音楽情報科学 24－9
(1998. 2. 20)

Sinusoidal modelを用いた楽音の接続について

榊原 健一　　小坂 直敏

NTT 基礎研究所

〒243-0198 神奈川県 厚木市 森の里若宮 3-1
0462-40-3657　　0462-40-3655
kis@theory.brl.ntt.co.jp　　osaka@idea.brl.ntt.co.jp

あらまし　本稿では，Sinusoidal model を用いた任意の音色の短い楽音の接続について述べる．楽音の接続は，定常楽音のサンプルに音楽情報を付与し，楽句を合成することを目的とした研究の要素技術の一つと位置付けている．本研究では，接続時点の部分音の最適な対応関係を，瞬時周波数以外の属性も用いて見い出し，マッチングした二つの部分音を線型補間することで接続を行なう．より一般的な有限集合上の狭義単調な局所探索マッチングを問題を解くアルゴリズムの概略も論じる．また，得られた接続音について，原音の定常部分とのスペクトルの親近度についての評価結果についても報告する．

キーワード　Sinusoidal model，DP，音色補間，混成音合成，信号モデル，音楽情報付与

On concatenation of musical sounds using a sinusoidal model

Ken-Ichi Sakakibara　　Naotoshi Osaka

NTT Basic Research Laboratories

3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa, 243-0198, Japan
+81-462-40-3657　　+81-462-40-3655
kis@theory.brl.ntt.co.jp　　osaka@idea.brl.ntt.co.jp

Abstract　This paper describes a method for concatenation of short stationary musical sounds in arbitrary timbre using a sinusoidal model. We place concatenation of musical sounds as one of the important study items for the goal of musical information attachment on musical sound samples to create musical phrases. Concatenation was done by two steps: to find optimal correspondences among partials at the junction using various characteristics other than their instantaneous frequencies and to linearly interpolate spectral data between matched partials. Evaluation result was reported using spectral closeness between a concatenated portion of synthesized sound and a stationary part of its original sounds.

key words　Sinusoidal model, DP, Timbre interpolation, Sound hybridization, Spectral model, Musical information attachment

1 はじめに

筆者らは現代における作曲，コンピュータ音楽やマルチメディアコンテンツの創作に必要とされる音合成研究の一つとして，信号モデルによる音色の制御の研究を行なってきた．それらの研究によりコンピュータを用いた従来の音楽創作の支援のみならず，新しい音の合成・制御の方法を提示することで新たな音楽の創作の方法を開拓することを目指している．

それらの音合成の研究の一環として，任意の音色による短い定常的なサンプル音に音楽情報を付与し楽句を合成すること目的とした研究を行なっている．これを実現する為の要素技術のうち重要なものの一つとして，Sinusoidal model を用いた二つの楽音の接続について検討を進めている．

本稿では，McAulay-Quatieri の方法 [2] による Sinusoidal model を枠組みとして用い，既存の楽器では不可能な連続的な接続を行なうことを目的に，サンプル音の定常部分を用いて，(i) 管楽器のグリッサンドの様なピッチの違う音の連続的ピッチの接続，(ii) 違う音色を補間する接続，などを試みた結果を報告する．

2 Sinusoidal model による分析 / 合成

Sinusoidal model の分析 / 合成は，品質の良さおよび聴覚にあった自然な表現形式から，コンピュータ音楽の分野ではしばしば用いられる．

その自然な表現形式から，原音の忠実な再現のみならず，分析したデータを変換し再合成することで音の連続的な制御，新しい音の創出に用いることができる (図 2.1).

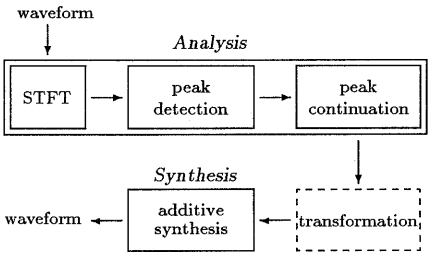


図 2.1: Sinusoidal Model の分析 / 変換 / 合成

McAulay-Quatieri は Sinusoidal model の手法を考案した初期から，音声の分野で分析 / 変換 / 合成の試みを行なっており [3]，

近年では，コンピュータ音楽の分野において，モーフィングへの応用が [4]，[8] に，楽音の混成音合成への応用が [5] に報告されている．

以下，[2] による Sinusoidal model の分析 / 合成の概略を説明する．

2.1 分析

Sinusoidal model の分析は，STFT による原音の分析により得られた各フレーム毎のスペクトルグラム上のピークを取りだし，それらの瞬時振幅，瞬時周波数，瞬時位相の情報から，時間的にピークをたどり部分音を決定していくという方法で行なわれる．

McAulay-Quatieri[2] の方法においては，一定の周波数幅で次のフレームのピークを探し接続して部分音を決定している．

他のピークの接続方法としては，数フレーム分の情報を参照する [7]，部分音のクロスを許す HMM を用いた [1] などがある．

2.2 合成

合成は分析で得た各フレームごとのスペクトルのピークのデータからフレーム内で線型に補間して各部分音の合成を行っていく．部分音の第 k フレームの瞬時振幅，瞬時周波数，瞬時位相をそれぞれ A_k，を ω_k，θ_k，フレーム長を T，フレーム内の時刻を t とする時，瞬時振幅は次のように線型に補間する．

$$A(t) = A_k + (A_{k+1} - A_k)\frac{t}{T}, \quad t \in [0, T]$$

位相函数 $\theta(t)$ は四つの値 $\omega_k, \omega_{k+1}, \theta_k, \theta_{k+1}$ から決定する為，以下のような三次式

$$\theta(t) = \beta t^3 + \alpha t^2 + \omega_k t + \theta_k$$

とし，境界条件から，

$$\alpha = \frac{3}{T^3}(\theta_{k+1} - \theta_k - \omega_k + 2\pi M_k) - \frac{1}{T}(\omega_{k+1} - \omega_k)$$

$$\beta = -\frac{2}{T^3}(\theta_{k+1} - \theta_k - \omega_k + 2\pi M_k) - \frac{1}{T^2}(\omega_{k+1} - \omega_k)$$

となる．ただし M_k は整数で，フレーム内の瞬時周波数の変動が最小になるようなものを選ぶ．簡単な計算により，

$$x = \frac{1}{2\pi}\left(\theta_k + \omega_k T - \theta_{k+1} + (\omega_{k+1} - \omega_k)\frac{T}{2}\right)$$

に最も近い整数を M_k とし，フレーム内の位相函数 $\theta(t)$ を決定する．

したがって，第 k フレームに存在している部分音数を P_k，第 p 部分音の位相函数を $\{\theta_p^k\}_k$ とする時，第 k フレームで n 番目のサンプルについて音信号 $s^k(n)$ は

$$s^k(n) = \sum_{p=1}^{P_k} A_p^k(n)\cos\theta_p^k(n)$$

の形で表される．

3 接続

Sinusoidal model で表現された二つの音の接続は，次の二つのステップで実現される (図 3.1).

Step 1.

接続時点で一般に数の違う二つの原音の部分音の集合をマッチングさせる．

Step 2.

マッチングした部分音それぞれを一本の滑らかな部分音とするように接続区間内をつなげる．また，マッチングしなかった接続前 / 後の部分音は，フェードアウト / インにより処理する．

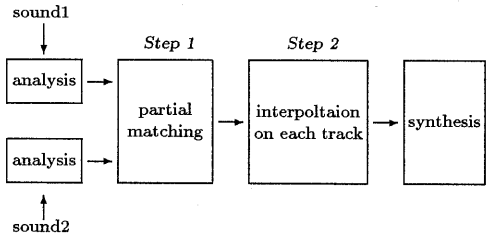


図 3.1: 接続の手順

Step 1. の部分音のマッチングについては，用いる Sinusoidal model のピーク接続と同様の方法で行なうのが最も簡単であるが，ピークの接続には連続的なスペクトルの変化を想定しているので，全く違うスペクトルを持つ二つの音の接続部分での対応を考えるとき，ピークの接続と同様の方法では多くの部分音がつながらないという可能性が大きい．この場合，接続時間がフレーム長の二倍より長ければ接続区間で音は途切れてしまう．そこで，本稿ではより一般的なマッチングの問題を [4] と同じく DP により解いていく．

Step 2. については，本稿では定常音同士の接続を考えているので，マッチングされた部分音については線型に補間することとした．

3.1 部分音のマッチング

接続時点での部分音のマッチングは，接続する二つの原音の接続時点の部分音の持つ瞬時周波数，瞬時振幅などの様々な属性に基づき決定していく．

[2] のピークの接続や [4] のモーフィングにおける部分音の対応は，部分音の持つ属性のうち瞬時周波数のみ考慮しているが，より一般的な部分音のマッチングを考える時，部分音の持つ周波数以外のさまざまな情報もあわせて考える必要がある．

そこで，ここでは多次元のデータを持つベクトルとしての部分音のマッチングに応用するため，より一般的な見地から二つの集合の間で相手が見付からない要素を場合を許す空コスト $d_\emptyset$ の定められた (狭義) 単調なマッチングを解くアルゴリズムについて概略を示す．

また，従来の考え方では，マッチングのコストおよびマッチングの対象を制限する窓函数は，双方とも周波数距離にのみに基づき定めていたが，一般にはコスト函数と，窓函数は独立に定めることが可能である．より一般的には，最適なマッチングを求めるアルゴリズムは，有限集合の整列順に相関させた窓函数と，任意のコスト函数で求めることができる．

例えば，本稿では瞬時周波数の小さい順に $\{x_i\}_i, \{y_j\}_j$ を並べ，x_i それぞれに周波数に関して窓を設け，コスト函数には瞬時振幅，瞬時周波数の両方を用いるというマッチングを使っている．その他，瞬時振幅順に並べ，瞬時振幅に関して窓を設け，瞬時周波数の差でコスト函数を考えるなど，設定は場合に応じていろいろ考えることも可能である．

ここでは窓の定め方によらず各々がマッチングする対象に制限のあるものを局所探索マッチングと呼び，制限のない場合とあわせて以下にアルゴリズムの概略を示す．

有限集合のマッチング

二つの有限集合

$$X = \{x_i\}_{1\le i\le N_X}, \quad Y = \{y_j\}_{1\le j\le N_Y}$$

上にコスト函数 $d : X \times Y \to \mathbf{R}$ および空コスト $d_\emptyset \in \mathbf{R}$ を与えるとき，X の要素から Y に相手が見付からない場合も考え，$\gamma : X \to Y \cup \{\emptyset\}$ なる対応 (写像) (以下，簡単の為，$\gamma(x_i) = y_j$ のとき，$y_j = y_{\gamma(i)}$ と記し，$\gamma(x_i) = \emptyset$ の時も同様に $\gamma(x_i) = y_{\gamma(i)}$ と記すこととする．

この時，総合コスト $C_{X,Y}(\gamma)$ を最小にするような対応，γ を見つけるアルゴリズムを求めたい．

$$C_{X,Y}(\gamma) = \sum_{i=1}^{N_X} d(x_i, y_{\gamma(i)})$$

今，γ が狭義単調な場合を考える．すなわち Y 上で，$i_1 < i_2 \Rightarrow \gamma(i_1) < \gamma(i_2)$ となっている場合について考える．今，$d(x_i, y_j) = d_{i,j}$ とする．この時，以下で DP における累積コスト $C[i](j)$ を考えればよい．

$$C[1](j) = \begin{cases} d_{1,1}, & \text{if } j = 1 \\ \min(d_{1,j},\ C[1](j-1)), & \text{if } 1 < j \le N_Y \\ \min(d_\emptyset,\ C[1](N_Y)), & \text{if } j = N_Y + 1 \end{cases}$$

$1 < i \le N_X$;

$$C[i](j) = \begin{cases} i \cdot d_\emptyset + d_{i,0}, & \text{if } j = 0 \\ \min(C[i](j-1),\ d_{i,j} + c_{i,j}), & \\ \qquad\qquad \text{if } 1 < j \le N_Y & \\ \min(C[i](j-1),\ d_\emptyset + C[i-1](N_Y)), & \\ \qquad\qquad \text{if } j = N_Y + 1 & \end{cases}$$

ただし

$$c_{i,j} = \min_{1\le k<i} (C[k](j-1) + (i-k-1)d_\emptyset)$$

局所探索マッチング

各 $i \in [1, N_X]$ について，区間 $[\phi^i{}_0, \phi^i{}_1] \subset [1, N_Y]$ が与えられ，Y の部分集合

$$\Phi(i) = \{y_j;\ j \in [\phi^i{}_0, \phi^i{}_1]\}$$

に属さない y_j については，コスト $d(x_i, y_j)$ が十分大きくなるようにコスト函数が定まっているとき，局所探索

なマッチングであると呼ぶこととする．局所探索なマッチングの場合，計算量を減らすことができる．

$$C[1](j) = \begin{cases} d_{1,\phi^1{}_0}, & \text{if } j = \phi^1{}_0 \\ \min(d_{1,j},\ C[1](j-1)), & \text{if } 1 < j \le \phi^1{}_1 \\ \min(d_{\emptyset},\ C[1](N_Y)), & \text{if } j = \phi^1{}_1 + 1 \end{cases}$$

$1 < i \le N_X$;
$j = \phi^i{}_0$;
$\phi^i{}_0 = 0$;
$C[i](j) = i \cdot d_{\emptyset} + d_{i,0}$
$1 < \phi^i{}_0 \le N_Y$;
$C[i](j) = d_{i,j} + \tilde{c}_{i,j}$

$\phi^i{}_0 < j \le \phi^i{}_1$;
$C[i](j) = \min(C[i](j-1),\ d_{i,j} + \tilde{c}_{i,j})$

$j = \phi^i{}_1 + 1$;
$C[i](j) = \min(C[i](\phi^i{}_1),\ d_{\emptyset} + C[i-1](\phi^{i-1}{}_1 + 1))$

ただし，$\tilde{c}_{i,j}$ は次のように決まる．$K_j = \{k;\ \phi^k{}_0 \le j \le \phi^k{}_1\}$ とおくとき，

$$c_{i,j} = \min_{k \in K_j}(d_{i,j} + (i-k-1)d_{\emptyset})$$

とし，$c_j(\emptyset) = C[\min_{k \in K_j}(k) - 1](j-1)$ とおく．ただし，任意の j について $C[0](j) = 0$ とする．このとき，

$$\tilde{c}_{i,j} = \min(c_{i,j}, c_j(\emptyset))$$

とすればよい．

3.2 部分音の補間

マッチングした二本の部分音は，接続時点の双方の瞬時振幅，瞬時周波数を接続時間内において線型に補間し，一本の部分音としてつなげていく．

接続された合成音の接続区間を第 n_0 フレームから第 n_1 フレームとすれば，原音の接続時点の瞬時周波数と瞬時振幅を用いて接続区間内の瞬時周波数は，

$$\omega_k = \omega_{n_0} + \frac{k-n_0}{n_1-n_0}(\omega_{n_1} - \omega_{n_0})$$

瞬時振幅は

$$A_k = A_{n_0} + \frac{k-n_0}{n_1-n_0}(A_{n_1} - A_{n_0})$$

と表される．

また，フレーム内では位相の時間微分 $\omega_k(t)$ は線型であるとする．このとき，位相函数は

$$\theta_k(t) = \frac{\omega_{k+1} - \omega_k}{2T}t^2 + \omega_k t + \theta_k$$

となり，瞬時位相は順に送って，

$$\theta_k \equiv \theta_{n_0} + \sum_{n_0 \le i < k} \frac{\omega_i + \omega_{i+1}}{2}T \pmod{2\pi}$$

線型に補間された部分音の接続区間の最終フレームで生じる位相のギャップは，その後に順に送ることとする．

今回は，マッチングしなかった接続前 / 後の部分音は，1 フレームの間でフェードアウト / インで処理しており，接続部分に残差を付け加えるなどの処理は行なっていない．

4 楽音への適用

接続はフルートとクラリネットの単音のモノラル音響信号 (16bit, 48kHz SF) を用いた．また，分析にはハミング窓を用い，FFT ポイント数は 8192，フレーム更新周期は 5msec とした．

4.1 マッチングをしない接続

まず，部分音のマッチングの問題を背負わずに，部分音の線型補間の特性を調べるために，1 秒程度の長さの音を分析した．

次に定常部分である中間部を切り取り同一の部分音同士の瞬時周波数，瞬時振幅，瞬時位相を線型に補間し分析音の復元を試みた．原音の再合成音について補間時間と補間部分の S/N のに関する特性を調べた．

結果を図 4.1に示す．

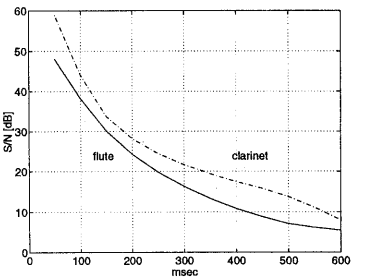


図 4.1: 時間 – 補間部分の S/N 特性

今回は，主な部分音のみの補間に対して特性を調べており，原音の接続部分に存在した残差は接続音からは取り除かれている．そのような条件の下でも，フルート，クラリネットとも，100msec 以下の接続時間であれば，S/N 40dB 程度の品質で補間できている．

4.2 異なる楽音の接続

接続時点の二つの部分音の集合 $\{x_i = (\omega_i, A_i)\}_i$，$\{y_j = (\omega_j, A_j)\}_j$ は，瞬時周波数に関して小さいものの順に整列させ，部分音のマッチングの際のコスト函数は，瞬時周波数 ω_k を基本周波数で割って正規化した $\bar{\omega}_k$ と対数パワーとを，接続時点から数フレーム分の平均値を取り，w_f, w_a で重み付けられた二乗和でより定めた．

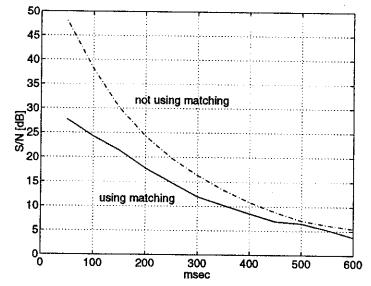
$$w_f(\bar{\omega}_i - \bar{\omega}_j)^2 + w_a\left(10\log_{10}A_i{}^2 - 10\log_{10}A_j{}^2\right)^2$$

マッチングの局所化するため，周波数に関して窓幅を一定値 δ で定め，$x_i = (\omega_i, A_i)$ に対しては，

$$\Phi(i) = \{y_j = (\omega_j, A_j)\ ;\ \bar{\omega}_i - \delta < \bar{\omega}_j \le \bar{\omega}_i + \delta\}$$

のみを接続対象とするという制限を設けた．また，接続の前処理として，接続時点でのパワーの小さい部分音をマッチングの対象から除外するなどの接続対象の部分音の絞り込みを行ない，重み w_f, w_a，空コスト $d_\emptyset$ の値を動かし接続を試みた．マッチングを行なう際の定数の設定値としては，$(w_f, w_a) = (1, 50)$，$1 \leq d_\emptyset \leq 10$ を主に用いた．

マッチングを背負わない線型補間と比較するため，原音の中間部分を部分音のマッチングを行なって接続し，S/N を調べた (図 4.2)．



$w_f = 1, w_a = 50, d_\emptyset = 10, \delta = 0.5$

図 4.2: 接続による補間部分の S/N 比：フルート a^1

部分音のマッチングを背負わない場合と比較すると，補間時間が短い場合は，50msec の時に 20dB 劣っているが，補間時間が長くなるにつれ，差は小さくなり 500msec を越えるとほとんど値に差はない．

二つの異なるサンプルについて接続を行なうとき，接続された合成音と比較すべき原音が存在していないので (波形に関する)S/N では接続の特性を評価することができない．そこで，二つの音のスペクトルの近さを比較するための尺度として以下のものを用いる．幅 ζ で周波数帯域を区切り，$[i \cdot \zeta,\ (i+1) \cdot \zeta]$ の区間に存在する二つの音の部分音のパワーの和をそれぞれ X_i, Y_i とするとき，S/N

$$10 \log_{10} \left(\frac{\sum_i {X_i}^2}{\sum_i (Y_i - X_i)^2} \right)$$

により $\{X_i\}$ からみた $\{Y_i\}$ の部分音の構造の親近度を測ることにする．

接続の難易度の評価をするため，同一サンプル内時点，サンプル，奏者，楽器，ピッチが異なる定常音について，それぞれをパワーで正規化した後，この尺度を用いてスペクトルの親近度を比較した (表 4.1)．

表 4.1 にあるさまざまに要因の異なる二音の接続については規範となる波形は存在しないので，結果を評価するため，この尺度により接続前音，接続後音それぞれの定常部分と接続部分の合成音との親近度を調べた結果を図 4.3 に示す．

	要因	平均	最大	最小
1	時点	21.32	37.96	17.12
2	サンプル	22.12	44.39	7.01
3	奏者	10.69	15.52	3.07
4	ピッチ $(a^1 - a^2)$	3.00	−8.42	−1.28
5	楽器 (fl. − cl.)	−0.02	−1.73	−3.43
6	ピッチ $(a^1 - c^2)$	−4.95	−3.85	−6.49

(dB)

$a^1 = 440\text{Hz}$, $c^2 = 523.25\text{Hz}$

表 4.1: 要因を変化させた二つの定常音の親近度

異なるピッチの二音について接続を行なった図 4.3 の 4, 6 については，接続区間の始めと終りにおいて急激にスペクトルがそれぞれの原音に近いものとなっていることが分かる．しかしながら，他の実施例については接続部分のスペクトルが，接続前音の定常的なスペクトルから後音の定常的なスペクトルへと滑らかに移行しているのがわかる．特に，時点の異なる同一サンプルの接続や奏者の異なる場合の接続については，図 4.3 の 1, 2 に見られるとおり接続区間のどこにおいてもスペクトルは定常性をほぼ保持していることがわかる．

今回の実施例では，スペクトルに隔たりがあるものについては途中に雑音や異音が乗るか，接続部分の音が痩せてしまうなどの問題が残るものの，スペクトルの近い音については滑らかな接続に成功しているといえる．

5 まとめ

スペクトルの近い音の接続は，本稿の方法は有用であり，定常部分の時間伸展などの応用も考えられる．

ただし，スペクトルに隔たりのある二音の接続に際しては，現状では雑音を避ける為には接続する部分音の本数を減らさざるを得ず，今後は音質の向上のため残差の接続部分への付加の方法を課題として検討していきたい．また原音の接続時点付近のスペクトルの変化を接続区間に反映させたより自然な接続方法も考えていきたい．

謝辞　本研究に関して御討論頂いた，NTT 基礎研究所の柏野 邦夫，柏野 牧夫，白木 善尚，引地 孝文の諸氏，並びに本研究を進める機会を下さった同所情報科学研究部 石井 健一郎 部長に感謝したい．

参考文献

[1] Ph. Depalle, G. García and X. Rodet, "Analysis of Sound for Additive Synthesis: Tracking of Partials Using Hidden Markov Models." In *Proceedings of 1993 International Computer Music Conference.* Computer Music Assosiation, San Francisco, pp.94–97, 1993.

[2] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on Sinusoidal Representation." *IEEE Trans. ASSP*, ASSP-34: 744–754, 1986.

[3] R. J. McAulay and T. F. Quatieri, "Speech Transformations Based on a Sinusoidal Representation."

IEEE Trans. ASSP, ASSP-34: 1449–1464, 1986.

[4] N. Osaka, "Timbre interpolation of sounds using a sinusoidal model." In *Proceedings of 1995 International Computer Music Conference.* Computer Music Assosiation, San Francisco, pp.408–411, 1995.

[5] X. Serra, "Sound hybridization based on a deterministic plus stocahstic decompostion model." In ıProceedings of the 1994 International Computer Music Conference. Computer Music Assosiation, San Francisco, pp.348–351, 1994.

[6] X. Serra, "Musical sound modeling with sinusoids plus noise." In *Musical Signal Processing*, Swets, Lisse. pp. 91–122, 1997.

[7] X. Serra and J. O. Smith III, "Spectral Modeling Synthesis: A Sound Analysis Synthesis System Based on a Deterministic plus Stochastic Decomposition." *Computer Music Journal*, 14-4: 12–24, 1990.

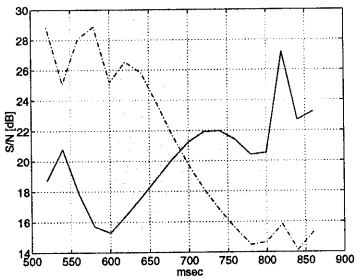
[8] E. Tellman, L. Haken and B. Holloway, "Timbre Morphing Using The Lemur Representation." In ıProceedings of the 1994 International Computer Music Conference. Computer Music Assosiation, San Francisco, pp.329–330, 1994.

接続区間: 600 — 800msec

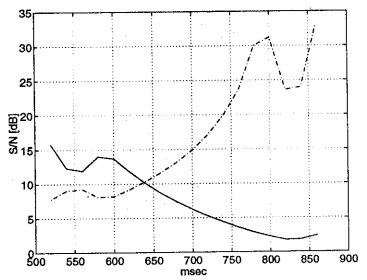
ζ = 100Hz (0 — 12kHz で測定)

—— : 接続前音の定常部分からみた親近度

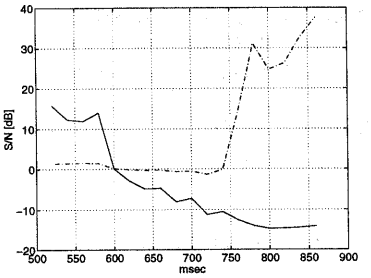
–·– : 接続後音の定常部分からみた親近度



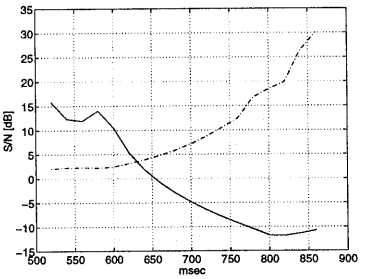
1. 同一音 (フルート a^1) の異なる時点



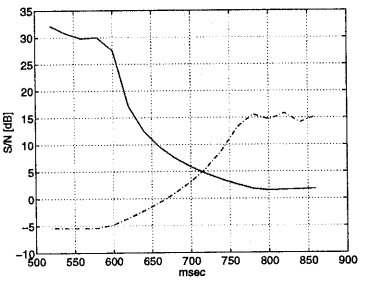
2. 別サンプル (フルート a^1)



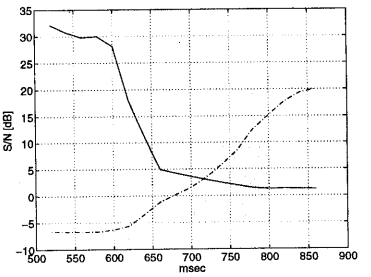
3. 別ピッチ (フルート a^1, a^2=880Hz)



4. 別奏者 (フルート a^1)



5. フルート a^1 ⟷ クラリネット a^1



6. クラリネット a^1 ⟷ c^2

図 4.3: 接続音の時間 — 親近度特性