録音音声データのパワーの正規化の検討*

村上仁一 （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

本論文では、録音音声の音量（パワー）を一定にするアルゴリズムを提案した。次に提案したアルゴリズムを用いて実験を行なった。最後に実験結果について定量的な報告をした。

大量の音声を録音した場合、同一話者でも発声する日時によって音量に違いがでる。しかし、録音編集方式による音声ガイダンスでは、録音された音声を組み合わせてガイダンスを作るため、音量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パワーを自動的に一定にする、録音音声パワー正規化アルゴリズムが必要になった。

2 パワー正規化方法

2.1 パワー分布

人間によってラベリングされたデータから、母音ごとにフレームのパワーの分布を調べた。パワーはフレーム窓長 25ms、フレーム間隔 5ms で計算した。調査したデータは ATR の Cset の中の 50 文女性話者 32 名合計 1600 文である。サンプルデータ中の/a/の分布を図 1 に示す。縦軸は頻度で横軸はパワーである。

この図から、パワーの頻度分布は 2 つのピークを持っていることがわかる。解析の結果、弱いパワーの領域はクロージャーや文末などであることがわかった。

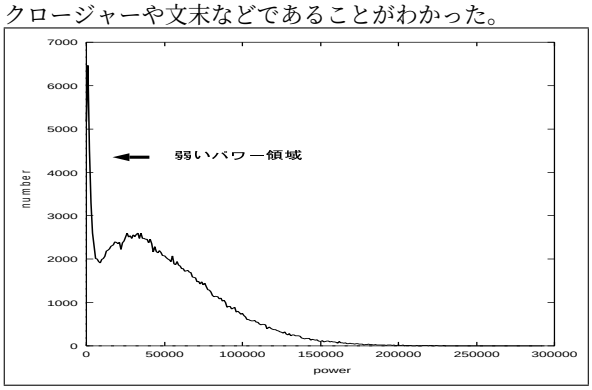

図 1: ラベル情報を利用したときの /a/ のパワー分布

2.2 パワー正規化アルゴリズム

図 1 調査の結果から、考案したパワー正規化アルゴリズムを以下に示す。

1. ATR の DB から第 2 層のラベルを使用して母音ごとの標準平均パワー (P_a) を計算する。
2. 正規化する音声 (w) に対し連続音素認識を行ない、母音区間を推定する。
3. 正規化する音声のフレームごとのパワー (p_{ia}) を計算する。
4. 認識された母音区間を利用して、フレームのパワー (p_{ia}) が標準平均パワー (P_a) になるように、フレーム正規化数 (T_{ia}) を求める。
5. フレーム正規化数 (T_{ia}) が、あまりにも小さい値 (T_L) あるいは大きな値 (T_B) の場合は削除して、平均値 (T) を求める。
6. 正規化する単語 (w) に対し平均値 (T) の逆数（正規化値 NP_w）を掛ける。

ATR の DB から第 2 層のラベルを使用して計算した各母音の標準平均パワー (P_a) を表 1 に示す。

表 1: 標準平均パワー

母音	標準平均パワー	母音	標準平均パワー
/a/	81076.7	/i/	65744.0
/u/	74762.6	/e/	81697.9
/o/	82552.1		

3 パワー正規化実験

提案した方式の有効性を定量的に把握するために、人間が検聴して求めた元のデータに対する倍率（以後 正規化値 NP_h）と提案した方式で求めた値（以後 NP_c）を比較して調べた。

3.1 サンプルデータ

サンプルデータは 4 つのグループ合計 200 単語で構成されている。表 2 にサンプルデータの概要を示す。表中のパワーの変動の項は、各単語のパワーの変動が、比較的大きいか小さいかを意味している。

表 2: サンプルデータ

グループ番号	話者	パワーの変動	発話単語数
1	複数話者 (7 名)	小	30 単語
2	複数話者 (7 名)	大	70 単語
3	同一話者 A	小	30 単語
4	同一話者 B	大	70 単語

発話者は女性のナレータで発話内容は日本の住所（地名）である。

3.2 検聴による正規化値

このデータを、3 人の被験者により音量が一定になるように調整して、正規化値（NP_h）を得た。この結果を

* “ A Study of Power Normalization for Speech Database ” by Jin’ichi Murakami (NTT Information and Communication System Laboratories) and Hirokazu Suzuki (NTT Advanced Technology)

図 2 に示す。図中横軸は単語、横軸は正規化値を意味している。

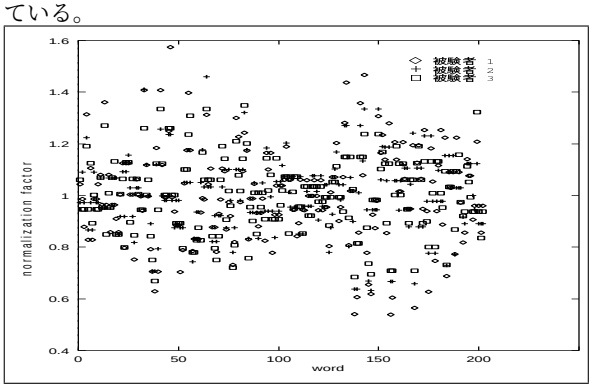

図 2: 検聴による正規化値 (NP_c)

3.3 本方式による正規化値

次に、提案した方式による正規化値（NP_c）を求めた。HMM の学習には HTK を使用した。学習条件を表 3 に示す。

表 3: 音素モデルの学習条件

音響モデル	4 状態 3 ループ混合分布型 HMM
混合数	10 混合 full covariance
音響パラメータ	log power + 12 次 FFT melcep + Δ log power + 12 次 Δ FFT melcep
学習データ	ATR C セット文発声データ
話者	女性 32 名
データ数	1600 文
フレーム長	5ms
フレーム窓長	25ms
sampling 周波数	16KHz

提案した方式による正規化値（NP_c）と検聴による値（NP_r）の比（NP_r）以後 $NP_r = NP_r/NP_c$）を図 3 に示す。ただし、閾値は（$T_L = 0.5, T_B = 2.0$）とした。

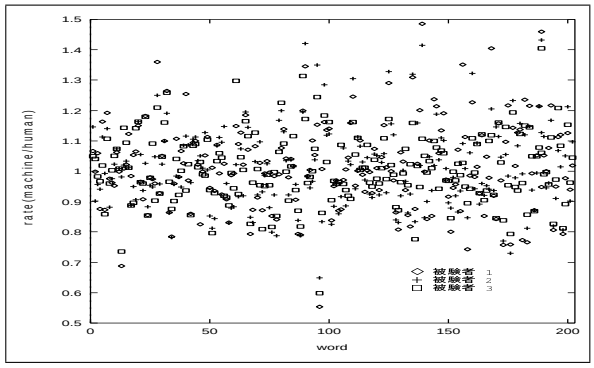

図 3: 提案した方式と検聴による正規化値の比 (NP_r)

この図から比（NP_r）が小さいことが分かる。したがって提案した方式は有効であることが分かる。

3.4 閾値の変化による正規化値の比の変化

次に閾値（T_L）および閾値（T_B）を変化させた場合の、提案した方式による正規化値と検聴による値の比（NP_r）の分散を図 4 に示す。

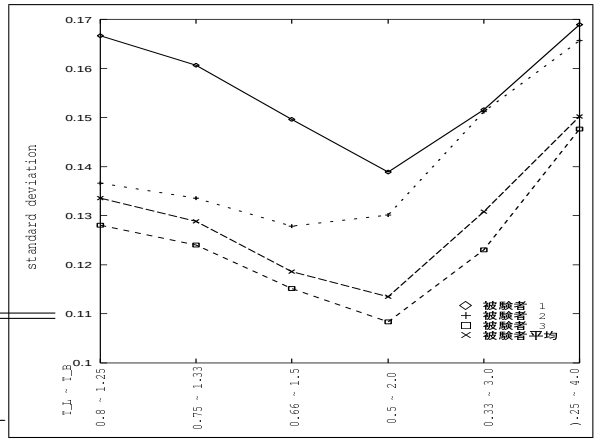

図 4: 閾値 (T_L, T_B) と正規化値の比の関係

この実験から、閾値（$T_L = 0.5 T_B = 2.0$）において最もよい結果（被験者の平均で、標準偏差 0.114）が得られることが示された。

4 考察

提案した方式による正規化値と検聴による値の比（NP_r）が大きなサンプルデータを調べたところ、計算するフレーム数が少ないことが示された。そこで平均フレーム数の 1/4 のサンプルデータを計算から削除して実験を行った。この結果を図 5 に示す。

この図から、良い結果（被験者の平均で、標準偏差 0.109）が得られることがわかる。

5 まとめ

本論文では、録音音声のパワーを一定にする技術について報告した。研究の結果、パワーの頻度分布は、2 つのピークを持っていること、そしてそこで、パワーの正規化の計算を行なう際、あまりにも小さい値はカットして、平均値を求める方法で、人によって得られた値と差が標準偏差で 0.109 と、実用的なパワーの正規化が可能であることを示した。

参考文献

[1] Cambridge University Engineering Department Speech Group and Entopic Research Lab Inc., “HTK:Hidden Markov Model Toolkit V1.5” (Sep. 1993).

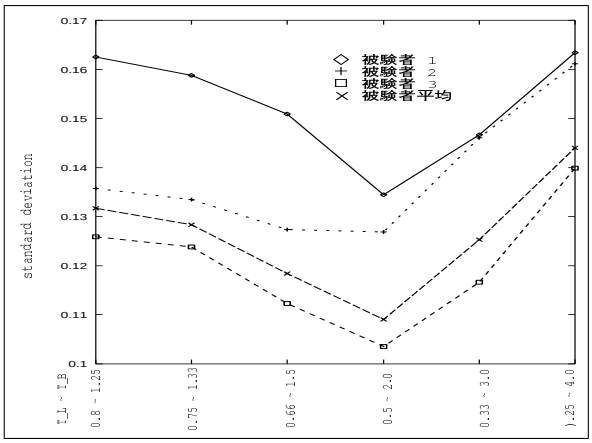

図 5: 平均フレーム数の 1/4 のデータを削除した場合