東京工業大学
Tokyo Institute of Technology

音声認識・音声情報処理

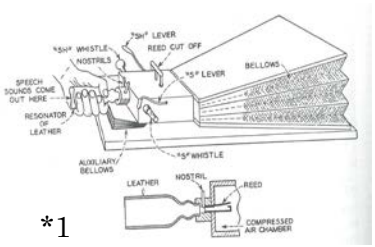
篠崎隆宏研究室
工学院情報通信系情報通信コース
すずかけ台キャンパスG2棟804室
www.ts.ip.titech.ac.jp

Shinozaki Lab.

音声技術の歴史

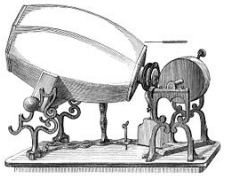
1800
1850
1900
1950
2000

1791
Kempelen's talking machine



*1

1857
Phonautograph


*2

1876
Telephone

Radio broad cast

1920s
Radio Rex

1939
Vocoder

1946
ENI
AC

1952
Speech recognizer
(Digit recognition)

1969
Internet

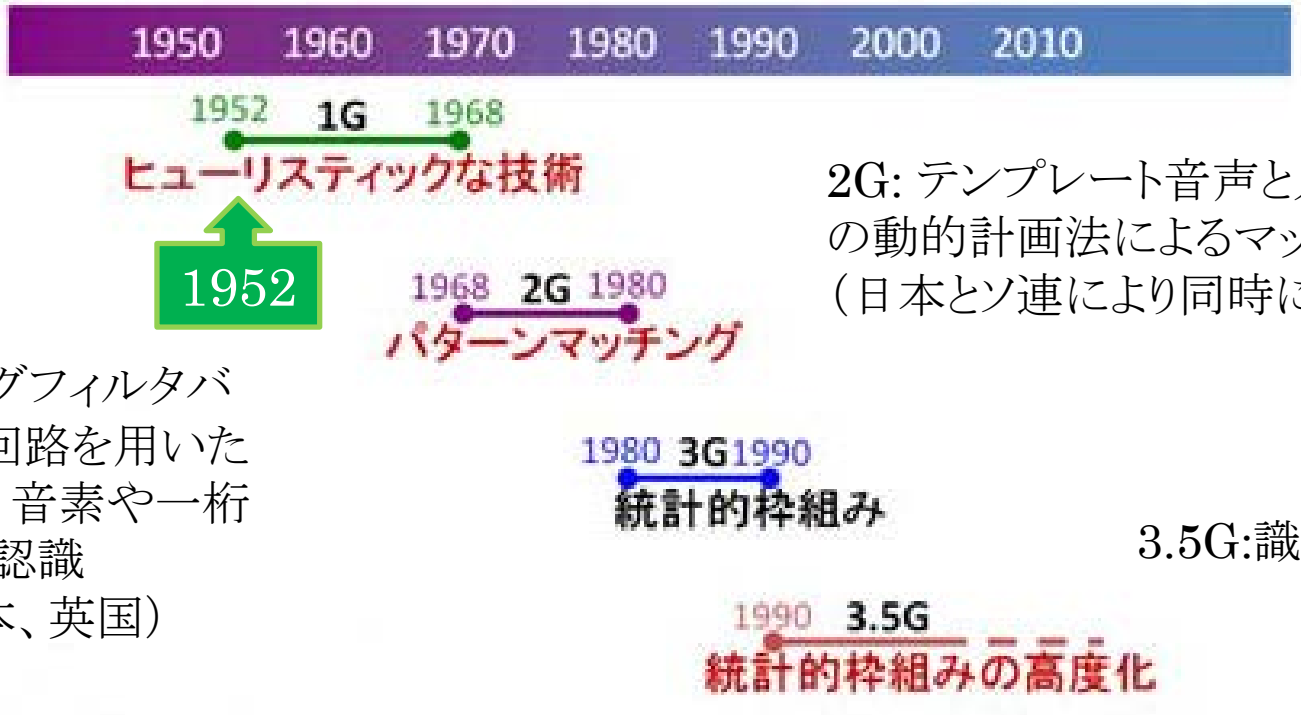
1990
www

1999
Seaman

2011
Siri on iOS

Sources:
1) Voice communication between human and machines, National academy press.
2) https://en.wikipedia.org/w/index.php?title=Phonautograph&oldid=676031521 (last visited Apr. 2, 2016).

音声認識の歴史



2G: テンプレート音声と入力音声の動的計画法によるマッチング（日本とソ連により同時に発明）

1G: アナログフィルタバンクと論理回路を用いた方法により、音素や一桁数字などを認識（米国、日本、英国）

3.5G:識別学習

3G: 隠れマルコフモデルやN-gramの利用
IBM, Bell研

4G:ディープニューラルネット

出典：古井貞熙 「コンピューターによる音声認識のこれまでと今後の展望」
日本音響学会2011年春季研究発表会1-31-1

音声技術の応用例

Transmit only speech segment to the server

Returned response

Speech output

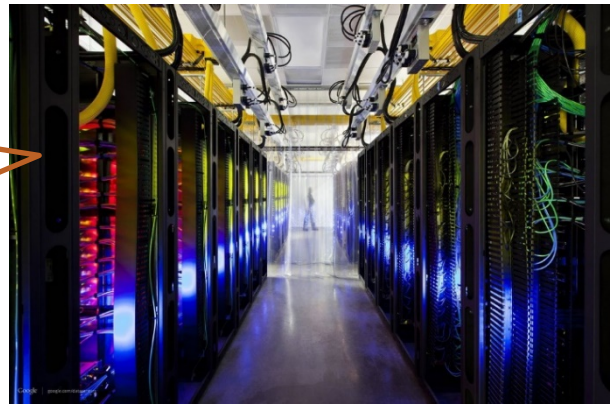
Speech input and voice activity detection

Speech output

Internet

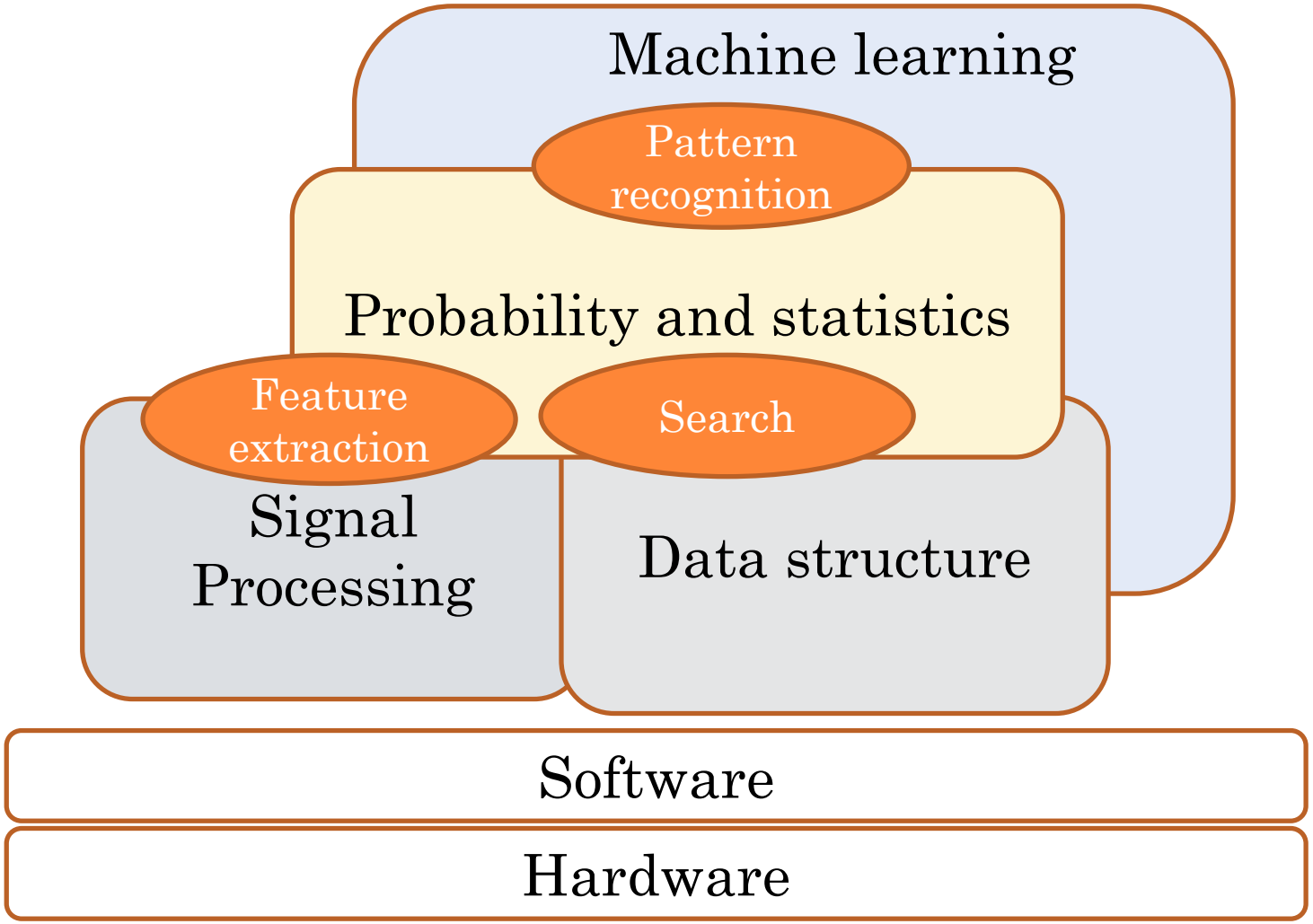
Speech input

Speech recognition, dialog management, and speech synthesis

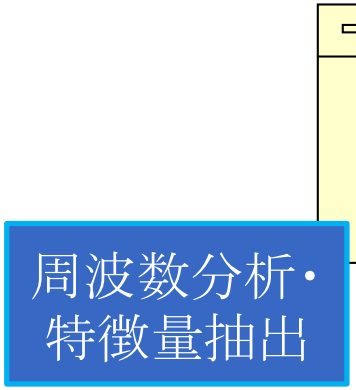


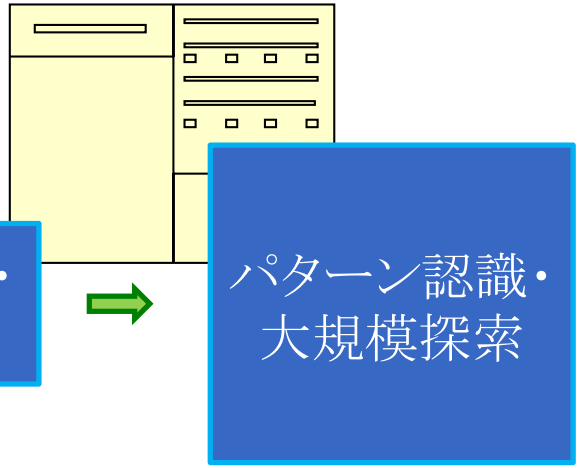
Server computers in data center

音声認識システムの実現に関わる主要分野



音声認識システムの仕組み



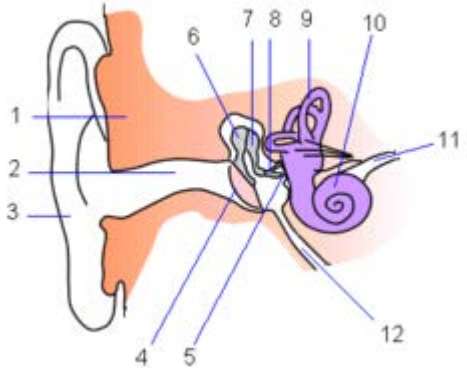


こんにちは

空気振動から電気信号への変換

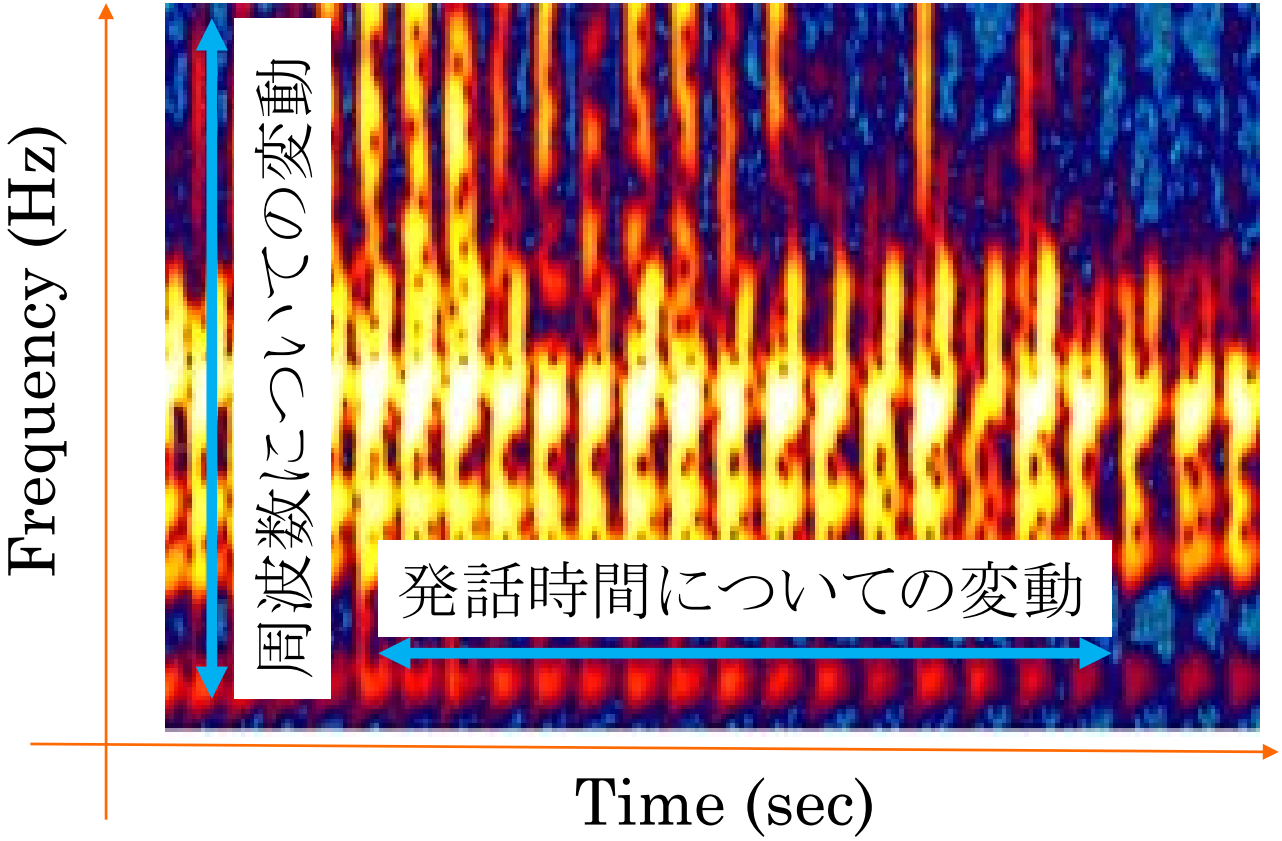
認識に有用な情報の抽出

音声言語の知識に基づいた大規模探索



認識を行う上での2つの問題

- 周波数および時間方向の変動
  利用できるのは希薄に分布した曖昧な手掛かり



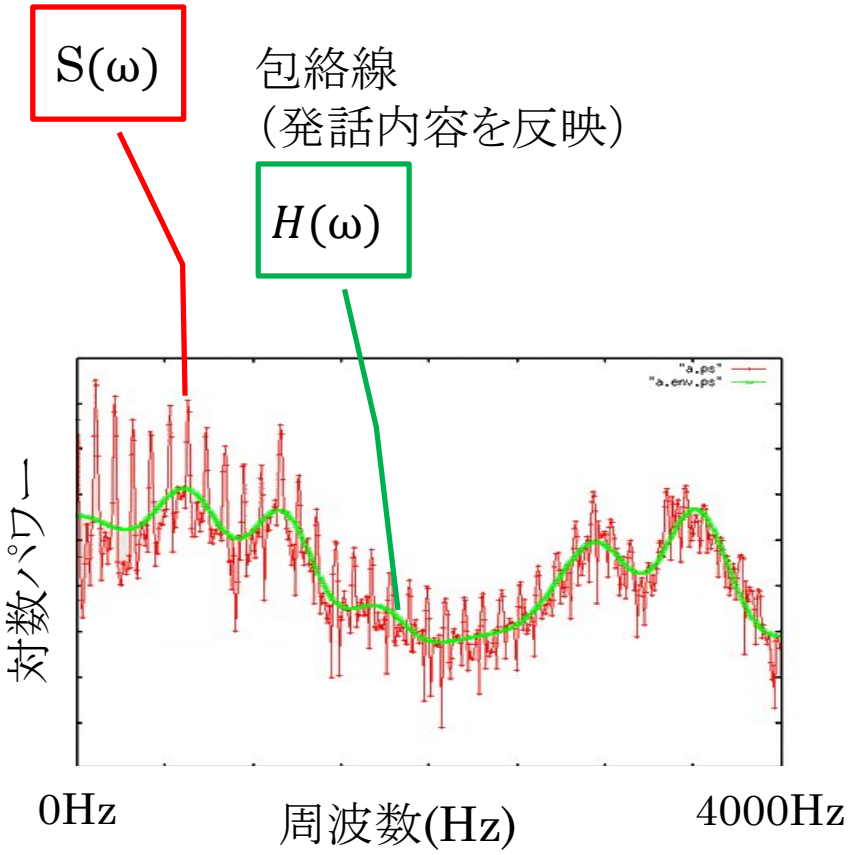
特徴量抽出

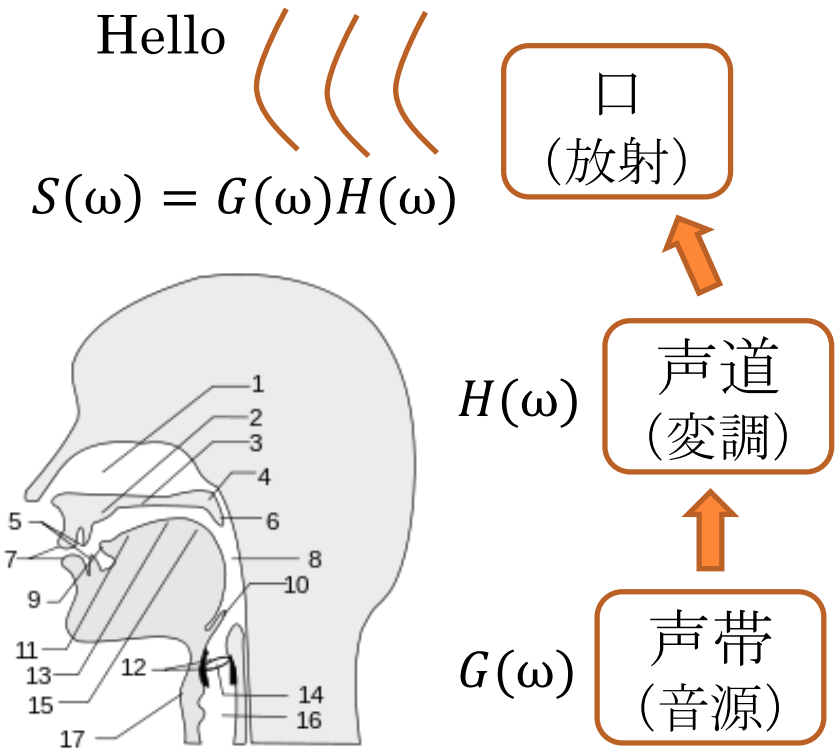
音声認識を行う上で有用な情報の取り出し

- 話者や雑音の影響等を抑え、認識性能を向上させる
- 不要な情報を捨てることで、後段での不要な計算を省く

音声に含まれている情報

実測信号
（細かい変動は声の高さに関係し、G(ω)に対応）



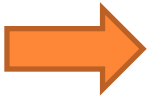


音源の取り換え実験

Original Sound $S(\omega) = G(\omega)H(\omega)$

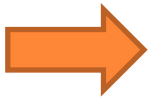
Webブラウザで再生されない場合は一度PCにダウンロードしてから開く等してください

Sawtooth wave G’ (100Hz)

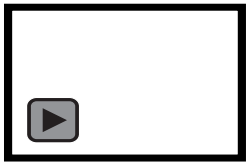


$G'(\omega)H(\omega)$

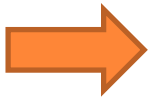
Sawtooth wave G’ (300Hz)



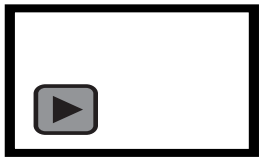
$G'(\omega)H(\omega)$



Acoustic11 G’ (Music) *1



$G'(\omega)H(\omega)$

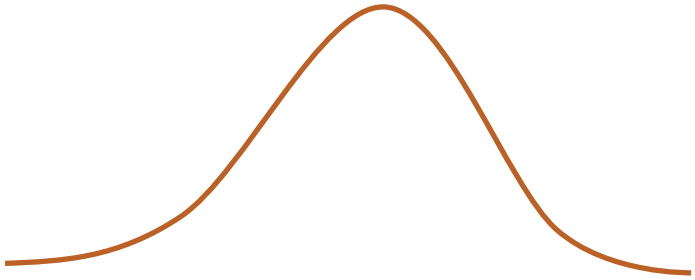


1) Source of the music file: http://maoudamashii.jokersounds.com/

音声の統計的モデル化

- 音声言語の知識を計算機で処理できる形で表現
- 多様な音声を高精度に認識するため、巨大で複雑な構造を持ったモデルを利用
- 認識システムの性能を決める心臓部
  - 様々なモデルが組み合わされて使用される

ガウス混合モデル(GMM)



ガウス分布

ガウス分布ではどんなに頑張っても山が一つで左右対称の分布しかモデル化出来ない

複数のガウス分布を
重ね合わせ

ガウス混合分布

定義：　$f(X)=\sum_i w_i N_i(X \mid \mu_i, S_i)$

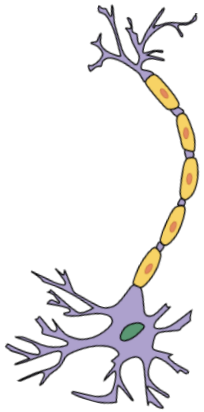
w_i　：　混合重み

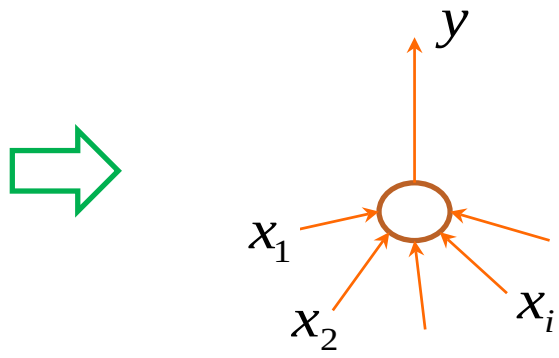
N_i　：　ガウス分布

より複雑な分布を表現できる

MULTI LAYER PERCEPTRON (MLP)

- 神経細胞の機能を数式で抽象化

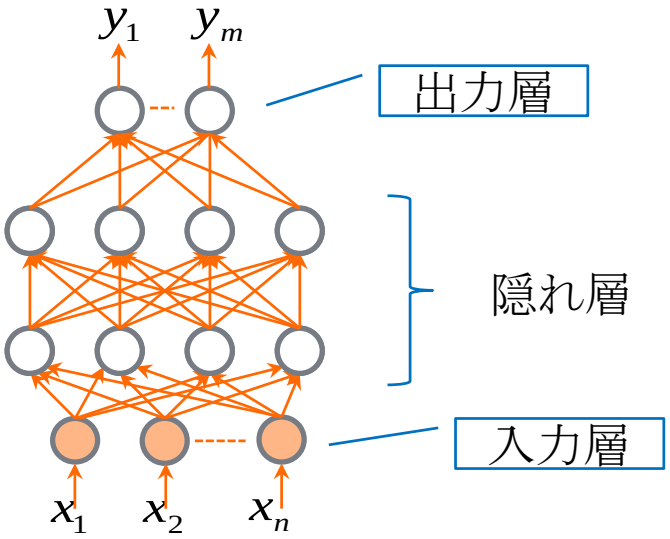




$$y = h\left(\sum_i w_i x_i + b\right)$$

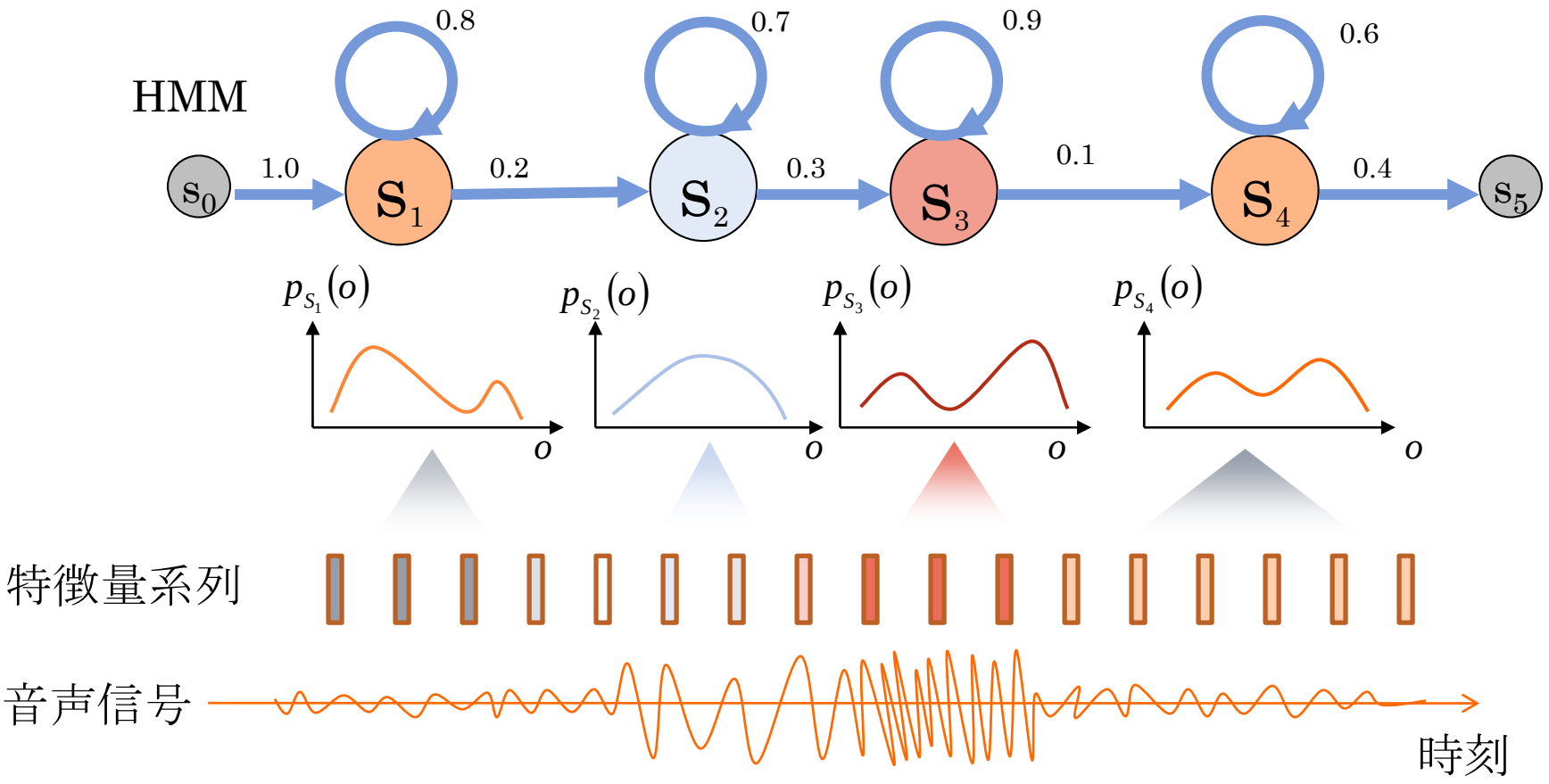
h: 活性化関数
w: 結合重み
b : バイアス

- 複数のユニットを階層的に結合した計算モデルがMLP



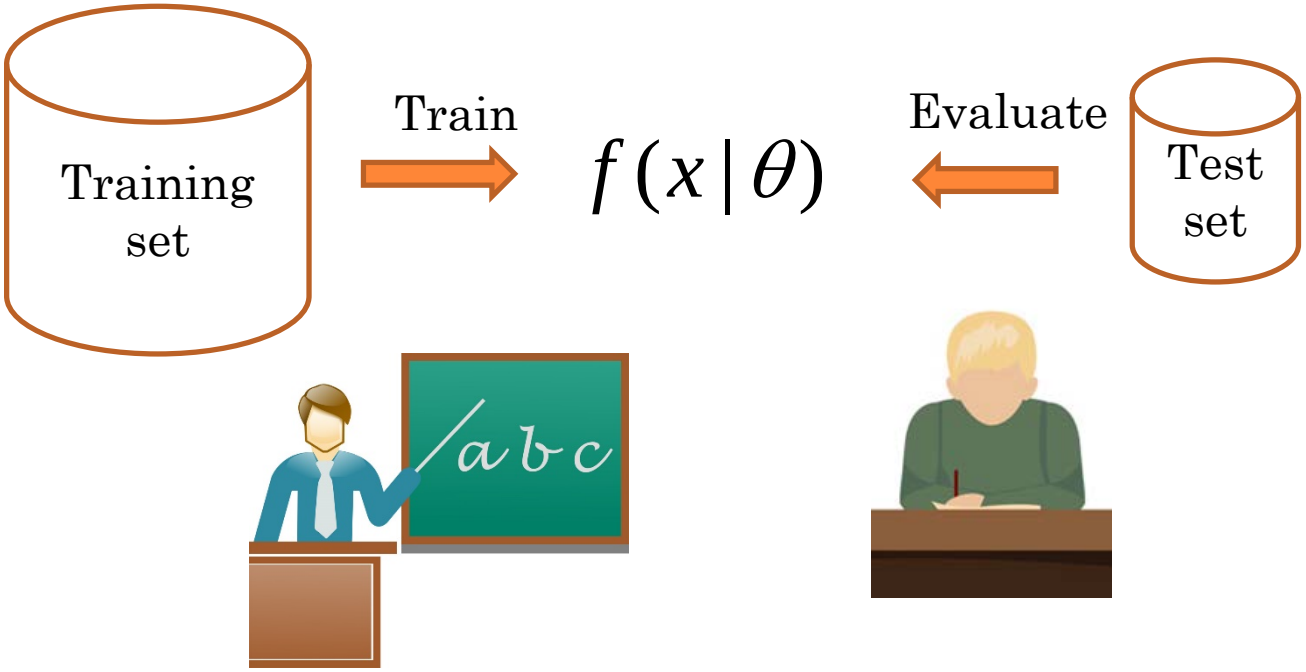
隠れマルコフモデル(HMM)

- グラフ構造(状態と状態遷移)＋確率分布



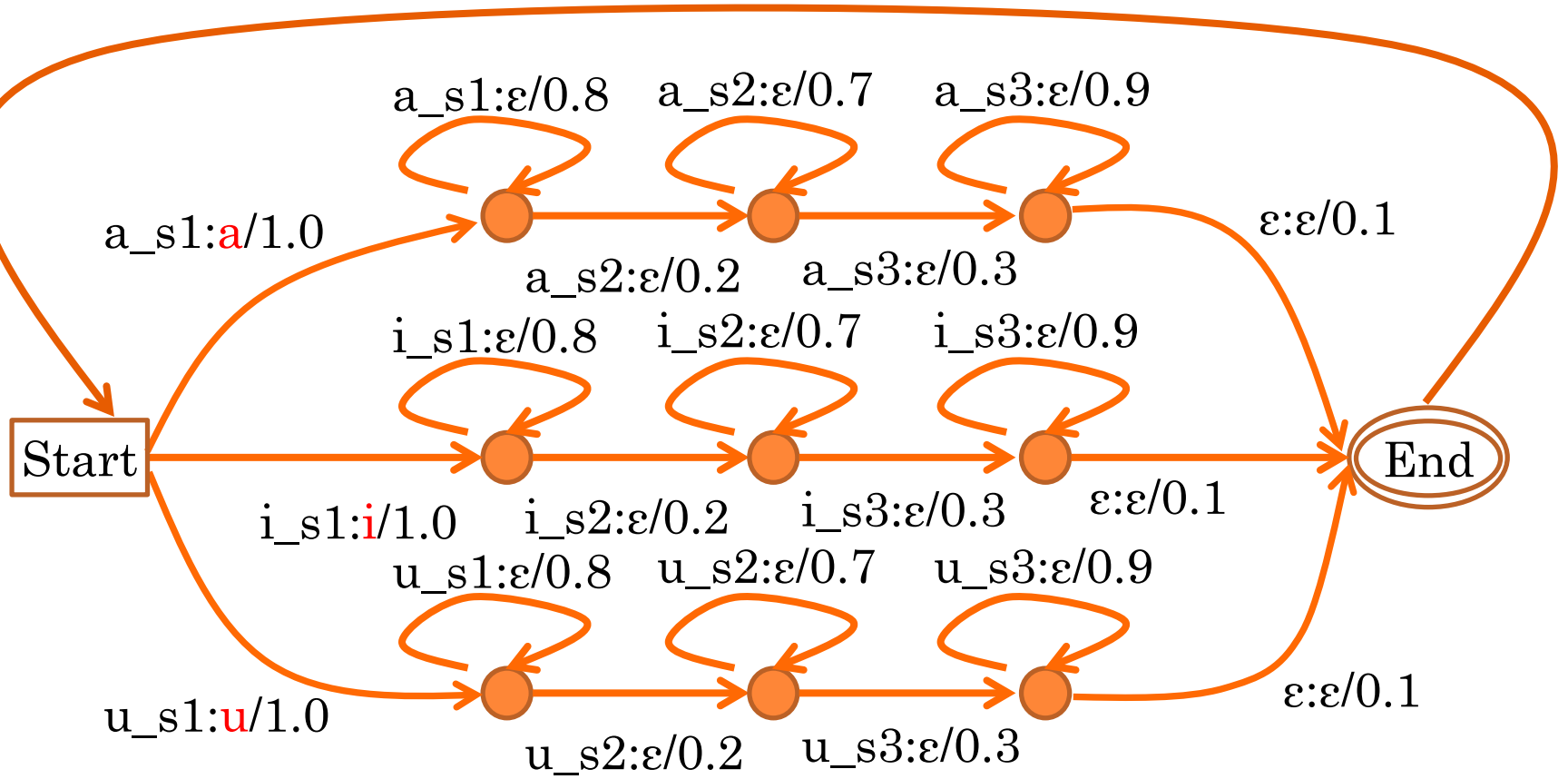
モデル学習と評価

- パターン認識器は入力パターンxから認識結果yを求める
  パラメタθにより規定される関数 y=f(x | Θ)
- パラメタは大量の学習データから推定
- 評価は、学習データとは独立した評価データを用いる



探索空間の構成

重み付き有限状態トランスデューサ(WFST)を用いた簡単な音素認識システムの例
（大語彙認識では数千万の遷移を持つ巨大ネットワークを使用）



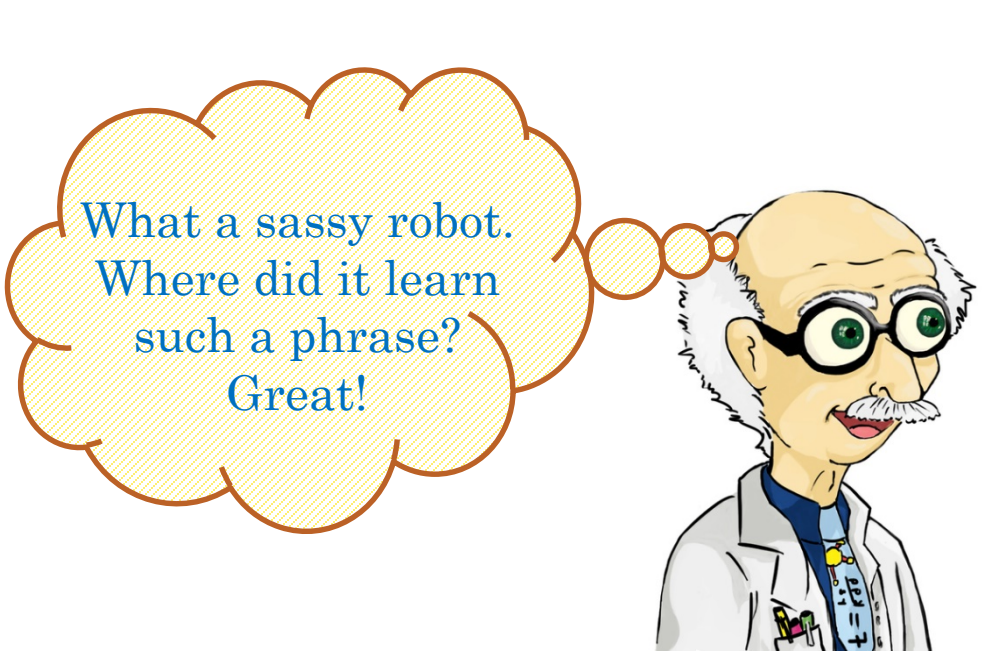
入力：HMM状態系列、出力：音素系列

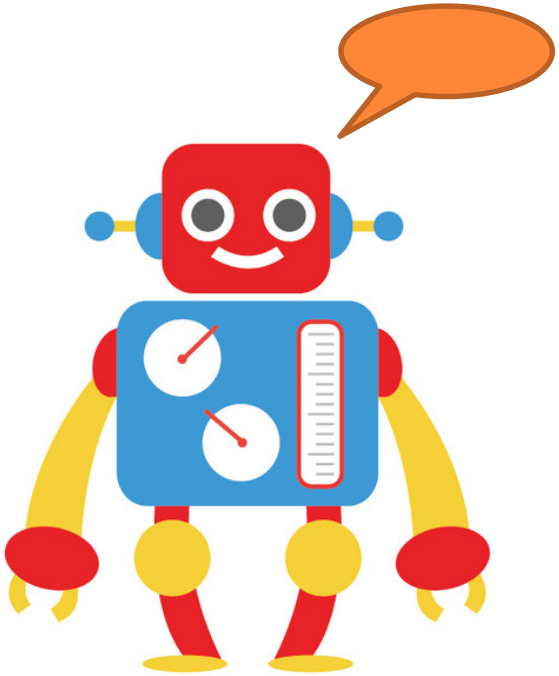
音声認識の現状と課題

- 現状
  - 音声対話アプリケーションがスマホに実装され実用化
  - テレビの自動字幕や、ビデオ録画に対する音声検索も一部実用化
  - 音声対話機能を備えたロボットが登場
- 課題
  - 話者や発話環境による認識性能のばらつきが大きい
    - スマホでの音声認識がすごく使えるという人と、誤認識ばかりで使えないという人で意見が分かれる
  - 人間と比べて学習・適応能力が貧弱
    - 高い性能を実現するためには膨大な開発費が必要で、経済的にペイするアプリケーションが限定される
    - 特定のタスクに対して高い認識性能を実現できても、少し違うタスクに応用すると性能が大きく劣化する
    - 日々登場する新しい単語や言い回しへの対応のため、継続的なメンテナンスが必要

将来展望

- 膨大な手間とコストのかかるマンツーマン（マンツーマシン？）の教育が不要で、勝手に学習するシステムの実現
- 音声認識から音声認識理解へ





当研究室の取り組み

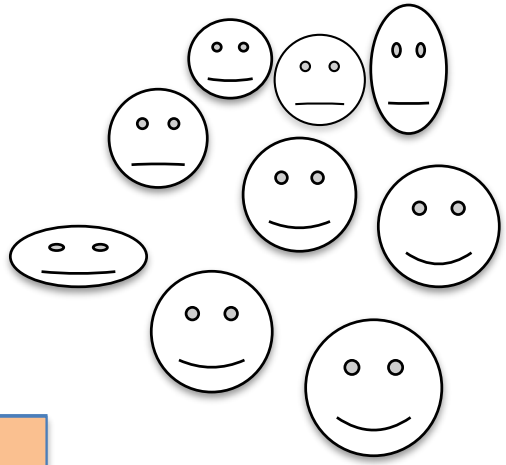
進化計算による認識システムの自律的高性能化

初期値 x_{init} を用意

初期集団 $\{x_1 \sim x_N\}$ を生成

DNNのメタパラメタ

$$x = \begin{pmatrix} Hidden\_layers \\ Units / Layer \\ \vdots \end{pmatrix}$$



各個体の認識性能$\{f(x_1) \sim f(x_N)\}$を評価
（モデルの学習，認識実験）

良い成績の個体をもとにガウス分布を更新

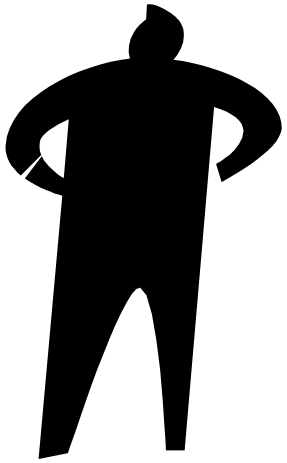
更新したガウス分布から
子供集団 $\{x'_1 \sim x'_N\}$ を生成

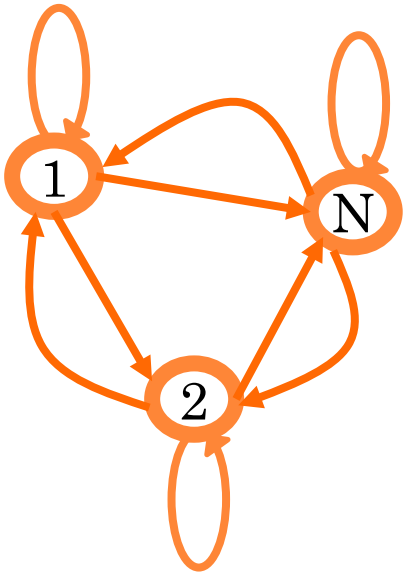
繰り返す

教師なし・半教師付き学習アルゴリズム

音声言語データ

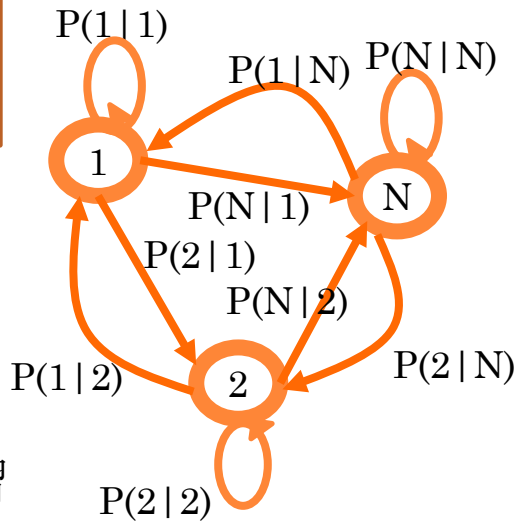
学習





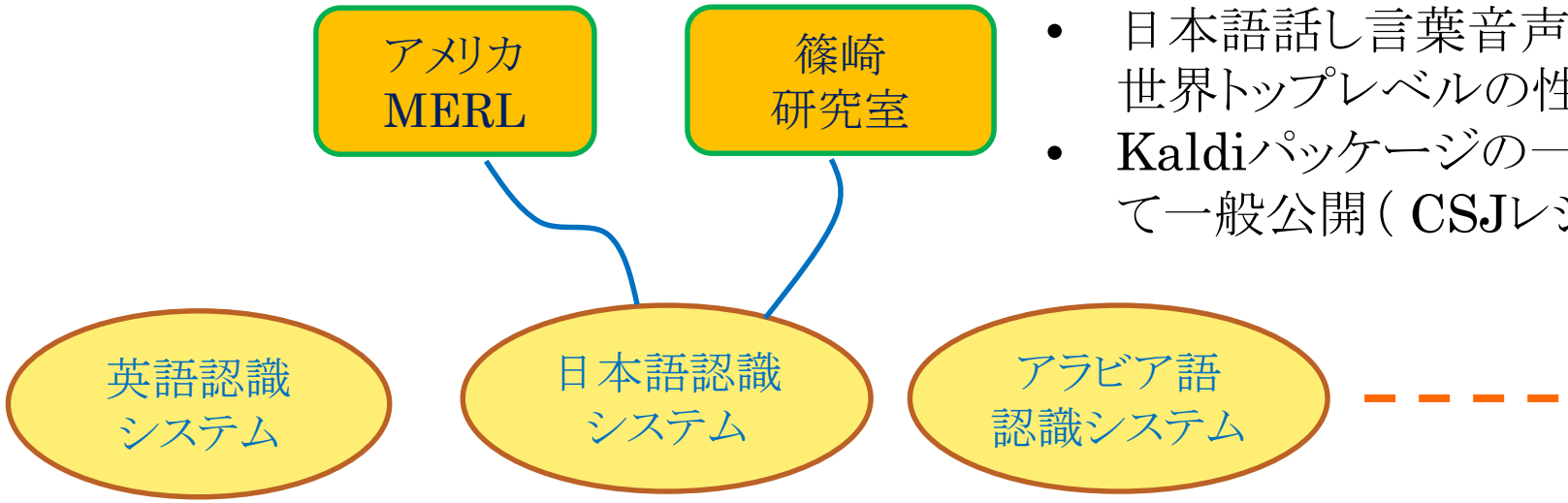
音声言語データ

学習



システムがより賢く自律的に学習する仕組みの実現

日本語大語彙認識システムの開発



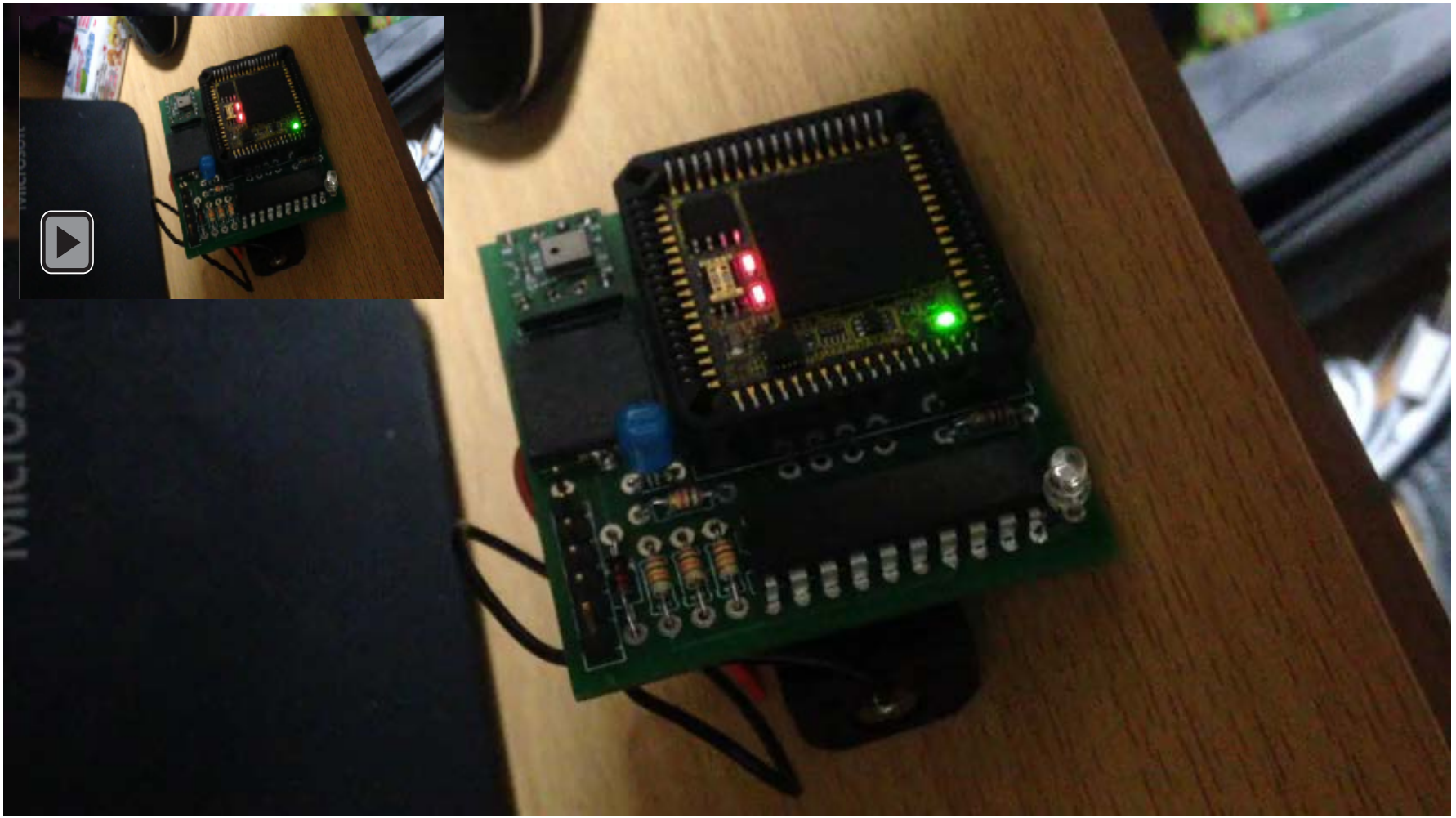
- 日本語話し言葉音声認識で世界トップレベルの性能
- Kaldiパッケージの一部として一般公開（CSJレシピ）

国際的な開発チーム
(githubを介して連携)

大学や研究所の教員、
研究者、大学院生

Kaldi Speech Recognition Toolkit

ユビキタス利用のための小型低消費電力音声認識



参考情報：研究用ソフトウエア

- Speech recognition toolkit
  - Kaldi
    - High performance
    - Modern software design
    - http://kaldi.sourceforge.net/index.html
  - HTK :
    - Traditional toolkit
    - “HTK Book” is a well written manual
    - http://htk.eng.cam.ac.uk/
- Sound visualization and manipulation
  - WaveSurfer
    - http://www.speech.kth.se/wavesurfer/
- Deep neural network
  - Theano Python library for neural network
    - http://deeplearning.net/software/theano
  - Chainer
    - http://chainer.org/