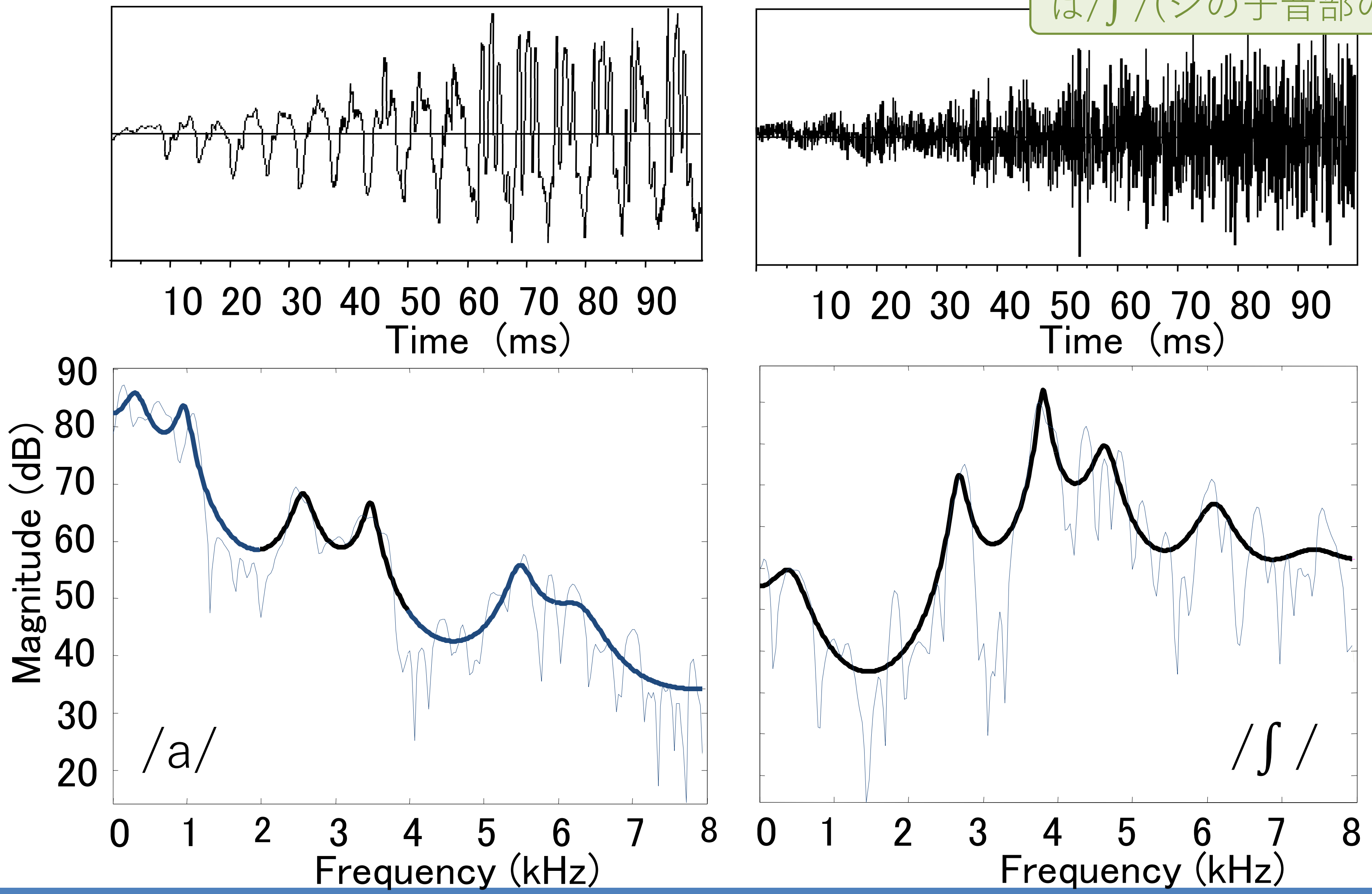
DTW(Dynamic Time Warping) と系列データの距離計算

本資料では，音声に代表される系列データの扱いについて，特に距離の計算とそこで必要になる動的な時間軸の変換法に焦点を当てて述べる。
そこでは，動的計画法が重要な役割を担う。

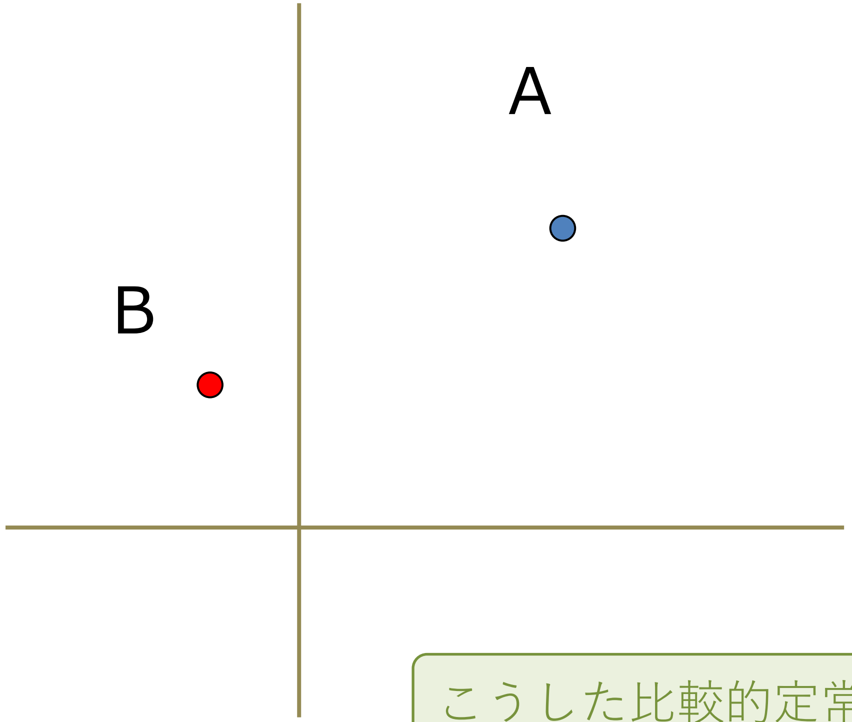
§ /a/と/s/の波形とスペクトル

上段は音声の波形，下段は対数振幅スペクトル，左側は音声/a/，右側は/ʃ/(シの子音部の音)である。



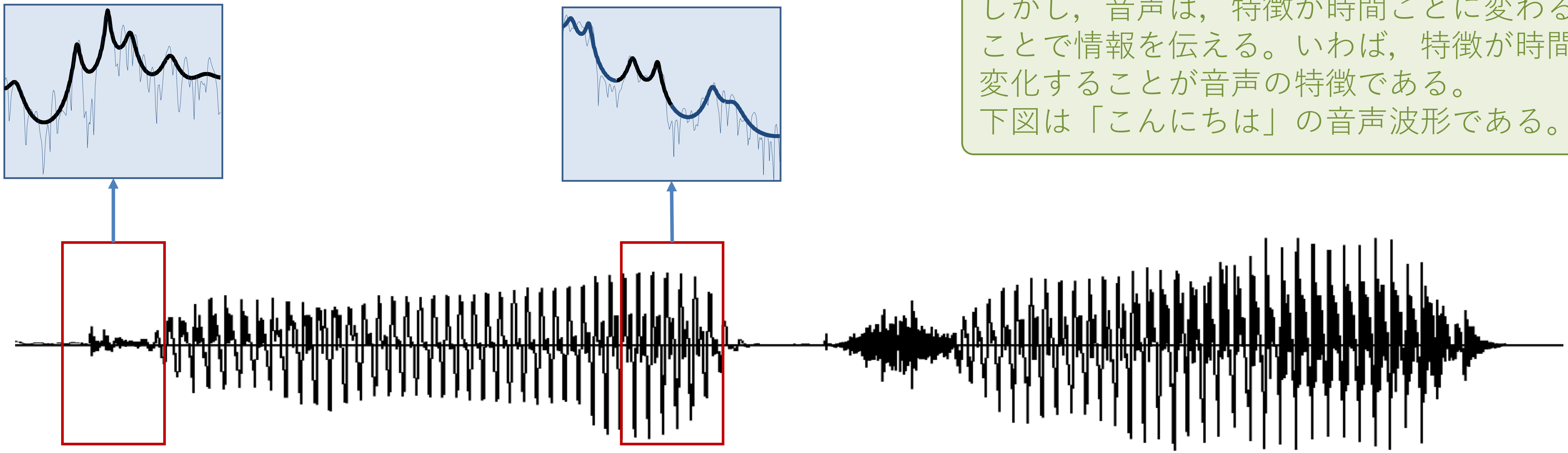
特徴空間上のパターン

静的なパターンであれば，データは特徴分析した後，特徴空間上の１点として表される。



こうした比較的定常的な音は，振幅スペクトルを特徴ベクトルとして，特徴空間上の一点で表現できる。

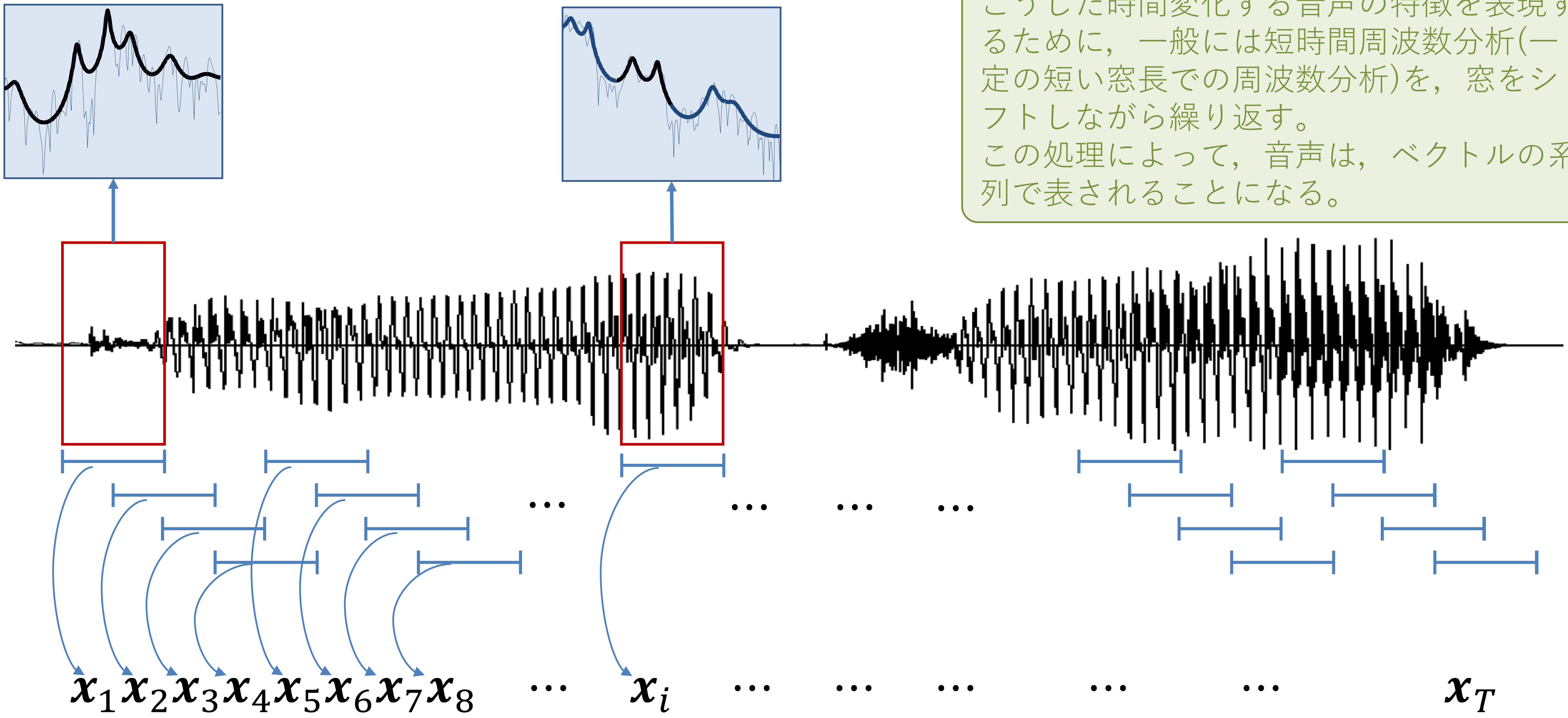
音声データのベクトル列での表現



しかし，音声は，特徴が時間ごとに変わることで情報を伝える。いわば，特徴が時間変化することが音声の特徴である。
下図は「こんにちは」の音声波形である。

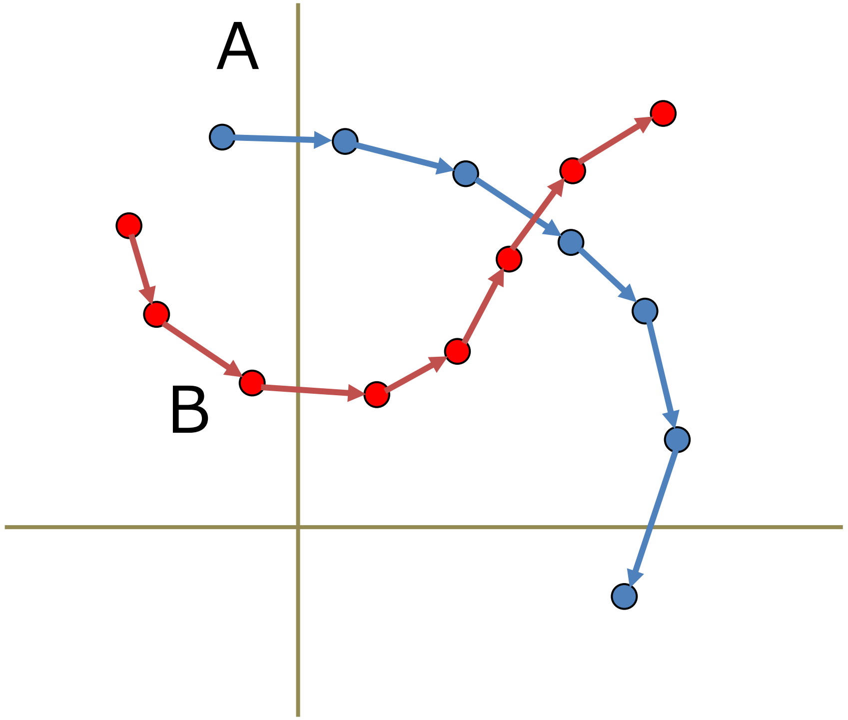
音声データのベクトル列での表現

こうした時間変化する音声の特徴を表現するために，一般には短時間周波数分析(一定の短い窓長での周波数分析)を，窓をシフトしながら繰り返す。
この処理によって，音声は，ベクトルの系列で表されることになる。



特徴空間上のパターン

動的なパターンであれば，データは特徴分析した後，特徴空間上のトラジェクトリ（軌跡）として表される。

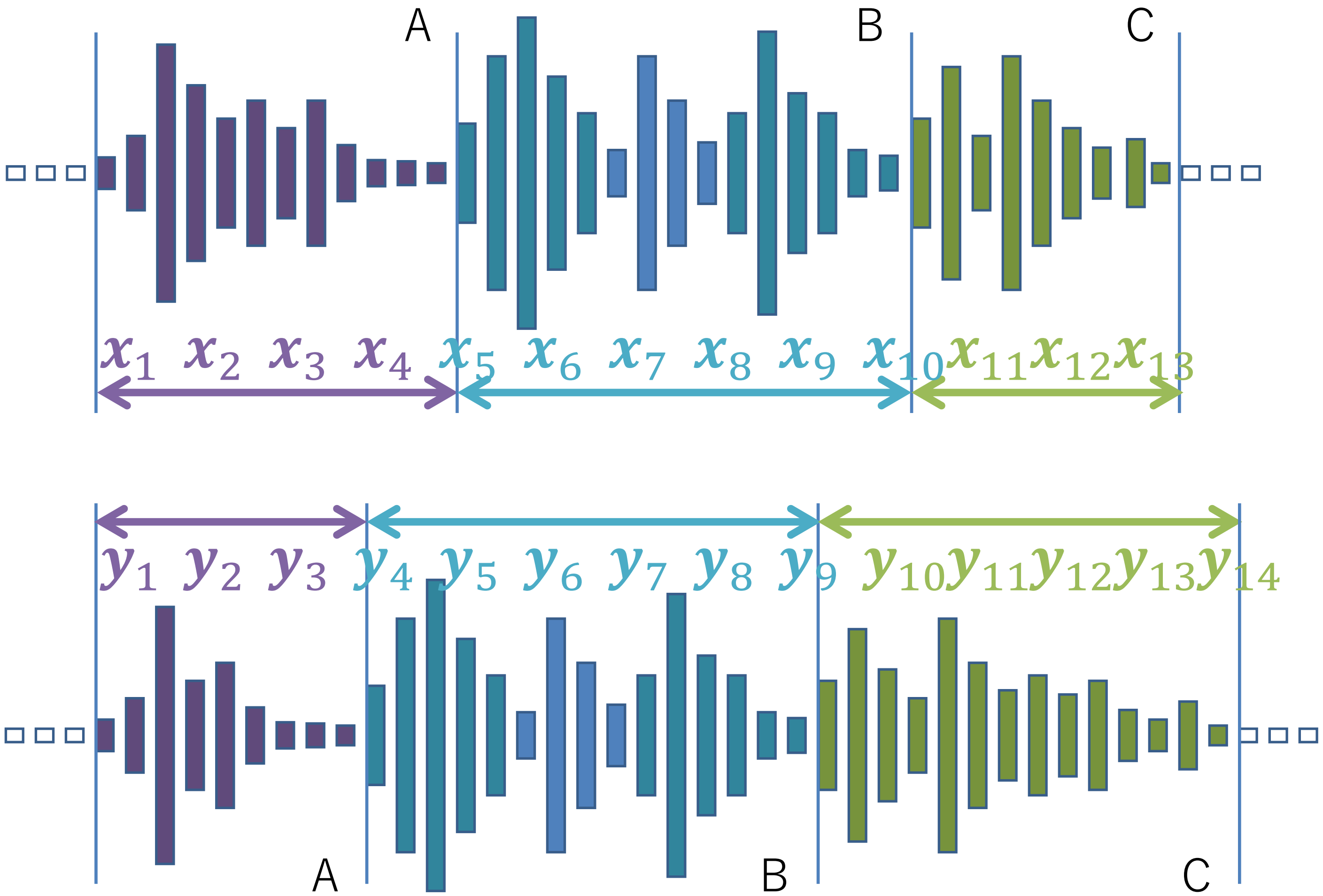


個々のデータが系列で表される＝特徴空間上のトラジェクトリとして表される＝とき，二つのトラジェクトリが似ているかどうか，すなわち，トラジェクトリの距離はどのように調べれば良いかが問題になる。

トラジェクトリが似ているかどうかは，どのように調べるべきか？

§ 時間軸の変換

問題は，異なる時間変化をする系列データに対し距離をどう定義するか。

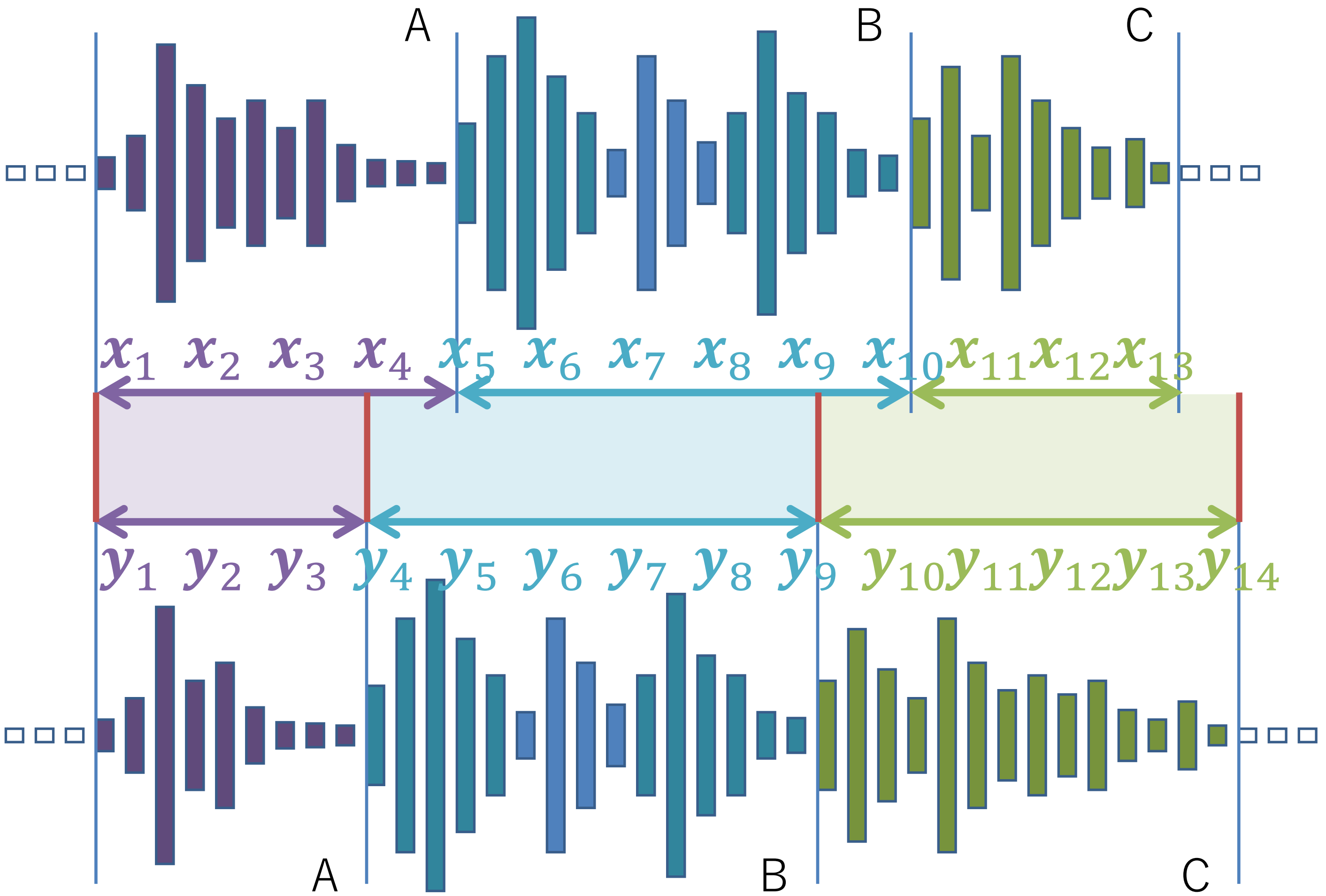


$D(\boldsymbol{x}, \boldsymbol{y})$??

時間変化の異なる系列データをどのように比較するか?

時間軸の変換

基本はフレーム毎の距離を累積すること。
しかし，時間長そのものが違うので対応がつかない。
よって，時間軸の変換が必要になる。



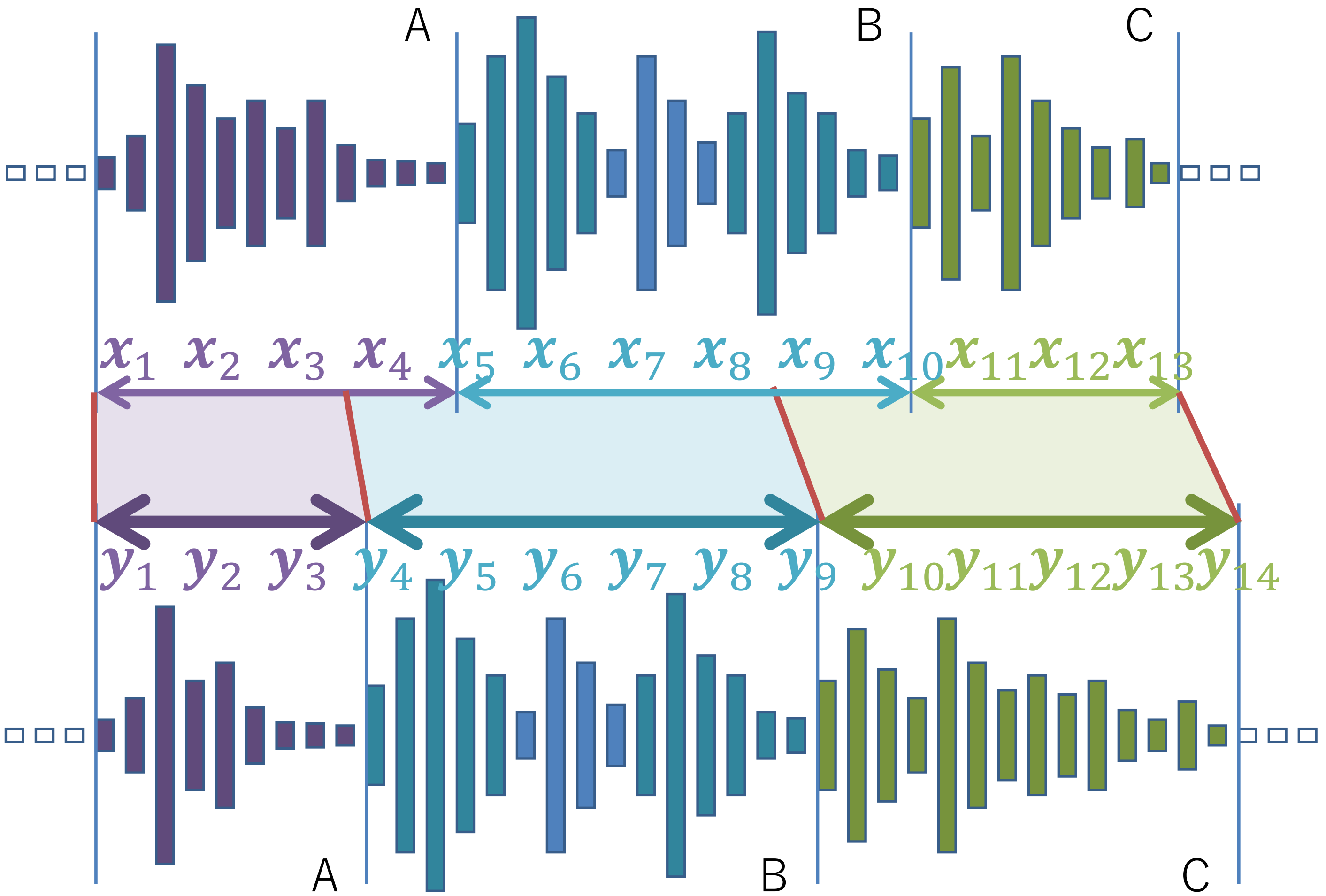
$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N}(x_i - y_i)^2$$

??

基本はフレーム毎の距離を累積。しかし一方の時間軸を変化させないと対応つかない。

時間軸の変換

最も単純な時間軸の変更は時間軸を線形に伸長／収縮させること。しかし，継続長の変化の割合は部分毎に異なるため，線形変換では対応できない。

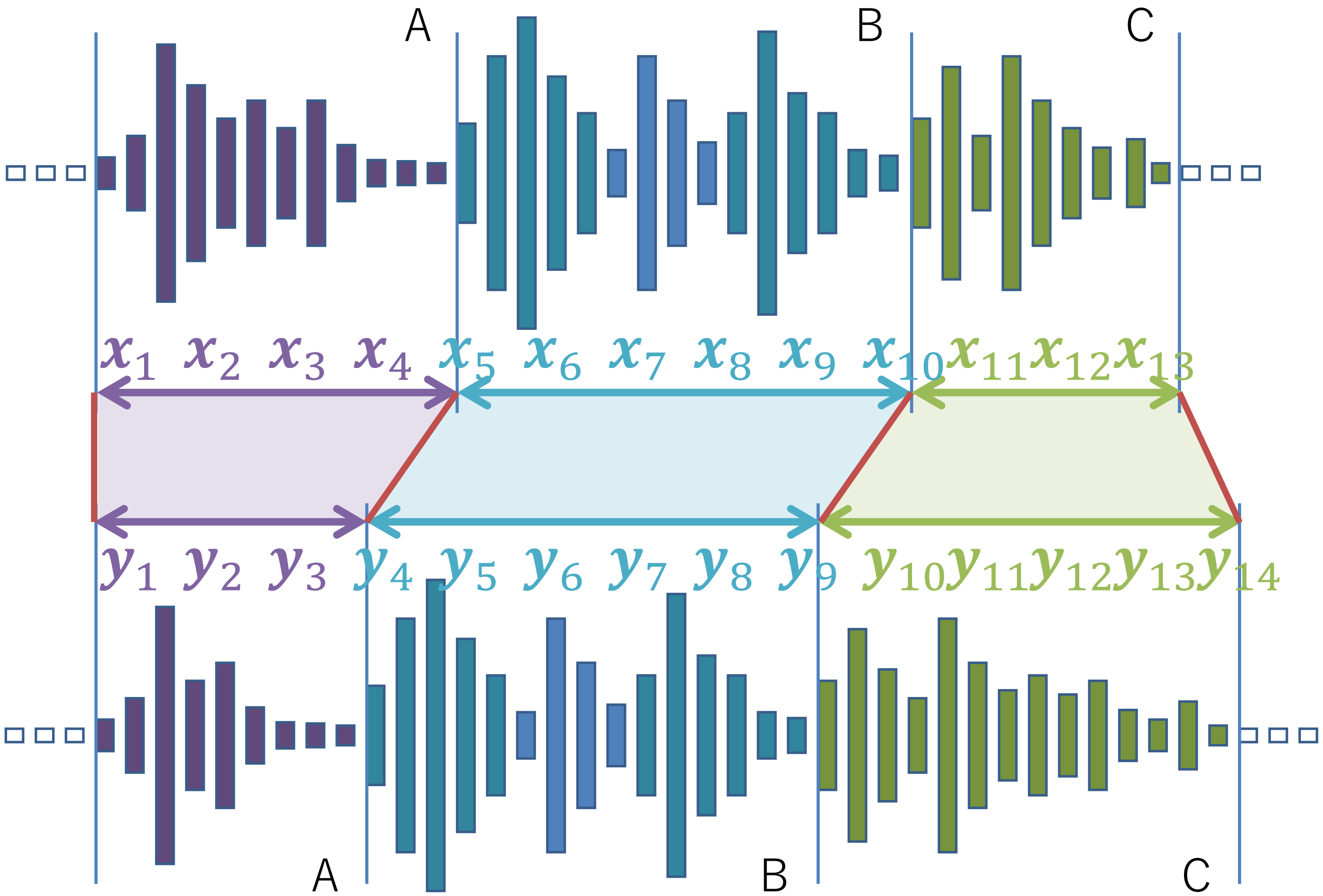


$$D(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} (x_{\alpha i} - y_i)^2$$

??

線形の対応づけは理に適わない（発話の速度変化は発話内で一定ではないから）

時間軸の変換

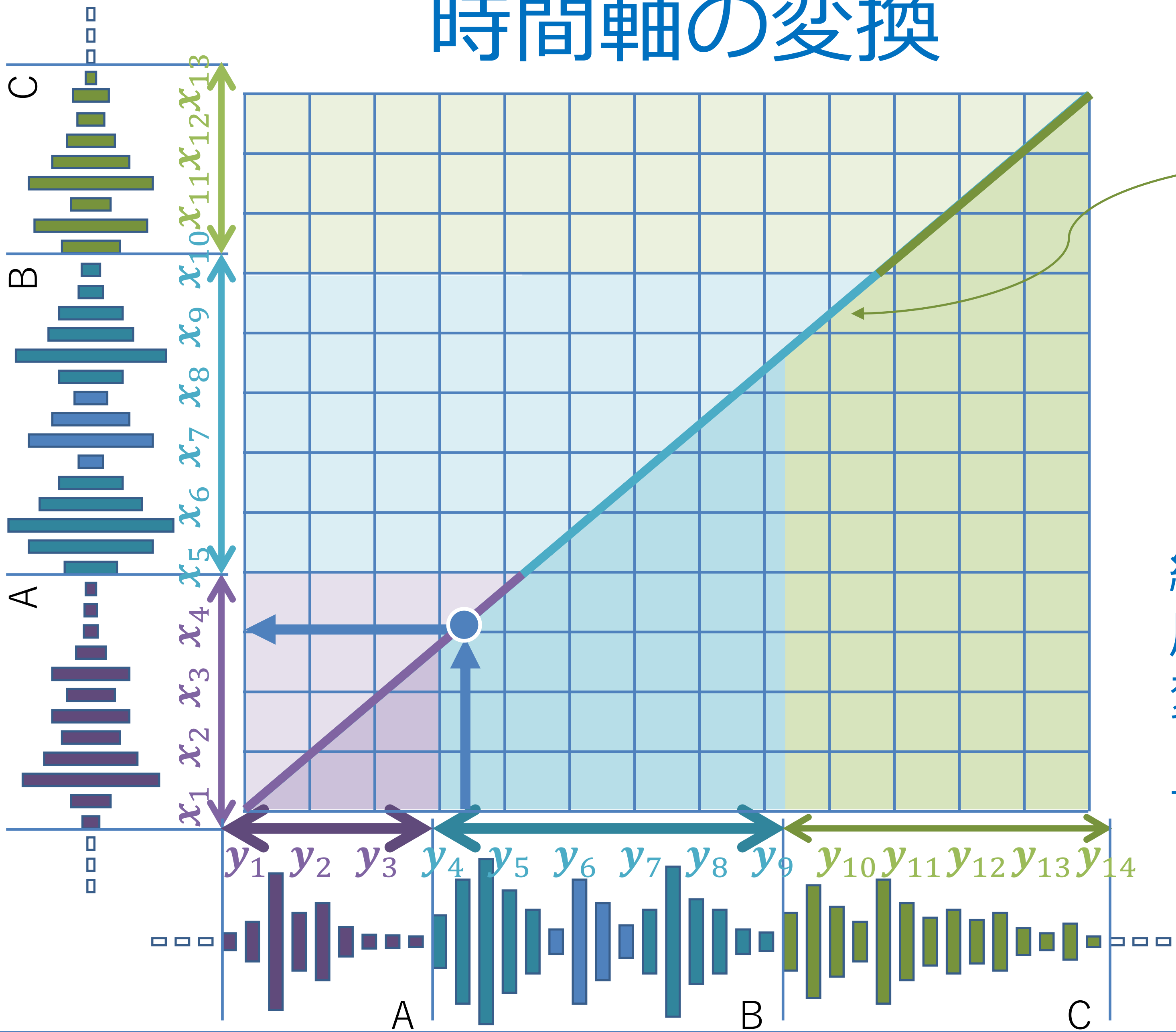


そこで，非線形の時間軸変換を行って，最もフレーム毎の距離の和が小さくなる対応付けを探し，この最小値をもって系列間の距離と定義する。この探索には，動的計画法を用いる。

$$D(\boldsymbol{x}, \boldsymbol{y}) = \min_{f} \sum_{i=1}^{N} (x_{f(i)} - y_i)^2$$

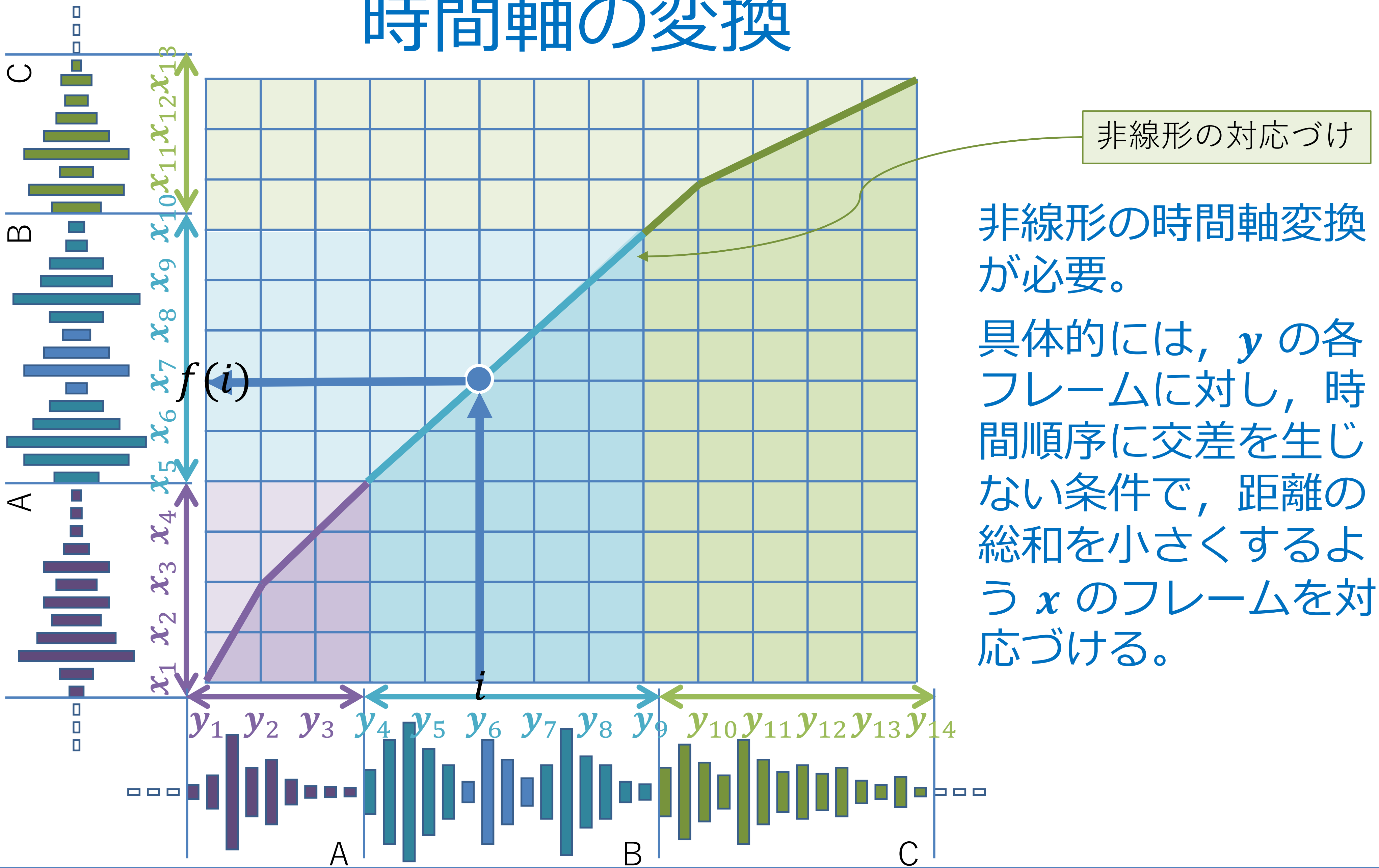
非線形に時間軸を歪ませて，特徴の似た部分を対応づける必要
（両者の距離を一番短くする時間軸変換をした上で，距離を決めるのが妥当）

時間軸の変換



線形の対応づけでは局所的な発話速度の変化に対応できない
（適切な対応関係を実現できない）

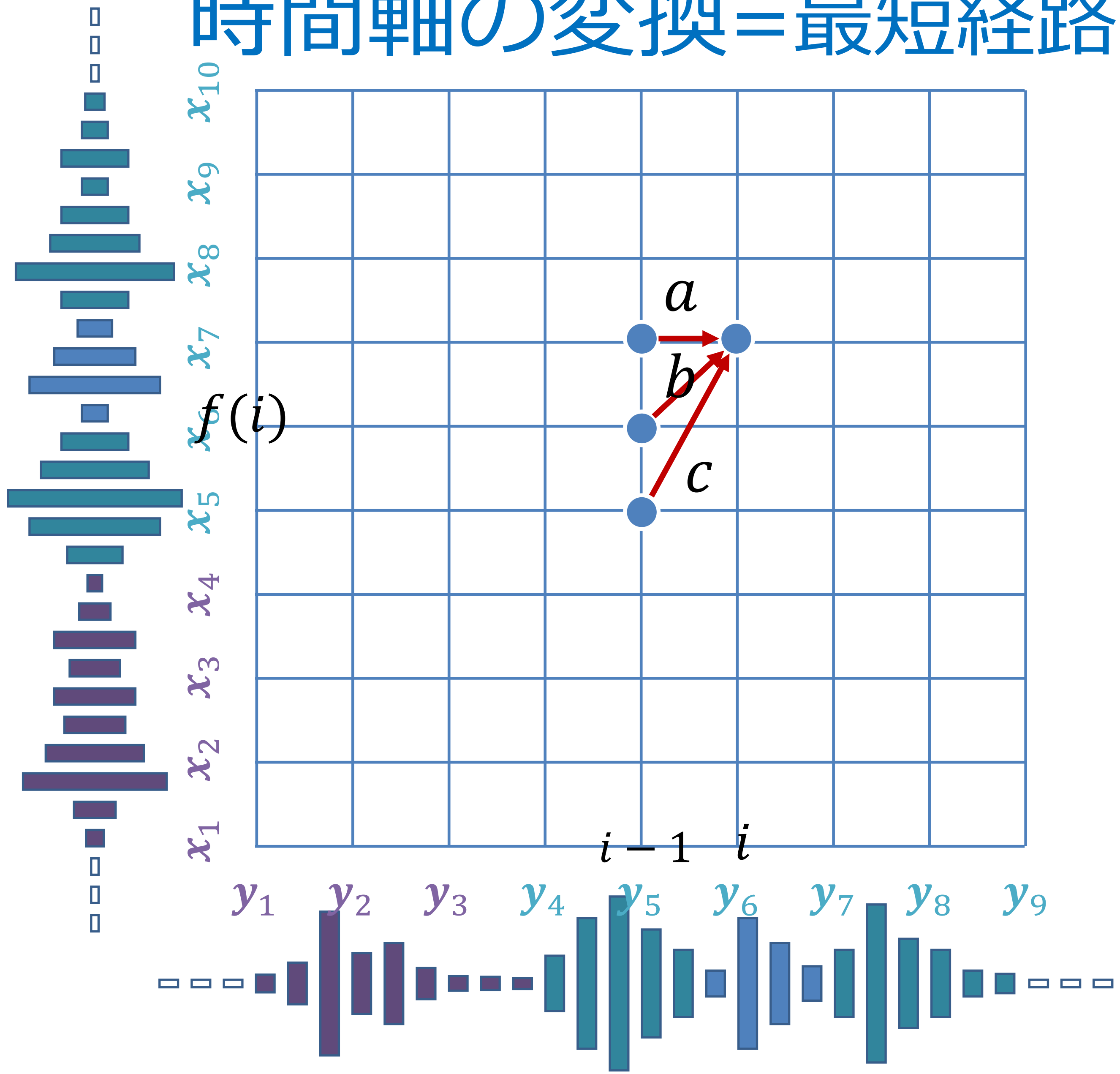
時間軸の変換



非線形の時間軸変換が必要。

具体的には，$\boldsymbol{y}$ の各フレームに対し，時間順序に交差を生じない条件で，距離の総和を小さくするよう $\boldsymbol{x}$ のフレームを対応づける。

時間軸の変換＝最短経路問題と等価



許容される経路の例：

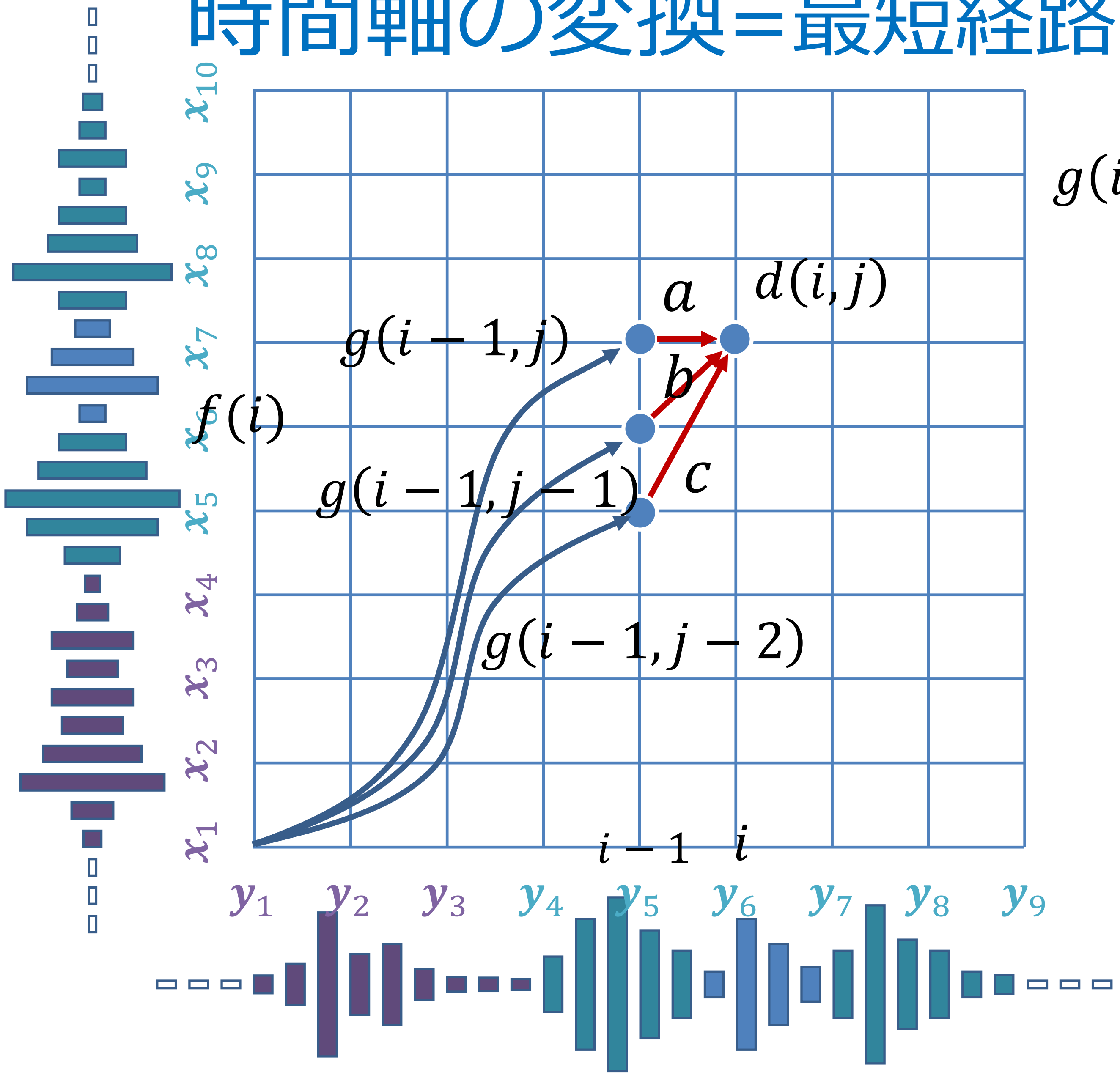
a: $\boldsymbol{y}$ を1フレーム飛ばして $\boldsymbol{x}$ と対応づける（$\boldsymbol{x}$ の1フレームに，$\boldsymbol{y}$ の2フレームを対応づける）

b: フレームを飛ばすことなく $\boldsymbol{x}$ と $\boldsymbol{y}$ を対応づける

c: $\boldsymbol{x}$ を１フレーム飛ばして対応づける

この問題は，最短経路問題と等価である。例えば上の局所経路を許容した上で，DPによって解く。

時間軸の変換＝最短経路問題と等価



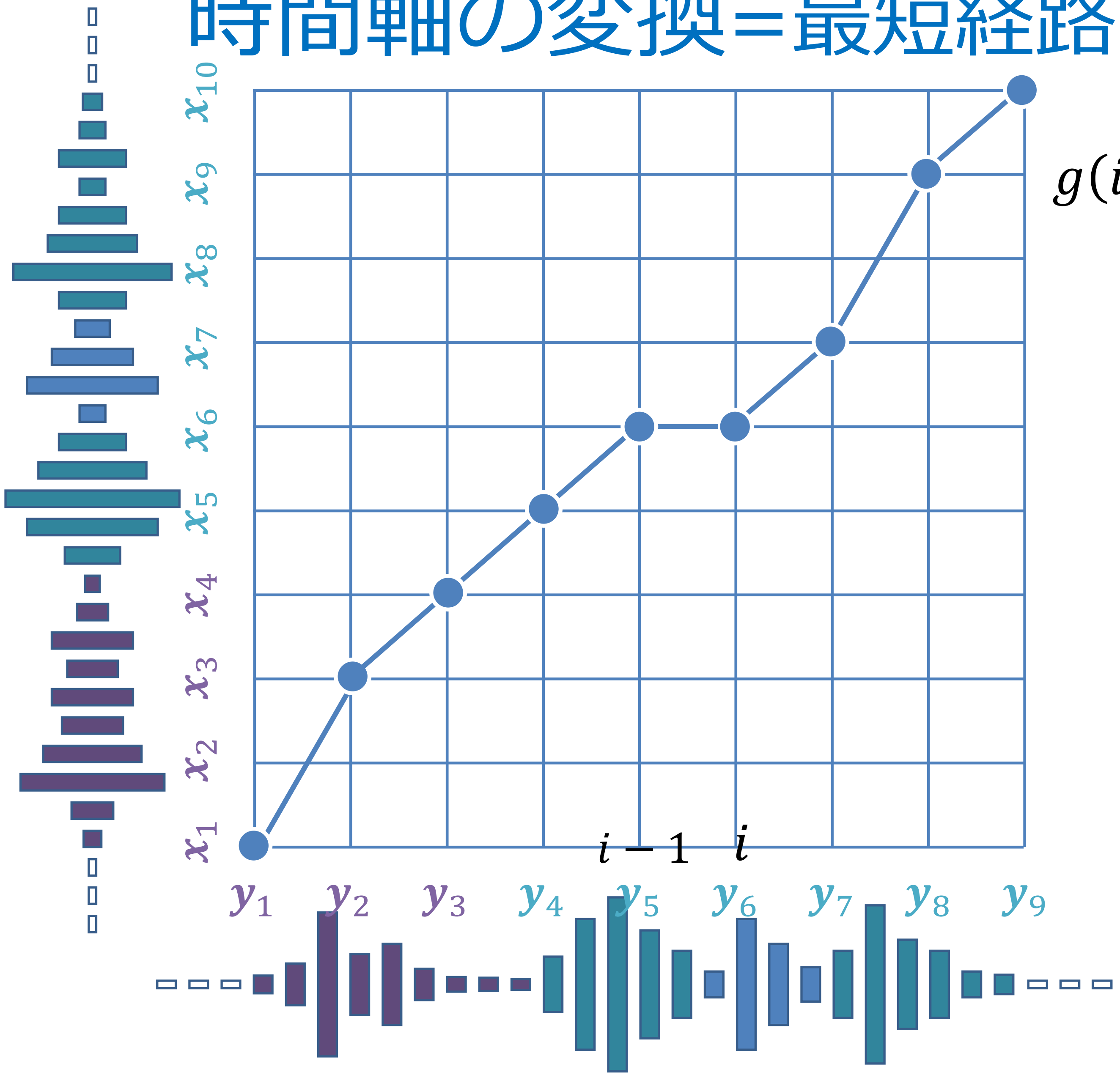
$$g(i,j) = d(i,j) + \min \begin{bmatrix} g(i-1,j) \\ g(i-1,j-1) \\ g(i-1,j-2) \end{bmatrix}$$

$d(i,j)$ ： $\boldsymbol{y}$ の第iフレームと$\boldsymbol{x}$の第jフレームとのフレーム間距離

$g(i,j)$ ：$\boldsymbol{y}$ の1～ iフレームと $\boldsymbol{x}$ の1～ jフレームとの間のフレーム間距離の総和の最小値

漸化式は，上の通りとなる。
よって，(i,j)を，(1,1)から順に，最終フレームまで繰り返せば・・・

時間軸の変換＝最短経路問題と等価



$$g(i,j) = d(i,j) + \min \begin{bmatrix} g(i-1,j) \\ g(i-1,j-1) \\ g(i-1,j-2) \end{bmatrix}$$

$d(i,j)$ ： $\boldsymbol{y}$ の第iフレームと$\boldsymbol{x}$の第jフレームとのフレーム間距離

$g(i,j)$ ：$\boldsymbol{y}$ の１～ iフレーム

$\boldsymbol{x}$と$\boldsymbol{y}$の各フレームの最適な対応づけができる。これができると，その時の距離を求めることができる。

このような形で，DPを用いて時間軸を変換することを，Dynamic Time Warping (DTW)と呼ぶ。
また，DPを用いて（従ってDTWを用いて）系列同士の距離計算を行うことを，DPマッチングと呼ぶ。

§ 距離計算による系列データのパターン認識

認識対象のリファレンスを用意することで，パターン認識ができる。

$$\tilde{w} = \min_{w} \mathrm{DP}(X, R_w)$$

X：入力データ系列

R_w：クラス w のリファレンスのデータ系列

$\mathrm{DP}(A, B)$：系列Aと系列BをDPマッチングしたときの距離

認識対象となるクラス毎にリファレンスを用意し，これらと入力データとのDPマッチングを行って，最小距離を与えるクラスを選べば，系列データのパタン認識ができる。

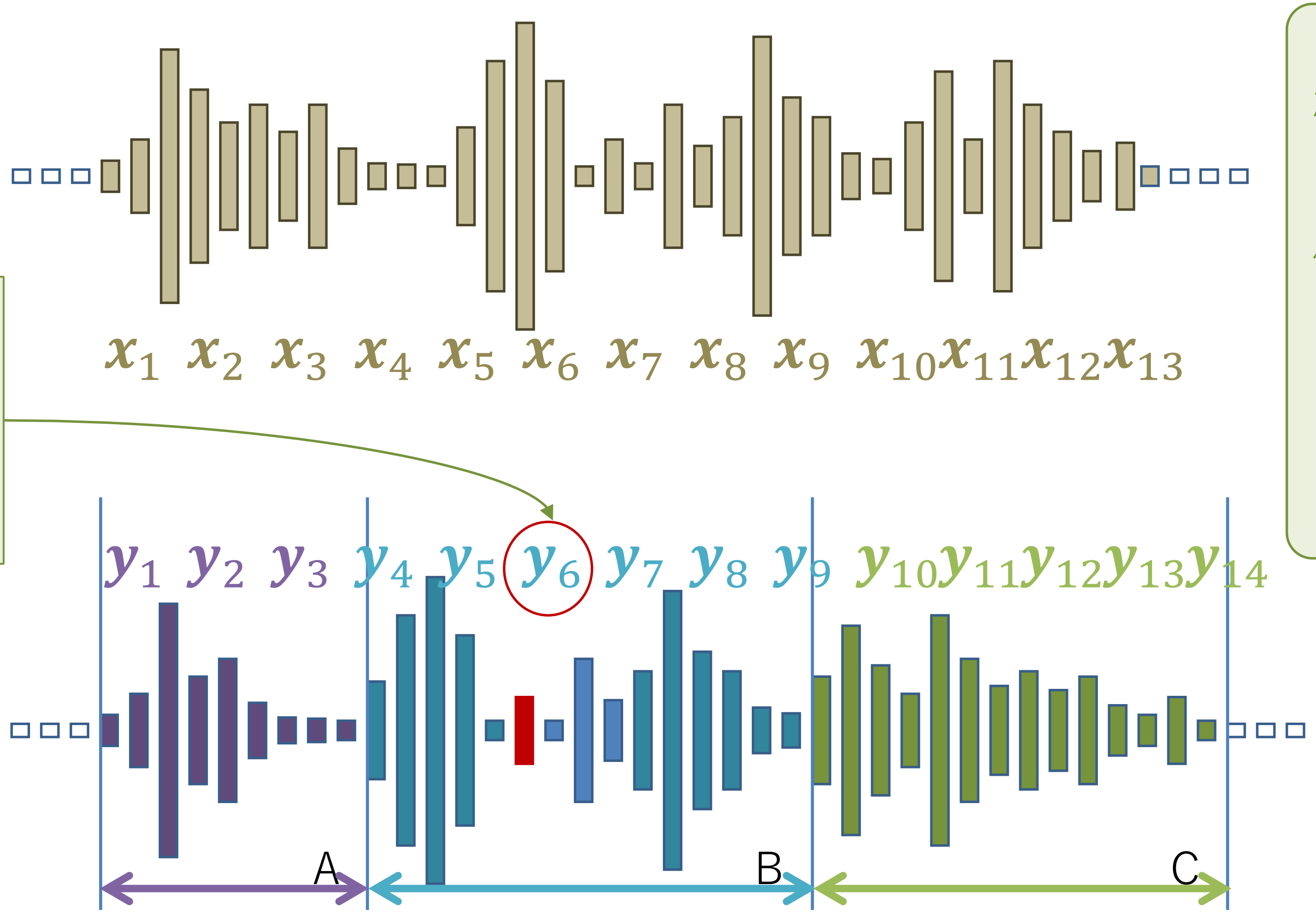
ただし，実際問題として単純なDPマッチングがパターン認識に用いられることはない。「リファレンスとの距離でパターン認識」という考え方自体が成立しない。

しかし，おおよそ全ての系列パターン認識は，なんらかの形でDPの要素を含む。

ただし，このような形で，DPマッチングを用いてパターン認識をすることはほとんどない。

（少数のリファレンスで，クラスのデータ分布を代表することは困難だから。）

§ 時間アラインメント（Time Alignment）



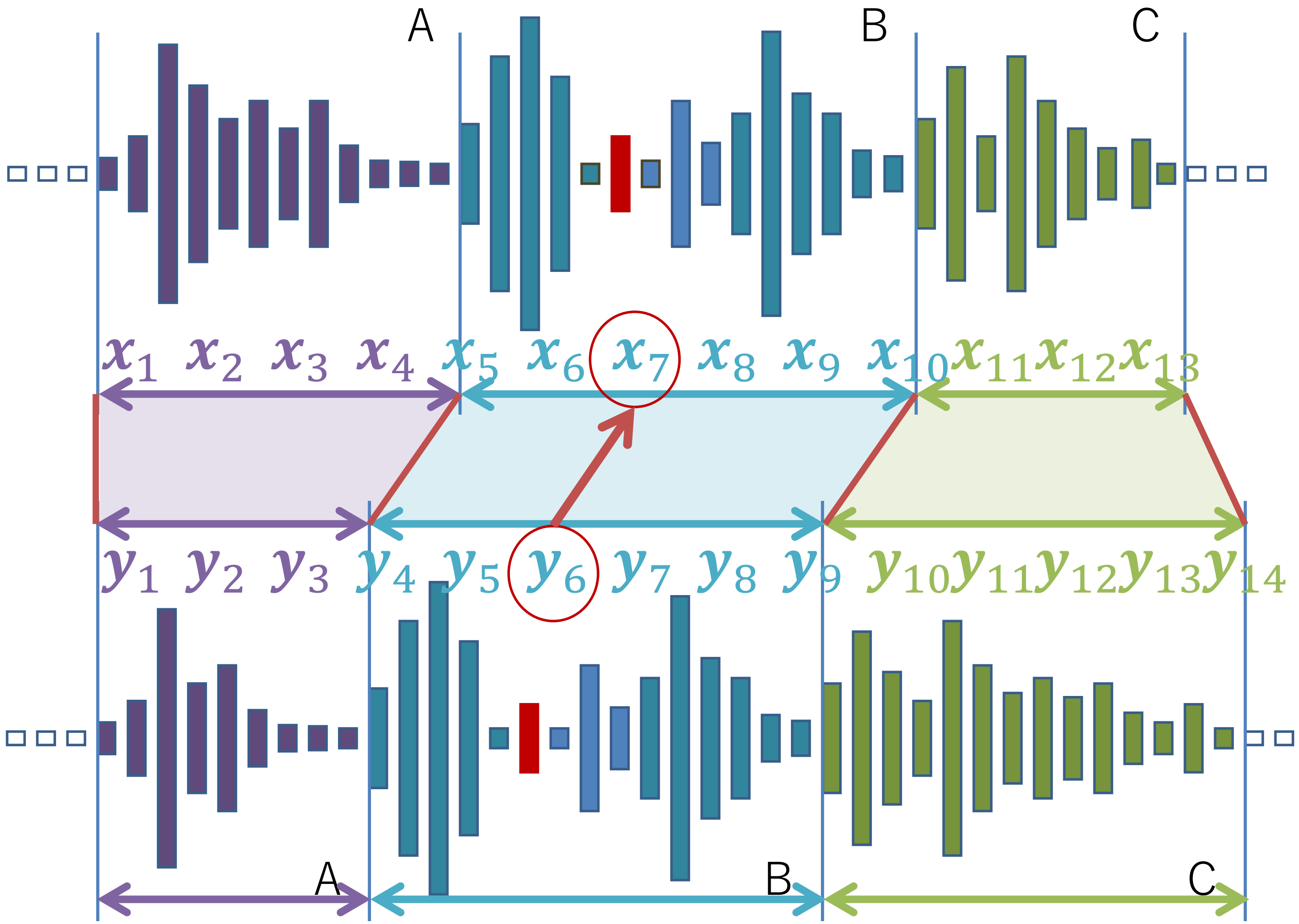
$\boldsymbol{y}$ の赤のフレームに対応するデータは，$\boldsymbol{x}$の中のどこにあるのか？

DTWは，特定のイベントの生起時間を検出するのに使うことができる。

例えば， $\boldsymbol{x}$ はラベルなしデータ。 $\boldsymbol{y}$ は検出対象(図中赤のデータ)に関するラベルつきのデータ とする。
また，$\boldsymbol{x}$は$\boldsymbol{y}$と同じ音素の並びと仮定する。

データ中の特定のイベントを探す問題

時間アラインメント（Time Alignment）



ここで２つの系列のDPマッチングを行えば，$\boldsymbol{x}$ と $\boldsymbol{y}$ のどのフレーム同士が対応するかがわかる。
よって，$\boldsymbol{x}$ の中でのイベントの位置を決めることができる。

このようにして２つのデータのフレームの対応関係を求め，時間構造を合わせることを「強制アライメント(Forced Alignment)」あるいは「時間アライメント」と呼ぶ。
強制アライメントによって，特定のイベントの出現位置を検出することができる。

§ まとめ

- 音声に代表される，時間毎に特徴が変化する系列データの距離計算には，非線形の時間軸変換が必要になる。この変換を Dynamic Time Warping (DTW) と呼ぶ。
- DTWは動的計画法によって実現される。
- DTWを用いて系列の距離計算を行うことをDPマッチングと呼ぶ。
- ラベルのない系列データと，ラベルありの系列データのフレームの対応関係を調べることで，ラベルなしデータにラベルを与える処理を，強制アライメント (Forced Alignment) と呼ぶ。
- 強制アライメントによって，ラベルなしデータ中の特定のイベントの出現位置を検出することができる。
- DPマッチングは，強制アライメントにも利用される。