人間の声？それともコンピュータ？
音声情報処理におけるディープラーニング最前線

国立情報学研究所　コンテンツ科学研究系
高木 信二

講義の構成

- 音声合成
  - 文章の原語情報の抽出
    - 音素，形態素，アクセント
  - 音声の音響特徴量の抽出
    - スペクトル，スペクトル包絡，F0
  - ディープラーニング
    - 言語情報と音響特徴量の対応付け
    - 様々なニューラルネットワーク
- 応用

音声情報処理技術

- 音声認識
  - 字幕付与，会議録の書き起こし
- 音声検索
  - Google voice search
- 音声対話
  - しゃべってコンシェルジュ，Siri, Amazon Echo
- 音声合成
  - スクリーンリーダ，ナレーション，ラジオ，ボーガロイド

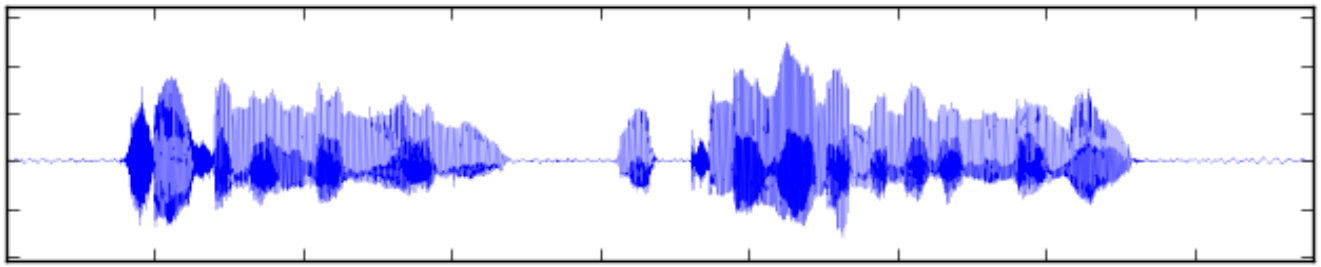
音声情報処理技術の普及
ディープラーニングによる性能向上

テキスト音声合成

- 入力テキストを音声へ変換

変換

小さな鰻屋に，熱気のようなものがみなぎる



- なぜ必要？
  - 情報の取得（カーナビ，携帯端末での音声対話）
  - 視覚・音声の障害（スクリーンリーダ，会話補助）
  - エンターテインメント（歌声，ナレーション）

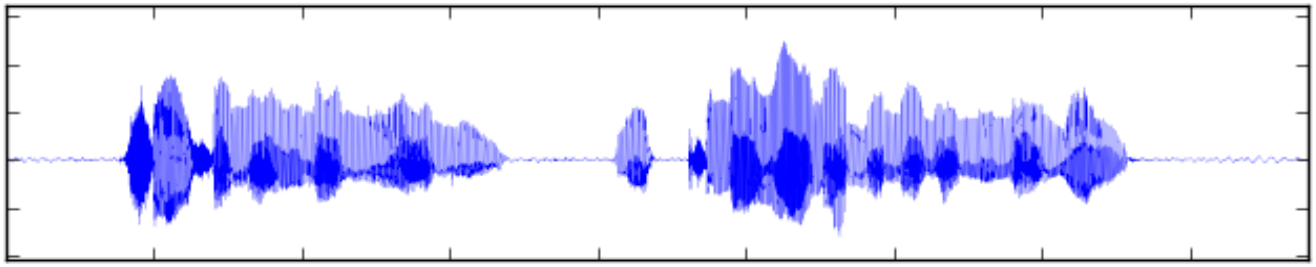
音声合成：文章と音声の対応付け

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出　　言語処理

言語的情報と音響的特徴量の対応付け　　機械学習

音声の音響的特徴量抽出　　信号処理



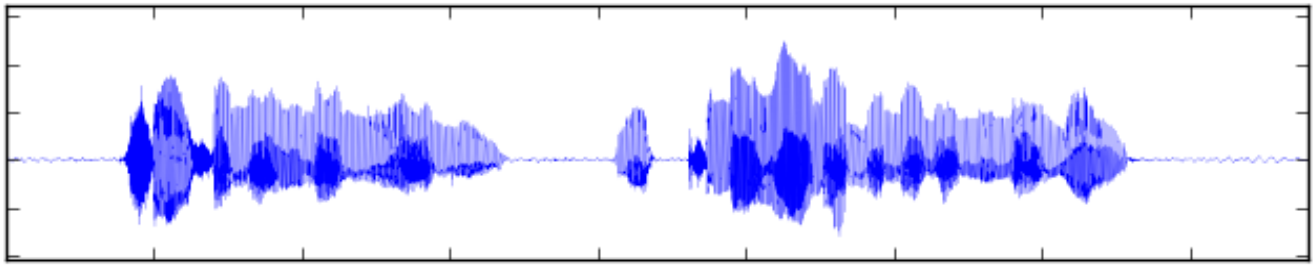
音声合成：文章と音声の対応付け

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出　　言語処理

言語的情報と音響的特徴量の対応付け　　機械学習

音声の音響的特徴量抽出　　信号処理



音素

- 単語よりも小さい音声の単位（sub-word）
- 単語の意味が区別される最小単位
  - 「滝（/t/ /a/ /k/ /i/）」と「柿（/k/ /a/ /k/ /i/）」
  - /t/ と /k/ で区別される. /t/ と /k/ は音素
- 言語的情報を分析する単位としてよく利用される

文		赤い空
句	アクセント句	あかいそ｜ら
語	単語	赤い 空
音節	ひらがな	/a/, /ka/, /i/, /so/, /ra/
音素	母音・子音	/a/ /k/ /i/ /s/ /o/ /r/

音素セット

- 日本語の場合（Wikipediaから引用）

母音	/a/, /i/, /u/, /e/, /o/
子音	/k/, /s/, /t/, /c/, /n/, /h/, /m/, /r/, /g/, /z/, /d/, /b/, /p/
半母音	/j/, /w/
特殊モーラ	/N/, /Q/, /H/

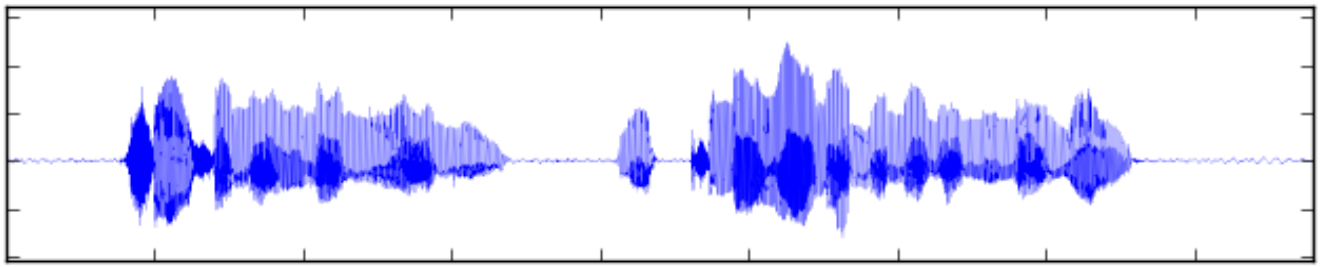
- 言語によって音素セットは異なる
- 発音辞書（単語と音素の対応付け）の利用
- 日本語の単語から音素への変換は比較的容易

形態素，アクセント

- 形態素解析
  - 名詞や動詞といった品詞
  - 単語のかかり受けや句の構造解析
- 形態素解析 → 更に多くの言語的情報を抽出
  - ポーズや呼気の位置
  - アクセント句境界・アクセント核の推定
    - 例：音声，合成，音声合成
    - ただつなげるだけでは不十分

変換

小さな鰻屋に，熱気のようなものがみなぎる

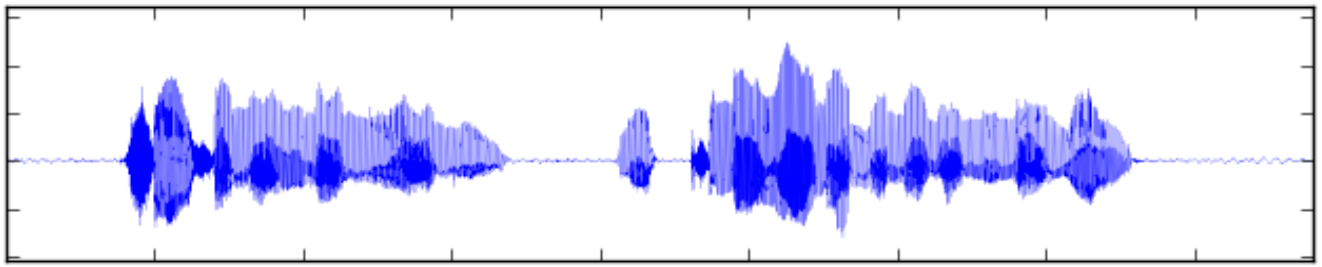


形態素，アクセント

- 形態素解析
  - 名詞や動詞といった品詞
  - 単語のかかり受けや句の構造解析
- 形態素解析 → 更に多くの言語的情報を抽出
  - ポーズや呼気の位置
  - アクセント句境界・アクセント核の推定
    - 例：音声，合成，音声合成
    - ただつなげるだけでは不十分

変換

音素，形態素解析，アクセント等の情報



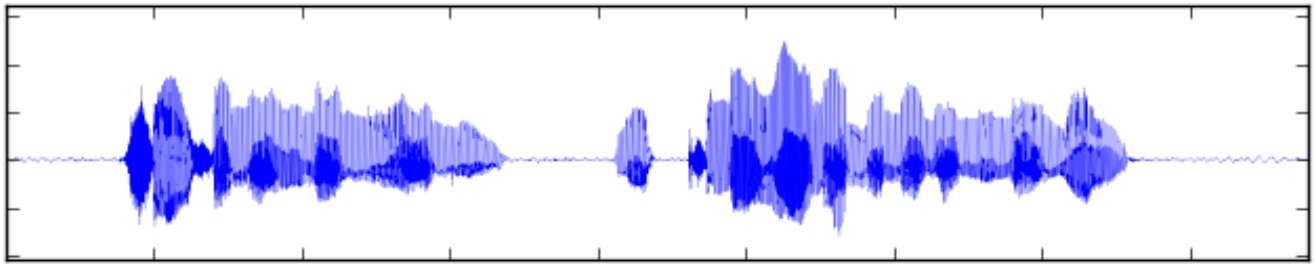
音声合成：文章と音声の対応付け

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出　　言語処理

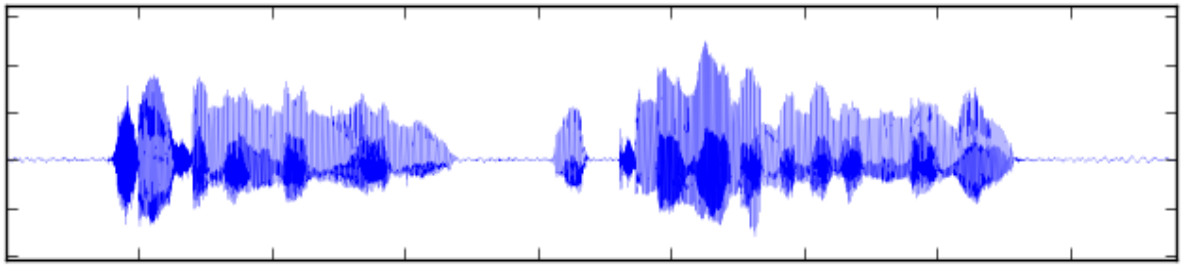
言語的情報と音響的特徴量の対応付け　　機械学習

音声の音響的特徴量抽出　　信号処理



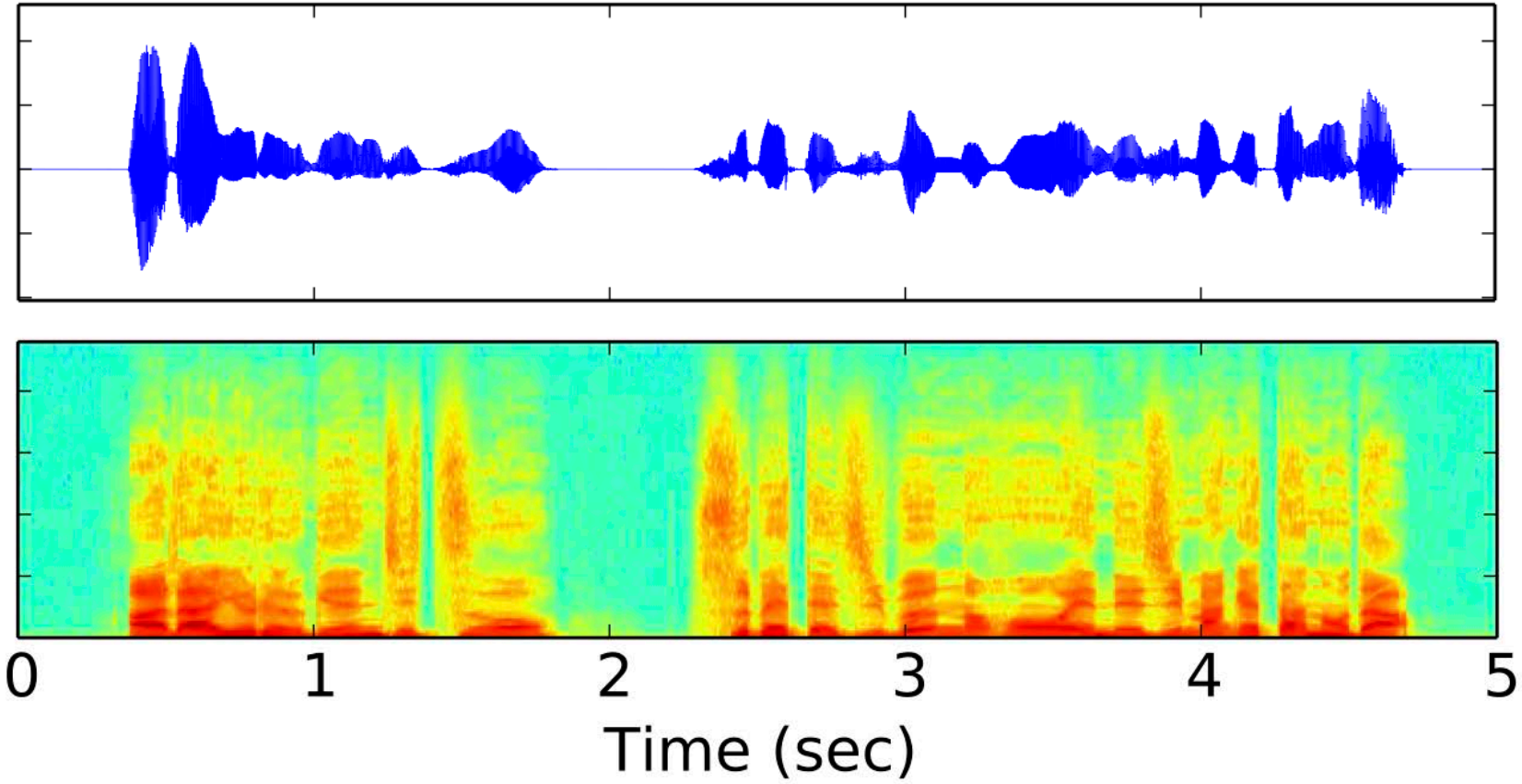
音声

- 空気の振動
- 音声の大きさ
  - 小さい，大きい
- 音声の長さ
  - 長い，短い
- 音声の高さ
  - 高い，低い
- 音声の音色（声色）
  - 太い声，低い声，子供っぽい声，大人っぽい声，・・・
  - 音素の違い（例：/a/ と /i/）
  - 大きさ，長さ，高さが同じでも異なる音に聞こえることも



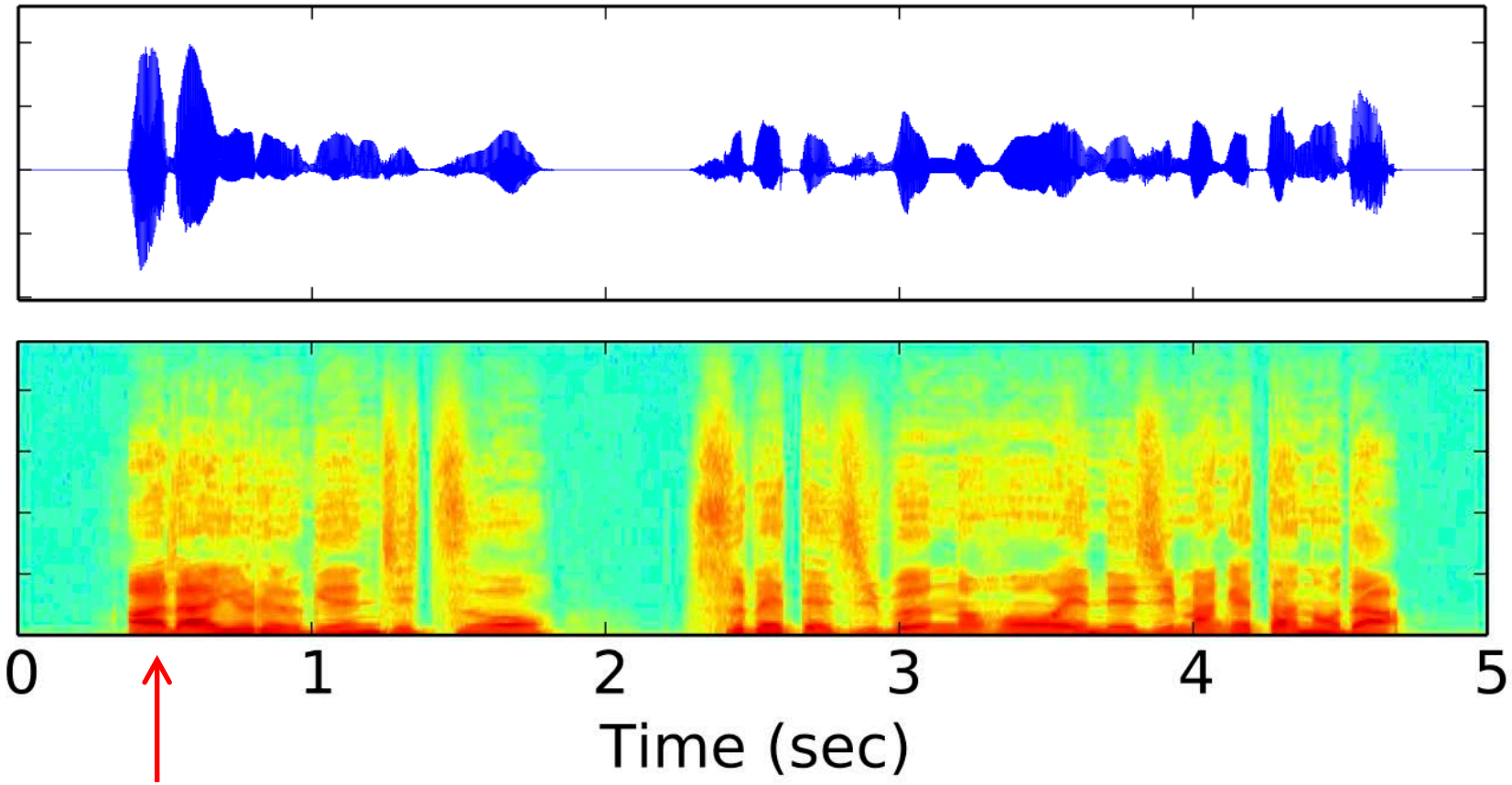
音声波形の周波数変換（1/2）

- スペクトル：波形信号を周波数表現したもの
- 周波数成分のピークが異なる



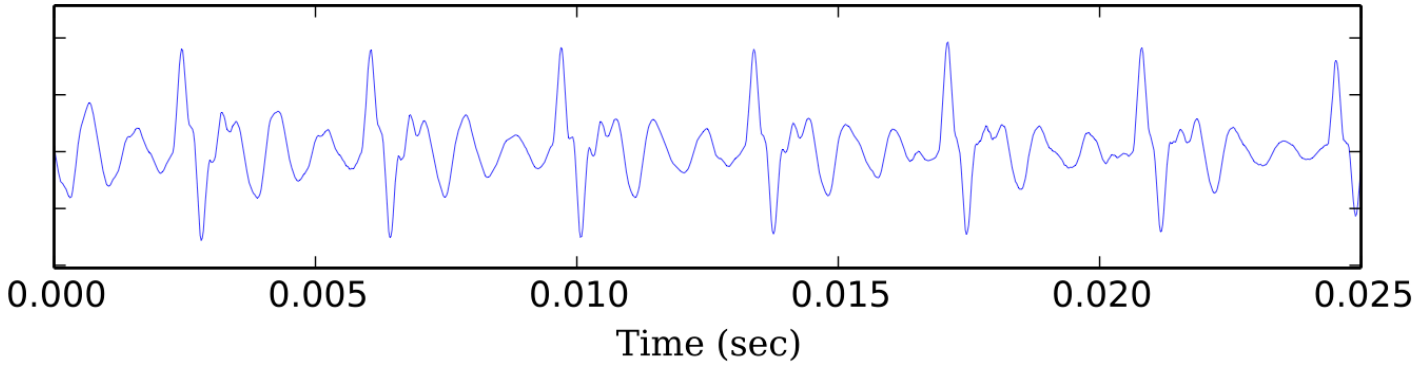
音声波形の周波数変換（1/2）

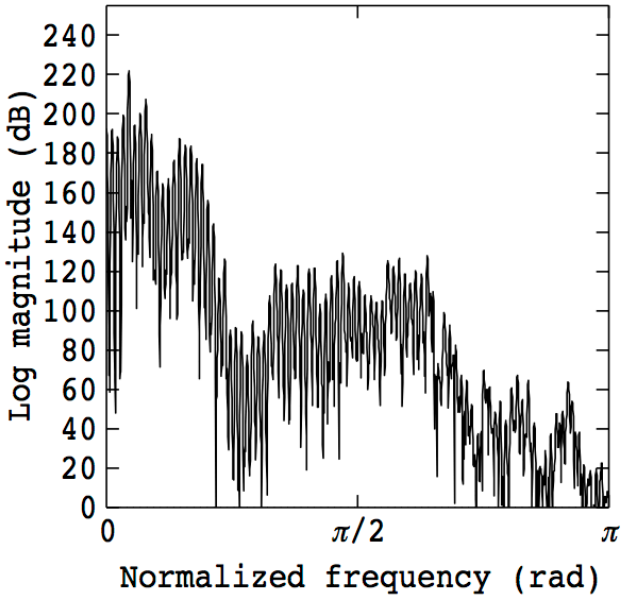
- スペクトル：波形信号を周波数表現したもの
- 周波数成分のピークが異なる

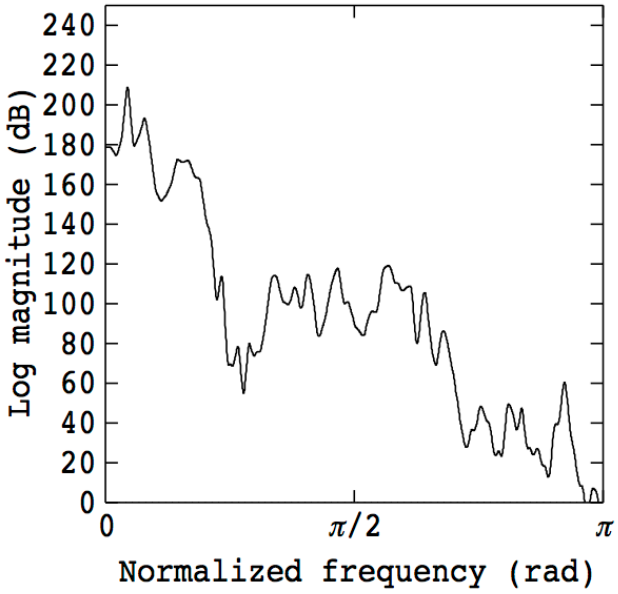


音声波形の周波数変換（2/2）

- 短い区間（この例では25ミリ秒）を分析



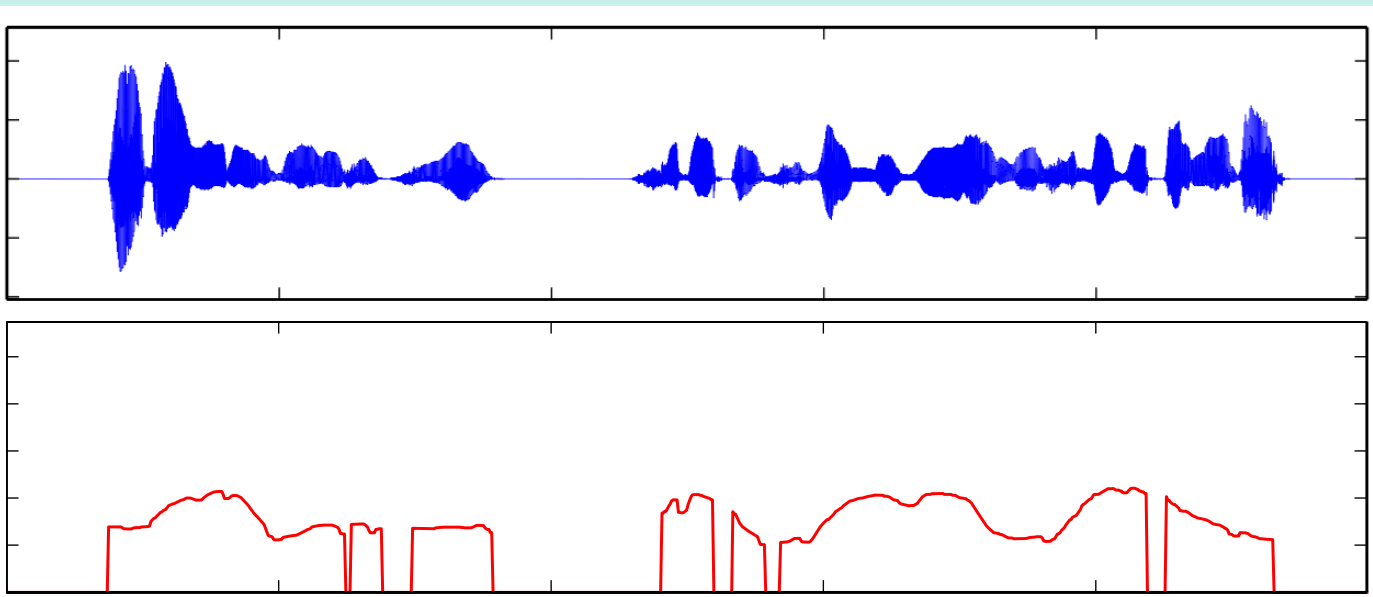




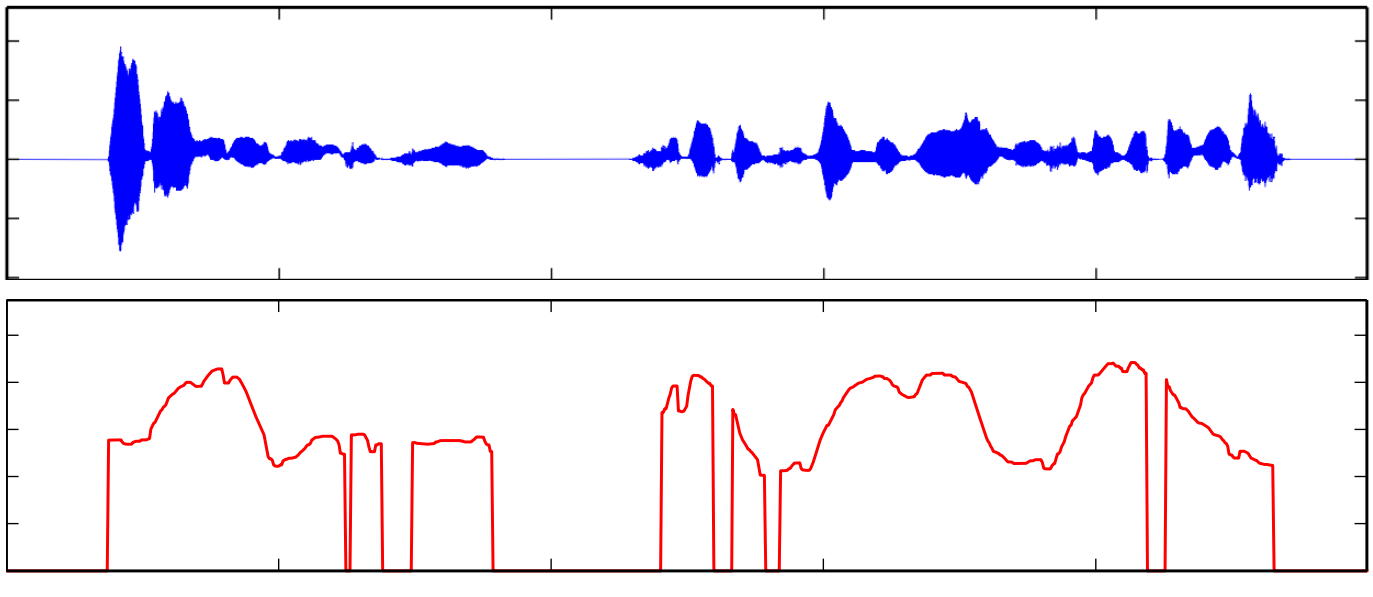
スペクトル：音声の情報を含む

包絡：音色の情報を含む

声の高さ（1/2）



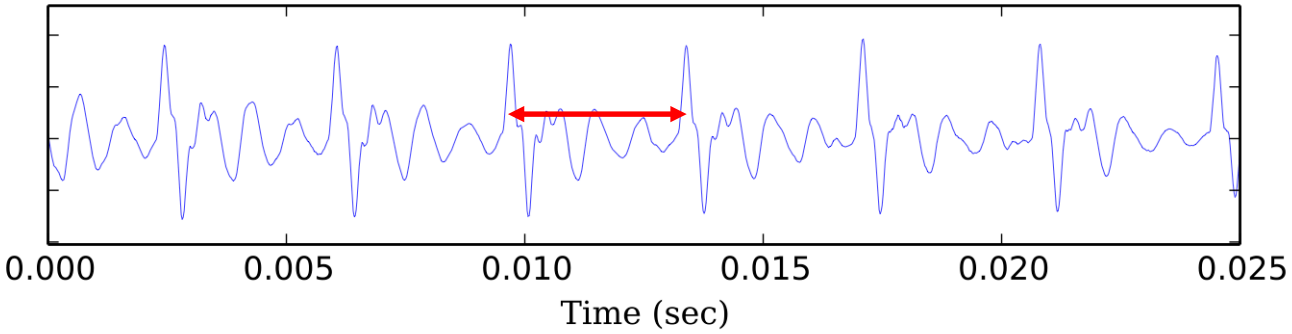
低い声



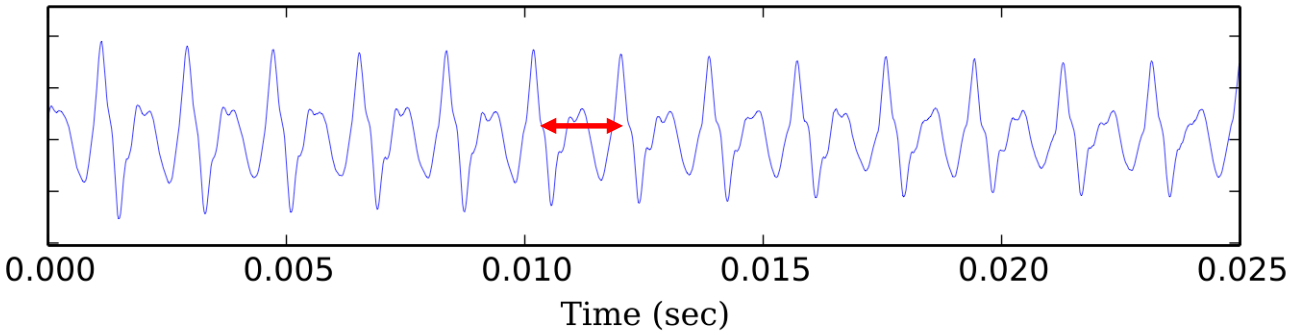
高い声

声の高さ（2/2）

- 似た波形が表れる間隔に注目
  - 音声の基本周期（T_0），基本周波数（$F_0 = 1/T_0$）
- 高い声の方が間隔が短い

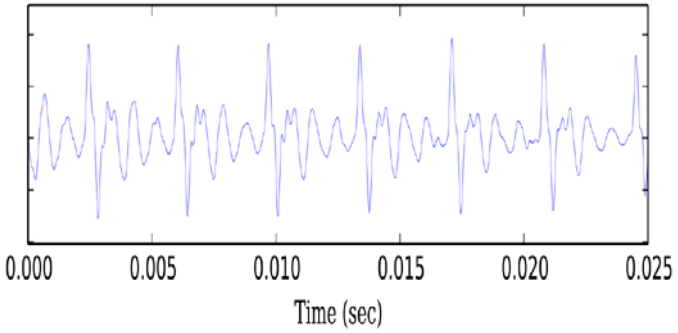


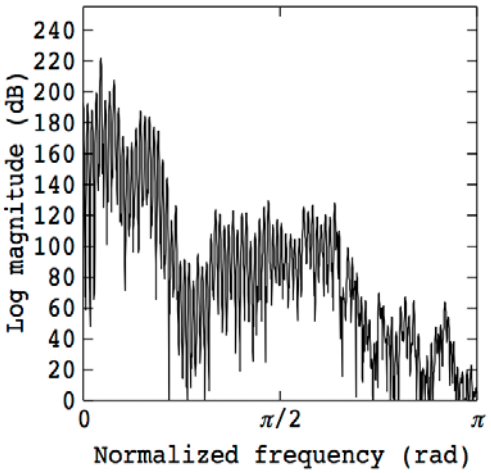
低い声

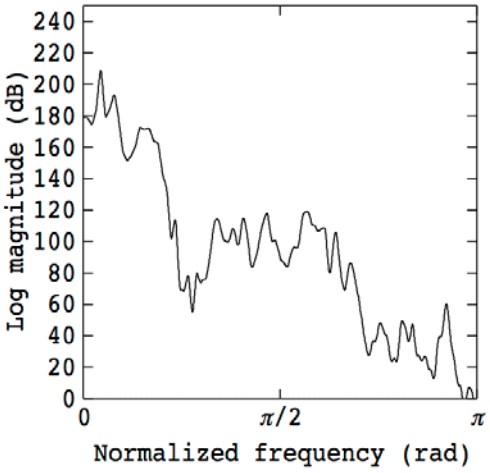


高い声

音声の特徴量まとめ







＋

基本周波数

波形
（扱いが難しい）

分解
（扱いが容易）

- 音声の特徴量は最終的に波形に変換される
  - 変換ミスが発生することも → 波形やスペクトルを扱う

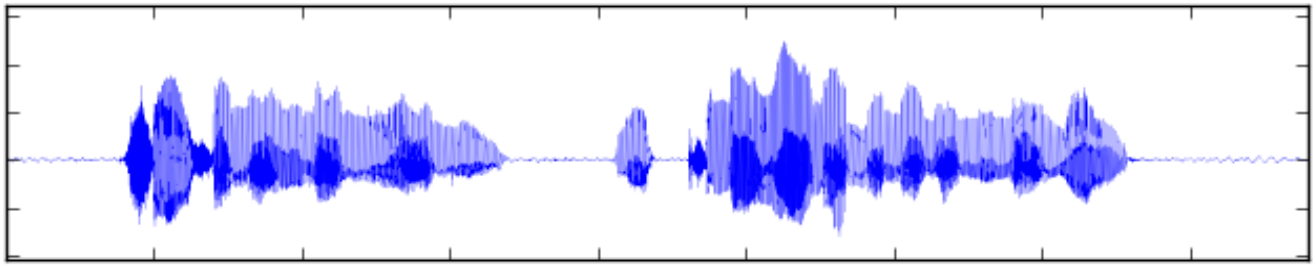
音声合成：文章と音声の対応付け

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出　　言語処理

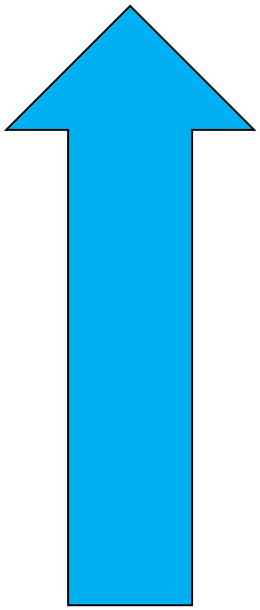
言語的情報と音響的特徴量の対応付け　　機械学習

音声の音響的特徴量抽出　　信号処理



ディープラーニングによる文章・音声の対応付け

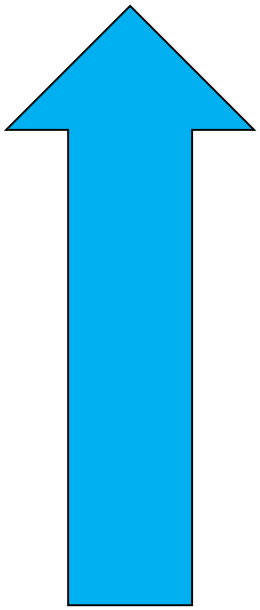
音声



テキスト

ディープラーニングによる文章・音声の対応付け

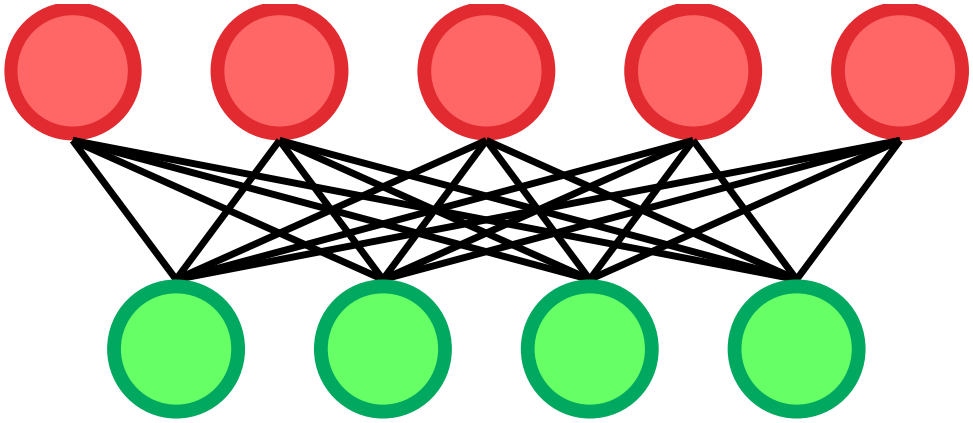
音声波形，または，音声特徴量



音素，形態素，アクセント等の情報

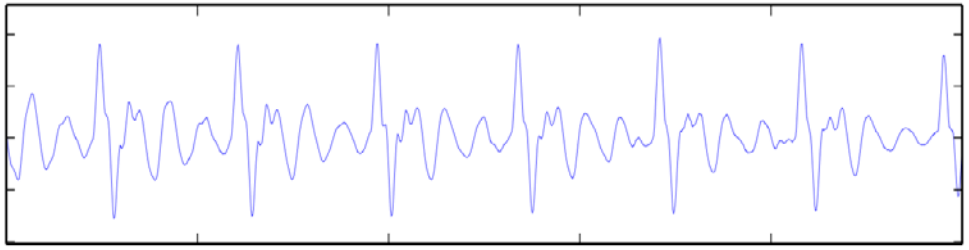
ディープラーニングによる文章・音声の対応付け

音声波形，または，音声特徴量



音素，形態素，アクセント等の情報

ディープラーニングによる文章・音声の対応付け



出力

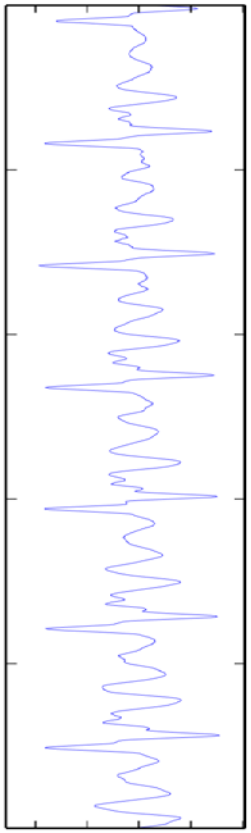
入力

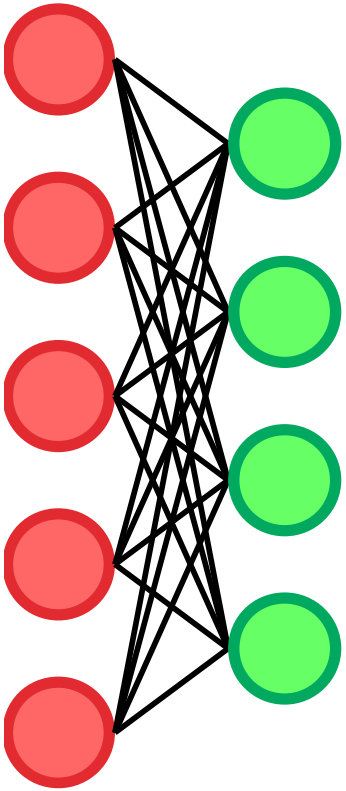
1 0 0 1

/a/ /i/ /u/ アクセント

簡単化した例（入出力の数はもっと多い）

ディープラーニングによる文章・音声の対応付け





1

0

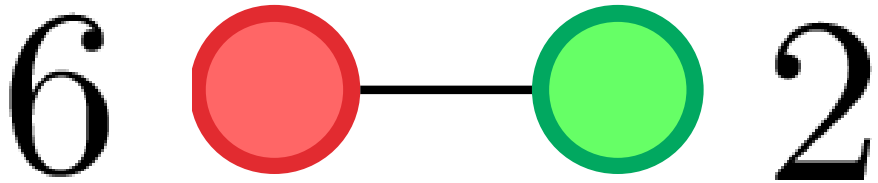
0

1

$$y = W\ '*'\ x$$

出力 ← 変換 ← 入力

ディープラーニングによる文章・音声の対応付け

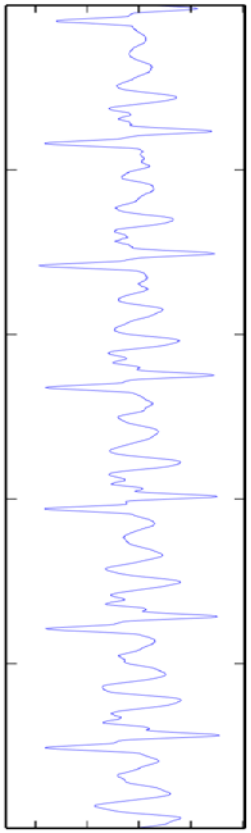


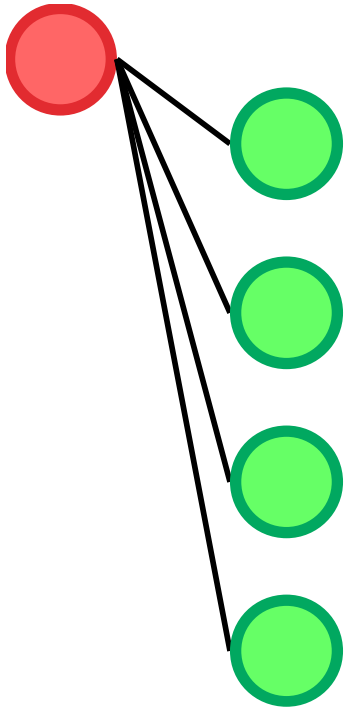
$$y = w \times x$$

出力 ← 変換 ← 入力

入力に重み付けして，出力に足す

ディープラーニングによる文章・音声の対応付け



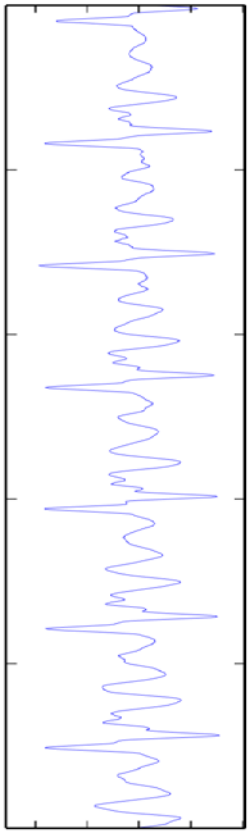


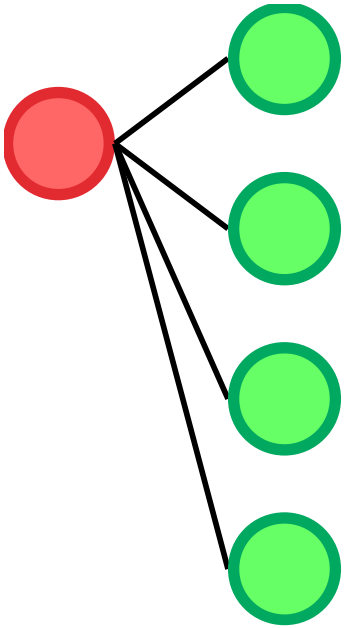
線毎に
異なる重み

$$y = W\ '*'\ x$$

出力 ← 変換 ← 入力

ディープラーニングによる文章・音声の対応付け



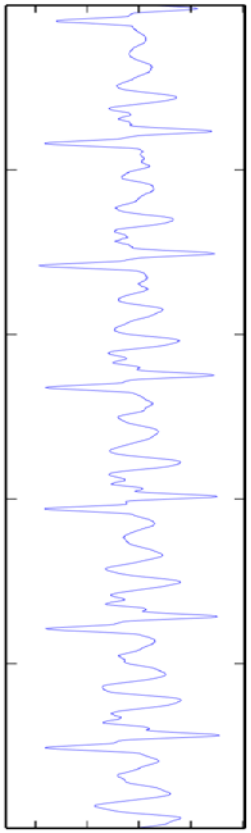


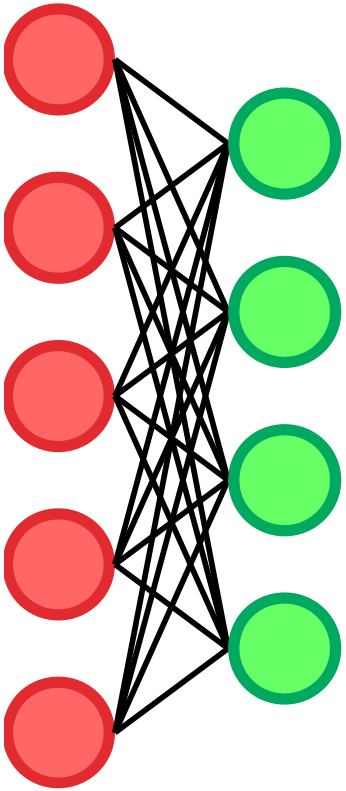
線毎に
異なる重み

$$y = W\ '*'\ x$$

出力 ← 変換 ← 入力

ディープラーニングによる文章・音声の対応付け





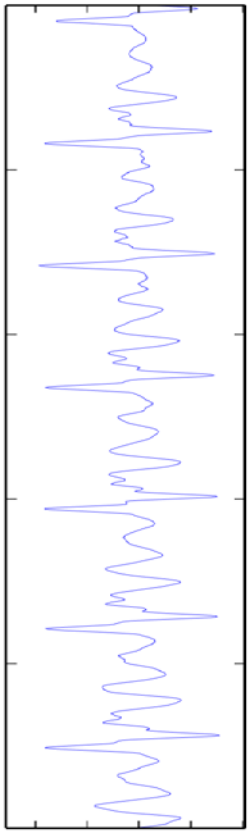
1
0
0
1

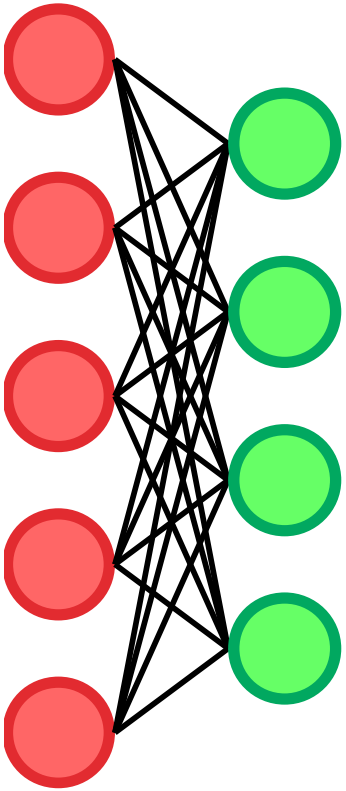
線毎に
異なる重み

$$y = W\ '*'\ x$$

出力 ← 変換 ← 入力

ディープラーニングによる文章・音声の対応付け





1

0

0

1

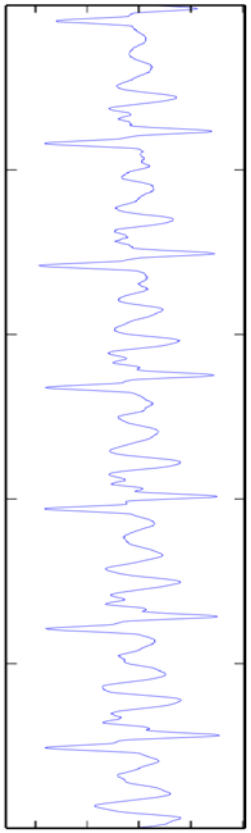
線毎に
異なる重み

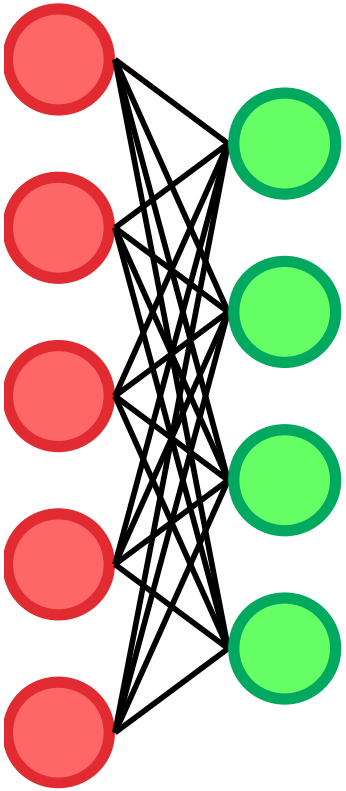
これについても
少しだけ．．．

$$y = W\ '*'\ x$$

出力 ← 変換 ← 入力

ディープラーニングによる文章・音声の対応付け





1

0

0

1

線毎に
異なる重み

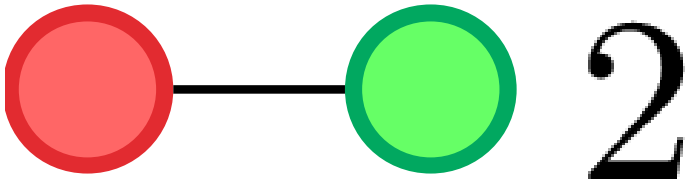
これについても
少しだけ．．．

$$y = W\ '*'\ x$$

音声合成の実現 → 適切なWを見つけ出すこと

ディープラーニングによる文章・音声の対応付け

- 文章と音声の関係の学習（適切なWを見つける）

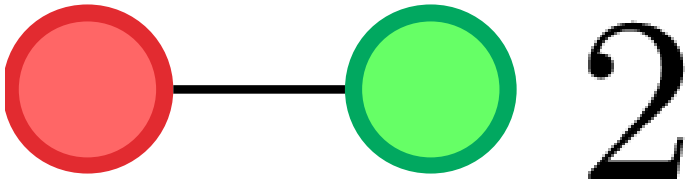


$$6 = w \times 2$$

既知　　？　　既知

ディープラーニングによる文章・音声の対応付け

- 文章と音声の関係の学習（適切なWを見つける）



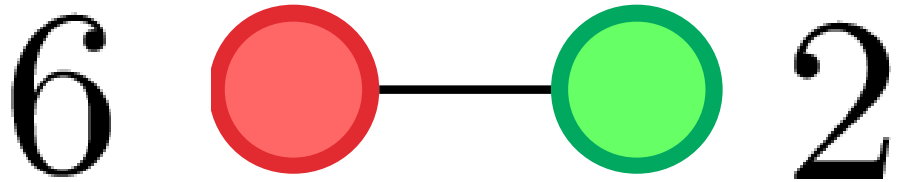
$$6 \quad 2$$

$$6 = w \times 2$$

既知　　3？　　既知

ディープラーニングによる文章・音声の対応付け

- 文章と音声の関係の学習（適切なWを見つける）



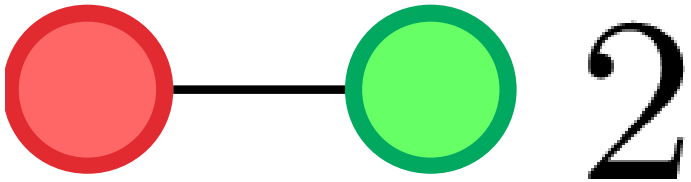
$$6 \quad 2$$

$$6 = w \; '*' \; 2$$

既知　　？　　既知

ディープラーニングによる文章・音声の対応付け

- 文章と音声の関係の学習（適切なWを見つける）

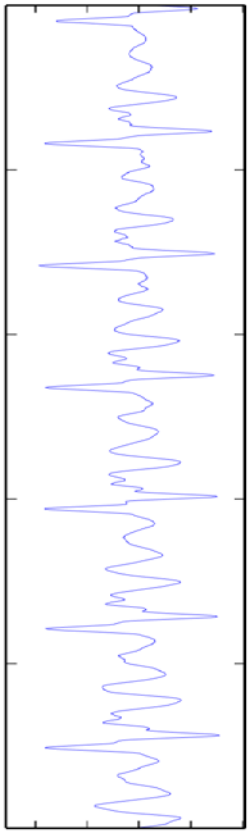


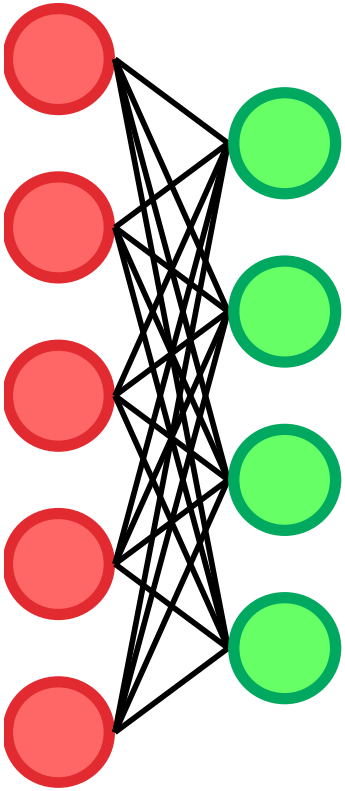
$$6 = w\ '*'\ 2$$

既知　　？　　既知

- 学習に一工夫が必要
- 誤差$\left|6 - w\ '*'\ 2\right|$を小さくする$w$の方向を得る
- wを少しずつ更新 → 適切な値を見つける

ディープラーニングによる文章・音声の対応付け





1

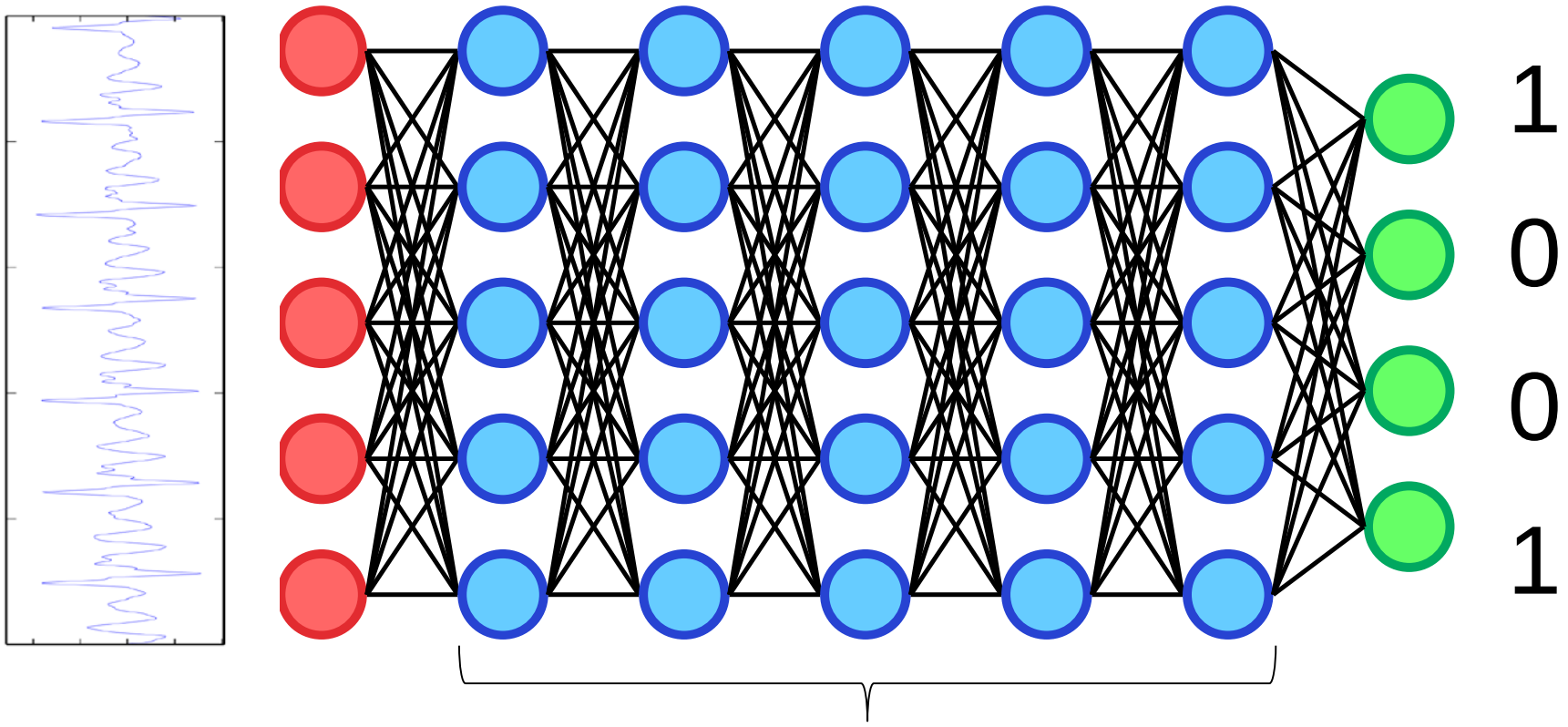
0

0

1

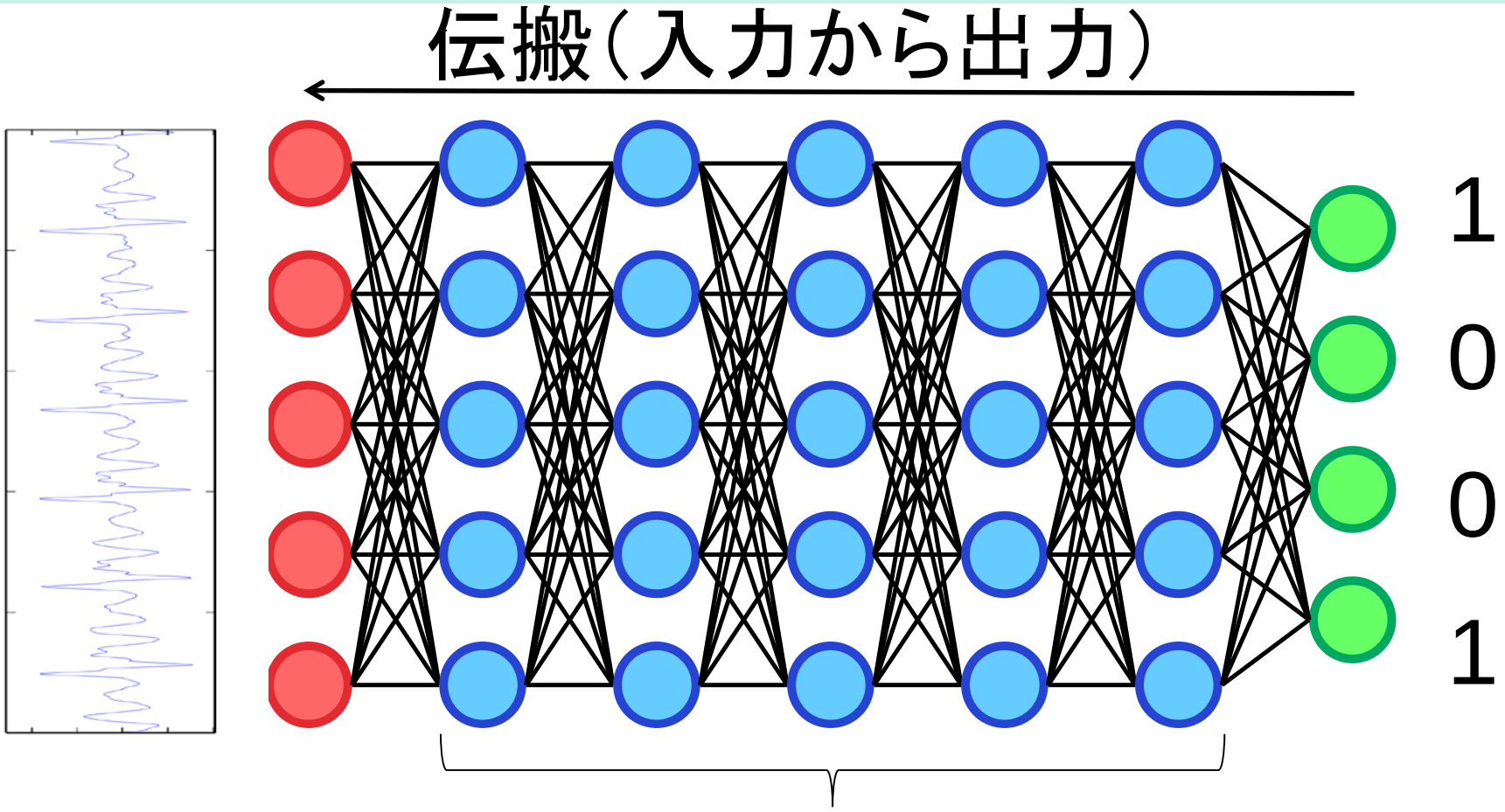
$$y = W\ '*'\ x$$

ディープラーニングによる文章・音声の対応付け



$$y = W_6 \,'*'\, W_5 \,'*'\, W_4 \,'*'\, W_3 \,'*'\, W_2 \,'*'\, W_1 \,'*'\, x$$

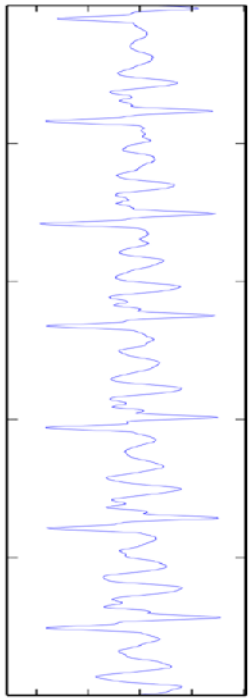
ディープラーニングによる文章・音声の対応付け

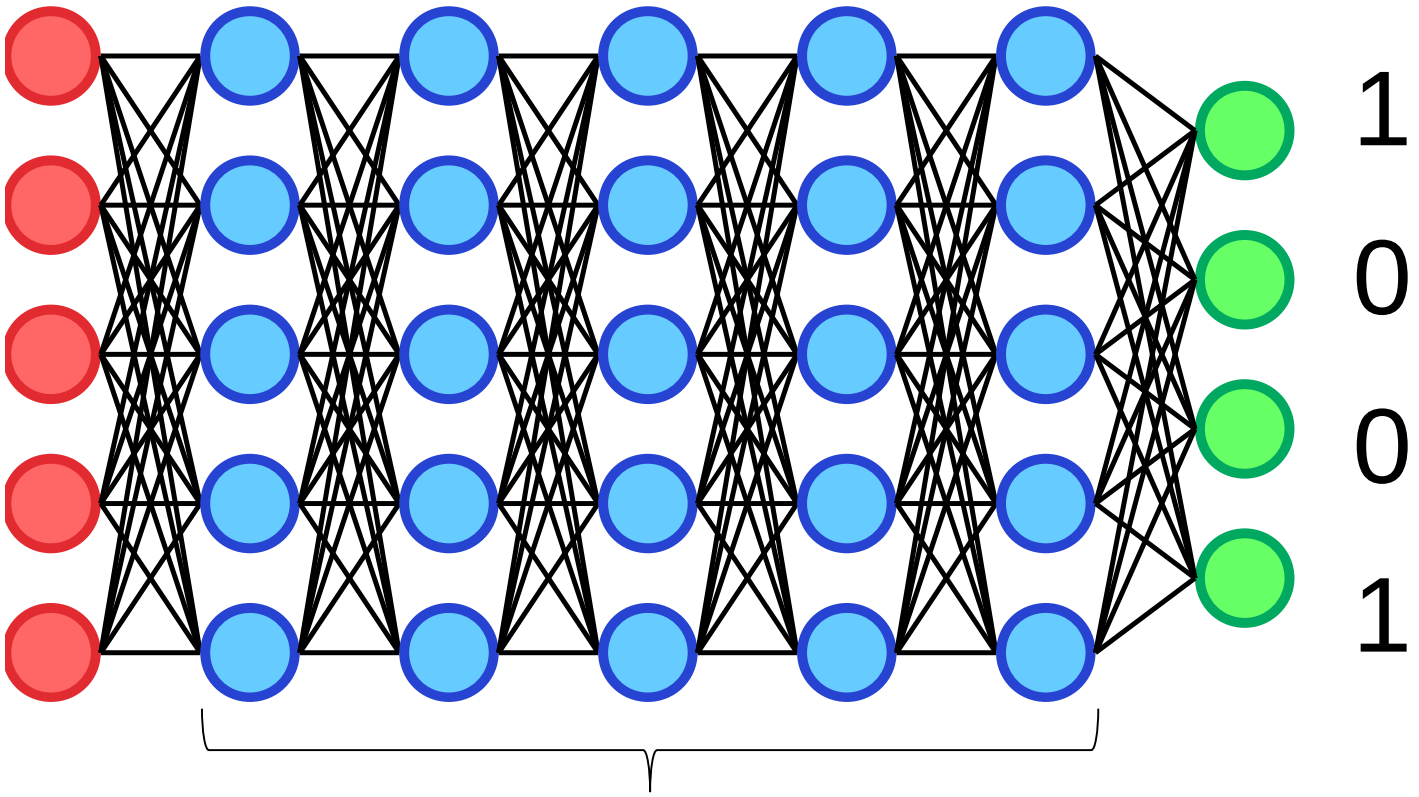


$$y = W_6 \,'*'\, W_5 \,'*'\, W_4 \,'*'\, W_3 \,'*'\, W_2 \,'*'\, W_1 \,'*'\, x$$

ディープラーニングによる文章・音声の対応付け

誤差伝搬（出力から入力）





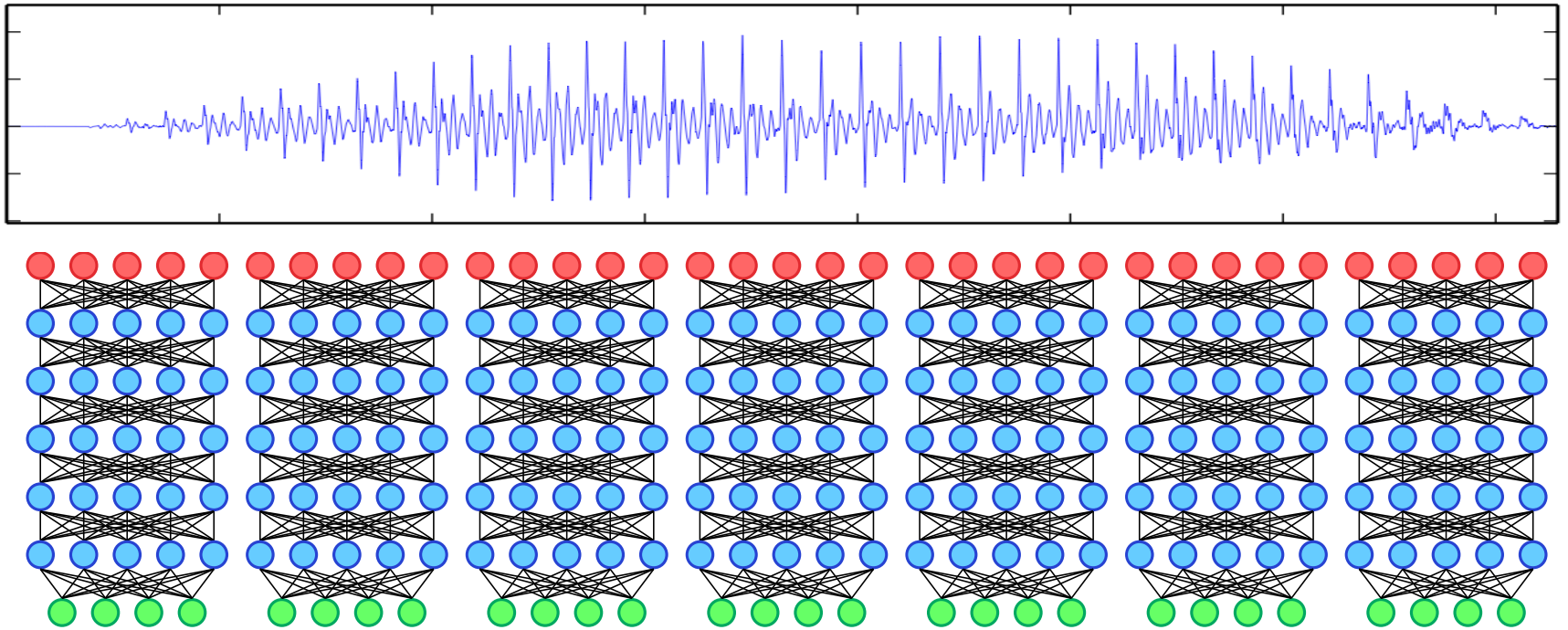
出力層　　中間層　　入力層

（自由に設計）

$$y = W_6 \,'*'\, W_5 \,'*'\, W_4 \,'*'\, W_3 \,'*'\, W_2 \,'*'\, W_1 \,'*'\, x$$

ディープは何故必要か？

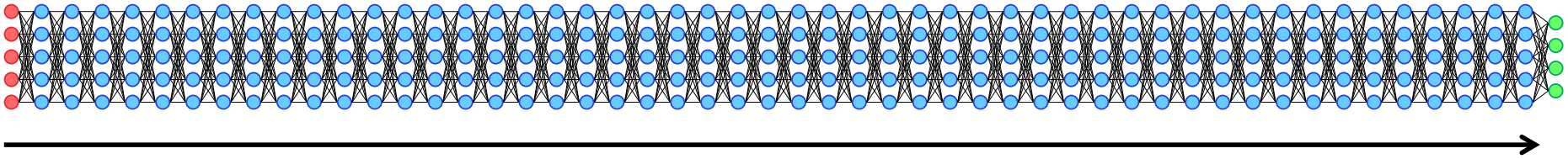
- 文章・音声の複雑な対応関係を学習可能！



- 1つの音素内でも刻一刻と波形は変化
- 様々な音素，形態素，アクセント情報の入力
  - 入力（言語情報，位置）が変化すると，出力が適切に変化

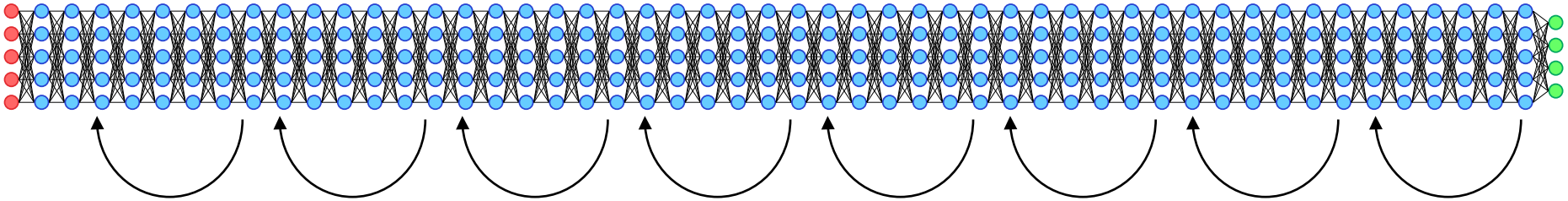
様々なネットワーク（多層）

- 中間層が3層くらいあればディープと呼んでも良い
- 層が多ければ多いほど良いのか？
- 多層のネットワークは学習が難しい



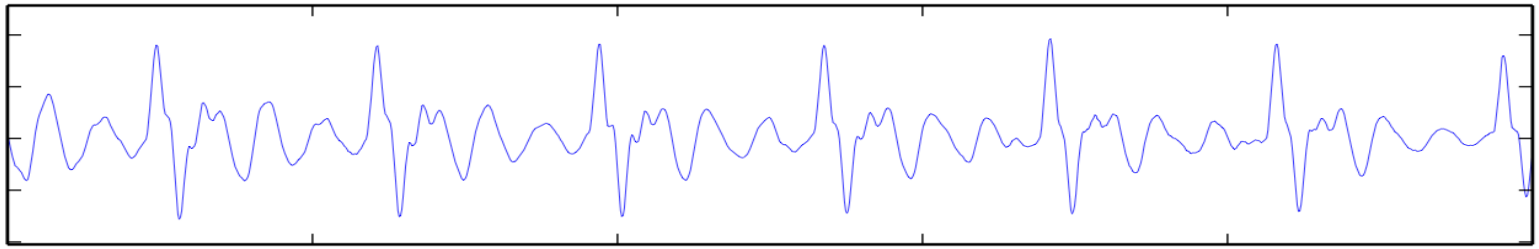
誤差が伝わらなくなる（勾配喪失）

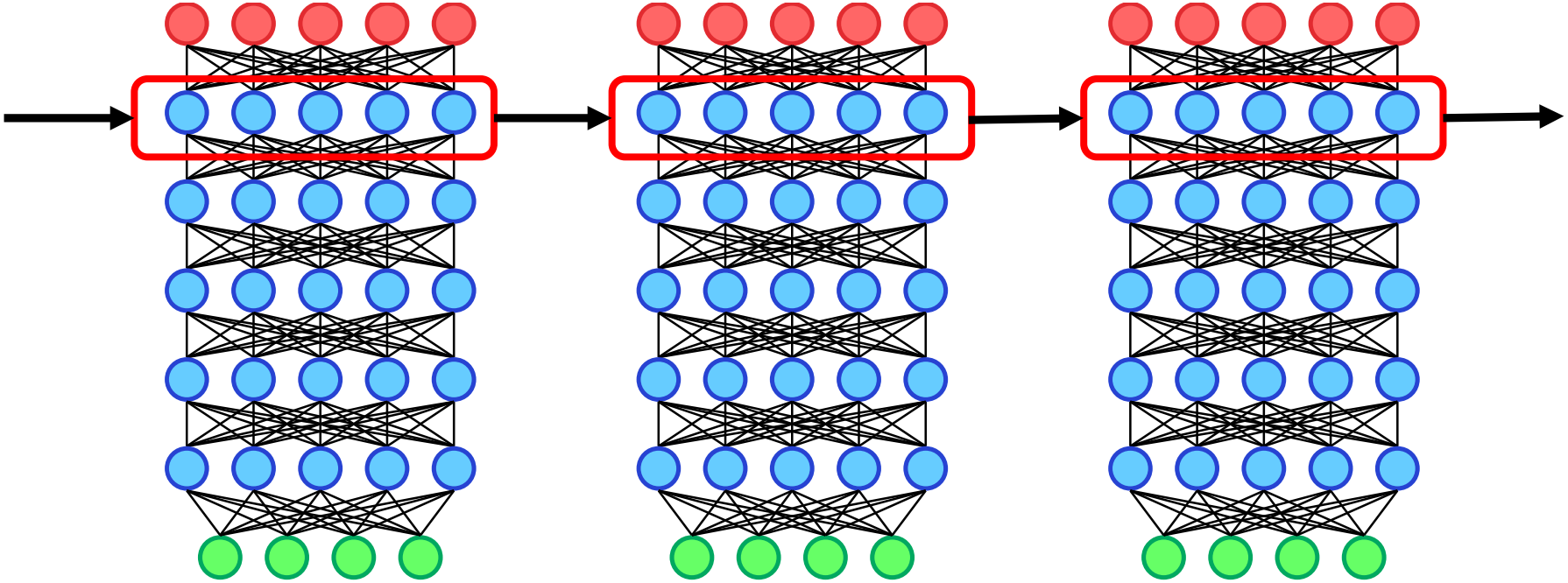
- しかし，最近では100層以上のネットワークも登場
  - ショートカットコネクションの導入



様々なネットワーク（時間方向での接続）

- 音声波形は時間とともに変化（時系列データ）
- 時系列データに適したネットワーク（RNN）





合成音声サンプル

- RNN

音響モデル	Feed-forward		RNN	
言語特徴量	quin-phone	quin-phone prosodic info.	quin-phone	quin-phone prosodic info.

- 多層

手法	ストリーム	構造	サンプル
DS	Single	Feed-forwad	
HS	Single	Highway	
HM	Multi	Highway	

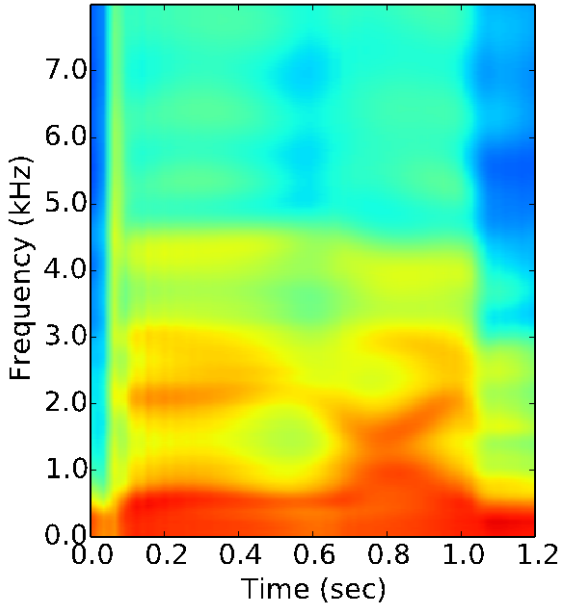
層数： 4 20 40 80

講義の構成

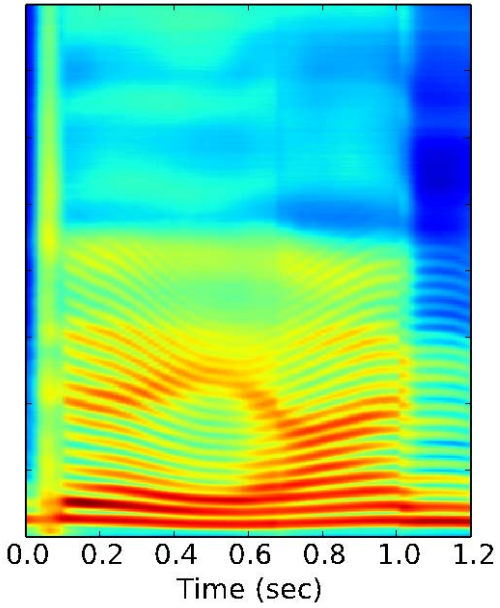
- 音声合成
  - 文章の原語情報の抽出
    - 音素，形態素，アクセント
  - 音声の音響特徴量の抽出
    - スペクトル，スペクトル包絡，F0
  - ディープラーニング
    - 言語情報と音響特徴量の対応付け
    - 様々なニューラルネットワーク
- 応用

出力の変更

- 高次元スペクトルの利用
  - 従来手法では扱いが困難
  - ディープラーニングにより高精度な出力が可能に



スペクトル包絡

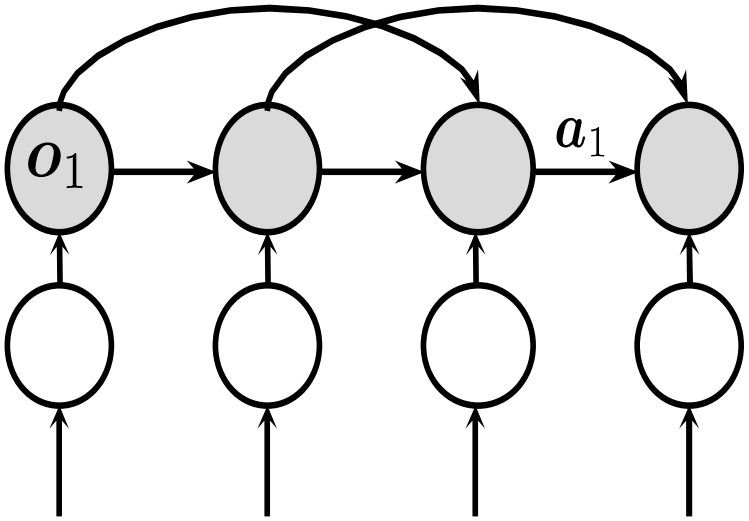


スペクトル

- 音声波形そのものの出力も登場（Google WaveNet）

複雑なネットワーク構造の検討

- 日々改良が行われています



様々な話者表現，様々な感情表現

- 複数話者データや複数感情音声データを利用

- 感情

感情制御

平静	楽しげ	悲しげ	安心	不安	励ます	怒る
						

- 話者

話者補完	ジェンダー補完
	

その他，合成音声サンプル

- 抑揚や韻律の劇的な改善

RNN	新手法
	

まとめ

- 音声合成の基礎
- 音声合成のためのディープラーニング

} 紹介

- 合成音声の品質，表現力が向上
  - 様々な話者
  - 様々な感情
  - 韻律改善

楽しい！音声情報処理
楽しい！ディープラーニング

参考

- 丸善ライブラリー「おしゃべりなコンピュータ」
  - 音声合成技術が分かりやすく解説がされています
  - ディープラーニングについては触れる程度
- 日本音響学会誌 “深層学習に基づく統計的音声合成”
  - 研究者向け解説論文