Vol.2009-SLP-77 No.20
2009/7/18

# コンテンツ活用のための報道番組自動書き起こしシステム

小林彰夫[†1] 奥貴裕[†1] 本間真一[†1]
佐藤庄衛[†1] 今井亨[†1] 都木徹[†1]

本報告では，放送コンテンツ活用のための報道番組自動書き起こしシステムについて述べる．本システムは，音声認識のための学習・評価データを収集したり，放送番組に付与するメタデータを抽出・制作するために，報道番組の音声を随時認識して，映像・音声とともに認識結果を蓄積する．音声認識は，音楽箇所・男女の発話区間の検出を行いながらリアルタイムでデコードを行い，デコードと並行して，話者識別により発話に話者タグを付与する．言語モデルは，ウェブ上のニューステキストを取得して随時更新される．また，音声認識により得られたラティスをコンフュージョンネットワークに圧縮し，キーワード検索用のインデックスを生成する．本システムを用いてニュース 53 番組の評価を行ったところ，単語誤り率は 9.2%となった．また，unigram クエリを用いたキーワード検索の結果，F 値は約 95 となった．

# A Broadcast News Transcription System for Content Application

Akio Kobayashi ,[†1] Takahiro Oku ,[†1]
Shinich Homma ,[†1] Shoei Sato ,[†1] Toru Imai [†1]
and Tohru Takagi[†1]

This paper describes a new transcription system for content application. The system archives broadcast news programs with their transcriptions and speaker tags with the aim of getting a collection of training and evaluation data for acoustic and language models. Besides it is also utilized for extracting and describing metadata for TV programs. The system has the functions of music and speech detection during dual-gender decoding, speaker diarization, and automatic language model updating for upcoming news shows. Trigram lattices are compressed into confusion networks that are indexed for known item retrieval. The system achieved a 9.2 % of word error rate and a 95 of F-measure in evaluation of known item retrieval for 53 Japanese broadcast news shows.

## 1. はじめに

これまで NHK では，ニュースを対象にした字幕制作を目的に，音声認識の研究を行ってきた[1)]．NHK の現在の音声認識研究の課題は，自由発話を含む放送番組の認識率の改善である．自由発話は，発声変形や口語特有の言い回しのために，発話が不明瞭になりやすく，ニュース読み上げの認識率と比べると，十分な認識率が得られていない．認識率を改善するためには，大量の放送番組を蓄積して認識率を評価したり，得られたデータを用いて統計的音響・言語モデルの学習を行うことが必要となる．この際，蓄積した放送番組に話者名などの情報が付与されていれば，音響モデルを話者適応することも可能になるため，認識率の改善に役立つと考えられる．一方，発話内容や話者名などの情報が放送番組に付与できれば，放送番組の検索やメタデータ抽出・作成[2)] などへの応用が期待できる．例えば，放送局の報道現場などでは，ニュース番組の検索に対するニーズが高く，話者名やキーワードなどをたよりに番組を検索できれば，効率的な番組制作に役立つと考えられる．

テレビやラジオなどの放送番組の検索を目的とした研究は，音声認識技術だけではなく，文境界検出やトピック検出などの要素技術を含めた，いわゆる音声ドキュメント処理の研究として行われてきた[3)–5)]．また，放送に限定しなければ，文献 6)，7) のようなウェブサービスも提案されており，音声から得られる情報を活用するための研究が活発に行われている．

本稿では，音声認識のための学習・評価データの収集やメタデータ制作など，放送コンテンツの活用を目的とした，報道番組自動書き起こしシステムを開発し，自動的に収集・書き起こされたニュース番組について，発話区間検出，音声認識，話者識別，発話内のキーワード検索 (Known Item Retrieval) による評価を行ったので報告する．

## 2. 報道番組自動書き起こしシステム

### 2.1 システム概要

報道番組自動書き起こしシステムは，放送中の番組をリアルタイムに音声認識して，認識結果を放送音声・映像とともに蓄積し，ブラウザを用いて閲覧・検索するシステムである (**図 1**)．現在は，NHK 総合テレビ/衛星第 1 の報道番組 (定時ニュース，クローズアップ現

†1 NHK 放送技術研究所
NHK Science and Technology Research Laboratories

Vol.2009-SLP-77 No.20
2009/7/18

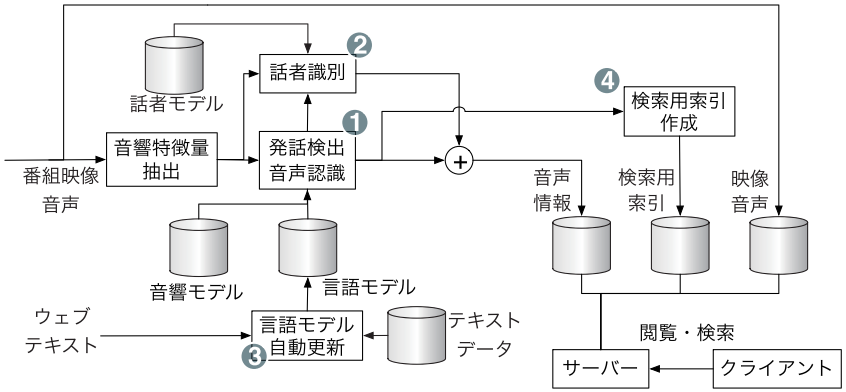


**図 1** 報道番組自動書き起こしシステム概要
Fig. 1 Broadcast News Transcription System

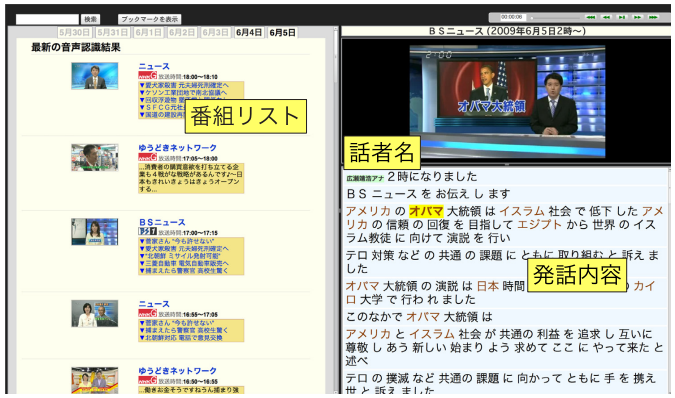


**図 2** クライアント画面
Fig. 2 Client Application

代，海外ネットワークなど) を対象にデータを収集している．図 1-①の音声認識は，番組音声から抽出された音響特徴量を入力として発話区間を検出し，音声認識結果を出力する．図 1-②の話者識別は，音声認識と同様の音響特徴量を入力として，音声認識と並行して話者を識別する．図 1-③の言語モデル自動更新は，ウェブ上のニュースから最新のニューステキストを取得し，言語モデルを逐次更新する．各ブロックから出力された発話区間，発話内容，話者名の各情報は音声情報として統合され，データベースに蓄積される．また，音声認識で得られたラティスはコンフュージョンネットワークに圧縮され，番組情報・発話時刻とともに索引化してデータベースに蓄積される (図 1-④).

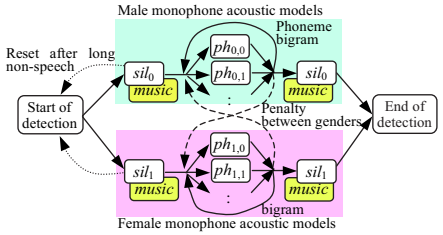


**図 3** 音素認識による発話区間検出と音楽検出
Fig. 3 Speech and Music Dtection

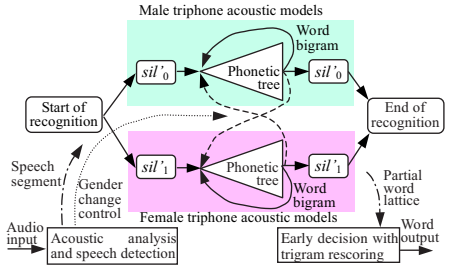


**図 4** 男女並列の連続音声認識
Fig. 4 Dual-Gender Speech Decoder

**図 2** に示すクライアントでは，ビデオ映像と同期して発話内容を閲覧したり，キーワードを入力して発話内容の検索を行う．

### 2.2 発話区間・音楽検出

背景音や男女の話者が混在した放送音声の自動書き起こしのための発話区間検出は，フレーム単位の細かな音声/非音声の判定よりも，多少の非音声区間を音声区間と誤ることはあっても，音声区間の欠落をできる限り抑え，音声を適度な長さの区間に切り出して，認識率の向上に寄与することが重要である．また，音声始終端検出までの遅れ時間はできる限り小さく，音声認識に不要なテーマ曲やジングル等の音楽検出も求められる．

本システムの発話区間検出は，音のパワーだけでなく周波数特性も考慮して，男女並列の性別依存音響モデルによる音素認識をエンドレスに実行し，その時の尤度から発話区間検出および音楽検出を行う (**図 3**)．音素認識は，男女間遷移が可能で枝刈り共通の男女並列音素認識を常時実行し，累積音素尤度の比を利用して発話の始端と終端を早期に検出する[8),9)]．

音楽の検出には，まず音楽専用 HMM(6 状態 4 出力・戻り遷移あり・32 混合モデル) を，各種報道番組で放送されるテーマ曲やジングル等 46 個の音楽データ (切り出し位置を 16 通りに拡張) から，最尤推定法で学習した．この音楽専用 HMM(music) を，無音・非音声モデル (sil) と並列に前記男女並列音素ネットワークへ加え (図 3)，累積尤度比に基づいて発話区間検出と同時に音楽区間も検出する．音楽と判定された区間の音声は，後段の男女並列連続音声認識には送られず，「♪〜」マークを音声認識結果として出力する．

### 2.3 男女並列連続音声認識[8),9)]

男女が混在した音声の認識は，性別非依存音響モデルの使用が簡易であるが，認識率が十分ではない．そこで，本システムにおける音声認識では，男女それぞれの性別依存音響モデルにリンクした単語発音辞書ネットワークを並列化し，発話区間検出の音素認識における最尤音素列の性別属性を利用して，1セグメント内で男女間の遷移を効率よく制御できるようにしている (**図 4**)．具体的には，男女のビームサーチの枝刈り閾値は共通とし，発話検出用音素認識の最尤音素列における性別属性が入れ替わった時刻に限って，男女間の単語の遷移を許し，一方の性別の探索が自然消滅していた場合でも，再度両性での探索を開始する．これにより，発話検出時の性別に限定することなく，計算量の増加を抑えて，より高精度な男女並列の連続音声認識が可能となる．

### 2.4 話 者 識 別

話者識別は，音声から「誰が、いつ」発話したかを検出する技術であり，音響モデルの話者適応化や，コンテンツ検索への利用が期待される．本システムの話者識別では，入力された音響特徴量から，話者の交替点を検出しつつ，あらかじめ登録しておいた話者モデル（番組のアナウンサーなど）を用い，話者の判定を行う．また，番組の放送終了後，できる限り早い話者識別結果の出力が望まれるため，話者交替点の候補としては，フレーム単位ではなく，発話区間検出時に得られる音素境界を用いることとする[10)]．話者交替点の検出および話者の判定は，ともに BIC(Bayesian Information Criterion; ベイズ情報量基準) に基づく ΔBIC を用いる[11)]．ΔBIC は２つの発話の特徴ベクトル列 $\boldsymbol{x}, \boldsymbol{y}$ に対して，それらが同一話者によるものかどうかを判定する基準であり，話者のモデルは特徴ベクトルの共分散行列 $\Sigma$ とフレーム数 $N$ で表現される．

$$\Delta BIC(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\left[N_{\boldsymbol{xy}} \log\left|\Sigma_{\boldsymbol{xy}}\right| - N_{\boldsymbol{x}} \log|\Sigma_{\boldsymbol{x}}| - N_{\boldsymbol{y}} \log\left|\Sigma_{\boldsymbol{y}}\right|\right] - \alpha P \qquad (1)$$

ここで，$\Sigma_{\boldsymbol{xy}}$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ が同一話者による発話と仮定した場合のモデルとし，$P, \alpha$ はそれぞれペナルティ項とその重み係数を示す．$\Delta BIC$ の値が負のとき，$\boldsymbol{x}$ と $\boldsymbol{y}$ は同一話者による発話と判定される．

話者交替点検出では，式 (1) の $\boldsymbol{x}$ は話者交替の候補点以前の発話，$\boldsymbol{y}$ は話者交替の候補点以降の発話にそれぞれ対応する．一方，話者の判定では，式 (1) の $\boldsymbol{x}$ は 1 つ前に判定した話者交替点から現在判定された話者交替点までの発話，$\boldsymbol{y}$ は登録された話者モデルの発話にそれぞれ対応する．$\Delta BIC$ により，入力された音響特徴量が登録された話者と判定されれば，その話者名を出力し，当該話者モデルの学習を行う．どの登録話者とも判定されなければ，新たに話者モデルの作成と登録を行い，その話者番号を出力する．

### 2.5 言語モデルの自動更新

ニュースを含む報道番組の特徴は，人名や地名，組織名などの新しい単語が毎日出現することである．したがって，高い認識率を得るためには，最新ニュースを使って言語モデルを適応化する必要がある．本システムでは，新しい話題に対応するために，ウェブ (NHK オンライン) を通じてニューステキストを随時取得し，言語モデルを自動的に更新する．更新の手順は，まず，ウェブから取得したニューステキストを形態素解析し，テキストに含まれる単語を発音辞書 (語彙) に登録する．次に新しく定めた語彙を用いて，あらかじめ蓄積しておいたニュース原稿から作成した trigram と，取得したニューステキスト中の trigram を重み付けして混合し，言語モデルを学習する[12)]．ただし，本稿では，形態素解析辞書のエントリを固定 (12.5M) したまま，自動的に言語モデルを更新して，音声認識を行うことにする．

### 2.6 キーワード検索のためのラティスのインデキシング

音声認識で生成されたラティスは，文献 13) の識別的言語モデルや，音素誤り最小化 (MPE; Minimum Phone Error) 学習[18)] に基づく識別的音響モデルを学習するためのデータとして蓄積される．

一方，ラティスをドキュメントとみなせば，ユーザーが与えたキーワードをクエリとして，蓄積された発話内容を検索することも可能である．ラティス上の単語仮説の事後確率を利用すれば，キーワードの検索を事後確率に応じて柔軟に行うことができる[14)]．ただし，ラティスは冗長な表現であるため，効率的なキーワード検索を行うためには，ラティスを圧縮する必要がある[15)]．そこで，本稿では，ラティスをコンフュージョンネットワーク[16),17)] に圧縮することにした．まず，音声認識から得た bigram ラティスを trigram ラティスに展開して，ラティスの各リンクの言語モデルスコアを trigram スコアに置換する．次に，文献 17) の pivot アルゴリズムに基づいてラティスをコンフュージョンネットワークに圧縮する．最後に，コンフュージョンネットワーク上の単語仮説をキーとして，事後確率と発話開始時刻を記録したインデックスを作成する．

## 3. 実験と考察

### 3.1 評価データ

2009 年 5 月 20 日から 23 日に放送された NHK 総合/衛星第 1 テレビの定時ニュース 53 番組を評価データとして実験を行った．発話者に応じて，評価データを「アナウンサー/記

**表 1** 評価データ
Table 1 Evaluation Data

| | 時間 (分) | 発話数 | 単語数 | パープレキシティ | 未知語率 (%) |
|---|---|---|---|---|---|
| 全体 (53 番組) | 532.2 | 8.4k | 105.7k | 28.8 | 0.51 |
| スタジオ | 465.1 | 6.8k | 92.9k | 22.6 | 0.35 |
| 中継 | 5.8 | 139 | 1.2k | 88.6 | 0.33 |
| 実況 | 0.7 | 46 | 182 | 932.6 | 2.74 |
| インタビュー | 56.6 | 1.4k | 11.4k | 172.2 | 1.88 |

者」と「その他 (一般の話者)」に分けるとともに，発話環境に応じて，以下のように分類した．

**スタジオ** アナウンサー・記者によるスタジオ発話

**中継** アナウンサー・記者による屋外での発話

**実況** アナウンサーによるスポーツ実況

**インタビュー** 話者「その他」による発話

評価データには，外国語箇所が 4.1 分 (発話箇所の 0.8%) 含まれているが，本稿では音声認識の評価の対象からはずした．

### 3.2 実験結果・考察

#### 3.2.1 音 声 認 識

音声認識では，第 1 パスで tree lexicon および bigram 言語モデルでデコードしたのち第 2 パスで trigram リスコアリングし，認識結果を逐次確定していく．

音響モデルの学習には，識別学習法のひとつである MPE 学習[18] を用い，認識率向上の妨げとなりやすい学習データを選択的に除外した[8]．各学習データ (男性 340 時間，女性 250 時間) の単語ラティスに対して，正解音素数の期待値をレファレンスの音素数で割って平均音素認識率を求め，これが閾値 0.5 未満のものを外れ値とみなして学習データから除外し，MPE の繰り返し学習を 10 回行った．音響特徴量は，12 次元 MFCC+対数パワーおよび 1 次・2 次の回帰係数の 39 次元ベクトルとした．

言語モデルの学習データはニュース原稿，放送書き起こし (ニュース以外の番組を含む) などの 660 万文 (202.3$M$ 単語) とした．適応化前の語彙サイズは 60$k$ とした (20 種類のフィラーを含む)．また，言語モデル更新のたびに，ウェブから取得したニューステキストに含まれる単語を語彙に追加した．

**表 2** に示す実験結果をみると，全話者を評価したときの単語誤り率は 9.2%となり，アナウンサー/記者に限ると単語誤り率 5.3%となった．発話環境別の結果 (**表 3**) をみると，アナウンサー/記者によるスタジオ発話は，単語誤り率 5.1%となった．字幕制作に音声認識を利用する際は，ウェブ上のニューステキストではなく，記者が作成したニュース原稿を用いるが，こちらも単語誤り率が 5%前後であること[19] から，ウェブ上のニューステキストを得て自動的に言語モデルを適応化しても，単語誤り率はほとんど変わらないといってよい．

アナウンサー/記者の単語誤り率を性別でみると，女性の方が，単語誤り率が大きくなっているが，これは，女性話者の発話にフィラーや言いよどみが多かったことが原因として考えられる．また，インタビュー箇所はさまざまな集音条件で収録された発話が多く，記者会見のようにマイクに向かって発話するだけでなく，カメラマイクが拾ったような不明瞭な発話まであるため，単語誤り率のばらつきが大きかった．

**表 2** 音声認識結果
Table 2 Overall Recognition Results

| 話者 | 発話数 | 単語数 | WER (%) |
|---|---|---|---|
| 全話者 | 8.4k | 95.8k | 9.2 |
| アナウンサー/記者 | 7.0k | 85.9k | 5.3 |
| その他 | 1.4k | 9.9k | 43.6 |

**表 3** 音声認識結果 (%)
Table 3 Recognition Results (WER, %)

| 話者 | 男性 | 女性 | スタジオ | 中継 | 実況 | インタビュー |
|---|---|---|---|---|---|---|
| アナウンサー/記者 | 4.2 | 6.5 | 5.1 | 6.1 | 91.1 | - |
| その他 | 40.2 | 74.3 | - | - | - | 43.6 |

**表 4** 話者識別結果 (%)
Table 4 Speaker Diarization Results (%)

| | DER | MS | FS | SE |
|---|---|---|---|---|
| 全体 | 13.4 | 0.1 | 0.5 | 12.8 |
| NHK 総合 | 5.2 | 0.1 | 0.5 | 4.7 |
| NHK 衛星第 1 | 15.3 | 0.1 | 0.5 | 14.7 |

#### 3.2.2 音楽・発話区間検出

音楽検出の実験結果は、誤棄却率 (FRR; False Rejection Rate) が 21.3%(115 箇所/540 箇所)，誤受理率 (FAR; False Acceptance Rate) が 26.0%(149 箇所/573 箇所) であった．記者会見や大相撲実況における雑音が，音楽と誤検出されることがあった．

**表 5** 登録済み話者の識別結果 (%)
Table 5 Speaker Diarization Results (Known Speakers, %)

| | FRR | FAR |
|---|---|---|
| 全体 | 32.2 | 7.7 |
| NHK 総合 | 19.0 | 12.3 |
| NHK 衛星第 1 | 35.1 | 7.7 |

**表 6** キーワード検索結果 (unigram, %)
Table 6 Known Item Retrieval Results (unigram, %)

| | unigram(高頻度) | | | unigram(低頻度) | | |
|---|---|---|---|---|---|---|
| 閾値 | 適合率 | 再現率 | F 値 | 適合率 | 再現率 | F 値 |
| 0.0 | 89.2 | 94.3 | 91.7 | 83.4 | 97.2 | 89.8 |
| 0.5 | 95.6 | 92.8 | 94.2 | 96.2 | 97.2 | 96.7 |
| 0.9 | 96.3 | 91.7 | 93.9 | 96.5 | 90.1 | 93.2 |

**表 7** キーワード検索結果 (bigram, %)
Table 7 Known Item Retrieval Results (bigram, %)

| | bigram(高頻度) | | | bigram(低頻度) | | |
|---|---|---|---|---|---|---|
| 閾値 | 適合率 | 再現率 | F 値 | 適合率 | 再現率 | F 値 |
| 0.0 | 94.3 | 90.6 | 92.4 | 87.2 | 79.0 | 82.9 |
| 0.5 | 94.6 | 90.5 | 92.5 | 92.0 | 78.4 | 84.7 |
| 0.9 | 94.6 | 89.2 | 91.8 | 94.3 | 70.0 | 80.4 |

音声認識結果における男女判定の正解精度を

$$\text{正解精度} = \frac{(\text{正解区間長} - \text{誤挿入区間長})}{\text{音声区間長}} \tag{2}$$

とすると，正解精度は 98.7%であった．

### 3.2.3 話 者 識 別

開発データとして 2009 年 4 月の NHK ニュースを使用し，式 (1) の重み係数 $\alpha$ を話者交替点において 0.75，話者クラスタリングにおいて 1.0 とした．登録しておく既知の話者モデルとして，2009 年 4 月の NHK のニュース番組の中から NHK 総合のアナウンサー 24 名，NHK 衛星第 1 のアナウンサー 11 名の話者モデルを作成した．

実験結果を**表 4** に示す．表 4 では，話者識別の評価指標として，NIST が提案する DER(Diarization Error Rate) を用いた[20]．DER の定義は

$$\text{DER} = \frac{\text{FS} + \text{MS} + \text{SE}}{\text{総発話時間}} \times 100 \tag{3}$$

である．ただし，FS(False Speech) は発話者なしの区間で発話と誤判定した時間，MS(Missed Speech) は発話者ありの区間で発話なしと判定した区間，SE(Speech Error) は話者を誤った区間の時間である．

また，本システムの話者識別では，アナウンサーなどの話者モデルをあらかじめ登録しておき，正しく識別できることが望まれるため，登録済み話者に対する誤棄却率と誤受理率についても評価を行った (**表 5**)．

表 4，表 5 をみると，NHK 総合よりも NHK 衛星第 1 の DER,FRR が悪化している．NHK 衛星第 1 のニュース番組は，NHK 総合のニュースと違って，音楽が番組冒頭から挿入されるため，これが話者識別の性能を悪化させ，誤判定を起こすものと考えられる．

また，NHK 総合において FAR が衛星第 1 よりも悪くなっているが，これは登録話者数が NHK 衛星第 1 よりも NHK 総合の方が多いことが原因のひとつとして考えられる．

### 3.2.4 キーワード検索結果

音声認識で切り出された発話単位をドキュメントとみなし，キーワード検索 (Known Item Retrieval) の実験を行った．この際，コンフュージョンネットワーク上の単語仮説の事後確率に応じて，あらかじめ定めた閾値を超える仮説を検索結果として出力することにした．評価は適合率 (precision)，再現率 (recall)，F 値 (F-measure) により行い，事後確率の閾値を変えて実験した．検索クエリは，評価データに含まれる固有名詞や固有表現の unigram,bigram とし，頻度に応じてそれぞれ 20 個ずつを定めた．例えば，unigram クエリは,「ナスダック」「橋下」,bigram クエリは「臓器/移植」「党首/討論」などである．

**表 6**, **表 7** に実験結果を示す．unigram をクエリとした場合，クエリの頻度に関わらず，事後確率の閾値が = 0.5 で F 値が 95 前後となった．一方，bigram をクエリとした場合は，unigram に比べ F 値が小さくなり，低頻度 bigram では閾値 = 0.5 で F 値が 84.7 となった．アナウンサー/記者の発話箇所で単語誤り率が 5.3%であることから，unigram のような単独のクエリで検索した場合は高い数値を示した．また，unigram クエリは，人名や固有名詞などの語であり，これらは，言語モデルの適応化によって認識されやすくなっていることも，F 値が高くなった原因として考えられる．

## 4. おわりに

放送コンテンツ活用のための報道番組自動書き起こしシステムについて報告した．ニュース番組による音声認識実験の結果，単語誤り率は 9.2%となった．また，キーワード検索実験の結果，unigram クエリの検索では F 値が約 95 となった．

本システムを学習・評価データの収集システムとしてみた場合，学習データとしての利用方法が課題としてあげられる．討論番組や情報番組は，ニュースよりも単語誤り率が大きいが，このような番組の単語誤り率を教師あり/なし学習でどのように削減できるのか，言語モデル・音響モデルの番組や話者への適応化を含めて，収集したデータをもとに検討していきたい．一方，検索システムとしてみた場合の課題は，放送番組のトピック分割手法である．一口に報道番組といっても，読み上げ中心のニュースから，討論番組や同時通訳の多い番組など，さまざまな形態があるため，番組に応じたトピック分割手法について検討していきたい．

## 参考文献

1) 本間真一，小林彰夫，奥　貴裕，佐藤庄衛，今井　亨，都木　徹：ダイレクト方式とリスピーク方式の音声認識を併用したリアルタイム字幕制作システム，映像情報メディア学会論文誌， Vol.63, No.3, pp.331–338 (2008).
2) 住吉, 柴田, 藤井, 後藤, 山田, 望月, 松井, 三須, 宮崎, 高橋, 河合, 三浦, 八木：CurioView : 情報検索を活用した新しい視聴スタイルの提案，映像情報メディア学会年次大会予稿集, 7-5 (2008).
3) Renals, S., Abberley, D., Kirby, D. and Robinson, T.: Indexing and Retrieval of Broadcast News, *Speech Communication*, Vol.32, pp.5–20 (2000).
4) Federico, M.: A System for the Retrieval of Italian Broadcast News, *Speech Communication*, Vol.32, pp.37–47 (2000).
5) Dowman, M., Tablan, V., Cunningham, H. and Popov, B.: Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News, *Proc. the 14th International World Wide Web Conference*, pp.225–234 (2005).
6) 後藤真孝，緒方　淳，江渡浩一郎：PodCastle の提案 : 音声認識研究 2.0 を目指して，情報処理学会研究報告 (2007-SLP-65)，Vol.2007, No.11, pp.35–40 (2007).
7) 緒方　淳，後藤真孝，江渡浩一郎：PodCastle の実現 : Web 2.0 に基づく音声認識性能の向上について，情報処理学会研究報告 (2007-SLP-65)，Vol.2007, No.11, pp.41–46 (2007).
8) 今井　亨，小林彰夫，佐藤庄衛，本間真一，奥　貴裕，都木　徹：放送用リアルタイム字幕制作のための音声認識技術の改善，第 2 回音声ドキュメント処理ワークショップ (2008).
9) Imai, T., Sato, S., Homma, S., Onoe, K. and Kobayashi, A.: Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News, *IEICE Trans. Information and Systems*, Vol.E90-D, No.8, pp.1286–1291 (2007).
10) Liu, D. and Kubala, F.: Fast Speaker Change Detection for Broadcast News Transcription and Indexing, *Proc. EUROSPEECH 99*, Vol.3, pp.1031–1034 (1999).
11) Chen, S. and Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the Bayesian information criterion, *Proc. DARPA Speech Recognition Workshop*, pp.127–132 (1998).
12) 小林彰夫，今井　亨，安藤彰男，中林克己：ニュース音声認識のための時期依存言語モデル，情報処理学会論文誌， Vol.40, No.4, pp.1421–1429 (1999).
13) 小林彰夫，奥　貴裕，本間真一，佐藤庄衛，今井　亨，都木　徹：単語誤り最小化に基づく識別的リスコアリングによる音声認識，電子情報通信学会研究報告，Vol.108-338, pp.225–260 (2008).
14) Chelba, C. and Acero, A.: Position specific posterior lattices for indexing speech, *Proc. the 43rd Annual Meeting on ACL*, pp.443–450 (2005).
15) Meng, S., Peng, Y., Seide, F. and Liu, J.: A Study of Lattice-Based Spoken Term Detection for Chinese Spontaneous Speech, *ASRU IEEE Workshop*, pp. 635–640 (2007).
16) Mangu, L., Brill, E. and Stolcke, A.: Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks, *Computer Speech and Language*, Vol.14, No.4, pp.373–400 (2000).
17) Hakkani-Tür, D., Bechet, F., Riccardi, G. and Tur, G.: Beyond ASR 1-best: Using Word Confusion Networks in Spoken Language Understanding, *Computer Speech and Language*, Vol.20, No.4, pp.495–514 (2006).
18) Povey, D. and Woodland, P.: Minimum phone error and I-smoothing for improved discriminative training, *Proc. ICASSP*, pp.I–105–108 (2002).
19) 今井　亨，尾上和穂，本間真一，佐藤庄衛，小林彰夫：番組音声とリスピーク音声の認識を併用した生字幕制作の検討，映像情報メディア学会年次大会講演予稿集，10–2 (2006).
20) http://www.nist.gov/speech/test/rt