法として、各周波数スペクトル合計値を求め、隣り合うフレームよりも値が小さくなるフレームを区切れ位置とした。その後、それぞれの補正方法に基づいて自動的に音声補正を行い、補正音声を出力する。

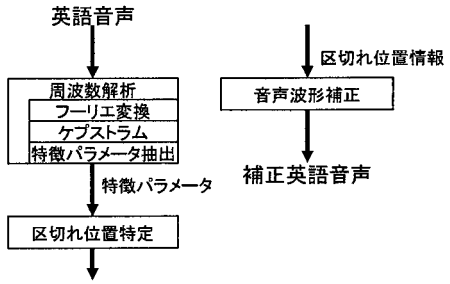

図 11　処理フロー

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	32bit
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

8. 評価実験

被験者 10 名に対し、5 種類 8 条件の音声を聞いてもらい、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～5:聞取りやすい、の 5 段階とした。結果を表 4 に示す。下線の引かれている補正が各音声で最高得点の音声である。

表 3　評価実験音声

条件	補正方法	(略称)
-	元音声	n
1	音節	s
2	音節、振幅平準化	s_a
3	リズム	r
4	リズム、振幅平準化	r_a
5	音節、リズム	s_r
6	音節、リズム、振幅平準化	s_r_a
7	タイムストレッチ	ts

表 4　評価実験結果

音声	n	s	s_a	r	r_a	s_r	s_r_a	ts
a	3.4	2.8	2	3.9	2.6	3.5	3.1	3.1
b	2.4	2.8	2	3.7	3	2.9	3.7	3.2
c	2.4	3	1.8	3.5	2.6	3.9	2.5	2.4
d	2.8	2.6	2.6	4.1	3.3	3.3	2.4	3.2
e	3.1	3.1	2.3	3.7	3.6	2.7	2.4	3.1
平均	2.8	2.8	2.1	3.8	3.2	3.1	2.8	3

評価実験の結果、リズムの違いを考慮した補正である補正 *r* が、全ての音声において比較対象であるタイムストレッチによる補正 *ts* より評価が高くなり、音声 c を除いて最も効果の高い補正方法という結果になった。また、全種類の合計での評価点も 3.8 と最も高い結果となり、次いで補正 *r_a*、補正 *s_r_a*、補正 *ts* という順になった。

この評価結果から元音声と各補正、タイムストレッチと上位 3 補正の評価得点に有意差があるか、ウィルコクソンの符号順位和検定により検定を行った。使用するデータはそれぞれの補正条件 5 種類の音声×10 名の 50 データにより算出を行った結果を表 5、表 6 に示す。有意水準 α を 0.05 とすると、元音声の評価に対して補正 *r* と補正 *r_a* は、補正の効果の有意差があるといえ、補正 *ts* に対して補正 *r* は有意差があるといえることがわかった。

表 5　元音声との検定結果(p 値)

	補正 r	補正 r_a	補正 s_r	補正 ts
p 値	0.00001	0.03702	0.22203	0.25801

表 6　タイムストレッチとの検定結果(p 値)

	補正 r	補正 r_a	補正 s_r
p 値	0.00001	0.19667	0.82703

以上の結果から、本提案方法であるリズムの違いを考慮した補正方法と、リズムの違いを考慮した補正と振幅平準化を組み合わせた補正方法が元音声より聞き取りやすさが向上するという結果が得られた。

音節の違いを考慮した補正である補正 *s* は元音声より評価が低くなる場合や、タイムストレッチより高くなる場合など、安定した結果が得られなかった。音声補正を組み合わせた場合の結果では、振幅平準化を行うと殆どの音声が平準化を行わない場合に比べて評価が下がった。特に補正 *s_a* は全種類で最低の結果となってしまい、元音声より聞き取りにくくなってしまった。これは、振幅平準化時にもノイズが含まれてしまうためと考えられる。補正 *s_a* が最低になってしまったのは音節の違いを考慮した補正でもノイズ低減が不完全であり、より一層ノイズが目立ってしまっての結果だと考えられる。しかし、音声 b では補正 *s_r_a* が補正 *r* と同点で最も聞取りやすい結果になったことから、全く効果がないわけではないと考えられる。

次に、リズムの違いを考慮した補正と音節の違いを考慮した補正を組み合わせた場合である補正 *s_r* が補正 *r* より評価点が低くなってしまったのは、音節の違いによる補正の際にノイズ低減がまだ十分にできていなかったからと考えられる。しかし、表 4 の結果より、元音声より聞き取りやすくなっている結果が得られた。また、音声 c では補正 *s* が元音声に比べ大きく評価点が高く、その場合に補正 *r_s* も評価点が補正 *r* を上回っていることから、1 単語中の音節数の違いによる補正の効果を向上させることで、リズムの違いを考慮した補正の効果を上回るのではないかと考えられる。

9. 今後の課題

評価実験の結果から、音節の違いによる補正と振幅平準化は改良の余地があると考えられる。両補正共に必要な改良として、ノイズの低減が挙げられる。補正を行った音声の区切れ位置の周波数成分を見てみると、全周波数帯においてスペクトル値が高くなっていることがわかった。今回のノイズ低減の方法は、音声波形そのものに手を加えていたが、周波数帯で手を加えるなどして、補正を加えた部分に限定したノイズ低減方法を考え、聞き取りやすさの向上につなげたい。

10. まとめ

筆者らは、これまでに日本語と英語の発話特徴の違いの中でも、1 単語における音節数の違い、発話リズムの違いに着目し、それぞれの違いに応じた補正方法によって英語音声を補正し、日本人にとって聞き取りやすくなる英語音声の作成法を提案してきた。しかし、ある程度の聞き取りやすさの向上に繋がったが、比較対象の補正方法であるタイムストレッチの音声と同程度の効果しか得られなかった。

そこで本研究では、補正方法の改良、振幅平準化による補正を加え、再度評価を行った結果、リズムの違いを考慮した補正が元音声、タイムストレッチによる補正より聞き取りやすさが向上し、十分な効果が得られた。また、リズムの違いを考慮した補正後、振幅平準化を行った音声も元の音声より聞き取りやすさが向上した。

今後は、音節数の違いによる補正と振幅平準化のアルゴリズムの改良を加え、各補正の組み合わせによる補正の聞き取りやすさ向上と、処理の高速化によるリアルタイムへの対応をしていきたい。

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

[5] 鹿野,伊藤,河原,武田,山本：音声認識システム,オーム社,1995