

輪 講 資 料

スペクトル包絡からの音声合成における
音源モデルに関する考察
(下書き)

1986年12月4日(木)

徳田 恵一

あらまし 改良ケブストラム法など⁽¹⁾⁻⁽⁵⁾で得られたスペクトル包絡により音声合成する際の音源モデルの振幅の与え方について考察する。

1. はじめに

有声音において、正しいスペクトル包絡を安定に抽出するためには、スペクトルの微細周波数構造のピークを通るスペクトル包絡をとればよい⁽¹⁾⁻⁽⁵⁾。本文では、このようにして得られたスペクトル包絡から、音声合成をする際の、音源モデルについてまとめる。また、無声音の場合にも、スペクトル包絡からの合成のための音源モデルについて考察する。

2. 有声音の音源モデル

音声信号は短い区間について考えれば、ある線形時不変システムつまり声道を周期的なインパルス列または白色雑音で励振することにより、生成されることができる⁽⁶⁾。従って、音声信号 $x(n)$ は、音源 $e(n)$ と、声道のインパルス応答 $h(n)$ が、畳み込みの形で結び付けられたもので表すことができる。

$$x(n) = e(n) * h(n) \quad (1)$$

ただし、 $*$ は畳み込みを表す。実際に音声信号を分析する際には、窓 $w(n)$ をかけるため、信号 $x(n)$ は、

$$x(n) = (h(n) * e(n)) \cdot w(n) \quad (2)$$

の形になる。ここでは、有声音について考えることにして、音源 $e(n)$ が

$$e(n) = K \cdot \sum_{k=-\infty}^{\infty} \delta(n - k \cdot N_p) \quad (3)$$

と表されるインパルス列であるとする。ただし、 $\delta(n)$ は離散信号の単位インパルスを、 N_p は音源の周期を表す。このとき、式(3)のスペクトル $E(\omega)$ は、

$$E(\omega) = K \cdot (2\pi / N_p) \sum_{k=-N_p/2}^{N_p/2} \delta_D(\omega - k \cdot \omega_p) \quad (4)$$

となる。ただし、 $\delta_D(\omega)$ はディラックのデルタ関数を表し、 $\omega_p = 2\pi / N_p$ である。従って、式(2)のスペクトルは、声道のインパルス応答 $h(n)$ のスペクトルを $H(\omega)$ 、窓のスペクトルを $WN(\omega)$ として、

$$\begin{aligned} X(\omega) &= (E(\omega) \cdot H(\omega)) * WN(\omega) \\ &= (K / N_p) \sum_{k=-N_p/2}^{N_p/2} H(k \cdot \omega_p) \cdot WN(\omega - k \cdot \omega_p) \end{aligned} \quad (5)$$

と表され、振幅スペクトルにおける微細周波数構造のピーク位置は $\omega = k \cdot \omega_p$ で与えられることになる。ここで、ある微細周波数構造のピーク $m \cdot \omega_p$ に中心を持つ窓のスペクトル $WN(\omega - m \cdot \omega_p)$ のサイドローブが、他の微細周波数構造のピークつまり $\omega = k \cdot \omega_p$ 、 $k \neq m$ で、十分減衰していると仮定する。このとき、微細周波数構造のピークについてのみ考えれば、

$$X(k \cdot \omega_p) = K \cdot (WN(0) / N_p) \cdot H(k \cdot \omega_p) \quad (6)$$

となる。ただし、

$$WN(0) = \sum_{k=-\infty}^{\infty} wn(k). \quad (7)$$

つまり、微細周波数構造のピークを通るように抽出されたスペクトル包絡は、 $K \cdot (WN(0) / N_p) |H(\omega)|$ を抽出していると考えられる。

従って、このことから逆に得られたスペクトル包絡が声道特性 $|H(\omega)|$ を表しているとするなら、音源 $e(n)$ の係数 K は、

$$K = N_p / WN(0) \quad (8)$$

と定められるべきである。つまり、微細周波数構造のピークを通るスペクトル包絡を用いる有声音の分析・合成では、音源 $e(n)$ は、

$$e(n) = (N_p / WN(0)) \sum_{k=-\infty}^{\infty} \delta(n - k \cdot N_p) \quad (9)$$

とモデル化されなければならない。

3. 無声音の音源モデル

無声音では、音源を白色雑音と考えるため、そのスペクトルも統計的性質しか知ることができず、微細周波数構造のピークを通るスペクトル包絡の物理的意味も明らかではない。

ここでは、白色雑音の短時間振幅スペクトルが完全に白色つまり平坦であると仮定することにする。このとき音源 $e(n)$ が、分散 $\sigma^2 = 1$ の白色雑音 $r(n)$ に係数 K を掛けたもの

$$e(n) = K \cdot r(n) \quad (10)$$

とすれば、その短時間スペクトル $E(\omega)$ は、未知の位相成分を $\theta(\omega)$ として、

$$E(\omega) = K \exp j\theta(\omega) \quad (11)$$

と考えることができる。従って、窓掛けされた音声信号つまり式(2)のスペクトルは、

$$\begin{aligned} X(\omega) &= (E(\omega) \cdot H(\omega)) * WN(\omega) \\ &= K \cdot (\exp j\theta(\omega) H(\omega)) * WN(\omega) \end{aligned} \quad (12)$$

となり、窓のスペクトル $WN(\omega)$ が、 $H(\omega)$ の変化の緩やかさに対し、十分ディラックのデルタ関数に近いとすれば、

$$X(\omega) = K \cdot WNN^{1/2} \exp j\theta(\omega) H(\omega) \quad (13)$$

とすることができる。ただし、

$$WNN = \sum_{k=-\infty}^{\infty} wn^2(k). \quad (14)$$

従って、微細周波数構造のピークを通るスペクトル包絡は $K \cdot WNN^{1/2} |H(\omega)|$ を抽出していると考えられるので、係数 K は前節と同様の考え方により、

$$K = WNN^{1/2} \quad (15)$$

と定められ、音源 $e(n)$ は、

$$e(n) = (1/WNN^{1/2}) r(n) \quad (16)$$

とモデル化することができる。

$$\rightarrow WN(\omega) = \sqrt{\sum_{n=-\infty}^{\infty} wn^2(n)} \cdot \delta_0(\omega)$$

↓
110.セバールの関係より

実際には無声音も微細周波数構造を持つため、微細周波数構造のピークを通るスペクトル包絡のレベルは若干高めに見積られることになる。しかし、このようなモデルで音声を合成した場合にも、聴感上不自然さはなく、大きな問題はないと考えられる。また、このようなオフセットは、実験的に評価できるので、得られた実験値により補償することも考えられる。

線形予測法⁽⁷⁾（自己相関法）では、窓掛けした音声信号と、合成音の平均パワーが一致するように音源のパワーが与えられる。有声部の分析では、音声信号と窓の相対的位置関係により、そのパワーは異なってくるため、合成音のパワーも、分析フレームごとに変動することになる。それに対し、本文で述べたような微細周波数構造のピークを通るスペクトル包絡からの有声音の合成では、このような分析フレームごとの変動は起こらない。従って、本モデルは、無声音での定式化が曖昧になりはするが、有声音での利点を考慮すれば十分妥当なものと考えられる。

4. 適用

以上より、音源モデルは、

$$e(n) = \begin{cases} (N_p / WN(0)) \sum_{k=-\infty}^{\infty} \delta(n - k \cdot N_p), & \text{有声} \\ (1 / WNN^{1/2}) r(n), & \text{無声} \end{cases} \quad (17)$$

となる。

ここでは、分散 $\sigma^2 = 1$ の白色雑音 $r(n)$ として、 ± 1 の値をとる M 系列 $mseq(n)$ を用いることにし、全体に $WNN^{1/2}$ をかければ、音源モデルは、

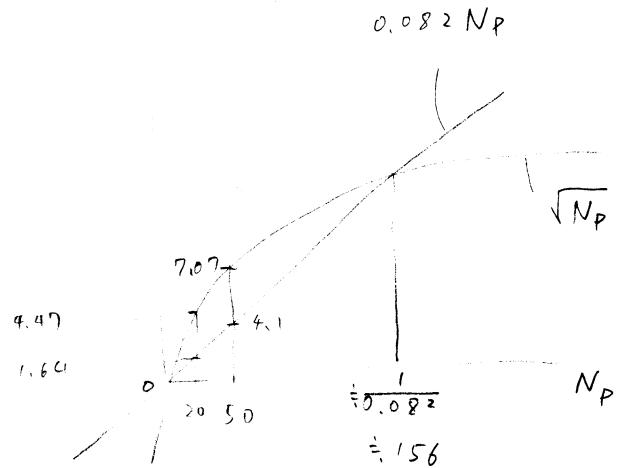
$$e(n) = \begin{cases} N_p \frac{(WNN^{1/2} / WN(0))}{0.08229} \sum_{k=-\infty}^{\infty} \delta(n - k \cdot N_p), & \text{有声} \\ mseq(n), & \text{無声} \end{cases} \quad (18)$$

と書くことができる。

いくつかの窓について $WNN^{1/2}/WN(0)$ の値を表 1 に示しておく．いずれも，窓長は 256 点で計算した．

窓の種類	$WNN^{1/2}/WN(0)$
ブラックマン窓	0.82289
ハミング窓	0.073065
方形窓	0.062500

表 1 $WNN^{1/2}/WN(0)$ の値



5. むすび

微細周波数構造のピークを通るスペクトル包絡による音声の分析・合成の際の音源モデルについて考察した．

その結果，有声音の場合，音源はピッチに比例した振幅を持つ必要があることが示される．これは，線形予測法において，パワの値が与えられた音源を用いて音声合成が行われるのと対照される．

無声音の場合の音源モデルについては，さらに詳しい検討が必要である．

文 献

- (1) 今井，阿部：“改良ケブストラム法によるスペクトル包絡の抽出”，信学論(A)，J62-A,4, pp.217-223(昭54-04)．
- (2) 徳田，小林，今井：“重み付き最小2乗法による対数スペクトル包絡の抽出”，電子通信学会総合全国大会(昭61-03)．

- (3) 徳田, 小林, 今井: "非一様なスペクトル荷重による対数スペクトル包絡の抽出", 信学技報, SP86-41(1986-09).
- (4) Paul, D.B.: "The Spectral Envelope estimation Vocoder", IEEE Trans. Acoust. Speech and Signal Process., ASSP-29, 4, pp.786-794(Aug.1981).
- (5) 三上, 大場: "スペクトル包絡より得られる群遅延特性を用いた有声音の極・零点の推定", 信学論(A), J68-A, 1, pp.86-91 (昭60-01).
- (6) Oppenheim, A.V. and Schafer, R.W.: "Digital Signal processing", Prentice-Hall, Englewood Cliffs, N.J. (1975)
- (7) Markel, J.D., Gray, A.H., Jr. 著, 鈴木 訳: "音声の線形予測" コロナ社(昭55)