2005－SLP－57（5）
2005／7／15

# 大語彙連続音声認識を用いた落語のリアルタイム字幕付与

**西光 雅弘　　秋田 祐哉　　河原 達也**

京都大学 情報学研究科 知能情報学専攻
〒 606-8501 京都市左京区吉田二本松町
e-mail: saikou@ar.media.kyoto-u.ac.jp

**あらまし**　本稿では落語を対象に劇場内においてリアルタイムで字幕を付与する方法を検討する．落語は演目ごとに基本的なシナリオ (台本) が決まっており，同一演者・演目の音声データとその書き起こしの収集が可能であることから，これを用いて当該演目専用の音響モデルと言語モデルを構築する．特に，台本からの逸脱への頑健性を保持しながら，言語的制約を強力に反映させるために，言語モデルの単位として文節を採用する．実際の落語 3 演目を用いて認識実験を行ったところ，3 演目平均で 90%に近い単語認識精度を得た．

# Real-time Captioning of Rakugo using Large Vocabulary Continuous Speech Recognition

**Masahiro Saikou　　Yuya Akita　　Tatsuya Kawahara**

School of Informatics, Kyoto University, Kyoto 606-8501, Japan
e-mail: saikou@ar.media.kyoto-u.ac.jp

**Abstract**　Automatic real-time captioning of Rakugo using large vocabulary continuous speech recognition is addressed. Rakugo is a Japanese traditional monologue show of story telling performed by a professional Rakugo-ka. Rakugo-ka follows a script, but does not read out it like drama. For automatic captioning, we construct a dedicated language model from the script and an adapted acoustic model. In addition, we adopt the phrase (bunsetsu) unit for language modeling. At this moment, we achieved word accuracy close to 90%.

# 1　はじめに

近年，テレビジョン放送などの公共サービスにおいては，障害者や高齢者に対しても情報伝達の機会を提供するためのユニバーサルデザインの取り組みが進められている．その一つとして音声字幕化があり，総務省は 2007 年までに可能なテレビジョン番組全てに字幕を付与することを義務づけている．このほか，講演などの様々な場面でも字幕付与の要望が高まりつつある．

従来，字幕は人手によって音声を書き起こし，要約することで作成されてきた．これらのコストは膨大であるため，字幕付与の拡充にはコストの削減が強く望まれる．また，ライブ放送ではリアルタイムの字幕付与が求められるが，日本語の入力には仮名漢字変換が必要であることから，人手によりリアルタイムの字幕付与を行うことが非常に困難である．

これらの問題に対して，音声認識技術を用いて字幕を自動生成・付与する研究が行われている．従来研究は，リアルタイムに字幕を生成するオンライン方式と，事後的に付与するオフライン方式に大別される．オンライン方式の研究としては，ニュース番組を対象としたリアルタイム字幕制作システム [1] があり，すでに実用化されている．ただし，高い音声認識精度を求められるため，対象としている音声は基本的にアナウンサーによる明瞭な発声に限られている．会話調の自発的な発話 (いわゆる話し言葉) に対応するために，リスピーカーと呼ばれる字幕作成用のアナウンサーを利用する方法 [2] もあるが，字幕付与のさらなる拡充を考えた場合，リスピーカーを常時用意することはそれほど容易でなく，コストも高いと考えられる．一方，オフライン方式の研究としては，事前に文単位の字幕を用意して音声と自動同期させる枠組が一般的である [3, 4, 5]．これは事前に用意した字幕と音声の同期に音声認識を利用することから，高い音声認識精度を求められず，雑音の重畳した話し言葉音声なども対象としている．一方で，事前に用意した字幕以外の発話には対処することができない．

本研究では，落語を対象に，劇場 (寄席) 内においてリアルタイムで字幕を付与する方法を検討する．落語では演目ごとに基本的なシナリオ (台本) が決まっているが，落語家は観客の反応などを考慮して，セリフを常にアレンジすることから，セリフは上演ごとに異なる．また，このアレンジ自体が落語の魅力の一つでもある．このため，固定のシナリオにアライメントを行う従来のオフライン方式を適用することはできない．これに対して，本研究ではシナリオ外の発話に対しても柔軟に字幕を付与できることを目標として，大語彙連続音声認識に基づくオンライン手法を採用する．ただし，十分な音声認識精度を確保するために，過去に演じられた同一演者・演目の音声とその書き起こしテキストを利用して，当該演目に専用の音響モデルと言語モデルを構築する．さらに，言語モデルの単位として形態素や単語でなく文節を用いることで，高精度化を図る．

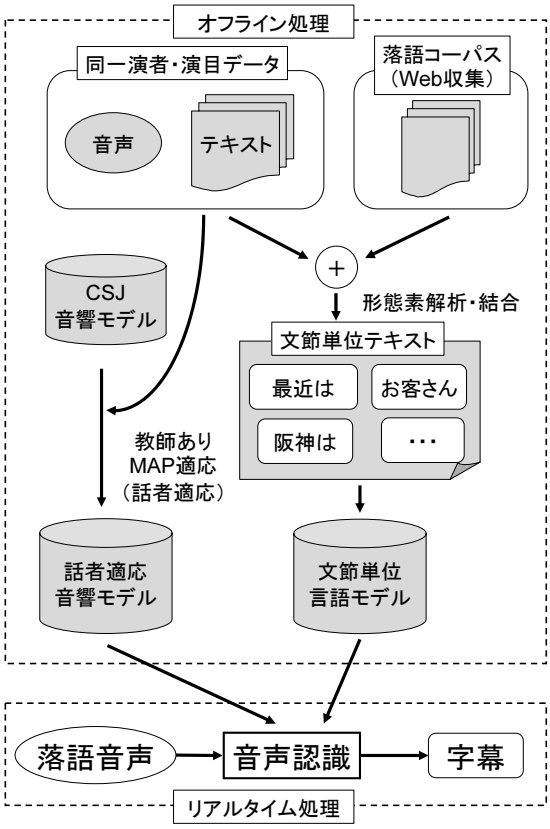


図 1: 提案手法の概要

# 2　提案手法の概要

本研究では，過去に上演された当該の演者・演目の音声と書き起こしが利用できることを前提として，それに適応した音響・言語モデルを事前に構築する．落語家は同一演目を何度も演じたり，事前にリハーサルを行うのが通常であるので，この仮定は妥当で

あると考えられる．提案手法の概要を図 1 に示す．

ベースラインの音響モデルは十分な落語音声を用いて学習することが望ましいが，そのような学習用コーパスは現実的に得られないので，本研究では『日本語話し言葉コーパス』(CSJ) に基づく音響モデル [6] を用いる．このベースライン音響モデルに対して，同一演者の音声データと書き起こしテキストを用いて話者適応を行った音響モデルを作成する．

言語モデルも同一演目の書き起こしに基づいて作成するが，柔軟な字幕付与を実現するために，落語の書き起こしを収集したコーパス (落語コーパス) も利用する．これらのテキストを形態素解析した後に，文節単位に結合し，文節単位の言語モデルを作成する．従来より言語モデルの単語単位には形態素や単語が利用されているが，本研究では文節単位の制約を利用することにより高精度化を図る．詳細は 4 章で述べる．

これらのモデルを用いて，音声認識をリアルタイムで行い，認識結果を字幕とする．

# 3　落語音声データ

## 3.1　落語音声の特徴

落語音声は会話調の独話であり，口語表現や発音変動などの話し言葉音声の特徴を持つ．さらに，これまでの話し言葉音声認識の対象とされてきた講演や討論などとは異なり，これらの現象に加えて，演技の上でイントネーションやリズムが意図的に変化・強調されて発声されることが多い．落語には複数の人物が登場するため，落語家は声質を巧みに変化させて人物を演じ分けるが，通常よりも誇張して個性や感情を表現する．例えば，酒に酔っている人物を演じる時は通常よりもなまけて発声する．

強調表現は単語の発音自体にも影響する．例えば「ちゃんと」「すごい」という単語が誇張されて発音された時,「ちゃ〜んと」「すご〜い」のように発音が変化する．このような発音の変化は従来の音声認識では想定されていない．

落語における演技は発話速度にも表れる．例えば慌てている人物を演じる時は通常よりも早口で発声し，落ち着いている人物を演じる時はゆっくりと発声する．また，人物の感情は逐一変化するため，同一のターン内でも発話速度は大きく変化する．

酔っ払いの客:
「ちょっと の 飲ん 飲ん ちょっと 飲みすぎ」
酔っ払いの客:
「ぐう (いびきの音)」
タクシー運転手:
「おい 寝たら困りまっせ，ちょっと」
タクシー運転手:
「ちょっと お客さん」

図 2: 落語データの例 (「深夜のタクシー」より抜粋)

## 3.2　本研究の実験試料

本研究では実験試料として,『桂三枝大全集〜創作落語 125 撰〜』より 3 演目 (「人情ラーメン一夢屋」,「念ずれば花ひらく」,「深夜のタクシー」) を用いた．落語はダジャレが頻出するものやラップ調のものなど多種多様であるが，これらはオーソドックスな会話中心に展開されるものである．実験試料の仕様を表 1 に示す．これらは 23〜30 分の音声であり，音声中には落語家の発声以外にも観客の笑い声，お囃子などが重畳している．また落語家自身が演技上，音声以外の音 (ラーメンをすする音，泣き声など) を発することもある．本研究では実験試料を落語家が演じる役柄に基づいて人手で分割を行った．データの一部を図 2 に示す．

# 4　文節単位言語モデル

## 4.1　文節の構成

大語彙連続音声認識において利用されている N-gram のような統計的言語モデルでは，その単位を大きくするほど表現できる制約は強くできるが，語彙エントリの数が増大するとともに，コーパス中の出現頻度が小さくなるので，信頼できる統計量の推定が困難になる．そのため，日本語の言語モデルの単位には形態素や単語が一般に用いられている．

ただし，落語では基本的なストーリーが決まっており，本研究では同一演目の書き起こしを利用することから，ほぼ定型的な単語連鎖の出現が十分に予測できると考えられる．そこで，本研究では言語モデルの単位として文節を採用する．文などの単位では，多少の逸脱に対応できなくなるが，文節の単位では変形は少なく，途中にフィラーやポーズなどが

表 1: 落語音声データの仕様

| 演目 | 「人情ラーメン一夢屋」(以下，nin) | 「念ずれば花開く」(以下，nen) | 「深夜のタクシー」(以下，shin) |
|---|---|---|---|
| 時間長 | 29 分 26 秒 | 22 分 56 秒 | 29 分 49 秒 |
| 登場人物総数 | 8 | 24 | 10 |
| 認識発話総数 | 481 | 281 | 465 |
| 発話速度 (モーラ/秒)[平均, 標準偏差] | 8.08, 2.79 | 8.39, 1.90 | 8.03, 2.73 |

テキスト:「ちょっと西宮まで行ってくれる」
形態素:
「ちょっと / 西宮 / まで / 行っ / て / くれる」
文節:
「ちょっと / 西宮まで / 行ってくれる」

図 3: 文節の構成例

挿入されることもほとんどないと考えられる．また，字幕の読み易さという観点においても，ある程度意味をなす文節のような単位の認識結果が得られる方が望ましい．

文節の構成にはサポートベクトルマシーンに基づいた日本語係り受け解析器である Cabocha [1] [7] を用いる．Cabocha は，京大コーパス [8] を学習データとして文節区切りと係り受け解析を実現するテキストチャンカーである．したがって，本研究で用いる文節は京大コーパスの定義に準拠したものとなっている．京大コーパスにおける文節の基本的構成は，「接頭辞＋自立語＋付属語・接尾辞」となっており，接頭辞，付属語・接尾辞は複数の場合とない場合がある．また，複合名詞などでは自立語が複数の場合もある．形態素を結合するこれらの規則は基本的に日本語構文解析システム KNP [2] に基づいている．文節の例を図 3 に示す．

### 4.2 言語モデルの比較

ここではテストセットの書き起こしを用いて，形態素単位と文節単位の言語モデルを作成した．言語モデル作成の際の形態素解析には ChaSen 2.3.3 (IPADIC 2.7.0) [3] を用いた．本実験に用いる落語には方言 (主に関西弁) が多く含まれているので，形態素解析及び文節への結合に少なからず誤りがみられたが，それらはすべて人手で修正している．各々の言語モデルの仕様を表 2，3 に示す．

表 2: 形態素単位言語モデルの仕様

| 演目 | nin | nen | shin | 平均 |
|---|---|---|---|---|
| 総形態素数 | 5929 | 5986 | 6974 | 6296 |
| 異なり形態素数 | 1162 | 1070 | 1259 | 1164 |
| パープレキシティ(形態素単位) | 4.08 | 3.88 | 4.15 | 4.04 |

表 3: 文節単位言語モデルの仕様

| 演目 | nin | nen | shin | 平均 |
|---|---|---|---|---|
| 総文節数 | 3309 | 3124 | 3937 | 3457 |
| 異なり文節数 | 1479 | 1345 | 1648 | 1491 |
| パープレキシティ(形態素単位) | 2.35 | 1.95 | 2.27 | 2.19 |

単位として文節を用いることによって語彙エントリ数 (異なり語数) が増加しているもののパープレキシティは大幅に (45.8%) 削減できている．これは，文節単位言語モデルの有用性を示している．なお文節に含まれる平均形態素数は 1.82 である．

## 5 評価実験

### 5.1 音響モデル・言語モデル・デコーダ

音響モデルは 2 章で述べたように，CSJ を用いて学習したモデルをベースラインとして，同一演者の音声データと書き起こしを用いて教師あり MAP 適応を行った．音響モデルの仕様を表 4 に示す．ただし，本研究では過去に演じられた同一演者・演目のデータが入手困難であったことから，テストデータを用いて話者適応した音響モデル (closed) と，テス

[1] http://chasen.org/~ taku/software/cabocha/
[2] http://www.kc/t.u-tokyo.ac.jp/nl-resource/knp.html
[3] http://chasen.naist.jp/hiki/ChaSen/

表 4: 音響モデルの仕様

| | |
|---|---|
| 学習データ | 日本語話し言葉コーパス (CSJ)<br>310 時間 (男性) |
| 特徴量 | MFCC(12) ΔMFCC(12)<br>ΔEnergy(1) 計 25 次元 |
| 音素数 | 42 |
| 状態数 | 3,000 |
| 共有混合分布数 | 16 |

表 5: 落語コーパスの仕様

| | |
|---|---|
| 総演目数 | 339 |
| 総文数 | 88K |
| 総形態素数 | 1.1M |
| 総文節数 | 760K |
| 異なり形態素数 | 33.8K |
| 異なり文節数 | 156K |

表 6: 母音の長音化に対処した単語の例

| 単語 (標記形) | 標準の読み | 登録した読み |
|---|---|---|
| ちゃんと | ch a N t o | ch a: N t o |
| それ | s o r e | s o: r e: |
| やっぱり | y a q p a r i | y a: q p a r i |
| びっしょり | b i q sh o r i | b i: q sh o r i |
| ごめん | g o m e N | g o: m e: N |

トデータを除いた残り 2 演目を用いて話者適応した音響モデル (open) の 2 種類を用意した．

言語モデルは，4 章で作成した各演目書き起こしモデルと落語コーパスより学習したモデルを用いる．落語コーパスの仕様を表 5 に示す．本実験では，書き起こしに落語コーパスを混合しない言語モデル (closed) と落語コーパスを混合した言語モデル (mix) の 2 種類を用意した．言語モデルの混合には，相補的バックオフを用いた手法 [9] を適用した．混合比は事後的に各演目 0.99，落語コーパス 0.01 とした．

次に落語音声への対処について述べる．3.1 節で述べたように，落語音声は話し言葉の特徴を有しており，発音変動などの対処が不可欠と考えられる．そこで，可変長文脈を用いた音素列の確率的変動モデル [10] を適用した．また演技上の強調のために起こる発音変動はこの手法では対処できないため，特に誇張して表現されていると思われた単語 (37 単語) は人手により変形発音を発音辞書に登録した．これらの大半は母音の長音化に伴うものである．登録した単語の例を表 6 に示す．

デコーダは Julius[4] rev.3.4.2 を利用した．

## 5.2 文節単位言語モデルの効果

まず，文節単位言語モデルの効果を確認するために，音響・言語モデルともに closed モデルを使用して認識実験を行った．実験結果を図 4 に示す．単語認識精度は形態素を単位として計数している．

図 4 より，言語モデルの単位として文節を用いることによって認識精度が大きく向上していることがわかる．文節単位の言語モデルを利用した場合，実験試料 3 演目平均で 86.3%の単語認識精度となり，聴衆の笑い声などの雑音が重畳していない音声のみ (全 1227 発話中 899 発話) では 88.2%の単語認識精度を得た．形態素単位の言語モデルを利用した場合には，それぞれ 72.3%，74.9%であるので，文節単位言語モデルを利用することにより，10%以上の改善を得ている．これは短い形態素の認識が困難であったことによると考えられ，言語モデルの単位として文節が有効であったことを示している．

音声認識誤りを解析すると，落語音声の特徴である個性や感情が誇張されている表現の前後で誤りが連続的に起こっていた．これらの誤りは本実験で用いた音響・発音モデルと大きく異なるためと考えられる．音声認識誤りの例を図 5 に示す．

## 5.3 音響モデルと言語モデルの条件の比較検討

次に，提案手法の実用化にむけた検討を行うために，2 種類の音響モデル (closed・open) と 2 種類の言語モデル (closed・mix) を組み合わせて実験を行った．言語モデルの単位は文節である．実験結果を図 6 に示す．

図 6 より，音響・言語モデルとして closed でないモデルを用いた場合に，単語認識精度がそれぞれ 5%程度低下することがわかる．テストデータ以外の音声データで話者適応した open 音響モデルと落語コーパスを混合した mix 言語モデルを用いた場合には，3 演目平均で単語認識精度が 7%程度低下した．

本実験では，open 音響モデルの適応データとして

[4] http://julius.sourceforge.jp/

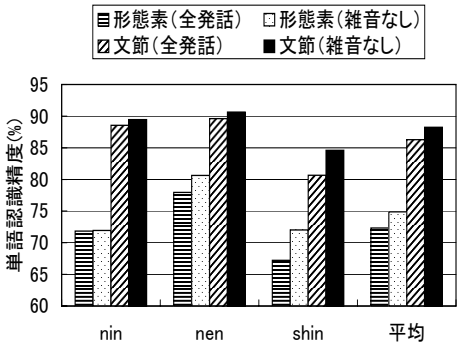


図 4: 形態素単位と文節単位の比較結果

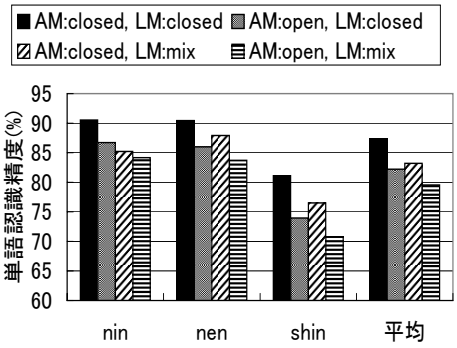


図 6: 音響モデルと言語モデルの比較・検討

(a) 叫び声:
「止めといてや」
(b) 表記通りに発音されない擬音擬態語:
「ぐう」(いびきの音)
(c) 酔っ払いの声:
「気分が悪い」

図 5: 音声認識誤りの実例

異なる演目のデータを用いたが，落語音声は演目ごとに登場人物が異なり，それに伴って落語家の声質も異なることから，適応データとして同一演目データを用いることにより，認識精度の低下をこの実験結果よりも抑えられることが期待できる．

## 6 おわりに

本稿では大語彙連続音声認識を用いた落語のリアルタイム字幕付与について検討した．提案手法は，同一演者・演目のデータを用いて非常にマッチした音響・言語モデルを構築するもので，言語モデルの単位として文節を利用することで制約の強化を図った．

実験の結果，言語モデルの単位として文節が有効であることを確認し，理想的な条件の下では 90%に近い認識精度が得られることを確認した．認識誤りとなった音声は，個性や感情が強調されたもの，表記通りに発音されない擬音擬態語を含むものであり，さらなる認識精度の向上にはこれらへの対応が必要である．

## 参考文献

[1] 安藤彰男, 今井亨, 小林彰夫, 本間真一, 後藤淳, 清山信正, 三島剛, 小早川健, 佐藤庄衛, 尾上和穂, 世木寛之, 今井篤, 松井淳, 中村章, 田中英輝, 都木徹, 宮坂栄一, 磯野春雄. 音声認識を利用した放送用ニュース字幕制作システム. 信学論, Vol. J84-DII, No. 6, pp. 877–887, 2001.

[2] 松井淳, 本間真一, 小早川健, 尾上和穂, 佐藤庄衛, 今井亨, 安藤彰男. 言い換えを利用したリスピーク方式によるスポーツ中継のリアルタイム字幕制作. 信学論, Vol. J87-DII, No. 2, pp. 427–435, 2004.

[3] 丸山一郎, 阿部芳春, 江原暉将, 白井克彦. ワードスポッティングと動的計画法を用いたテレビ番組に対する字幕提示タイミング検出法. 信学論, Vol. J85-DII, No. 2, pp. 184–192, 2002.

[4] 門馬隆雄, 沢村英治, 福島孝博, 丸山一郎, 江原暉将, 白井克彦. 聴覚障害者向け字幕付きテレビ番組の自動制作システム. 信学論, Vol. J84-DII, No. 6, pp. 888–897, 2001.

[5] 西沢容子, 杉山雅英. 字幕表示のための音声とテキスト間の自動対応付け手法とその評価. 信学技報, SP2003-176, 2004.

[6] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui. Benchmark test for speech recognition using the corpus of spontaneous Japanese. In *Proc. SSPR*, pp. 135–138, 2003.

[7] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情処学論, Vol. 43, No. 6, pp. 1834–1842, 2002.

[8] 黒橋禎夫, 長尾真. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会, pp. 115–118, 1997.

[9] 長友健太郎, 西村竜一, 小松久美子, 黒田由香, 李晃伸, 猿渡洋, 鹿野清宏. 相補的バックオフを用いた言語モデル融合ツールの構築. 情処学論, Vol. 43, No. 9, pp. 2884–2893, 2002.

[10] 秋田祐哉, 河原達也. 『日本語話し言葉コーパス』を用いた汎用的な発音モデルの統計的学習. 情処学研報, 04-SLP-53-3, 2004.