

## ISTCにおける 音声認識ソフトウェアの開発状況

河原達也 (京大)  
李 晃伸 (名工大)

## これまでの経過

- 1995～1997 IPSJ/SLP傘下 WG
  - JNASコーパスの設計
- 1997～2000 IPAプロジェクト
  - 「日本語ディクテーション基本ソフトウェア」の開発
- 2000～2003 連続音声認識コンソーシアム CSRC
  - 認識ソフトウェアの改善
  - 音響・言語モデル等の充実
- 2003～2008 e-Society基盤ソフトウェア
- 2003～2006 音声対話技術コンソーシアム ISTC
  - 対話システムを指向した音声認識の改善

## ISTCでの主要開発目標

- 音声認識エンジンJuliusの性能改善・機能追加
  - Julius 3.5 2005年11月リリース
- 音声認識エンジンJuliusのSAPI/SALT対応
  - 2003年度
- 音声認識ソフトウェアのカスタマイズを容易に
  - 2004/2005年度
  - <http://htk.ar.media.kyoto-u.ac.jp/julicus/>
- 多様なパッケージ
  - 英語版、雑音対応版

## Julius の開発状況

Rev. 3.5	(2005/11/11) - ISTC2005
Rev. 3.5.1	(2006/3/31)
Rev. 3.5.2	(2006/7/31)
Rev. 3.5.3	(2006/xx/xx) - ISTC2006

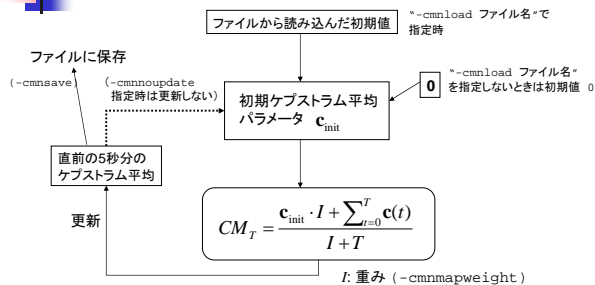
## 主要な変更点

- Rev. 3.5.1 (2006/3/31)
  - MAP-CMNの導入
  - HTK で可能なMFCC抽出条件をフルサポート
- Julius-3.5.2 (2006/7/31)
  - 単語グラフの精度改善
  - Windows版の音声入力遅延改善
  - 音響モデルのメモリ管理の最適化
- Julius-3.5.3 (2006/xx/xx) (予定)
  - 文法関連のツールを追加: dfa\_minimize, slf2dfa
  - メモリ管理の改善 (特に第2パス)

## MAP-CMN (3.5.1)

- マイク入力時のCMNを改善
- 従来法
  - 直前の5秒間のケプストラム平均
    - ⇒ 次の発話のケプストラム平均として適用
  - 話者交代時にミスマッチ
- MAP-CMN
  - 初期ケプストラム平均パラメータを用意
  - 初期フレームは上記のパラメータでCMN
  - 入力が伸びるにつれて発話自身のCMNに近づける
  - 直前の発話で初期パラメータを更新

## アルゴリズムとオプション



## 単語グラフ生成の改善 (3.5.2)

- アルゴリズムの修正
  - 第2パスで仮説マージ時に、スコアが高いパスの探索を優先
  - 後処理でグラフの深さによるカットオフを行う
  - 後処理の境界確定でループの打ち切り値を設定
  - 同一位置でスコアの異なる単語をマージしないようにできる
- デフォルト
  - カットオフ、ループ打ち切りはdefault で on
  - 3.5.1 以前と出力されるグラフが異なるので注意

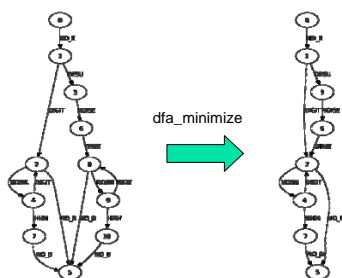
## おまけ: 1パスの単語グラフ生成

- 2-gram を用いた1パスのグラフ生成をサポート
    - 第1パスで単語トレリスの代わりにグラフを生成する
    - 第1パス終了時に単語グラフを出力
    - 実質上単語対近似が必須
- `configure --enable-word-graph --enable-wpair`
- 第2パスはグラフ上でのみ行われる
    - 第1パスの単語グラフを制約として探索
    - グラフのリスコアリングと等価

## DFA最小化ツール dfa\_minimize

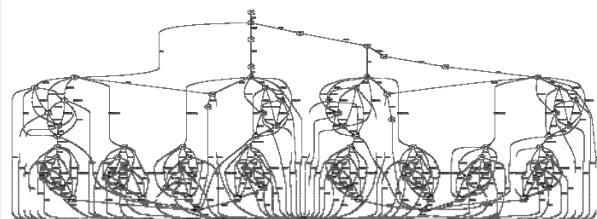
- 文法の有限オートマトン (DFA) を最小化
  - 現在の mkdfa.pl の出力は冗長な状態を含む
  - 最小化により、冗長な状態を削除
  - 認識処理の効率の改善
- 3.5.3 に含める予定

## サンプル文法 digit



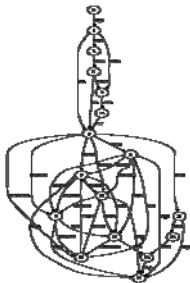
## サンプル文法 price

87 nodes, 267 arcs



## 最小化後

17 nodes, 39 arcs



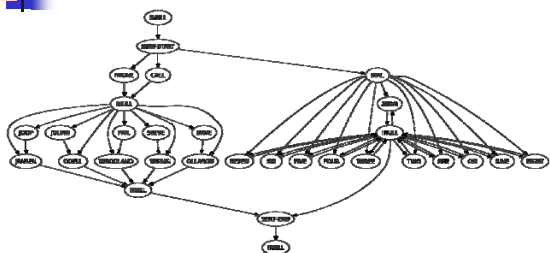
## SLF2DFA

- HTK の文法・辞書を Julian 形式へ自動変換
- 3.5.3 と同時にリリース

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
        SIX | SEVEN | EIGHT | NINE | OH | ZERO;
$name = [ JOOP ] JANSEN |
        [ JULIAN ] ODELL |
        [ DAVE ] OLLASON |
        [ PHIL ] WOODLAND |
        [ STEVE ] YOUNG;
(SENT-START (DIAL <$digit> | (PHONE|CALL) $name) SENT-END)
```

HTKの文法記述の例

## Standard Lattice Format



HTK の文法ネットワーク

## HTK SLF と Julian DFA の相違点

- Moore型
- NULL遷移可能
- Left-to-right
- 辞書から文法にある単語のみ読み込んで認識
- 単語単位の制約

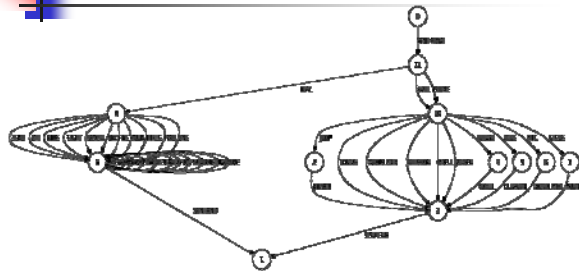


- Mealy型
- NULL遷移不可
- Right-to-left
- 辞書の単語を全部読み込んで認識
- カテゴリ単位の制約  
制約が同じ単語集合をよりコンパクトに扱える

## 変換処理のステップ

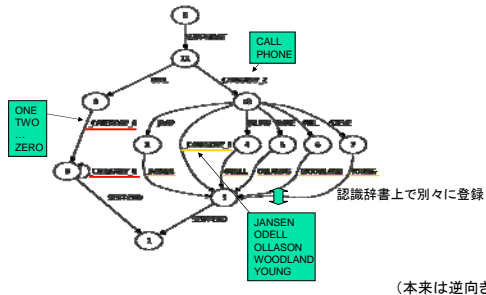
- SLFからNULL遷移を除去
- SLFをMealy型に変換
- ファイル形式をSLFからDFAに変換
- DFAを逆向きにする
- 決定化
- 最小化
- 辞書から単語を抽出して認識用辞書(DICT)を構成
- 同一遷移のシンボル集合からカテゴリ候補を検出
- 単語をカテゴリに置換し、不要な単語をDICTから除去

## DFAへの変換例



(本来は逆向き)

## 単語カテゴリの検出と置換



(本来は逆向き)

## ポイントとメリット

- HTKの文法制約を等価な Julian 用文法に自動変換
  - Julianユーザ: 正規表現による簡潔な文法記述が使える
  - HTKユーザ: Julian への移行が簡単
- 遷移パターンからの単語カテゴリ自動検出
  - 自分でカテゴリを記述する手間が省ける
  - 数詞・地名のように同じクラスの単語が多い場合に有効

## 英語版音声認識キット

- Julius/Julian
- Linux/Windows

## 英語版音声認識キット

- WSJコーパスでモデル学習
  - 音響モデル...EDAZ, 5641x16
    - Si284set (37K発話, 62時間)
  - 単語辞書...39(CMU phoneset)+sp+sil
  - ディクテーション用言語モデル...20K trigram
- Juliusの設定
  - マルチパス音響モデル
  - ショートポーズの処理

## WSJ評価結果

- ARPA Hub2, Nov. 1993
  - 5K: 215発話、7.3秒
  - 20K: 213発話、7.0秒

	Julius	HTKデコーダ
5K	91.4 /5RT	90.8 /10RT
	88.7 /RT	84.3 /RT
20K	82.4 /5RT	NA
	80.0 /RT	

## Julian用サンプル文法(英語版)

- digit: 連続数字
- number: 数字
- date: 日付
- time: 時間
- persons: 人数
- price: 値段
- yesno: Yes/No
- spell: 音素タイプ
- attendant: 受付
- fruit: 果物注文

## 用語の説明

- Julius: 大語彙連続音声認識のフリーソフトウェア
  - 当初はディクテーション向けのN-gram言語モデル対応のものだったが、現在は下記Julianも統合
- Julian:
  - (人手による)記述文法対応の認識プログラム
- SAPI: Speech API
  - マイクロソフト社策定のAPI、Windows XPに標準搭載
- SALT: Speech Application Language Tags
  - マイクロソフト社などが策定している、HTMLブラウザで音声認識・合成を行うためのタグ
  - Speech Application SDKに含まれる
- SRM: Speech Recognition Module
  - JuliusのGalatea Toolkitのためのインタフェース

## 関連Webページ

- Julius
  - 最新版(SAPI版含む)フリーダウンロード
    - <http://julius.sourceforge.jp/>
- 連続音声認識コンソーシアム(CSRC)
  - 最終版を一般頒布中(有償)
  - <http://www.lang.astem.or.jp/CSRC/>
- 日本語ディクテーションツールキット
  - 最終版は「音声認識システム」(オーム社)の付録CD-ROM
  - <http://www.ar.media.kyoto-u.ac.jp/dictation/>
- リーディングプロジェクト e-Society基盤ソフトウェア
  - <http://cif.iis.u-tokyo.ac.jp/e-society/>
- Microsoft Speech Application SDK
  - SALTには必要
  - <http://www.microsoft.com/speech/>