# サーチ・LVCSRシステム

名古屋工業大学
李　晃伸

# 第1回シンポジウム（1999年12月）

- 特別講演1「音声認識研究の課題」
  中川聖一(豊橋技科大)
  - 音響モデルと認識アルゴリズムが重要
- 特別講演2 「日本語ディクテーションソフトの現状と今後の課題」西村雅史(日本IBM)
  - 基本的枠組みの確立 (CD HMM + 3-gram)
  - 読み上げはほぼ完成
  - 自由発話認識へ

# 当時のシステム

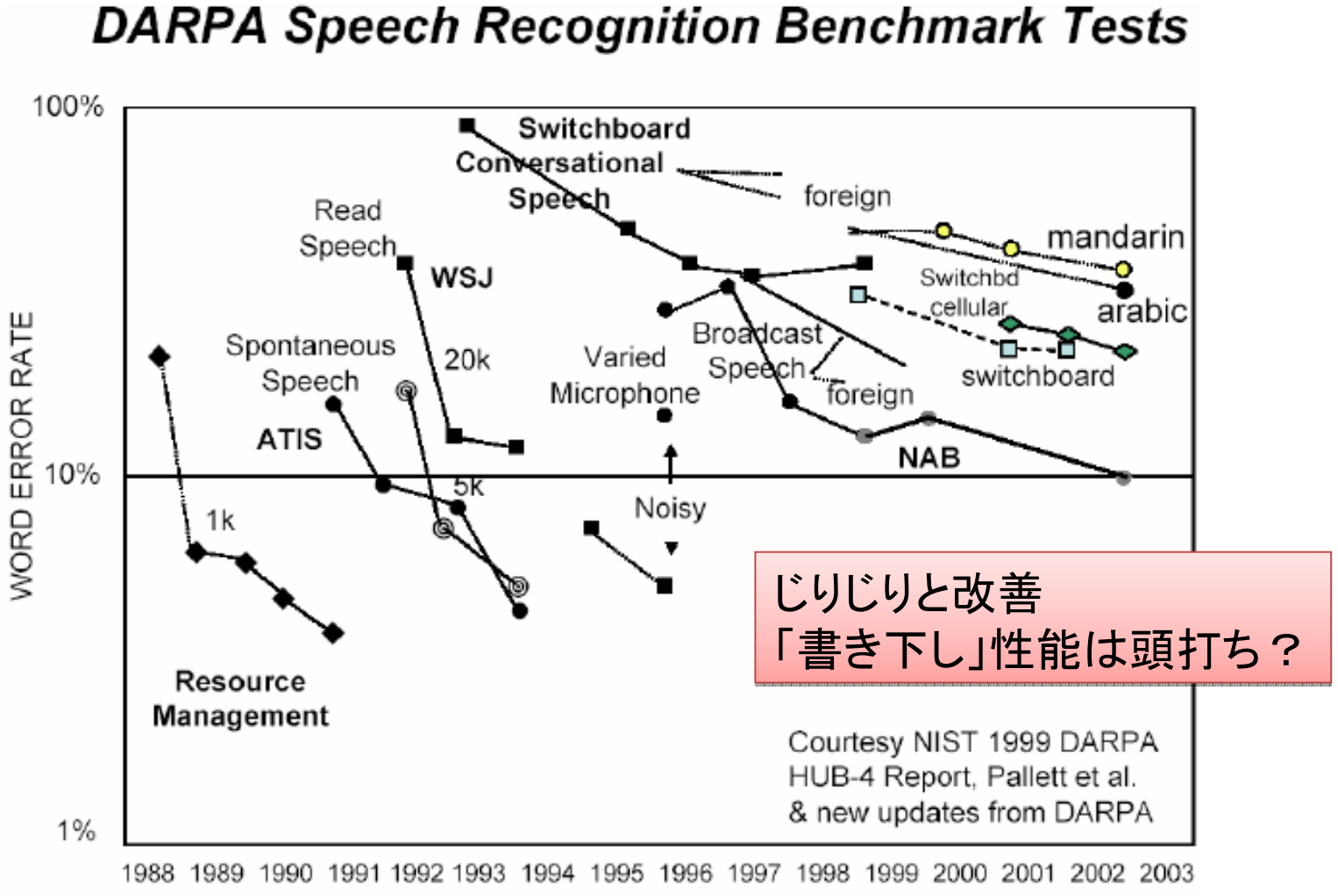
- 基本的なサーチアルゴリズムの確立
  - 単語 N-gram + 音素環境依存HMM
  - 1パス探索／マルチパス探索
  - ビーム探索／ヒューリスティック探索
  - 動的展開／静的展開
- 読み上げ音声DBでのベンチマーク
- ディクテーションの実現
  - IPA「基本ディクテーションソフトウェア」
  - SmartVoice (NEC), ViaVoice (IBM), etc…

# 音声言語データベース

- ～2000：読み上げ音声
  - DARPA (WSJ), SQALE, JNAS
- ～2004：話し言葉・Rich Transcription
  - CSJ：学会講演・模擬講演が中心
  - DARPA EARS Project：電話音声・ニュース音声・会議音声
- 現在：多言語データベース
  - DARPA GALE (2006～)
    - English, Arabic, Chinese　ニュース・会話等
    - 書き起こし・翻訳・要約
  - EU：TC-Star (2004～)
    - 欧州言語（＋中国）　会議・講演等
    - オープンソースツール (Moses, IRSTLM, etc.)
  - アジア言語：A-STAR コンソーシアム (2006～)

データ量の増大
学習理論の発展
多言語化

# 認識精度の変遷




じりじりと改善
「書き下し」性能は頭打ち？

From "*Automatic Speech Recognition - A Brief History of the Technolog*
By B.H.Juang and L.R.Rabiner, Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

# Recent Advances In The IBM GALE Mandarin Transcription System (ICASSP2008)

**Table 1.** Comparing CER of the baseline system (BASE) and the tonal system (TONE) on *eval*'06 and *dev*'07.

| | | | BASE | TONE |
|---|---|---|---|---|
| SI: | ---- | *eval*'06: | 24.6 | 23.5 |
| | ---- | *dev*'07: | 18.9 | 18.5 |
| SA:ML | +fMLLR | *eval*'06: | 21.9 | 21.3 |
| | | *dev*'07: | 15.7 | 15.4 |
| | +MLLR | *eval*'06: | 21.7 | 21.2 |
| | | *dev*'07: | 15.4 | 15.0 |
| SA:MPE | +fMLLR | *eval*'06: | 18.6 | 18.1 |
| | | *dev*'07: | 12.4 | 12.0 |
| | +MLLR | *eval*'06: | 18.5 | |
| | | *dev*'07: | 12.2 | |

| | *dev*'07 | *eval*'06 | *eval*' |
|---|---|---|---|
| BASE: | 12.2 | 18.5 | 11.( |
| TONE: | 11.8 | 17.9 | 10.( |
| BASE × TONE: | 11.5 | 17.6 | 10.1 |
| Topic LM + BASE × TONE: | 10.9 | 17.1 | 9.5 |

学習データ:
AM:1321 時間
LM:5GB テキスト

モデル:
AM:quintphone
(10k state, 300k Gauss.)
LM:4-gram (107k-word)

不特定話者・適応無し:24.6
話者適応:18.5
Tone modeling: 17.6
Topic LM: 17.1
(*eval'06*)

# デコーダ技術の進展

- サーチの効率化
  - 基本アルゴリズムはほとんど変化無し
  - 実装の洗練（パラメータ共有，最適化），大規模モデル対応
- 計算量削減
  - GS，固定小数点化，マルチスレッド，GPU，省電力チップ
- 機能拡張
  - 逐次確定
  - 端点フリー認識
  - デコーダベース VAD
  - マルチデコーディング
- 多様な出力・統合法
  - 単語グラフ・Confusion Network・信頼度
  - ROVER・CNC
  - DSR

# Julius の開発

- 1.0 (1998.02)
- 2.2 (1999.10)
  - 探索アルゴリズム改善，32k，第1パス単語間triphoneの近似計算
  - 第2パスにビームを導入
- 3.0 (2000.02)
  - PTM，Gaussian pruning，64k 対応，アラインメント
- 3.1p2 (2001.02)
- 3.2 (2001.08)
  - ショートポーズセグメンテーション
- 3.3 (2002.09)
  - モジュールモード，複数文法同時認識，SS，マルチパス版
- 3.4 (2003.10)
  - サーチ調整，信頼度計算，クラスN-gram，バイナリHMM
- 3.4.2 (2004.5) CSRC最終版
- 3.5 (2005.11)
  - GMMによる棄却，単語グラフ出力，信頼度枝刈り
  - MFCC　HTK互換性拡張，MAP-CMN (3.5.1)
- 3.5.2 (2006.07)
  - 計算速度 20% 前後アップ，文法の最小化，SLF2DFA
- 4.0 (2007.12)
  - Julian 統合，モジュール化，ライブラリ化，マルチデコーディング
  - 任意長N-gram，孤立単語認識，CN, GMM-VAD, Decoder VAD,
- 4.1 (2008.10)
  - プラグイン対応，マルチストリーム対応，MSD-HMM対応，CVN，VTLN
  - ドキュメント刷新，Juliusbook 作成

基本的アルゴリズムの確立

IPAツールキット (2000)

サーチの細かい改善
逐次音声認識
単語信頼度

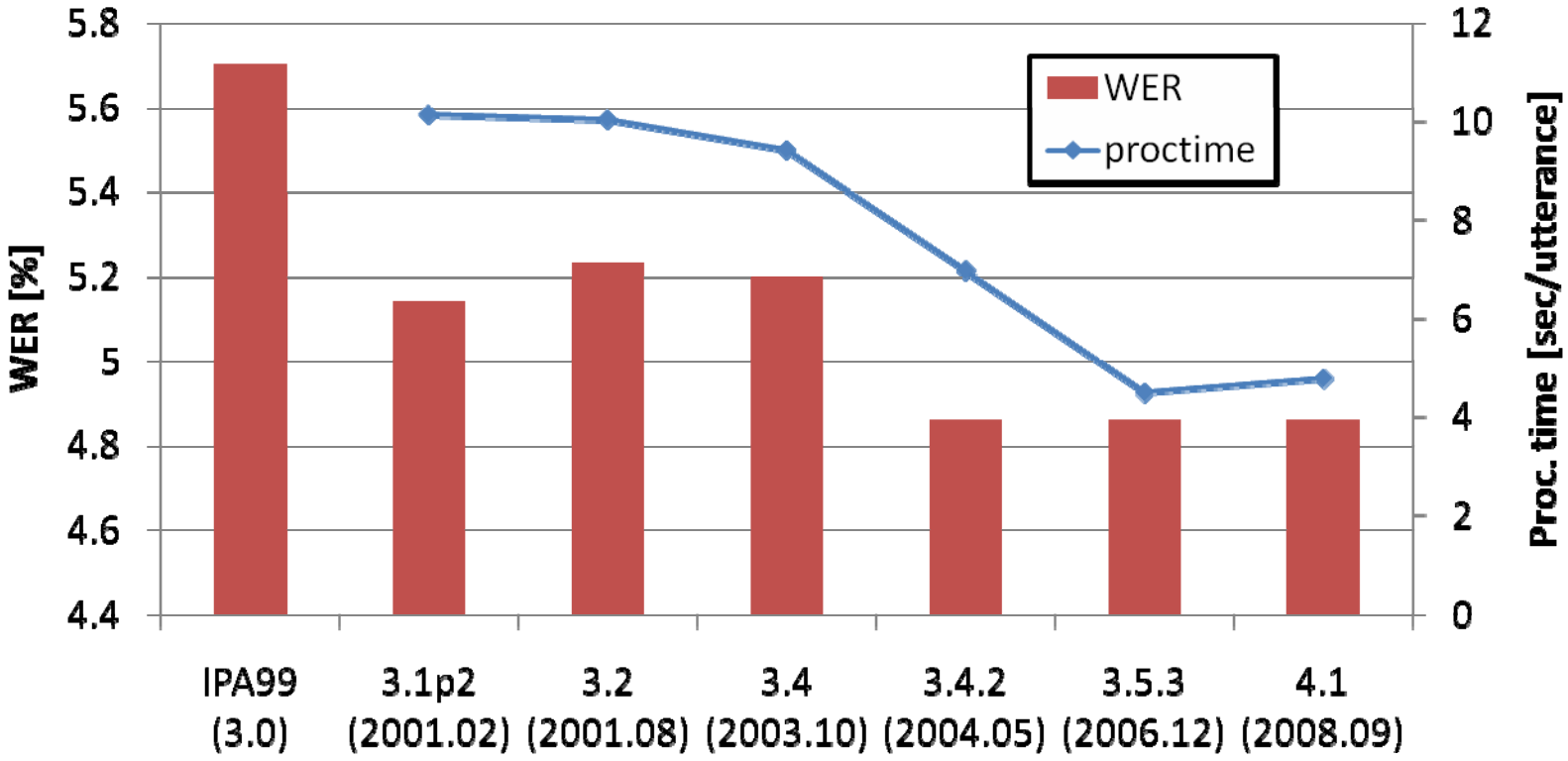
CSRC最終版 (2003)

単語グラフ出力
入力棄却
互換性向上
各部の最適化

ISTC最終版 (2008)

再構築
モジュール化
言語制約拡張
マルチデコーディング

# Juliusの性能変遷




Model: Gender-dependent triphone (2000x16), 20k-word 3-gram
Test set: JNAS IPA98 Test set (23 male, 23 female, 200 utterances)
CPU: Pentium 3GHz

# 10年でできたこと

- サーチ・デコーディングの進歩
  - アルゴリズムの精査，キャッシュやアクセスの最適化
  - wFST デコーダの登場
  - 組み込みから1000万単語まで
- 音声認識システムの実用化
  - 識別学習など大量データを前提とした学習方法の発展
  - 発話対象や発話様式をある程度絞った分野で実用化
  - ディクテーション
    - 医療カルテの口述筆記
    - リアルタイム字幕付与（NHK）
  - 音声検索
    - Voice search (Microsoft, Google, etc.)
  - 音声翻訳
    - 旅行会話ドメインなど（ATR）

# 10年でできなかったこと

- LVCSR システム
  - 自由発話認識のブレイクスルー
    - 不特定話者認識の単語誤り率：20% 前後
    - データを集めるだけでは突破できない？
- デコーダ
  - 他制約・モデルとの新たな密統合を可能とする柔軟なエンジンアーキテクチャ（⇔ブラックボックス化）
    - wFST は一つの解

# 今後の展開

- ソフトウェアとしての音声認識エンジン
  - 大規模システム用とコンパクト用に2極化
  - コモディティ化
- デコーダの柔軟化
  - 自然言語処理との密統合（音声言語処理の高度化）
  - 信号処理との密統合（ロバスト音声認識）
- LVCSR システム
  - 言語理解と融合したシステムへ
  - システム構築支援・（半）自動構築手法の研究
- 日本語LVCSRの研究用共通基盤の整備？
  - Julius の共同開発WG