音声認識を用いたラジオ放送の実時間情報表示システム
Real-time Radio Broadcasting Information Display System Using Speech Recognition

廣近 理希
Masaki Hirochika
法政大学情報科学部ディジタルメディア学科
E-mail: masaki.hirochika.df@stu.hosei.ac.jp

Abstract

This paper shows how to construct a system that extracts keywords by using speech recognition from radiobroadcast, display real-time topical constructive method information and products. Recently, Internet Protocol simulcast radio is used only internet delivery, but it has not been informed ongoing topic of the text type. Therefore, by using speech recognition, a content reception device that performs keyword extraction of radiobroadcast was developed. In speech recognition of radiobroadcast with BGM and noise environments, when the signal to noise ratio is lower than 10 dB, a recognition rate falls. Based on the knowledge, noise suppression was possessed. As a result, keyword ' s recognition rate improve from 37.8% to 53.3%. Based on proposed method, a prototype system that picks up words used as a search keyword and shows product and topic in radiobroadcast was constructed.

1 まえがき

IP サイマルラジオ放送によるインターネット経由での AM/FM ラジオ放送の聴取が定着しつつある。IP サイマルラジオ放送とは、地上波放送そのままの内容をほぼ同時刻に、株式会社 radiko[1] や NHK が運営するサイトから、音声配信を行うサービスである。インターネットを利用しているため、AM/FM 受信機を用いるより、手軽にラジオを聴取できる利点を持つ。しかし、SNS との連携といった音声以外の情報を発信できる利点を活かしたシステムは、いまだ少数である。

そこで、本研究では、IP サイマルラジオ放送の音源から、連続音声認識システムを用いて得られた発話の内容をもとに、話題や製品といったキーワードを抽出するプロトタイプシステムの構築を目指す。アーカイブ保存する際のデータ管理補助や、番組内容に関連するバナー広告、話者や商品に関係する情報表示への利用を目的とする。

システム構築にあたり、出力するキーワードを推定する必要がある。そこで、実際の番組で発話された全文データを作成し、キーワード候補の抽出と重みづけを行った。先行研究 [5] を参考に TF-IDF 手法により、候補群に含まれた単語をランク付けし、上位をキーワードとして推定した。また、発話情報を得るため、ラジオ音声に登場する人物や場面を音声認識に対応させることが課題となる。具体的な問題点として、ラジオ放送の演出上多く登場する BGM や背景雑音の影響が認識性能の悪化 [2] につながる点があげられる。従来手法 [2,3,4] を用いて、キーワード認識率を算出したところ、雑音除去後に音声成分の削除がある場合、認識率が下がる傾向が見られた。そこで、実際に放送されたラジオ音源で SN 比ごとにキーワード認識率を検証し、約 10dB に閾値があると評価した。SN 比 10dB 以下の場合に 11dB 以上となる処理を施し、キーワード認識率の向上をはかった。

2 システムの設計

2.1 システムの概要

本研究では、出演者が発した話題や商品を表す単語を抜き出し、ラジオ番組に登場した実時間情報としてキーワードの形で提示する。これにより、アーカイブ保存の際には、キーワードがその放送や番組を表し、ブラウザ上ではキーワードが商品や話題の情報を入手するために機能する。また、本システムでは、商品や話題の情報を表示するため、web 検索と商品検索の API を用いた。

実際にキーワードを提示する画面として、ラジオ放送を聞きながらシステムの開始と終了、キーワード更新をユーザが画面上で操作可能な設計にした。これにより、ユーザのタイミングで欲しいキーワード情報の更新を行うことが可能となる。また、画面上では変換誤りをユーザが訂正できるよう、テキストボックス形式で提示する。

2.2 システムの構成

実時間によるキーワード表示システムの構成図を以下に示す。

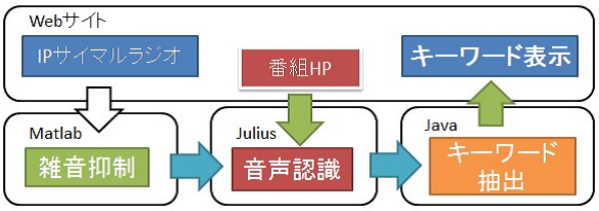


図 1:システムの構成図

IP サイマルラジオ放送の音声を録音し、音声認識を行う前処理として、BGM や雑音に対し雑音抑制処理を行う。処理後の音声データを、ソケット通信を用いて、サーバ側となる連続音声認識システムへ送り、認識結果を XML 解析から得る。認識結果を分析し、キーワード候補の抽出とランク付けから、キーワードを推定し画面へ表示する。音声ファイルは、IP サイマルラジオ放送の録音時に wav 形式 (サンプリング周波数 16kHz) へと変換する。

2.3 実装への課題

ラジオ音源の音声認識を行う具体的な問題点として、BGM や背景雑音の影響による認識性能の悪化 [2] や、音声認識を行うために発せられていない音声の認識の難しさ、CM や音楽を持つ各場面から認識を行うシーンの選択があげられる。また、認識結果から得られた情報の分析に関して、登場した単語がキーワードかどうかの判定、未知語処理 [7] 等があげられる。本システムでは、出力画面へキーワードを表示することが目的であるため、発話内容からキーワード候補となる単語を抜き出し、キーワードを推定することは必須である。また、放送の演出上ラジオ放送の中で多く登場する BGM や雑音の環境へ対策をとることで、認識率の向上が見込まれる。本研究では、これらに重点を置き研究を進めた。

Supervisor: Prof. Katunobu Itou

3 キーワード推定

3.1 先行研究

キーワード抽出の先行研究は、数多くされてきた。本研究でキーワードとする単語は、ラジオ放送の話題, 商品, 話者情報に関わる単語であり、検索エンジンを用いて商品や情報の検索を目的としている。従来研究として、ニュース記事から最新の話題語を抽出する研究 [8]、放送内容を示す字幕情報のキーワード検出 [10]、web 検索で用いる検索語推定の研究 [5] があげられる。各研究で広く用いられているキーワードの重みづけ手法として、TF-IDF 法があげられる。この手法は、ある語句の重みを、コンテンツの中で登場した回数、アーカイブ内である語句が使われているコンテンツ数の逆数により決定している。これにより、語句がそのコンテンツを示す内容であるかの指標を作成することができる。ラジオ放送でも話題となる単語は、そのラジオの番組や番組の回を表現しており、また話題となる単語は放送内で多く発話される。web 検索で用いる検索語推定を参考に、ラジオ放送に適応させることで、語句の重要度を示す指標として評価した。

3.2 番組のキーワード候補

具体的に 1 番組を例にキーワード抽出を行った。抽出した番組は、2011/8/26 の 25:00 から 27:00 に放送された TBS ラジオ、金曜 JUNK バナナマンのバナナムーン GOLD。番組で発話された内容から、話者, 話題, 商品を表す単語を手作業で抜き出し、15 単語のキーワードをあげた。また、登場した人物の情報を話者情報、話題を表現している単語を話題情報、話に登場する製品, 映画, 本, などを商品情報として、分類しどの情報が多く上げられるかを検証した。

話者情報: バナナマン, 設楽, 日村, ももいろクローバー Z
話題: ももクロ, 夏休み, ベガス, シルクドソレイユ,
　　ブランド, 松井, 本音, ライブ, ダンス, アイドル
商品: 「バトルアンドロマンス」

結果として抜き出すキーワードが登場するシーンの傾向を観察する。まず、ラジオ放送の傾向として、出演者自身が自己紹介を行う話者情報区間、1 区間の話題について出演者が語る話題区間の後、ゲストやボイスコメントが登場するコンテンツ区間に突入し、コンテンツ区間終了後、出演者がまとめを行い、CM, ブレークへと入るといった流れがある。その際、メールやホームページ紹介のタイミングは、話者情報後もしくは、まとめ後に挿入される。一般的に、コンテンツや話題が変化しながら、この流れを繰り返し番組が構成される。その中で、キーワードが多く登場する場面は、話題区間とコンテンツ区間であり、検証した番組のキーワードでも、番組冒頭の話者情報を除いてすべて話題区間とコンテンツ区間に属していた。

3.3 キーワード候補のランク付け

キーワード候補としてあげた各名詞群に対し、キーワードとなる単語を推定するため、式 (1) に示す TF-IDF 法を用いて重み付けを行った。

$$w_{i,j} = tf_{i,j} \log(\frac{N}{df_i}) \quad (1)$$

i : 対象とするキーワード
j : キーワードを含む現在のテキスト
tf : キーワード候補の認識結果中の出現回数
N : 全てのドキュメント数
df : キーワードを含むドキュメント数

先行研究 [6] を参考に、N を全ての単語の場合において、Yahoo!の検索結果数最大値である 252.7 億とし、キーワード候補となる各名詞を用いて、df を Yahoo!の Web 検索 API での返答検索結果の件数、tf をキーワード検出する区間に出現した回数とした。

この TF-IDF 手法により、認識結果後に行うキーワード抽出のため、ラジオ放送の全文データを作成し、形態素解析で名詞と判断された 345 単語に対し、TF-IDF 法でキーワード候補の重みづけを行った。TF-IDF の降順順位とその順位までに含まれたキーワードの数を表 1 に示す。

表 1:TF-IDF 順位のキーワード含有数

TF-IDF 順 (1〜X 位)	5	6	10	33	41
キーワード数 (単語)	5	6	6	12	15
–キーワード数 (話者)	3	3	3	4	4
–キーワード数 (話題)	2	3	3	8	10
–キーワード数 (商品)	0	0	0	0	1
precision(%)	100	100	60.0	36.6	36.6
recall(%)	33.3	40.0	60.0	80.0	100

345 単語のうち TF-IDF 法による重みづけ降順順位 50 位 (上位約 12%) の単語までを観察すると、80%(15 単語中 12 単語) のキーワードを網羅している。また、50 位までの単語のうち、キーワードになりにくい「俺」「ほんと」や「これ」「あれ」といった単語が 15 単語含まれているため、実質 33 位までに 80% のキーワードを含んでいる。そして、キーワードになりにくい単語を除くと上位 6 位はすべてキーワードであった。名詞のみを抜き出すときに、代名詞を直接指定して除外すると、精度が上がることが観測できる。よって、精度上げた TF-IDF 法によるランク付けを行い、上位約 10% の単語に絞ることにより、キーワード候補の 8 割が抜き出せる。認識結果に話題区間の全文データを用いた場合、手法はシステムに対しキーワード候補を絞る点において有効に働くと評価できる。

また、345 単語の候補群に対して 41 位までに recall 値による再現率 100% が達成された。本システムでは、最終的にキーワードをユーザが選択するシステムとなっている。実装するにあたり、ユーザがほしいキーワードを含むことを重要視し、キーワードの表示を行う。

4 ラジオ放送の雑音除去

4.1 雑音抑制の従来手法

雑音除去を行う方法は、数多く存在する。本研究では、雑音抑制により、騒音だけでなく BGM や残響といった多岐に渡る雑音も抑制したい。また、実時間での情報表示を実現するため、できるだけ短時間で組み込みやすいシンプルな処理が行える雑音抑制を用いる必要がある。そこで、ラジオ音源に応用しやすい比較的単純な雑音抑制手法である、スペクトル減算法 [2]、ウィーナーフィルタリング法 [3]、ランニングスペクトル解析法 [4]、の 3 手法を候補とし、実際のラジオ放送で処理後の結果を比較した。

4.1.1 スペクトル減算法

スペクトル減算法 (以下 SS 法) は、雑音が付加された音声信号のパワースペクトルから、別途推定した雑音のパワースペクトルを引き算し、そのパワースペクトルを逆フーリエ変換することで、雑音除去済みの音声信号を得る方法である。比較的シンプルなアルゴリズムで、効果的に雑音除去が行える手法 [1] である。雑音が定常であることと、発話中も同じ雑音の定常な性質を持つことが、前提となっている。

4.1.2 ウィーナーフィルタリング法

画像にも用いられるウィーナーフィルタリング法 (以下 WF 法) は、音声と雑音の違いとして、相関の有無を利用した方法である。本来の音声信号と推定した音声信号の平均二乗誤差を最小にする線形フィルタを作成し、音声信号を得る。時間領域では、時系列データ、周波数領域ではスペクトルの平均二乗差のどちらの領域でもフィルタを作成可能である。今回は、周波数領域で作成したフィルタを用いた。

4.1.3 ランニングスペクトル解析法

ランニングスペクトル解析 (以下 RSA 法) は、雑音と音声の性質の違いとして、音声は毎時に特徴が異なり、雑音は同じ音が発せられている部分に注目している。これにより、音声スペクトルは時間と共に変化するのに対し、雑音のスペクトルは時間が変化しても大きく変化しない状態が続く。これに着目し、スペクトルのある周波数に注目して、その時間変化を 1 つの信号とする。この信号から変調スペクトルを求めると雑音箇所は、変化が少ないので、超低周波に集中し、音声信号は雑音に比べて少し高い周波数に分布する。その変調スペクトルにあらわれる低周波を取り除き、雑音を除去する手法。

本研究で用いた RSA 処理は、変調スペクトル上の 2Hz〜7Hz が音声 [4] という知見に基づき、それ以外の成分を雑音として削除した。また、変調スペクトルに現れた雑音箇所の削除量を、

研究室内の実験で最も認識率が高かった上端下端両方 75% として処理した。

4.1.4 ラジオ音声における雑音処理比較

雑音抑制を比較するためサンプルとして、ラジオの野球中継音声を用いた (短文「四回の表が終わった段階で」)。この野球中継音声には、場内の声援やウグイス嬢の声が雑音として入り混じっており、実際に放送中に起こりうる雑音に対し、各雑音除去が有効であるかを考察した。録音データは、IP サイマルラジオ放送 wav ファイル、16kHz のモノラル。

スペクトログラム表示による比較図 (図 4-1 から図 4-4) を以下に示す。

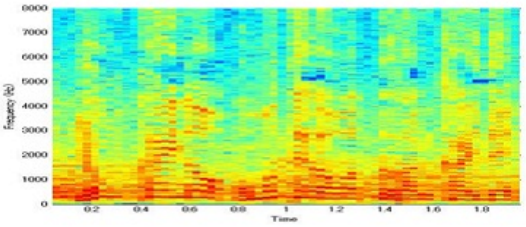


図 4-1 元のスペクトログラム

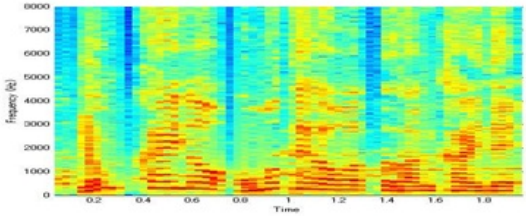


図 4-2 SS 後のスペクトログラム

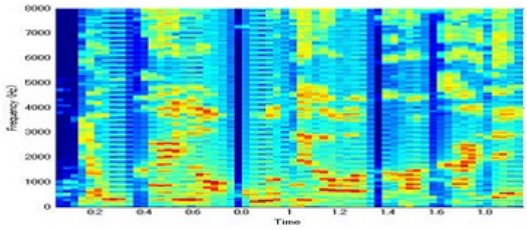


図 4-3 WF 後のスペクトログラム

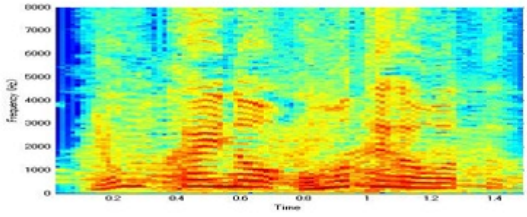


図 4-4 RSA でのスペクトログラム
(変調スペクトル 2-7kHz,25%)

まず、計算時間の違いとして、WF 法のフィルタは周波数領域で作られ、SS 法の減算も周波数領域で行われるため、高速フーリエ変換 (以下 FFT) はそれぞれ 1 度行われる。しかし、RSA 法は、雑音の変化率も計算する為 FFT を 2 度行っている。そのため、RSA の処理時間が 2 つと比べ長くかかる点が相違点としてあげられる。

各スペクトログラムを比較すると、WF 法が元の雑音混じりの音声から、削除されている要素の割合が高い。次に高いのが RSA 法で、最後が SS 法である。この削除率が高いと、実際に必要な音素まで削除してしまう可能性も高くなる。また、実際に音声を聞いた印象でも、WF 法は雑音と共に子音の音素も多く除去されてしまったうえ、人工的な声となっている。RSA 法と SS 法でも、WF 法に比べ音素は残っているが、消えてしまっている箇所もあり、特に RSA 法では音声データの最後の部分の声も雑音として削除されている。

次に、各処理後の音声に対し、連続音声認識システム julius による音声認識の出力結果での考察を行った。野球中継での結果例と、AM/FM ラジオ音源データからランダムに選んだ短文 100 サンプルの単語正解精度で求めた認識率の平均、キーワードを含む文 60 サンプルのキーワード認識率を以下に示す。

認識単語「四回の表が終わった段階で」

表 3:各雑音抑制手法の比較

	結果	認識率	キー認識率
原音	今回 の もの は 段階。	20.3%	26.7%
WF	お に お 会。	17.5%	6.7%
SS	四回 の 表 の 段階。	25.5%	38.3%
RSA	ある だけ の もの。	21.3%	31.7%

野球中継の認識結果例を観察すると、最も正解に近い出力結果は SS 処理後の結果であった。早口に感じる「終わった」箇所以外、正しく出力されている。WF 法,RSA 法は、アライメント表示により認識区間を確認すると、音素が消えた部分で発音が無いと判断されていた。また、キーワード認識率,100 サンプルの単語正解精度の平均をみても、SS 法の認識率が最も高かった。

そして、雑音と判断され削除された音素をもつ単語は、その音素が削除された形で認識結果に反映された。雑音を誤認識する場合より、音素が消えた部分の認識結果が認識率を左右すると考える。認識率平均と音素削除の割合を比較した場合、音素が消えている処理後ほど認識率平均が低くなっている。この結果から、音声認識を行う際、雑音を除去しながらも必要な音声成分を残すことが、キーワード認識率の向上につながると考察できる。

本研究では、スペクトログラムによる比較で音声成分の削除量が一番少なく、処理時間も比較的早く、またキーワード認識率が最も高かった SS 法を評価した。そして、課題である音声成分の削除量を減らすことにより、認識率の向上を試みる。

4.2 SN 比とキーワード認識の関係

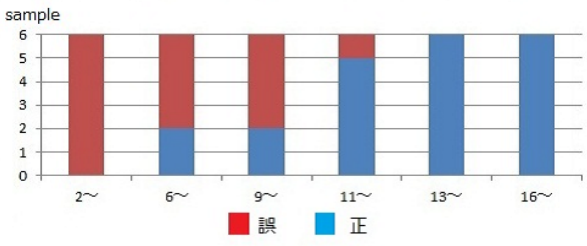
前項より、雑音の誤認識より音素が消える箇所がキーワード認識率に影響していることが観察された。そこで、SNR(小数点切り捨て) ごとにキーワード認識の正誤を、ラジオ音源の発話者の影響ができるだけ少ないと推測される短文 36 サンプルによって行い、SNR とキーワード認識の関係を示した。SN 比の計算方法は、先行研究 [7] を参考に無発話区間を雑音と仮定して計算する手法で行った。

1. 音声 $s(t)$ 雑音 $n(t)$, とし各平均パワーを $\hat{P}_s$,P_n と表す.
2. 発話音声中の音声信号を $x(t) = s(t) + n(t)$ とおく.
3. 音声信号の平均パワー $\hat{P}_s$ を推定し、SN 比を算出するため、$\hat{s(t)} = x(t) - n(t)$ より、$\hat{P}_s = P_{s+n} - P_n$ を求める.
4. 各値を用い、$SNR = 10log_{10}(\hat{P}_s/P_n)$ 式より算出する.

この手法の先行研究 [7] では、定常雑音を前提に算出している。しかし、ラジオは音声のみの情報であり、BGM が話者の声より大きくなることはほぼ無い ($s(t) > n(t)$) と考えられ、会話や発話中は、コンプレッサ等の放送通信機器により BGM の音量のばらつきが出来るだけ少なくなるよう放送されているため、この算出は有効であると評価した。

実際の計算と 36 サンプルによるキーワード認識の変化結果は表 4 のようになった。

表 4:SNR とキーワード認識率 (dB)



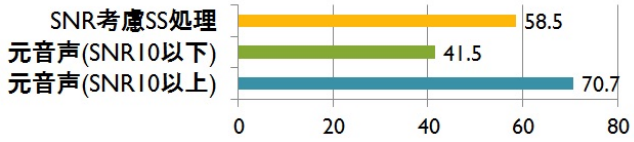
結果より、SN 比が約 10dB となる部分の前後に、キーワー

ド認識率の差がみられた。特異なデータを細かく観察すると、SNR=11.5dB で誤認識を起こしたサンプルの特徴として、キーワードの前後に言い淀みを含んでいる点、SNR=6.1dB のデータでは、アナウンサーによる話者紹介の発話で、SNR は低いが明瞭に聞こえる点があげられる。これらのデータに関して、SNR で示す BGM や雑音の影響より、出演者の発話に影響している。よって表 4 より、SN 比が 10.0dB 以上で、雑音の影響は受けにくいと推測でき、消去されすぎる音声成分と雑音除去のバランスを表す閾値として評価した。

4.3 SNR に基づく雑音抑制の評価

前項の閾値決定より、ラジオ音声では、SN 比が推定値 10.0dB 以上で BGM, 雑音の影響を受けにくいという結果が得られた。それに基づき、SNR が 10dB 以下の場合にのみ、SNR が 11dB 前後まで改善するような SS 処理を行い、キーワード認識率によって性能を評価した。ランダムに選んだラジオ音声のうち、SNR=10.0 以上 45 サンプル、SNR=10.0 以下 45 サンプルのキーワードを含む短文を用意し、それぞれのキーワード認識率の結果と、SNR が 10 以下の場合にのみ提案する処理を行った結果を表 5 に示す。

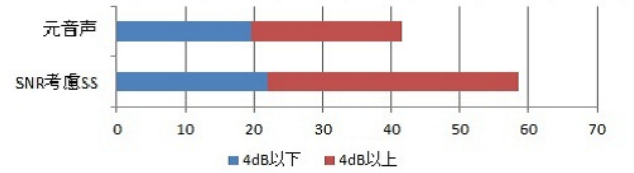
表 5:SNR 考慮 SS 処理のキーワード認識率 (%)



認識結果から、提案手法は雑音抑制を行わなかった音声に比べ、17.0% キーワード認識結果の向上が確認できた。また、SNR10dB 以上のデータは、雑音の影響が少ないと考え、誤認識が雑音以外の要因であると仮定した場合、SNR10dB 以上のデータと比較して、提案手法は雑音の影響を約 82% まで軽減していると分かる。

しかしながら、依然認識率が低いことが課題としてあげらる。特に、10dB 以下のラジオ音声について BGM が SNR4dB 以上と以下の場合で、認識率向上の差が見られた。表 6 に示す。

表 6:4dB 閾値別 SNR 考
S 処理のキー認識率 (%)



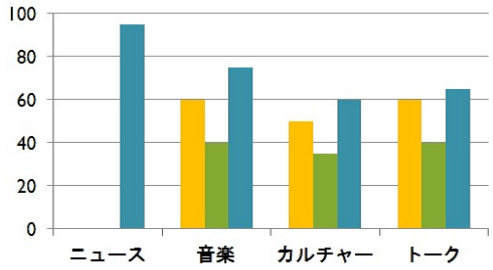
10dB 以下の音声のうち、SNR4dB 以上の BGM, 雑音においてキーワード認識率が向上したが、SNR4dB 以下の BGM, 雑音にはあまり効果が見られなかった。原因の一つとして、SNR4dB 以下のラジオ音声において BGM 中に声が入っている場合、BGM 中の音声も要素とともに戻され、誤認識を引き起こす場合があった。4dB 以下の場合に、新たな処理を加えることで、キーワード認識率の改善を考えなくてはならない。

4.4 番組別キーワード認識率

表 5 のデータのうち、ラジオ放送で最も多い 3 ジャンルの番組とニュース番組を対象に、番組別のキーワード認識率の観察を行った。SNR10dB 以上と SNR10dB 以下の短文を用いて、SNR10dB 以下の場合に SNR 考慮 SS 処理とそのままの音声を比較した。対象とした 3 ジャンルは、音楽, カルチャー, トーク、SNR ごとに 20 サンプルずつ各番組で行った。また、ニュース番組に関して、SNR10dB 以上でキーワードを含む音声は極稀であるため、対象外とした。結果を表 7 に示す。

どのジャンルでも、キーワード認識率が向上している結果となった。特に雑音区間終了後すぐの発話し始めの子音部分において、音声成分の消去対策となっており、子音の要素欠落による認識誤りを起こすデータに有効であった。特にトーク番組の認識率が向上した。これは、キーワードの短文を含む SNR の平均値が約 7.8dB(5 番組平均) であり、提案する処理が効果的に働いたからといえる。また、カルチャー番組の話題区間に登場したキーワードは、4dB 以下の状況が多く存在していたため、認識率の向上が少ない傾向が見られた。
それぞれ、番組単位でシーンを検出し、処理を行うことで認識率の向上につながると考えられる。

表 7:番組別キーワード認識率 (%)



5 あとがき

本研究では、SNR に基づく雑音除去手法の作成を行い、SNR10dB 以下の BGM, 雑音環境下にあるラジオ音声をそのまま音声認識にかける場合より、約 17.0% キーワードを多く入手することが可能な、実時間情報表示のプロトタイプシステム構築に成功した。

また、処理を行ったデータと SNR10dB 以上と比較した場合、82.5% 雑音の影響を軽減し、特に雑音が 4dB 以上 10dB 以下の場合が多いトーク番組において効果が見られた。

しかし、向上はみられたがキーワード認識率が依然低い問題点がある。雑音抑制の面では、SNR 計算の雑音量推定、SS 処理の雑音推定の誤差があげられる。SNR の計算は、雑音区間に推定したパワーの平均が、発話区間でも変わらない仮定のもと計算を行っているため、誤差が生じている。また、実験により評価した SS 処理は元来、定常雑音に対しての手法であるため、定常的でない BGM 環境も多くあるラジオ音声に適応させた場合、雑音の推定に誤差が生まれる。そして、本研究では、多くの番組に汎用的に利用できるよう、出演者や話者の特徴を限定しなかった。特定の番組や話者に絞って学習し、認識実験を行った場合、現在よりも認識率の向上が期待できると推測される。

参考文献

[1] IP サイマルラジオ協議会,"IP サイマルラジオの本格実用化に向け株式会社 radiko を設立", 2010 Nov

[2] STEVEN.Boll" Spectral Subtraction of acoustic noise in speech usin spectral subtraction "IEEE Trans.Acoustics Speech, and Sigbal Processing,VOL.ASSP-27,No.2, pp.113-120,1979

[3] T.Yamada,M.Kumakura,N.Kitawaki, " Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," subtraction " IEEE Trans. Audio Speech, and Language Processing,VOL.14,No.6, pp.2006-2013,2006

[4] 早坂 他, " ランニングスペクトルフィルタを用いた雑音にロバストな音声認識 " ,2003,Jun

[5] 阿部 他,"Web の閲覧履歴を情報源としたソーシャルブックマークにおけるタグ推薦の提案,"DEIM Forum,A2-5,2011,Feb

[6] 西崎 他,"未知語を考慮したニュース音声記事の検索", 音声言語情報処理,Vol.123,pp.171-176,2001,Dec

[7] 古谷 他,"定常雑音下における音声区間検出と SN 比の推定", 近畿大学九州工学部研究報告,Vol32,2004,Mar

[8] 佐藤 他,"時系列ニュース記事における最新話題語抽出方法", 電子情報通信学会技術研究報告,Vol105,2005,Jul

[9] 本間真一,"生字幕制作のための音声認識,"NHK 技研 RandD,No.122,2010,Jul

[10] 岡本昌之,"映像ブックマーク検索における字幕からの検索キーワード推定", 情報処理学会研究報告,Vol43,p35-42,2007,May