エネルギースペクトル密度を利用した音声解析[1]

1N−02

大條 雅彦[2]
中央大学大学院理工学研究科

鈴木 寿[3]
中央大学理工学部情報工学科

1 はじめに

現在, 実用化されている音声認識システムは単語認識が基盤である. しかし, 基本辞書と呼ばれるデータベースに依存するために通常会話に対応するシステムには到っていないのが現状である. 音声認識のさらなる汎用性向上のためには, 単音節認識, 特に子音認識が必要である. しかし, 従来のパワースペクトル密度(Power Spectral Density : PSD)の解析法では子音解析が困難である. そこで本稿では, 音声解析法としてエネルギースペクトル密度(Energy Spectral Density : ESD)を利用する. 開発したサウンドスペクトログラムのツールを用いて, ESDとPSDを利用した単音節の解析をおこない, 比較・検討をおこなった. 尚, PSDの算出には高速フーリエ変換(FFT)を, ESDの算出にはプロニー法を用いた.

2 ESD の算出

ESDでは, N個のサンプルからなる時系列データ x_i を,

$$x_i \approx \sum_{m=1}^{p} b_m z_m^i \quad (i = 0, 1, \cdots, N-1)$$

に近似する. b_m, z_m は

$$b_m = A_m e^{j\theta_m}, \quad z_m = e^{(\alpha_m + j2\pi f_m)\Delta t}$$

である. A_mは振幅, θ_mは位相, α_mは減衰率, f_mは周波数を表し, Δt は標本化時間間隔である. 今, 複素数 z_kとその共役複素数 z_k^*からなる$2p$次の代数方程式を線形予測係数 a_k $(k = 0, 1, \cdots, 2p)$ を用いて置きかえる$(a_{2p} = 1)$.

$$f(z) = \prod_{k=1}^{p}(z - z_k)(z - z_k^*) = \sum_{k=0}^{2p} a_k z^k$$

この代数方程式の解 z_m $(m = 1, 2, \cdots, p)$ と任意の整数lを用いて$f(z) = 0$は,

$$f(z) = \sum_{k=0}^{2p} a_k z_m^{k+l} = z_m^l \sum_{k=0}^{2p} a_k z_m^k = z_m^l \cdot 0 = 0$$

と表せ,

$$\sum_{k=0}^{2p} a_k x_{k+l} \approx \sum_{m=1}^{p} b_m \sum_{k=0}^{2p} a_k z_m^{k+l} = 0$$

が導出される. N個のサンプル値x_iを用いて自己相関法により求めたa_kを用い, 高次代数方程式より求めた z_m, z_m^*の一方をz_mとする. $b_m(m = 1, 2, \cdots, p)$は近似式にz_mを代入することで, 求められる.

b_m, z_mより, 振幅A_m, 位相θ_m, 減衰率α_m, 周波数f_mは,

$$A_m = |b_m|,$$

$$\theta_m = \tan^{-1}\frac{\mathrm{Im}\, b_m}{\mathrm{Re}\, b_m},$$

$$\alpha_m = \frac{\ln|z_m|}{\Delta t},$$

$$f_m = \frac{1}{2\pi\Delta t}\tan^{-1}\frac{\mathrm{Im}\, z_m}{\mathrm{Re}\, z_m}$$

となる. これより,

$$x_k \approx \sum_{m=1}^{p} A_m e^{j\theta_m}\{e^{(\alpha_m + j2\pi f_m)\Delta t}\}^k$$

$$= \sum_{m=1}^{p} A_m e^{\alpha_m k\Delta t} e^{j(2\pi f_m k\Delta t + \theta_m)}$$

となり, $k\Delta t = t$ とすると,

$$x(t) = \sum_{m=1}^{p} A_m e^{\alpha_m |t|} e^{j(2\pi f_m t + \theta_m)}$$

と表せる. このフーリエ変換は,

$$X(f) = \sum_{m=1}^{p} A_m e^{j\theta_m} \cdot \frac{-2\alpha_m}{\alpha_m^2 + 4\pi^2(f_m - f)^2}$$

となる. よって, ESDは$|X(f)|^2$として求められる.

3 解析のための設定

解析のための処理の際, フレーム区間, フレーム移動, 次数を最適な設定にしておく必要がある. 実はこの段階でESDによる処理は既存のPSDよりもこれらのパラメータの違いにより結果が大きく変わってしまうことが分かった. 以上のパラメータを変化させながらサウンドスペクトログラムを目視により確認し, 最適な設定値(固定値)を順次決定した. 結果として, フレーム区間では, PSDと同等の設定値まで区間をとらなければならなかった. また, フレーム移動では, 極短時間による変化を観測するために最小の値をとらなければならなかった. また, 次数では前述のパラメータを変化させながら冗長性を考慮し決定することができた. 最適とした判断基準は特徴(主にピーク点)の安定性である. 以下の表にそれぞれの設定値をあらわす.

---

[1] Voice Analysis using Energy Spectral Density
[2] Masahiko Ooeda, Graduate School of Science and Engineering, Chuo University
[3] Hisashi Suzuki, Faculty of Science and Engineering, Chuo University

表3.1 設定値

	ESDの設定値	PSDの設定値
フレーム区間	16[ms]	16[ms]
フレーム移動	0.125[ms]	0.125[ms]
次元数	10	設定の必要なし

5 音声解析

5.1 母音の解析

日本語母音/a-o/に対し，解析をおこなった．以下の図5.1に母音/a/の解析結果を図示する．PSDでは，1000[Hz]付近に高いピークが現れ，その周辺にもピークが散在している．また，2500[Hz]付近にもピークが見られる個所がある．PSDの特性としてスペクトルが一様に現れるのでピークの断定が困難である．一方，ESDでは500[Hz]付近と1000[Hz]付近の二ヶ所にピークが鋭く現れた．他の母音/i/から/o/に対しておこなった母音解析の結果，PSDではホルマントのスペクトルピークに変動があり，ホルマント周波数の同定および，近くに存在するホルマント同士の正確な分離が困難であることがいえる．ESDの解析では，/a/と同様に他の母音も鋭いピークが現れた．

5.2 子音の解析

破裂音で無声子音である音素/p/を含む日本語子音/pa-po/に対して解析をおこなった．以下の図5.2に単音節/pa/に対する結果を図示する．PSDでは，子音のスペクトルが現れず，母音の特徴しか現れない．わたり部あたりからスペクトルの特徴が出始めているので完全に子音の解析できないことがわかる．ESDでは，子音，母音および，わたり部のスペクトルの特徴がはっきりと現れている．音声の開始直後とわたり部では変動が非常に大きいが，その間の区間では，1000[Hz]と2000[Hz]付近に一定の値を保つスペクトルピークが確認できる．これより，子音部に相当する区間で安定したスペクトルピークが得られたことがいえる．/pa/から/po/のESDの解析結果では，2000[Hz]あたりに共通してピークが現れる．そして，このピーク以外に現れるピークは，後続母音のホルマントピークがある周波数に現れていた．これより，/pa/から/po/までの単音節のそれぞれの子音部/p/には，音素/p/として共通するピークとそれぞれ後続母音の影響を受けて現れるピークの二種類の特徴があると考えられる．ここで，後続母音の影響といっても実際に母音の発声は子音の発声後におこなわれるので，ここでは後続母音の発声に備える口唇の形状が関係しているといえる．PSDの解析結果では，/pa-po/についてほとんど子音の情報を得られないことがいえる．

6 おわりに

ESDの解析でPSDの解析では検出が困難であった子音のスペクトル情報を検出することができた．これより，子音の特徴抽出にESDを利用した解析法が有効であることがいえる．ESDでは母音のホルマントが鋭く検出されることが確認でき，これを従来のPSDのホルマント情報と同様に扱うことで，母音認識へ適用することができるといえる．以上より単音節単位での音声解析にESDを利用した方法が有効であるといえる．

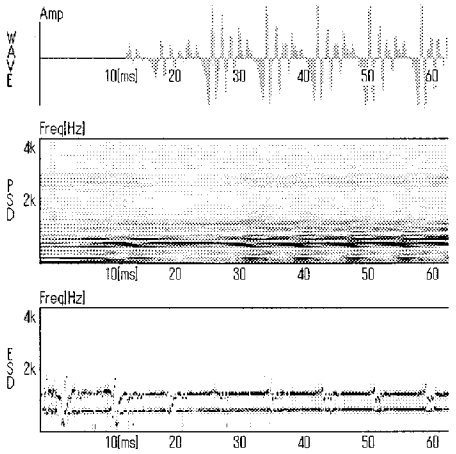

図5.1 /a/の解析結果

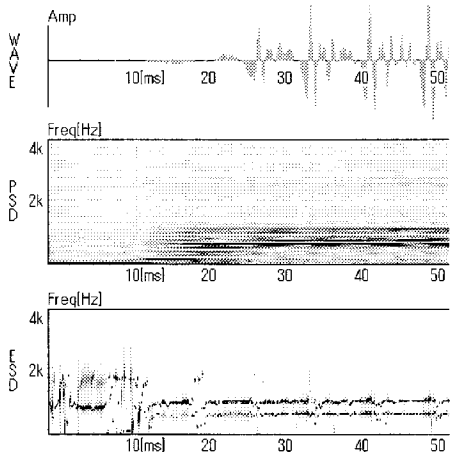

図5.2 /pa/の解析結果

参考文献

[1] 松永光輝:“Prony ESDによる音声認識の研究(Study of Speech Recognition using Prony ESD)”，中央大学大学院理工学研究科情報工学専攻修士論文，2000.

[2] Hisasi Suzuki and Mitsuteru Matsunaga, “Effective Recognition of Consonants Pa, Pi, Pu, Pe, Po Using Prony Energy Spectral Density”, International Symposium on Information Theory and Its Applications, 2000.