大規模進化計算による音声認識システム開発の自動化

篠崎隆宏
東京工業大学　情報通信系
www.ts.ip.titech.ac.jp

GTC Japan 2016

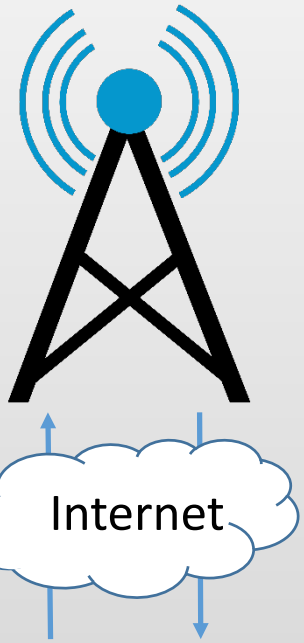
音声認識の普及

音声入力・音声区間検出

音声出力

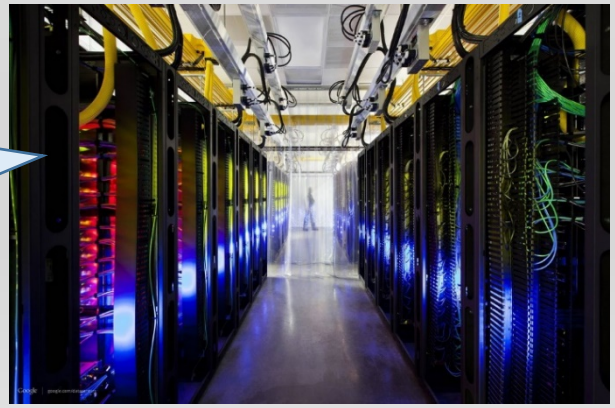
音声区間をサーバーへ

応答音声を端末へ



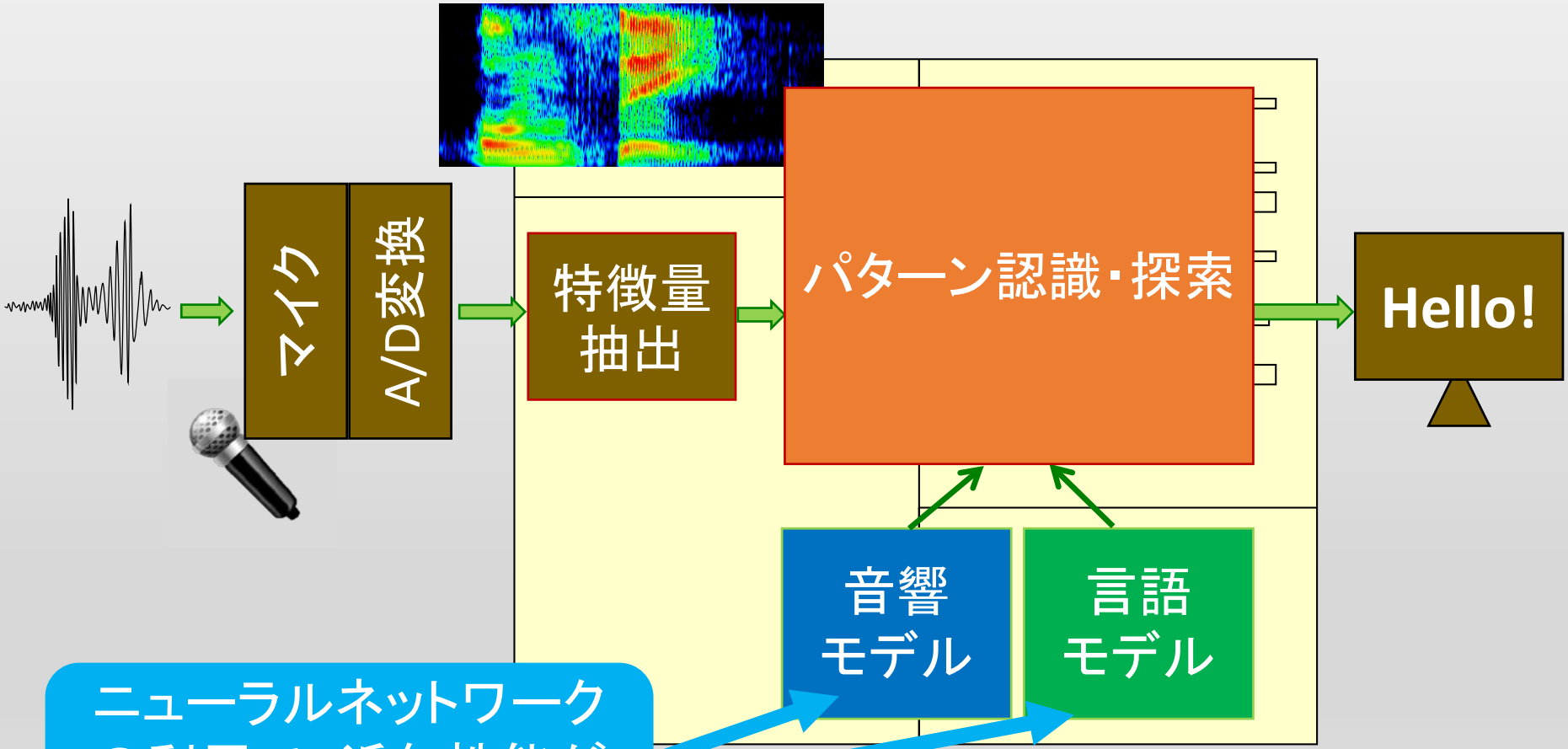
Internet

音声認識、応答文生成、音声合成



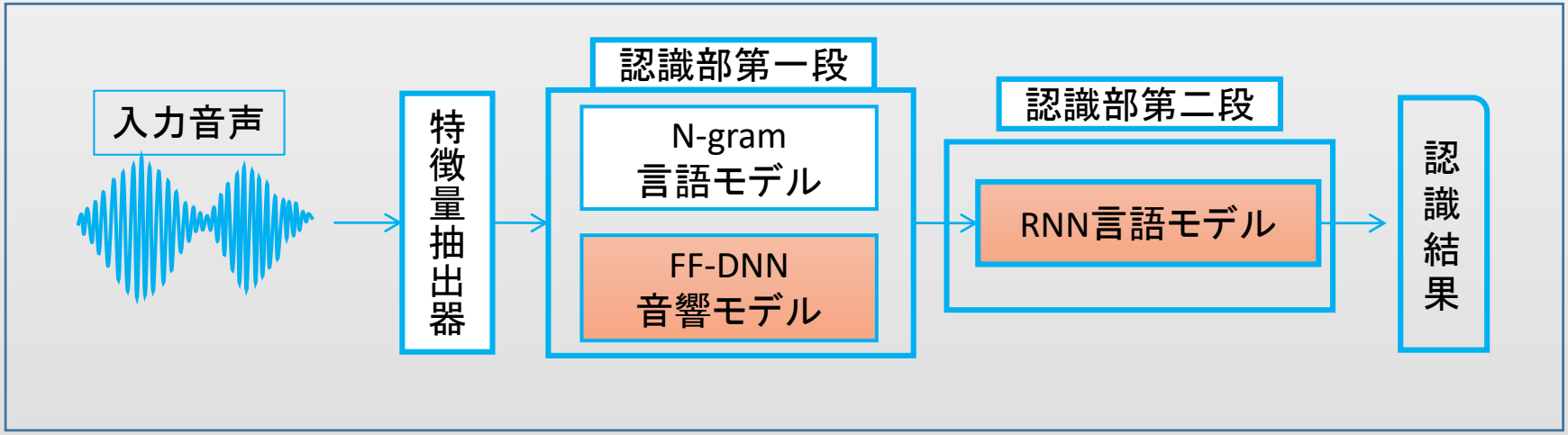
データセンター

音声認識の仕組み

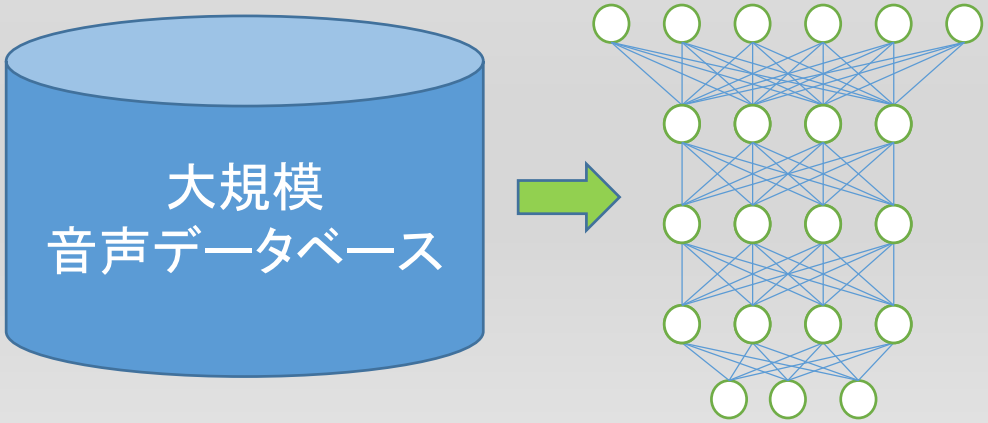

マイク
A/D変換
特徴量
抽出
パターン認識・探索
Hello!
音響
モデル
言語
モデル
ニューラルネットワーク
の利用で、近年性能が
大幅に向上

音声認識システムの構築

① システム設計



② 音響・言語モデルの学習



③ 認識性能の評価

単語誤り率(WER)

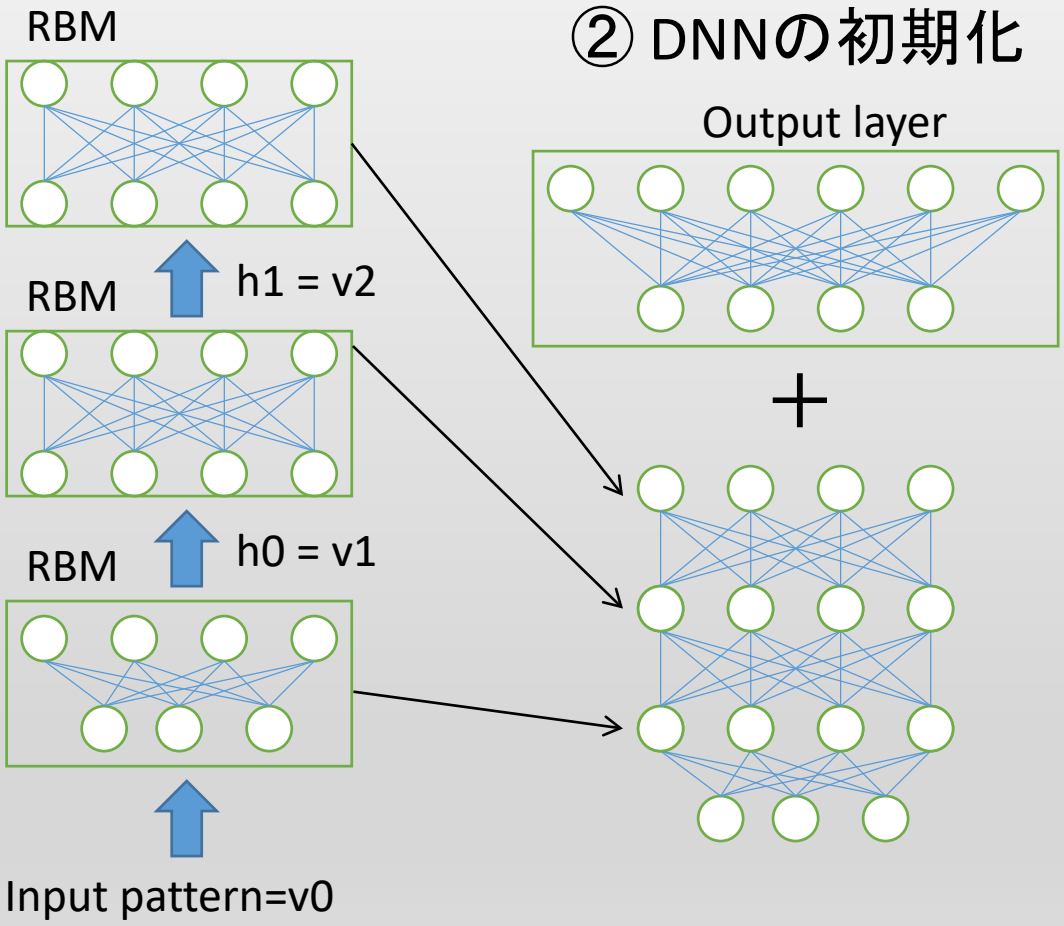
$$= \frac{\text{置換誤り} + \text{挿入誤り} + \text{削除誤り}}{\text{正解単語数}}$$

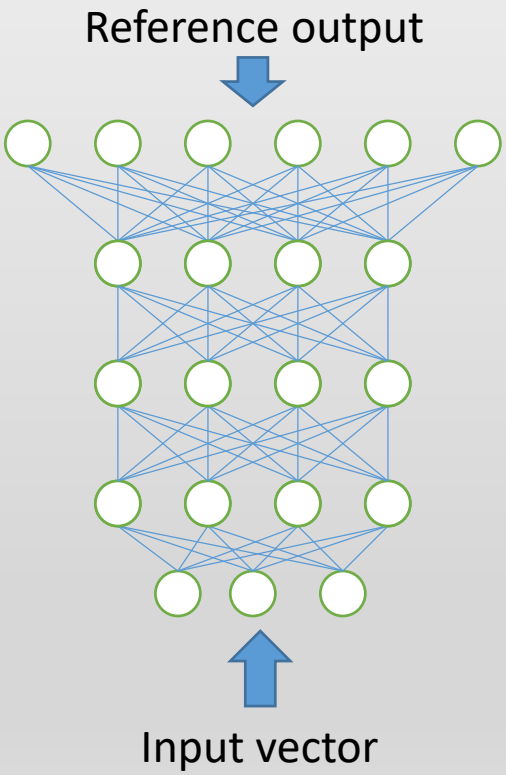
DNN音響モデルの学習プロセス

① RBMの教師なし学習

② DNNの初期化

③ バックプロパゲーションによる教師付き学習



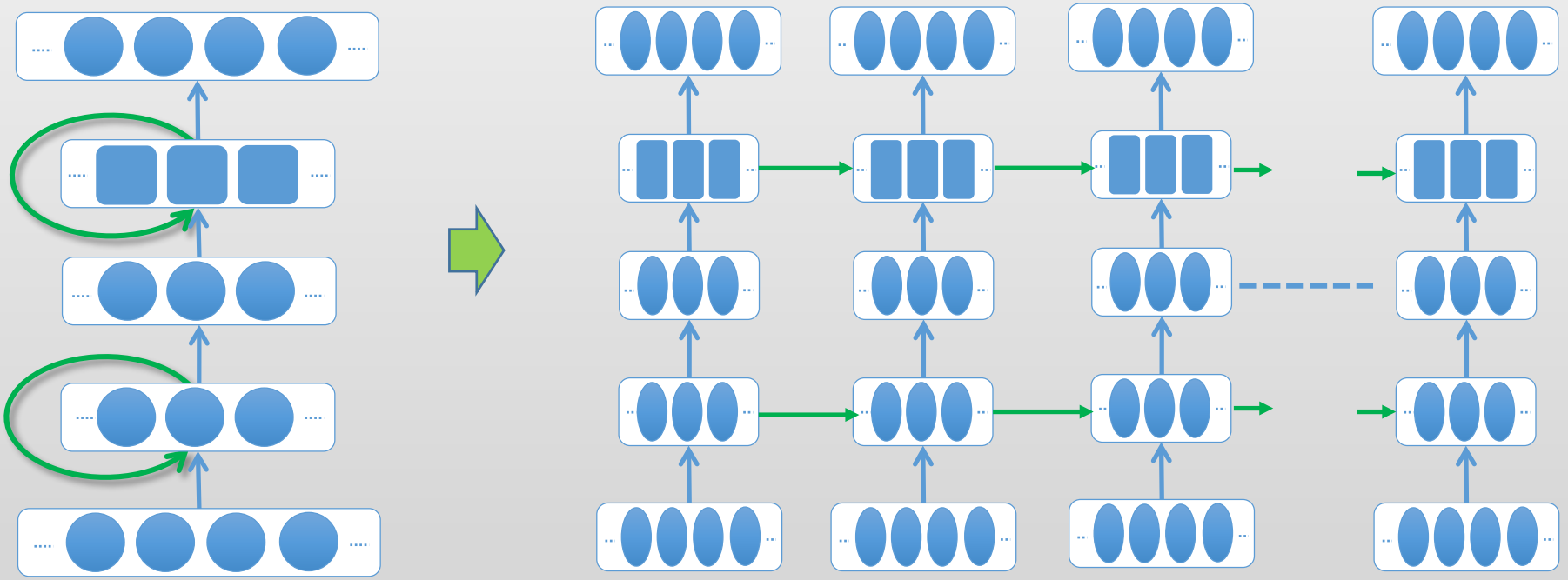


プレトレーニング

ファインチューニング

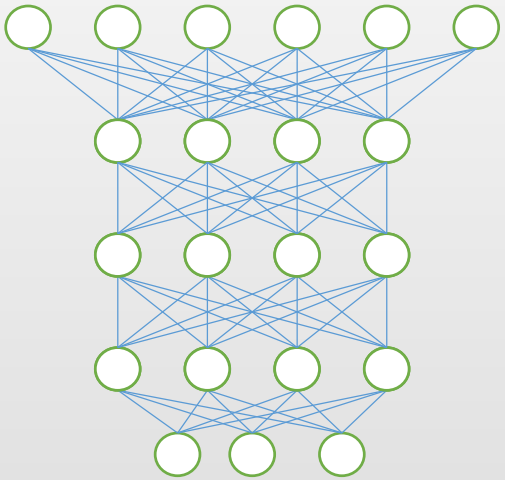
RNN言語モデルの学習プロセス

① RNNのループを
時間方向に展開



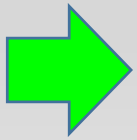
② 時間展開バックプロパゲーション
（BPTT）により教師付き学習

音響・言語モデルの学習の計算規模



大規模な音声データ
（数百～数千時間）

大規模な
ニューラルネットワーク
（数百万～数千万ユニット）

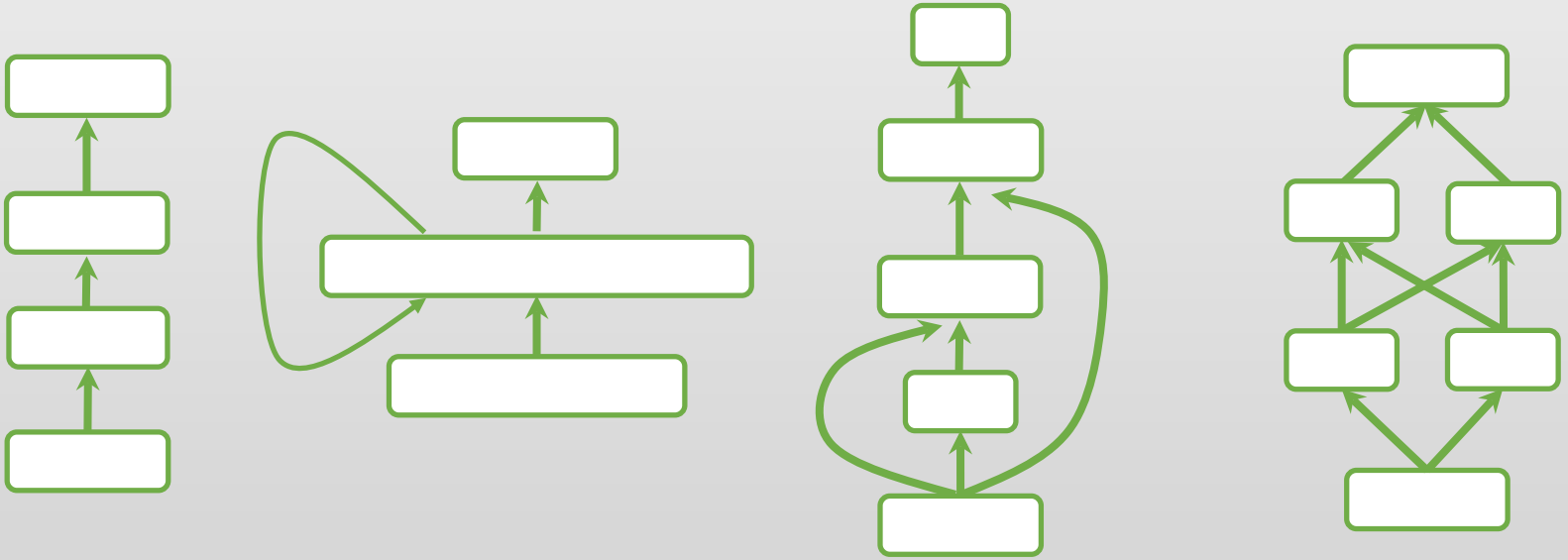


最新GPGPUを用いて数日～数か月

音声認識システム構築の問題点

相互依存する多数の調整要素（メタパラメタ）が存在

・ニューラルネットワークの構造
（隠れ層数、各層のユニット数、ユニット種別、接続構造等）



・学習条件
（学習率、モーメンタム、ミニバッチサイズ等）

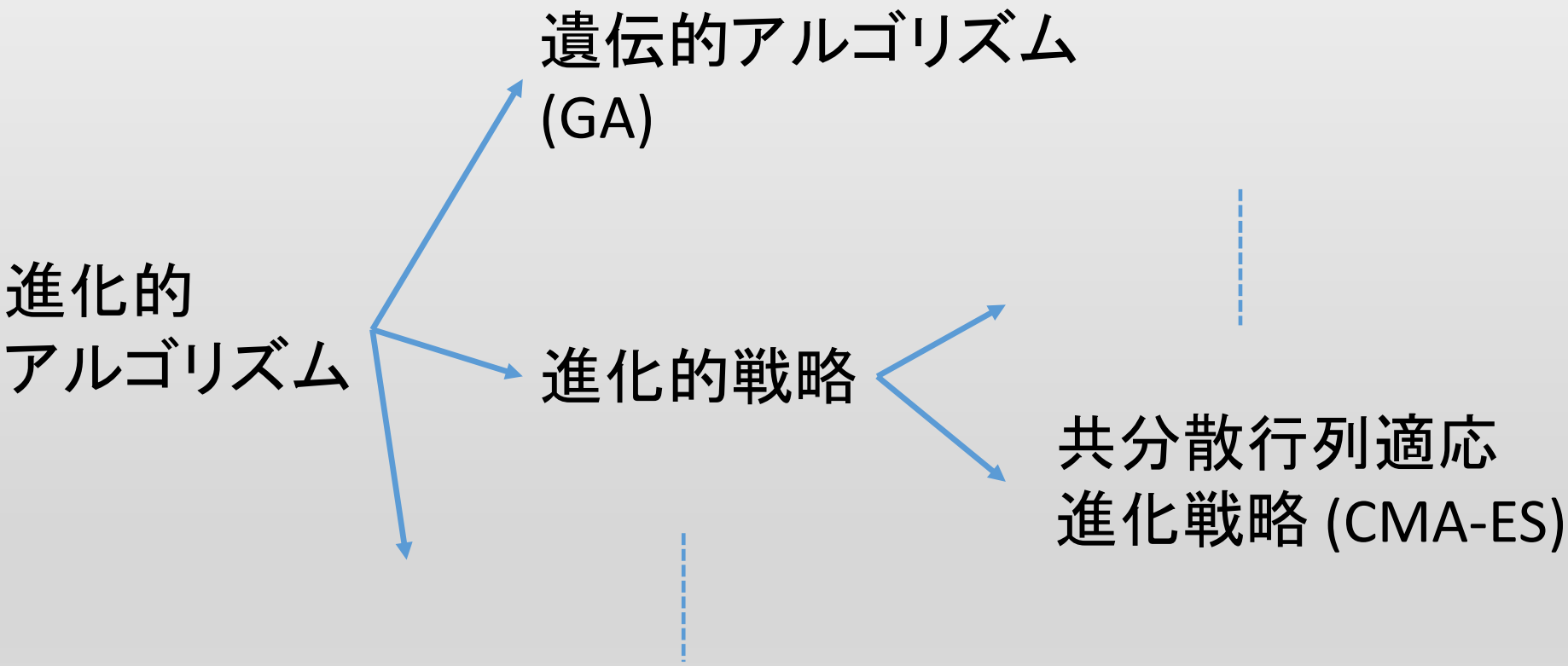
ノウハウと試行錯誤の世界

従来アプローチと提案アプローチ

従来アプローチ
専門家によるチューニング

提案アプローチ
進化計算による自動チューニング

進化的アルゴリズムの種類



CMA-ESによる音声認識システムの自動最適化

①メタパラメタを遺伝子（実数値ベクトル）にエンコード

初期メタパラメタ

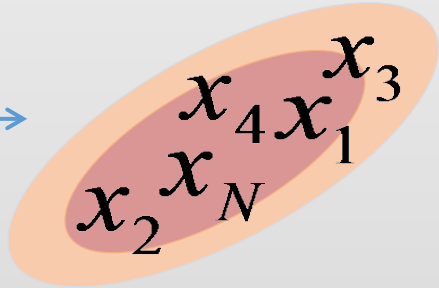
- 隠れ層の層数
- 各層のユニット数
- 学習係数

etc.

② 遺伝子分布をガウス分布で表現

ガウス分布

③ サンプリングにより遺伝子個体群を生成



繰り返す

④ 各遺伝子から個体を実体化・評価

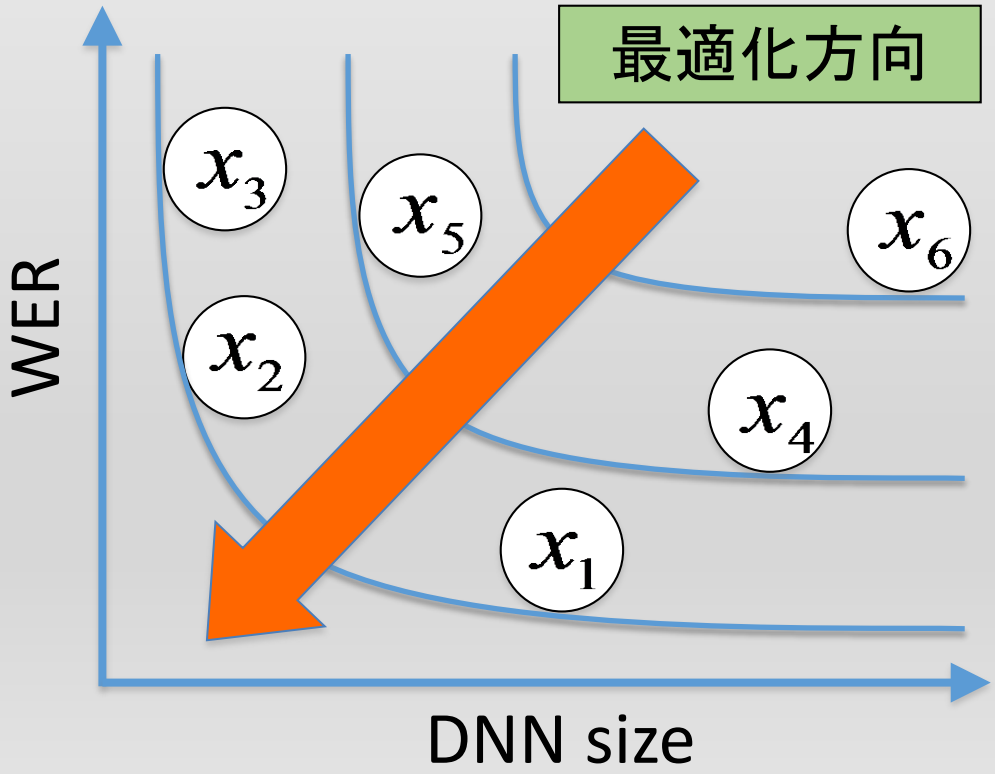
評価結果

⑤ 評価結果をもとに分布を更新

パレート最適を用いた多目的最適化

認識性能と計算時間など、複数の目的関数を同時最適化

$$\begin{cases} f_j(\boldsymbol{x}_k) \leq f_j(\boldsymbol{x}_{k'}) \ \forall\ j = 1, \ldots, J \\ \text{and} \\ f_j(\boldsymbol{x}_k) < f_j(\boldsymbol{x}_{k'}) \ \exists\ j = 1, \ldots, J \end{cases}$$



各個体の順位付け

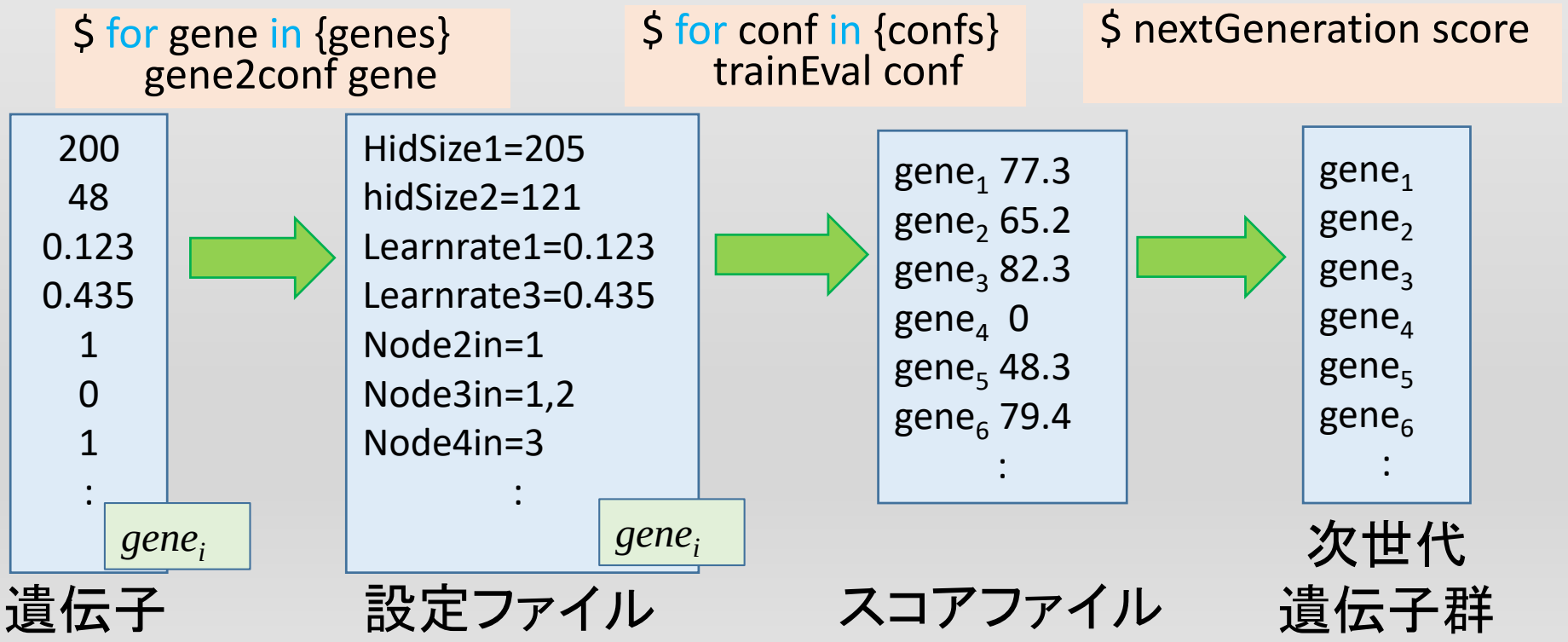
Ranking

1. x_1
   x_2
   x_3
2. x_4
   x_5
3. x_6

目的関数として使用

進化フレームワークの実装

- トップレベルスクリプト
  - gene2conf: 遺伝子を設定ファイルに変換（遺伝子の実体化）
  - trainEval: モデル学習と評価（個体の評価）
  - nextGeneration : ガウス分布の更新と次世代遺伝子群の生成



個体群の並列評価

1個体の評価　＝　ニューラルネットワークの学習と評価

全個体の評価には長時間の計算が必要

しかし、世代内においては、各個体の学習・評価は
完全に独立して実行できる

1個体に1GPUを割り当てて、
並列計算

DNN音響モデルの進化最適化

実験条件

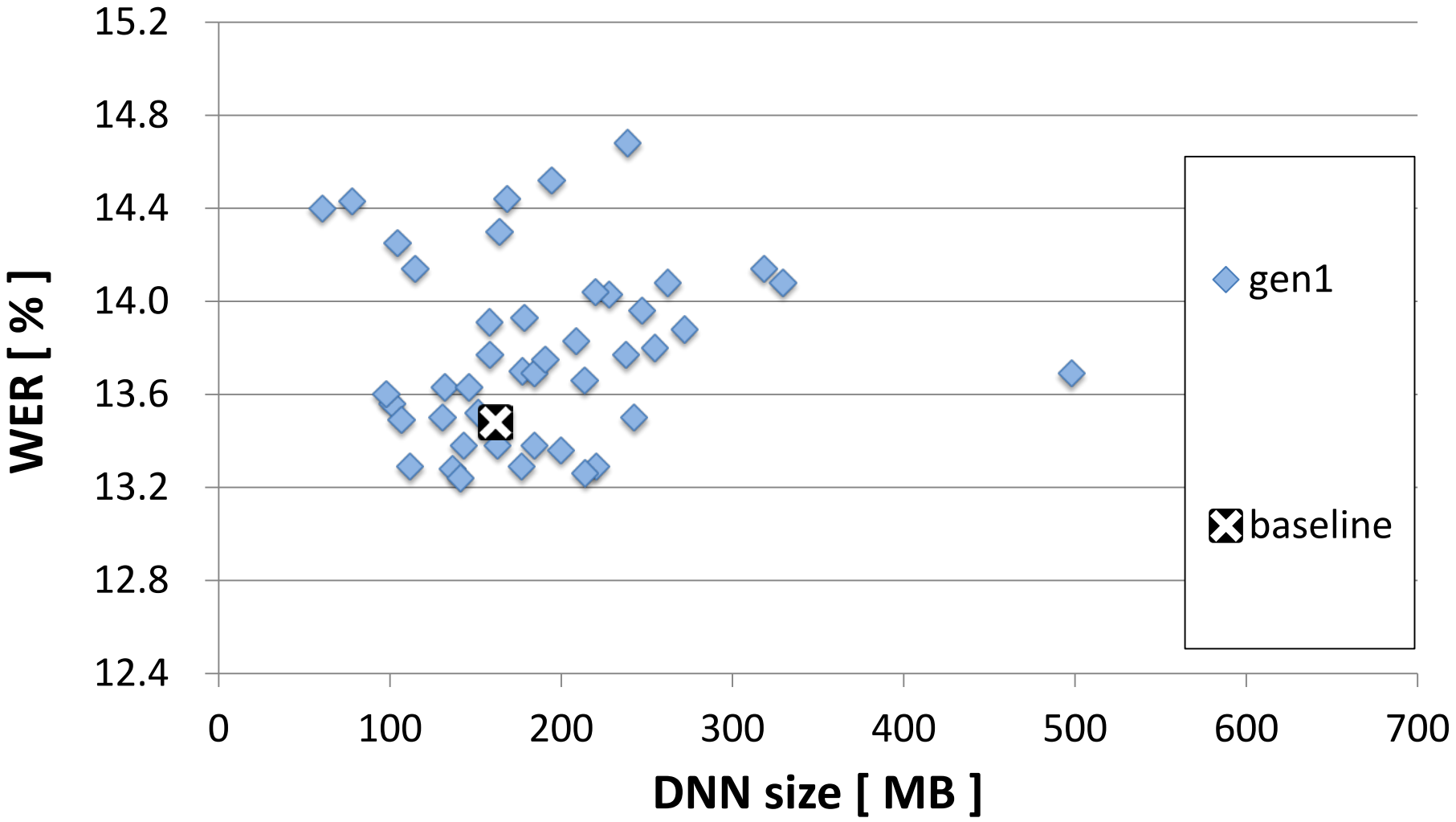
データベース	日本語話し言葉コーパス（CSJ）
認識システム構築	Kaldi Speech Recognition Toolkit
学習データ	240 時間
個体（並列）数 / 世代	44
使用計算機	TSUBAME 2.5 （GPU : Tesla K20X）
開発・評価セット	それぞれ2 時間（10講演）の音声データ

進化の評価尺度	**単語正解精度 ＋DNNファイルサイズ**

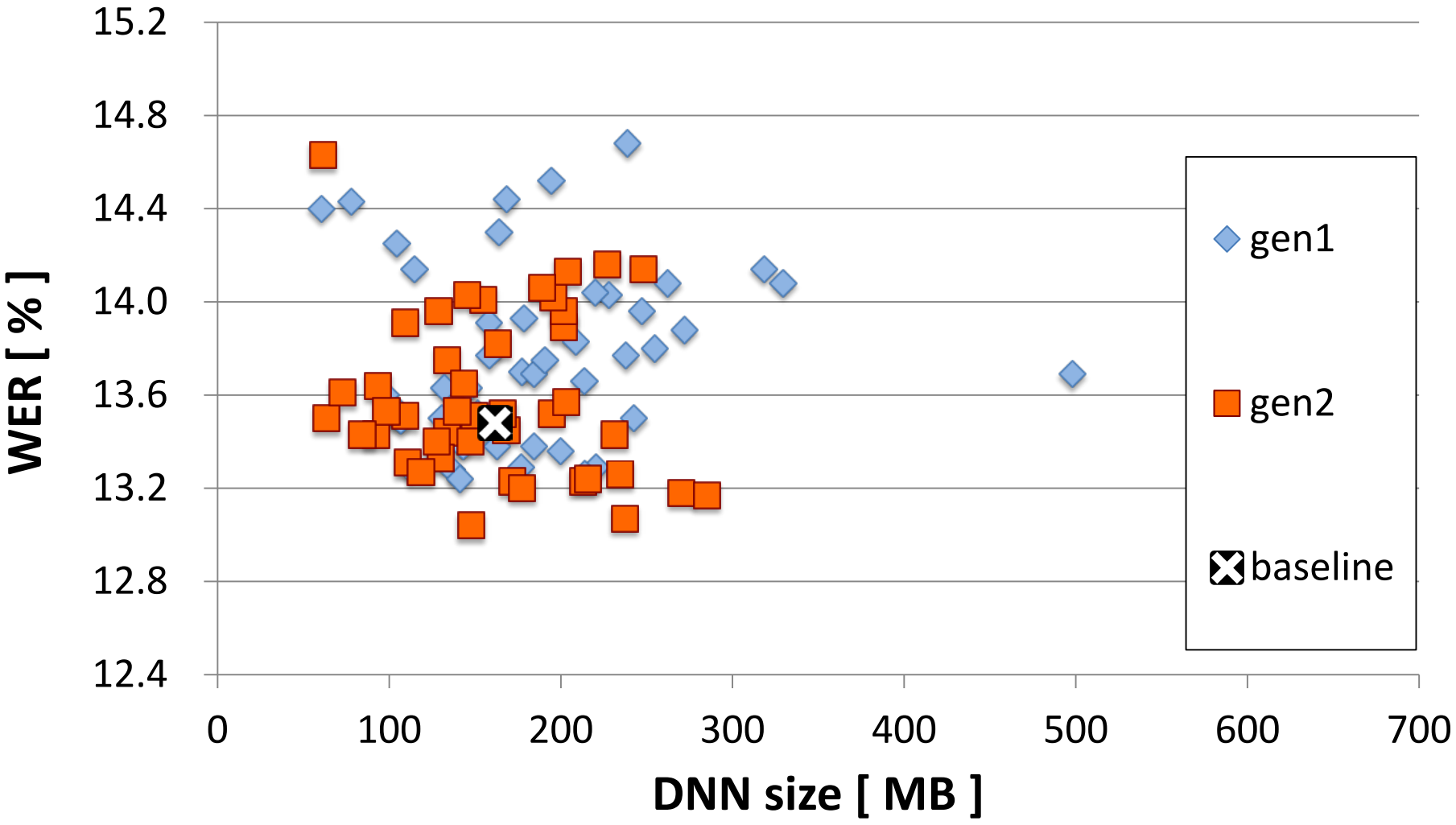
最適化対象のメタパラメタ

カテゴリ	メタパラメタ名	メタパラメタの詳細	初期値 x_{init}
特徴量	feat_type	特徴量の種類(MFCC,PLP,FBANK)	MFCC
DNN構造	splice	入力特徴量のフレーム幅	5
	nn_depth	中間層の数	6
	hid_dim	中間層1層あたりの素子数	2048
RBM学習パラメタ	param_stddev_first	パラメタ初期化の分散値	0.1
	param_stddev		0.1
	rbm_lrate	学習率	0.4
	rbm_lrate_low		0.01
	rbm_l2penalty	L2正則化係数	0.0002
DNN学習パラメタ	learn_rate	学習率	0.008
	momentum	モーメンタム	0.00001

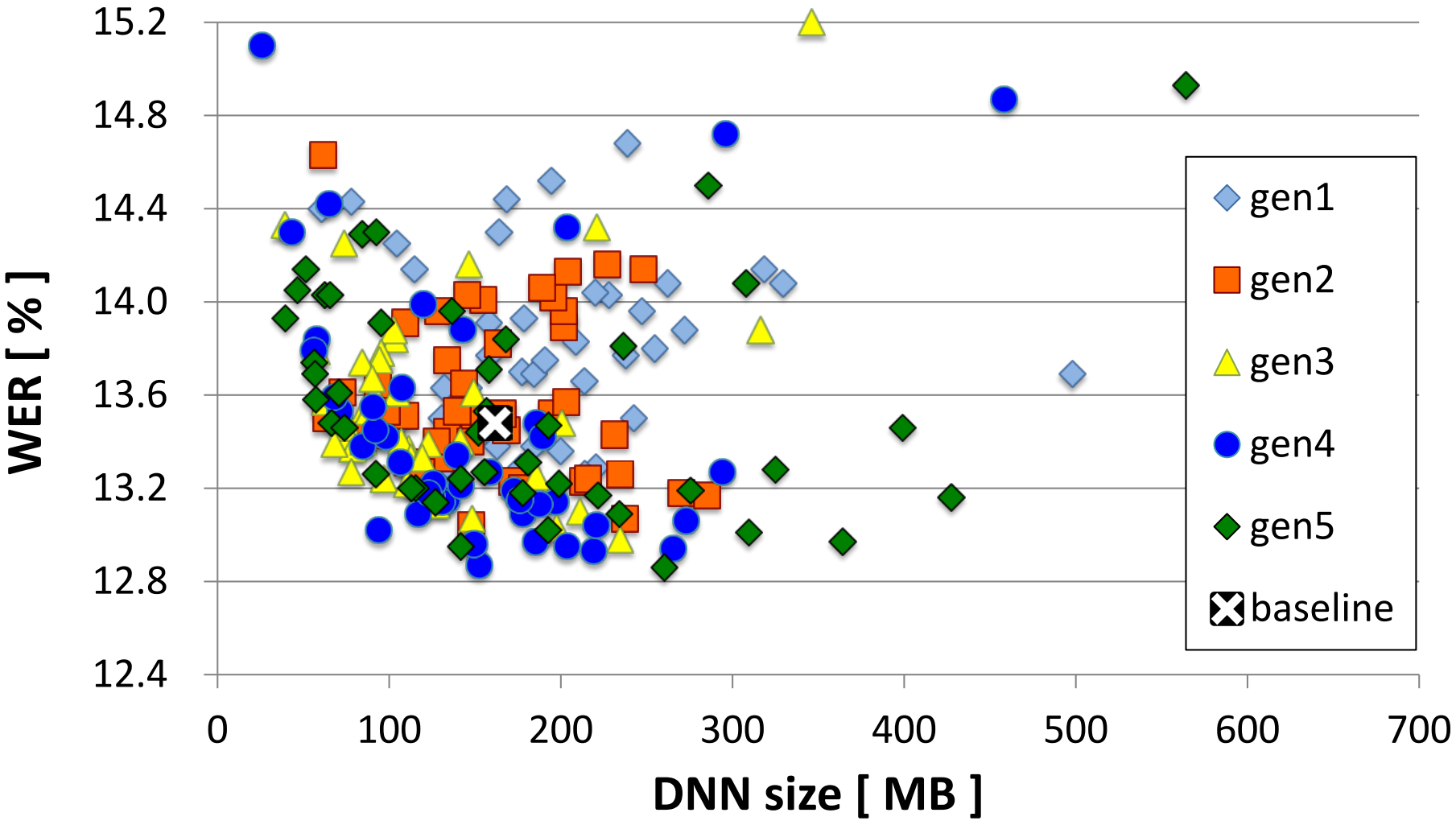
各世代の性能分布


15.2
14.8
14.4
14.0
13.6
13.2
12.8
12.4
WER [%]
0
100
200
300
400
500
600
700
DNN size [MB]
gen1
baseline

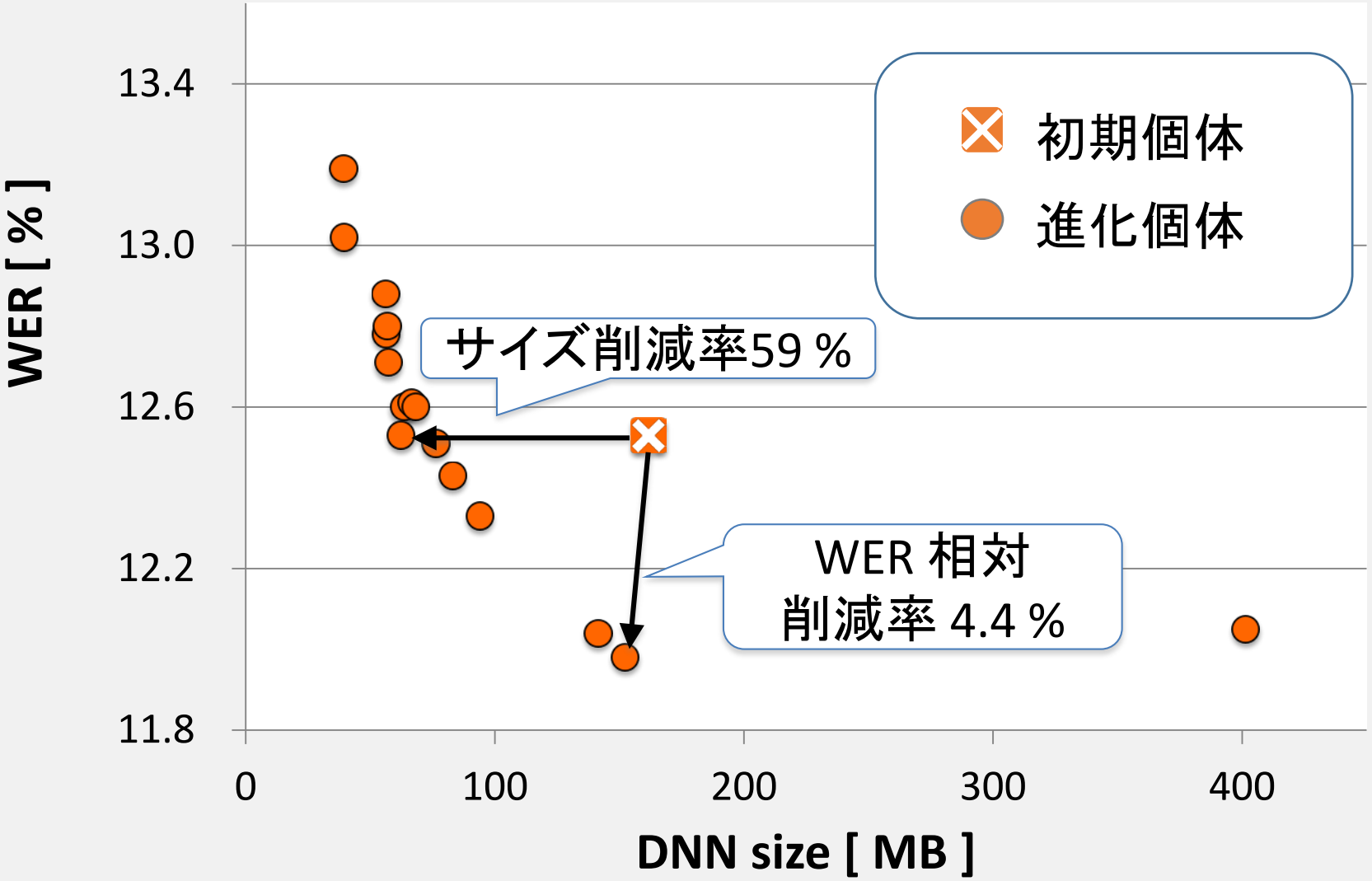
各世代の性能分布


15.2
14.8
14.4
14.0
13.6
13.2
12.8
12.4
WER [%]
0
100
200
300
400
500
600
700
DNN size [MB]
gen1
gen2
baseline

各世代の性能分布


15.2
14.8
14.4
14.0
13.6
13.2
12.8
12.4
WER [%]
0
100
200
300
400
500
600
700
DNN size [MB]
gen1
gen2
gen3
gen4
gen5
baseline

全個体によるパレートフロント



RNN言語モデルの進化最適化

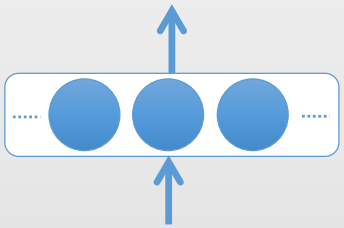
最適化対象のメタパラメタ

- ネットワーク構造
  - 隠れ層の数，各層のユニット数
  - 語彙サイズ（= 入力層，出力層のユニット数）
  - 各層のユニットタイプ (LSTM，RNN，　FF)
- 各種調整値
  - 学習率の初期値，モーメンタム
  - ミニバッチサイズ
  - 学習器の種類とそのメタパラメタ

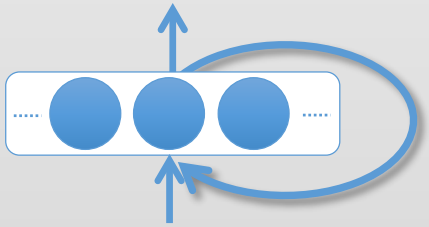
  etc.

隠れ層のユニットタイプ

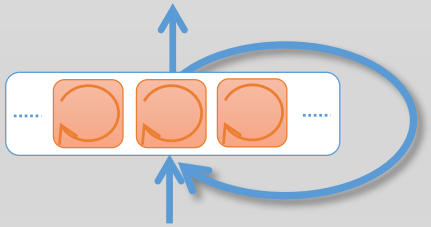
層毎のユニットタイプを可変にして進化により選択



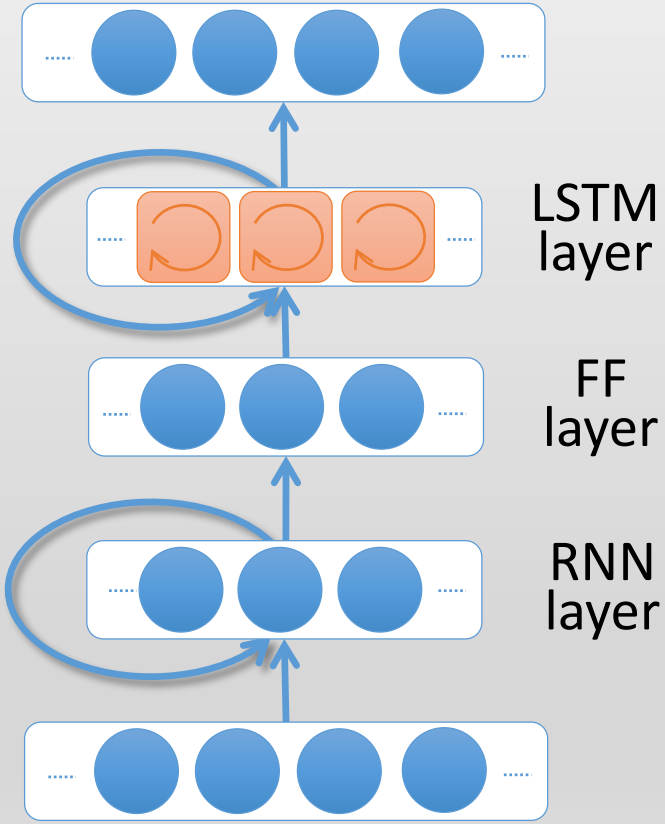
FF layer
最も単純な構造



RNN layer
再帰構造により過去の
文脈情報を利用可能



LSTM layer
RNN layer より長い
文脈情報が利用可能



例

*LSTM: Long Short-Term Memory

実験条件

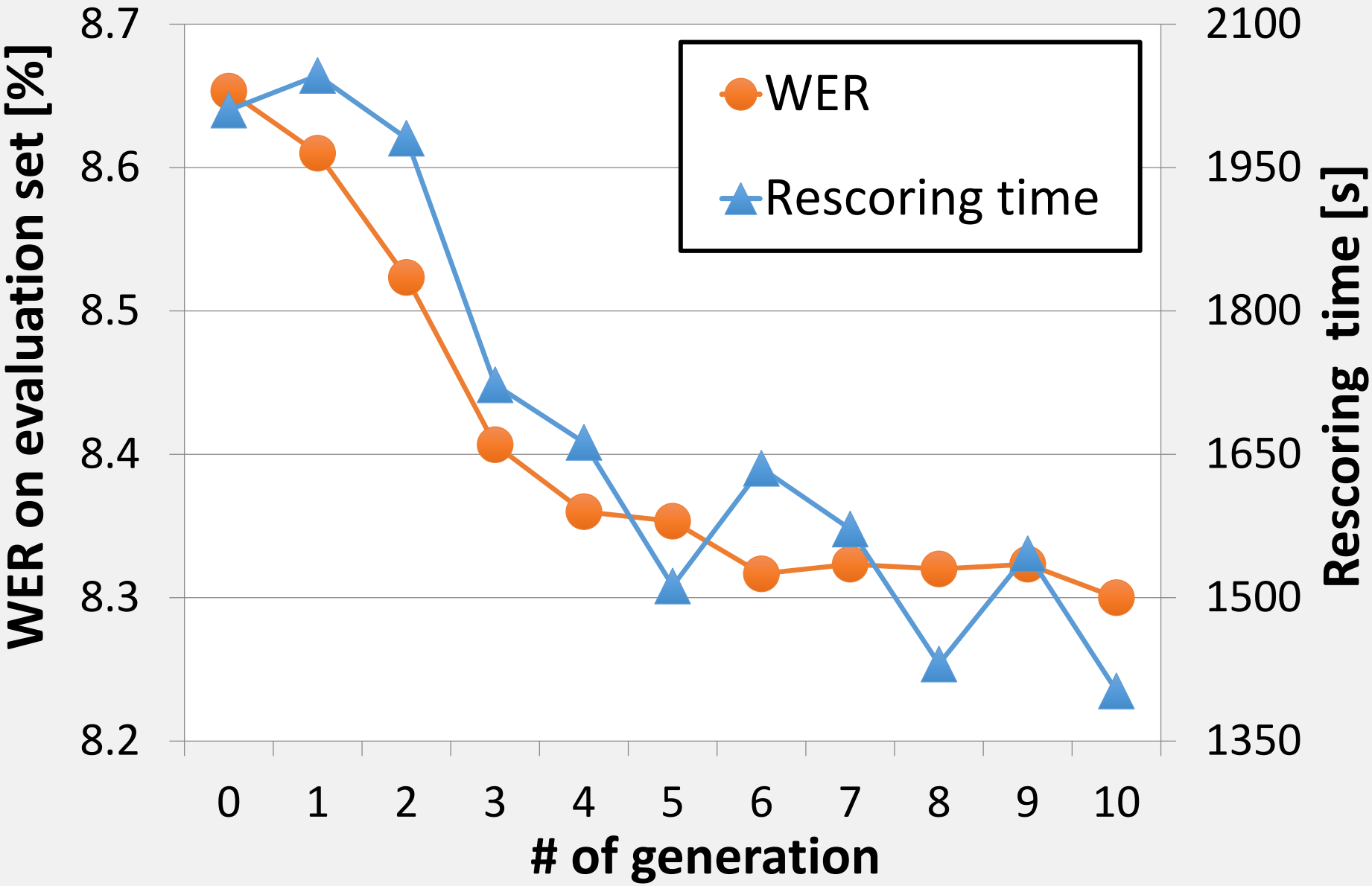
- 全般条件

データベース		日本語話し言葉コーパス（CSJ）
学習セット	音響モデル	520 時間
	言語モデル	750万 単語
評価セット		6 時間
認識デコーダ		Kaldi

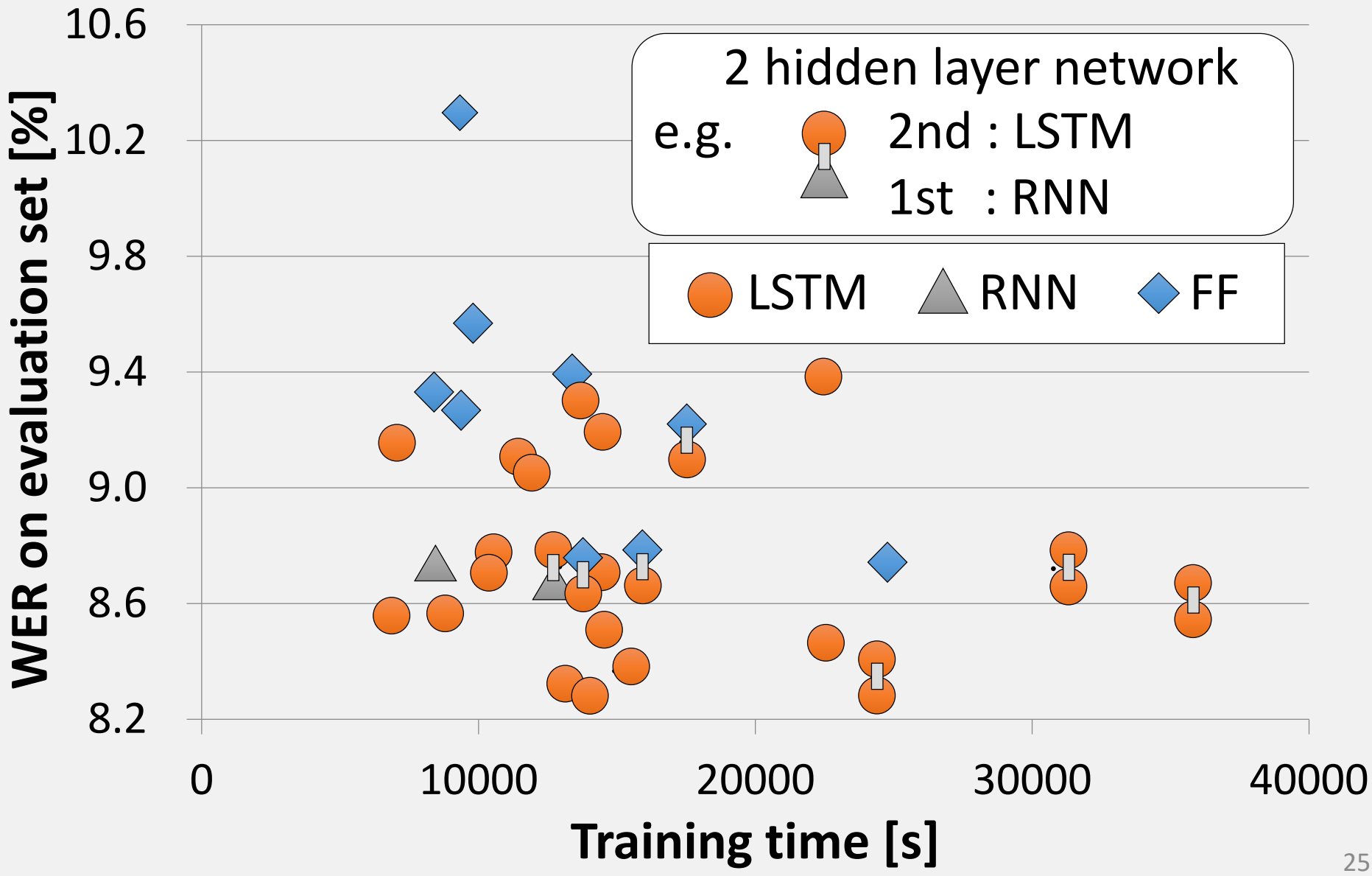
- 進化条件

開発セット	約 30 時間
世代数	10 世代
個体数	30 個体
評価尺度	単語正解精度＋ RNN言語モデルの計算時間

進化結果


WER
Rescoring time
WER on evaluation set [%]
Rescoring time [s]
# of generation
8.7
8.6
8.5
8.4
8.3
8.2
2100
1950
1800
1650
1500
1350
0
1
2
3
4
5
6
7
8
9
10

モデル構造と性能の分布（10世代目）



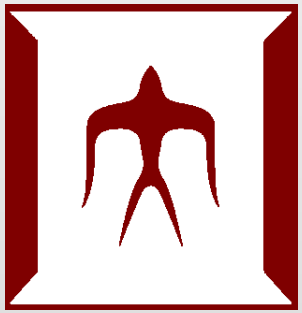
まとめと今後の課題

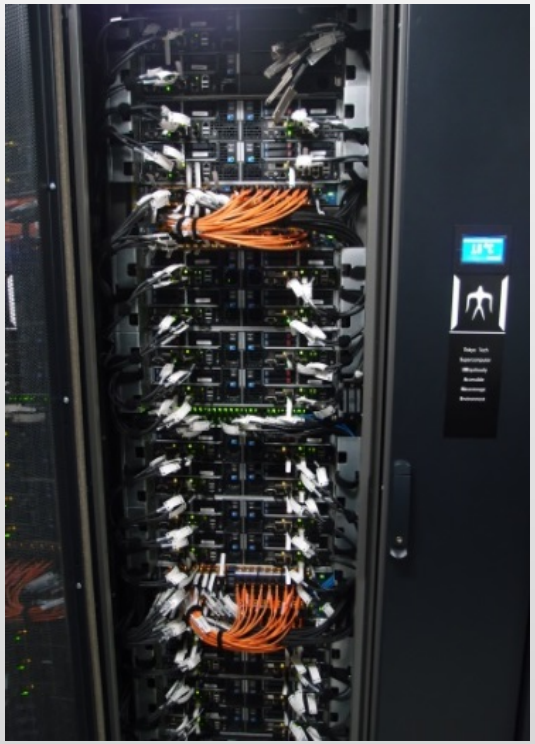
まとめ

- 進化的アルゴリズムにより、音声認識システムのシステム開発でボトルネックとなっているメタパラメタチューニングを自動化
- パレート最適を用いることで、単語誤り率WERと計算時間など複数の尺度を同時最適化
- 開発した高精度日本語音声認識システムは、Kaldiパッケージに収録され一般利用が可能　http://kaldi-asr.org/

今後の課題

- より複雑なネットワーク構造や多様なユニットタイプを対象とした進化実験
- 進化効率の向上と超大規模データへのスケールアップ

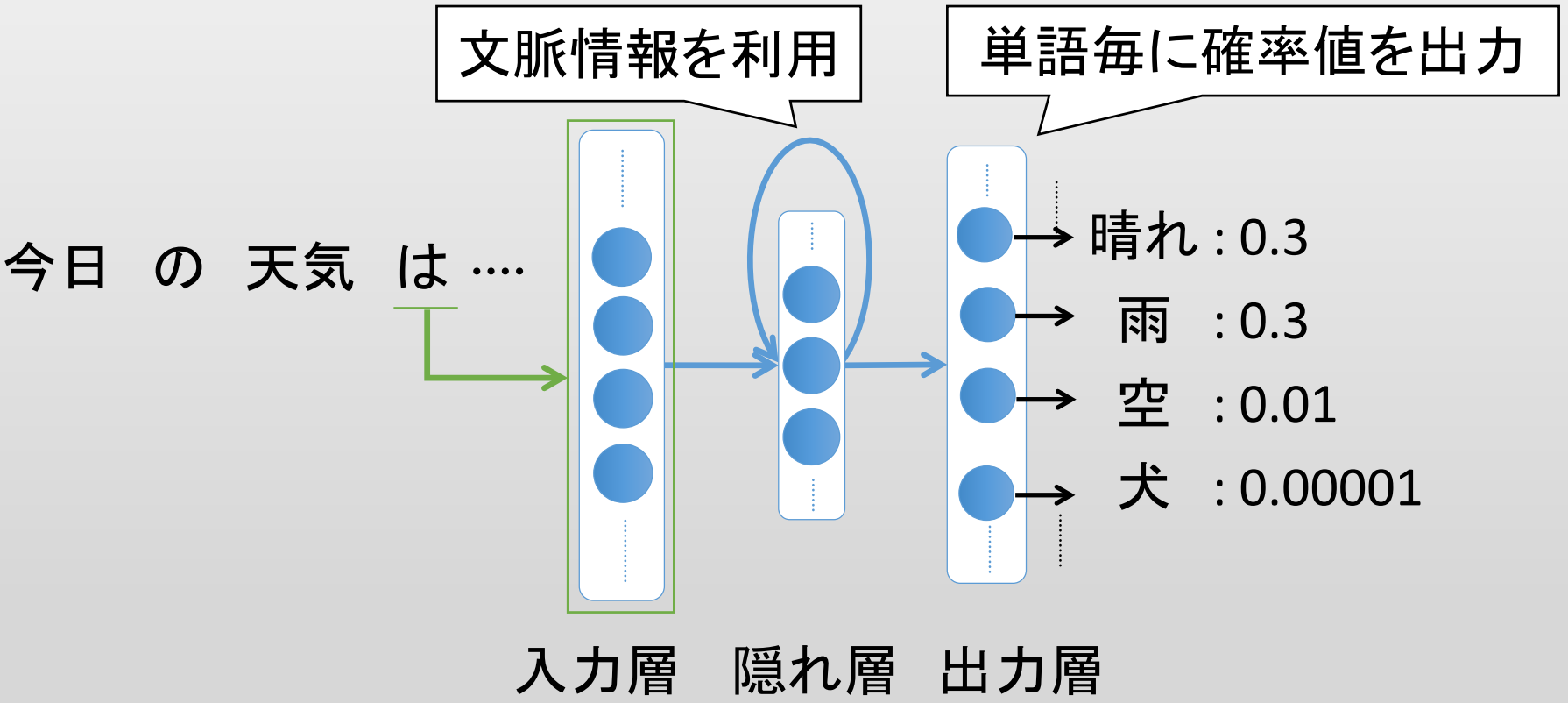




END

補足資料

RNN言語モデルの基本構造



LSTM layer

- RNN layer の各ユニットをLSTMブロックに置換
- RNN layer より長い文脈情報を利用可能