録音音声データの平均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（平均パワー）を一定にするアルゴリズムを提案し、その実験結果について報告する。

大量の音声を録音した場合、話者が同じでも発声する時間によって音声の平均パワーに違いがでる。しかし、音声を利用したサービスでは録音された音声を組み合わせてガイダンスを作り出すため、音量の違いは違和感になる。

天気予報などのサービスでは語彙が少ないため、単語ごとに人間による平均パワーの正規化が可能である。しかし、録音単語数が多いサービス（例えば、PB 電話機を利用した電話番号案内実験システム [1] では住所 18 万件、姓名 18 万件を録音）では、自動的に録音音声の平均パワーを一定にする、録音音声パワー正規化技術が必要になる。

2 平均パワー正規化の方法

平均パワー正規化のための情報として以下の 2 つが考えられる。

1. F0
   基本的に、平均パワーは F0 に依存すると思われる。しかし現在の技術では F0 情報を取り出すためのピッチ抽出の信頼度が低いという問題点がある。
2. 音素
   平均パワーは母音区間のフレーム毎のパワーの平均値に比例すると思われる。本論文では、この情報を利用した。

3 パワー分布

3.1 ラベル情報を利用したときのパワー分布

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。調査したデータのなかで/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、この図において弱いパワーの領域（以後、弱パワー区間）はクロージャーや文末などであることがわかった。

3.2 連続音素認識を利用したときのパワー分布

ここでは、母音の切り出しに連続音素認識を利用したときのパワーの分布を調べた。認識には HTK を用いた。HMM の学習条件を表 1 に示す。

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

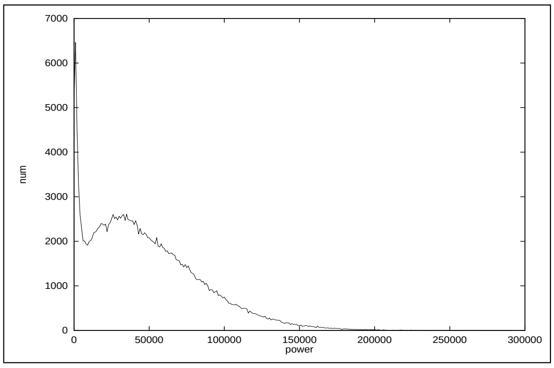

図 1: ラベル情報を利用したときの /a/ のパワー分布

表 1: 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

学習した HMM を用いて Viterbi サーチを行ない母音区間を切り出した。このときの/a/のパワーの分布を図 2 に示す。

図 2 から、連続音素認識をしても、パワーの分布が 2 つのピークを持つことがわかる。

3.3 弱パワー区間を削除したときのパワー分布

ATR のデータでは、ラベルファイルの第 2 層にパワーに関する記述があり、弱パワー区間を知ることができる。この範囲を削除した場合のパワーの分布を調べた。図 3 に、調査した母音のなかの/a/の分布を示す。

この図から、弱パワー区間を削除することにより、パワーの分布が 1 つになることがわかる。

4 パワー正規化手法

以上 3 つの調査から、弱パワー区間を削除して正規化をすることにより平均パワーが一定になることが予期される。そこで、フレームごとにパワーを計算し、あまりにも小さい値や大きな値は削除して、平均値を求めることでパワー正規化することを試みた。

本論文で提案するパワー正規化のアルゴリズムを以下

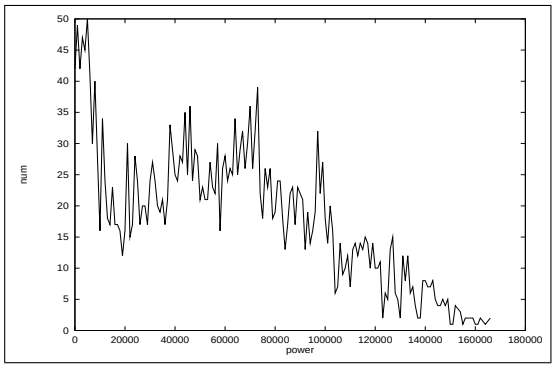

図 2: 連続音素認識を利用したときの/a/のパワー分布

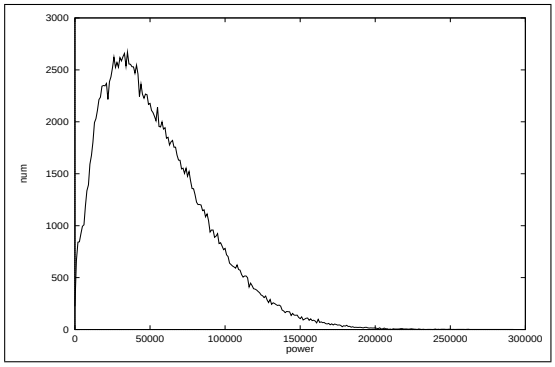

図 3: /a/のパワー分布 (パワーの小さい区間を削除)

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を表 2 に示す。

表 2: 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		

5 実験

約 50 万の単語発声のデータ（話者 7 名）の中から 30 単語を選択し、接続させた音声を 1 サンプルとし、12 サンプルでパワー正規化実験を行なった。また、閾値 (T_L) に 0.5、閾値 (T_B) に 2.0 を選んだ。そして、人間による聴取実験の結果、原音より平均パワーのバラツキが小さくなることを確認した。

6 考察・音素環境による正規化

ここでは、音素環境によって弱パワー区間を削除することを試みた。音素環境ごとのパワーを計算し、これと第 2 層のラベルと比較して弱パワー区間になりやすい音素環境を調べた。この結果の一部を表 3 に示す。

表 3: 弱パワー区間になりやすい音素環境

音素環境	弱パワー区間になった数 / 調査数	比率
a , zh	54 / 244	0.221311
i , z	42 / 223	0.188341
o , b	48 / 259	0.185328
e , h	16 / 95	0.168421

弱パワー区間になりやすい音素環境を除いてパワーの分布を調べた、この結果を、図 4 に示す。

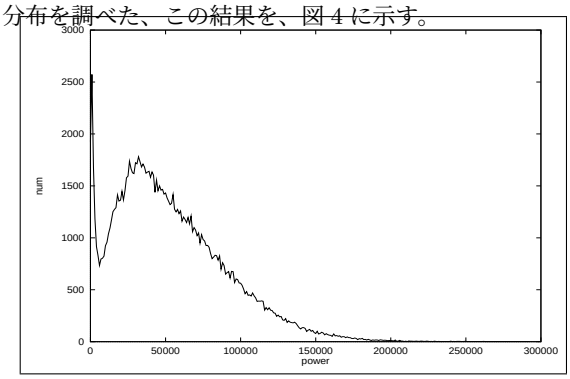

図 4: 音素環境によるパワー正規化方法 /a/

この図から、パワーの正規化方法として音素環境は利用できないことがわかる。

7 まとめ

本論文では、録音音声の平均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実用的なパワーの正規化が可能であることを示した。今後の課題として定量的な評価をすることがあげられる。

参考文献

[1] M. Higasida, ”A Fully Automated Directory Assistance Service that Accommodates Degenerated Keyword Input via Telephones”, Pacific Telecommunications Conference pp,175-179 (Jan. 1997).

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