大規模録音音声データーベースにおける ピッチ正規化の検討*

村上仁一 （鳥取大学工学部） 水澤紀子（NTT 情報通信研究所） 鈴木博和 （NTT アドバンステクノロジ）

1 まえがき

大量の音声を録音した場合、同じナレータでも発声する日時によって声の高さ (pitch) に違いがでる。しかし、音声ガイダンスを録音編集方式で作成した場合、録音された音声を組み合わせて音声を作るため、各単語の声の高さの違いは品質の劣化になる。そこで各々の録音された音声の声の高さを自動的に一定にする pitch 正規化アルゴリズムが必要になる。

本論文では、録音データベースにおいて pitch を一定にするアルゴリズムを考案し、基礎的な実験を行なった。この結果について報告する。

2 pitch 正規化方法

2.1 基本的な考え方

録音された単語の声の高さは、基本的に bitch 周波数に依存すると思われる。従って、声の高さを揃えた録音音声データベースを作成するには、各単語の平均 pitch 周波数を一定にすれば良いと考えられる。しかし精度の高い pitch 周波数を計算することは困難である。また発話内容によって平均 pitch 周波数が異なる可能性がある。

一方、録音データベースを作成する場合、同一のナレータでも発声する日時によって声の高さが異なる。しかし同一時間帯に録音された音声は、同じ声の高さであることが多い。

そこで、同一時間帯に発声された N 個のデータを 1 グループとし、グループごとにサンプリング周波数を変えて平均 pitch 周波数を一定にすることで、声の高さが揃った録音音声データベースを作成できると考えた。

2.2 pitch 分布

始めに pitch 周波数の分布を調べた。同一時間に録音されたナレータ（女性）の 400 単語を esps[1] を用いて pitch 周波数を計算した。図 1 に結果を示す。縦軸は頻度で横軸はピッチ周波数である。この図から、pitch の分布は 3 つのピークを持っていることがわかる。

図 1: pitch の分布

*“ Pitch Normarization for Speech Data Base ” by Jin’ichi Murakami (Tottori University) , Noriko Mizusawa (NTT Information and Communication System Lab.) and Hirokazu Suzuki (NTT Advanced Technology)

3　聴覚実験

3.1　実験データ

PB 電話番号案内実験サービス [2] は、ユーザーが音声ガイダンスを聞きながら PB ボタンを押すことにより電話番号を案内するシステムである。このシステムのために 3 月かけて 6 人のナレータを使用して全国の住所 20 万件を録音した (サンプリング周波数 16KHz)。この中からナレータごとに収録時間が同じ 400 単語を 1 グループとして、収録時間が異なる 3 グループを作成した。

3.2　実験方法

実験は以下のように行なった。

1. 始めに第 1 から第 3 グループの音声を ESPS を用いて平均 pitch 周波数を計算する。

2. 次に、試験者が、第 2 グループのサンプリング周波数を 50Hz ごとに変化させて、第 1 グループと同じ声の高さになると思われる再生周波数を聴覚的に捜す。

3. 同様なことを第 3 グループでも行なう。

試験者は 5 人で行なった。

3.3　実験結果

表 1 に実験の結果を示す。表中、bitch の欄はグループ 1 との平均ピッチの比を、試験者の欄はグループ 1 のサンプリング周波数と再生周波数との比を示している。

この結果から、平均ピッチの比と聴覚による再生周波数の比が比例関係にあることがわかる。しかしこの比の差はナレータによって異なる。例えばナレータ A は平均ピッチと再生周波数との比が同一であるのに対し、ナレータ E は約 2 倍ある。また試験者によって再生周波数の差が大きいこともわかる。

表 1: 実験結果

ナレータ	グループ	pitch	試験者 A	試験者 B	試験者 C	試験者 D	試験者 E
A	2	1.037	1.026	1.049	1	1.049	1.032
A	3	1.032	1	1.032	1.032	1.032	1.024
B	2	1.189	1.067	1.103	1.067	1.103	1.143
B	3	1.359	1.103	1.122	1.186	1.143	1.143
C	2	1.213	1.067	1.067	1	1.049	1.142
C	3	1.120	1.081	1.067	1.067	1.067	1.103
D	2	1.061	1.046	1.032	1	1	1.032
D	3	1.150	1.046	1.067	1.032	1.032	1.067
E	2	1.021	1.032	1.067	1	1	1.085
E	3	1.128	1.045	1.067	1	1	1.067
F	2	1.086	1.046	1.103	1.067	1.049	1.103
F	3	1.056	1.046	1.103	1.032	1.032	1.067
G	2	1.055	1.053	1.067	1	1.016	1.032
G	3	1.047	1.067	1	1	1.032	1

4　まとめ

本論文では、録音音声の声の高さを一定にするための基礎的な調査を行なった。研究の結果、平均ピッチと聴覚的な再生周波数が比例することが示された。そのため、自動的な pitch 正規化が可能であることが示された。 今後は、より精度の高い方法を研究する必要がある。

参考文献

[1] Entropic Research Lab. Inc., “Esps/waves+ Application Note” (1993).

[2] 東田 他, “オペレータレス自動電話番号検査システムの開発”, 自然言語処理研究会 98-NL-123-4 pp.25-32 (Jan. 1998).