ベイズリスク最小化音声認識を用いた音声検索システムにおけるクエリ生成方法の検討

古谷 遼[1,a)]　南條 浩輝[2,b)]

概要：音声入力型情報検索（音声検索）の研究を行う．音声検索システムは，1）音声認識，2）検索要求（クエリ）生成，3）情報検索の 3 つのモジュールを持つシステムである．このうち，1）の音声認識に関して，我々はこれまでに重要単語の誤りを最小化するための音声認識手法を提案している．しかし，重要単語誤り最小化音声認識と 2）の検索要求生成を組み合わせた研究はこれまでになかった．このような背景に基づき，本研究ではこれらの組み合わせ，すなわち 1) 重要単語の誤りを最小化する音声認識を行い，2) その結果得られる N-Best リストから検索要求を生成することを行った．提案手法が音声検索の精度向上に効果があることがわかった．

1. はじめに

音声認識をフロントエンドに持つ情報検索，すなわち音声入力型の情報検索（音声検索）システムについて研究を行う [1], [2], [3]．音声検索システムにおいては，バックエンドの検索システムが高精度であっても，音声認識で認識誤りが発生した場合，その影響を受けて検索精度が低下する．この問題に対し様々な研究が行われている．これまでに提案されている手法には，音声認識と情報検索を一体としてシステム化する方法 [4], [5]，対話形式を取ることで段階的な情報の絞り込みを行う方法 [6]，検索システムのキーワードの音声認識精度を高める方法 [7] などがある．これまでの音声検索システムの音声認識は，単語誤り率（WER: Word Error Rate）やキーワード誤り率（KER: Keyword Error Rate）で評価が行われており，WER や KER の削減を目的としていた．しかし，キーワードであっても，それぞれの重要度，すなわち音声認識誤りした際の検索性能への影響は異なる．このような点を考慮して音声検索のための音声認識が行われている研究は少ない．

この問題に対して，重み付き単語誤り率（WWER: Weighted Word Error Rate）とこれを削減する音声認識が提案されている [8]．WWER は各語句の重要度を考慮して音声認識の評価を行う尺度であり，情報検索での重要語句がどの程度正しく認識できているかを評価できる尺度である．WWER の削減は，ベイズリスク最小化（MBR: Minimum Bayes-Risk）の枠組みに基づいて音声認識を行うことで実現できる [8]．しかし，MBR 音声認識結果からの検索要求（クエリ）生成に関する研究は不十分である．実際に，従来の MBR 音声認識に基づく音声検索では，単に認識結果の最上位仮説をそのまま用いている [8], [9]．

これらの背景に基づき，本論文では MBR 音声認識を用いた情報検索について研究を行う．具体的には，1) 重要単語の誤りを最小化する MBR 音声認識（WWER 最小化）を行い，2) リスクの小さい順に並び替えられた N-Best リストから検索要求を生成する．このように，重要単語の誤り最小化音声認識と検索要求生成手法の両方を組み合わせてその効果を調査した研究はこれまでになく，新しい．

本論文の構成について述べる．2 章で WWER および WWER 最小化音声認識について述べる．3 章で WWER 最小化音声認識を用いたベースライン検索システムについて述べる．4 章で N-Best リストを用いた検索要求生成手法について述べ，本手法が情報検索の精度向上に有効であることを示す．5 章で結論を述べる．

2. 情報検索のための音声認識の枠組み

2.1 重み付き単語誤り率

情報検索においては，音声認識誤りに由来する検索性能

[1] 龍谷大学理工学研究科
Graduate School of Science and Technology, Ryukoku University
[2] 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku, University
[a)] furutani@nlp.i.ryukoku.ac.jp
[b)] nanjo@nlp.i.ryukoku.ac.jp

低下の影響の大きさは誤った単語によって異なる．すなわち，情報検索の視点から重要な単語とそうでない単語が存在する．このことから，音声認識結果の評価は誤り単語が情報検索に与える影響に基づいて行うべきである．このような評価を行うための評価尺度として我々は重み付き単語誤り率（WWER: Weighted Word Error Rate）を提案している [8]（式（1））．

$$WWER = \frac{V_I + V_D + V_S}{V_N} \tag{1}$$

V_I は挿入誤り単語の重要度の合計を，V_D は削除誤り単語の重要度の合計を，V_S は置換誤り区間の単語重要度の合計を，V_N は正解文の単語重要度の合計を表す．なお，誤り単語を同定する際には，単語誤り率（WER）を求める際と同様に DP マッチングの結果を用いるため，全ての単語の重みを等しく設定したとき，WWER は WER と一致する．

2.2 ベイズリスク最小化音声認識

音声認識誤りによる検索要求の情報損失を削減するためには，音声認識誤りを削減することが重要である．このための手法として式 (2) で定式化される，ベイズリスク最小化（MBR: Minimum Bayes-Risk）音声認識がある [10], [11].

$$\hat{W} = \arg\min_{W} \sum_{W'} l\left(W, W'\right)^{\lambda_1} P\left(W', X\right)^{\lambda_2} \tag{2}$$

ここで，$l\left(W, W'\right)$ は仮説 W' を仮説 W に誤った際の損失を求める損失関数を表し，$P\left(W', X\right)$ は入力信号 X と仮説 W' の同時確率（音声認識スコア）を表す．λ_1，λ_2 は損失関数および確率の重みパラメータである．この枠組みにおいて，損失関数 $l\left(W, W'\right)$ として 0/1 評価関数を用いると文誤り最小化（＝事後確率最大化）音声認識と等価になる．損失関数として WER もしくは WER の分子に相当する編集距離（Levenshtein Distance）を用いると WER 最小化音声認識を行えることが知られている [10], [11]．同様に，損失関数として WWER もしくは WWER の定義式（式（1））の分子を用いると WWER 最小化音声認識を行えることを示している [8].

また，我々は WWER 最小化音声認識機能をオープンソースの音声認識エンジン Julius [12] に実装し，MBR-Julius[13] を開発している．MBR-Julius では，N-Best リスコアリングの枠組みに基づき WWER 最小化音声認識を実現している．また，MBR-Julius を用いることが，音声入力型情報検索システムの精度向上につながることを示している [9], [14].

3. ベースライン音声検索システムの構築と評価

ベイズリスク最小化音声認識の効果を調べるために，実際にベースラインとなる音声入力型情報検索システムを構築し評価を行った．本研究で構築したベースライン検索システムは，音声認識結果の 1-Best から検索要求を生成するシステムである．

3.1 検索タスク

検索タスクとして，日本語音声ドキュメント検索テストコレクション [15] を用いた．これは，情報処理学会音声言語情報処理研究会の音声ドキュメント処理ワーキンググループが作成した音声ドキュメント検索評価用テストコレクションである．テストコレクションには研究者が共通で使える実験データセット（講演音声データ，検索課題とそれに対する正解データ）が用意されている．データセットの構成を以下に示す．

- 検索対象文書：CSJ の講演の書き起こしテキスト（2702 講演）
- 利用者の検索要求を記述した「検索課題」：39 課題
- 検索課題を満たす「正解文書のリスト」

テストコレクションには検索課題の読み上げ音声データが含まれていないため，検索課題の読み上げ音声データは，男性 10 名と女性 4 名の計 14 名に読み上げてもらい，合計 546 件を作成した [16].

3.2 音声認識システム

音声認識システムのデコーダには，Julius rev.4.1.5.1 にベイズリスク最小化機能を実装したもの [13]，すなわち MBR-Julius を用いた．音響モデルには，JNAS コーパスから学習した triphone モデル（CSRC2003 年度最終版 [17] に収録）を用いた．言語モデルには，CSJ[18] の講演 2702 件の書き起こしから学習した 3-gram 言語モデル（語彙サイズ約 20K）を用いた．

音声認識手法として，以下の 3 種類を用いた．

- 通常の音声認識である事後確率最大化音声認識
- 単語の誤り数を最小化する WER 最小化音声認識
- 検索の重要語の誤りを最小化する WWER 最小化音声認識

WER 最小化および WWER 最小化音声認識を行うための N-Best リストは，N = 100 とした．MBR 音声認識のためのパラメータは実験的に決定した．その他の音声認識パラメータは Julius rev.4.1.5.1 のデフォルト値を用いた．また，WWER 最小化音声認識を行うための単語重要度として，我々の提案する 2 種類の教師なし推定 [19] によって情報検索への影響が大きい単語に大きな重要度を与えた．なお，この単語重要度についての詳細は 3.5 節で述べる．

3.3 情報検索システム

情報検索システムは，ベクトル空間モデルに基づく文書検索システムとし，GETA[20] を用いて構築した．索引語には名詞と動詞の基本形を用いた．本研究では，検索要求

$\boldsymbol{Q}$ が与えられたとき，全ての文書 $\boldsymbol{D}_i$ について $\boldsymbol{Q}$ との類似度 $\mathrm{Sim}(\boldsymbol{Q}, \boldsymbol{D}_i)$ を算出し，類似度が高い順に上位 1000 件を出力する．

本研究では，ベクトルの類似度尺度として SMART[21] を用いた（式（3））．

$$\mathrm{Sim}\left(\boldsymbol{Q}, \boldsymbol{D}_i\right) = \mathrm{SMART}\left(\boldsymbol{Q}, \boldsymbol{D}_i\right) = \sum_{t=1}^{T} \left(Q_t \cdot D_{i,t}\right) \quad (3)$$

$$Q_t = \begin{cases} \dfrac{1+\log\left(\mathrm{qtf}_t\right)}{1+\log\left(\mathrm{avqtf}\right)} \cdot \log\dfrac{N}{n_t} & \text{if} \quad \mathrm{qtf}_t > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$D_{i,t} = \begin{cases} \dfrac{1+\log\left(\mathrm{tf}_{i,t}\right)}{1+\log\left(\mathrm{avtf}\right)} \cdot \mathrm{Norm} & \text{if} \quad \mathrm{tf}_{i,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{Norm} = \frac{1}{(1-\mathrm{slope}) \cdot \mathrm{pivot} + \mathrm{slope} \cdot \mathrm{utf}_i}$$

ここで，$\mathrm{tf}_{i,t}$ は $\boldsymbol{D}_i$ 中での索引語 t の出現数，avtf は $\boldsymbol{D}_i$ における索引語の出現数の平均を表す．pivot は 1 文書中の異なり索引語数の平均，utf_i は $\boldsymbol{D}_i$ 中の異なり索引語数を表す．slope は補間係数であり，本研究では 0.2 とした．qtf_t は，$\boldsymbol{Q}$ 中での索引語 t の出現数，avqtf は $\boldsymbol{Q}$ に含まれる索引語の出現数の平均を表す．N は検索対象の文書集合の全文書数を表し，n_t は索引語 t を含む文書の数を表す．T は索引語の総数を表す．

3.4 システムの評価尺度

音声認識の評価尺度として，WWER（式（1））を用いた．WWER の計算に用いる単語重要度は，3.2 節で述べた WWER 最小化音声認識で用いたものと同一とした．

情報検索の評価尺度として，11 点平均精度（11ptAP: 11-point Average Precision）[22] を用いた（式（4））．

$$\mathrm{11ptAP}_{Q_k} = \frac{1}{11}\sum_{i=0}^{10} IP_{Q_k}\left(\frac{i}{10}\right) \quad (4)$$

$$IP_{Q_k}\left(x\right) = \max_{x \le R_{Q_k}(t)} P_{Q_k}\left(t\right)$$

ここで，$R_{Q_k}\left(t\right)$ と $P_{Q_k}\left(t\right)$ は，それぞれ Q_k に関する検索順位 t における再現率と精度を表す．$IP_{Q_k}\left(x\right)$ は，再現率レベルが x 以上の精度 $P_{Q_k}\left(t\right)$ の最大値を表す補間精度である．

3.5 単語重要度の自動推定

本節では，WWER 最小化音声認識を行うための単語重要度の自動推定について述べる．まず，単語重要度の推定手法について述べる．次に，実際に推定した単語重要度の評価について述べる．

3.5.1 適切な単語重要度

単語の重要度は，その語の音声認識誤りが情報検索精度に与える影響の程度に基づいて設定することが重要である．具体的には，WWER と検索性能の低下が等しくなるように単語の重要度を設定することが重要である．ここで，情報検索の性能低下の評価尺度のために検索精度低下率（IRDR: Information Retrieval performance Degradation Ratio）を導入する（式（5））．

$$\mathrm{IRDR} = 1 - \frac{H}{R} \quad (5)$$

R と H はそれぞれ，書き起こしと音声認識結果の検索要求を用いた際の検索精度を表す．したがって，IRDR は音声認識誤りが情報検索に与える影響の割合を表す．この IRDR を WWER から予測できれば，WWER 最小化音声認識により検索性能の向上が得られると考えられる．このような WWER を定義できる単語重要度が適切である．

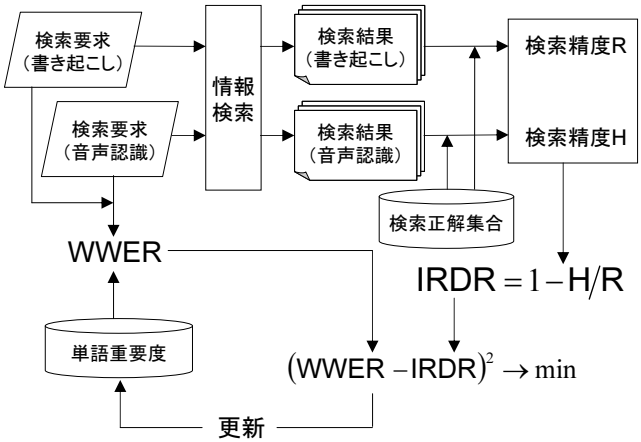

図 1 単語重要度の自動推定の概要

3.5.2 単語重要度の自動推定手法

WWER から IRDR を予測できるような単語重要度を人手で設定することは困難である．これに対し，本研究では，想定される検索要求の発話データ，その書き起こし，および検索の正解集合の 3 つのデータを用いて，単語重要度を自動推定する [19], [23]．これは，検索要求の音声認識結果 W'_m と誤りのない書き起こし W_m のペアを N 件用意し，各 $m\,(m = 1 \ldots N)$ に対して $\mathrm{WWER}_m\left(\boldsymbol{x}\right)$ と IRDR_m の平均二乗誤差 $F\left(\boldsymbol{x}\right)$ を最小化する手法である（式（6））．

$$F\left(\boldsymbol{x}\right) = \sum_m \left(\mathrm{WWER}_m\left(\boldsymbol{x}\right) - \mathrm{IRDR}_m\right)^2 \quad (6)$$

単語重要度の自動推定の概要を図 1 に示す．

なお，この手法には以下の 3 つのデータが必要である．

- 想定される検索要求の発話データ（の音声認識結果）
- 発話データの正しい書き起こし
- 各検索要求の正解集合

本研究では，この 3 つのデータを全て自動で作成する手法である，教師なし推定を用いて単語重要度の推定を行う [19], [23]．

3.5.3 単語重要度自動推定の実験条件

単語重要度の自動推定の実験を行った．単語重要度推定のために用いる検索システムは 3.3 節で述べたものを用い

元の検索要求:	サルサ を 踊れる よう に なる 方法 が 知り たい
テンプレート:	<> に なる 方法 が 知り たい
擬似検索要求:	農地 に なる 方法 が 知り たい

図 2 教師なし推定 1 の擬似検索要求生成の例

擬似検索要求 1:	研修生 の 歴代 の 文句 が 知り たい
擬似音声認識結果 1:	健在 の インテリア 文句 が 知り たい
擬似検索要求 2:	修道院 の 歴代 の 越後湯沢 が 知り たい
擬似音声認識結果 2:	修道院 の 歴代 の うんそう が 知り たい

図 3 教師なし推定 1 の擬似検索要求と擬似音声認識結果の例

た．IRDR 計算（式（5））のための情報検索の評価尺度として，11 点平均精度（式（4））を用いた．

本実験では，単語重要度推定の教師なし推定のための学習データの生成手法として，以下の 2 種類を用いた．

- 教師なし推定 1
- 教師なし推定 2

それぞれについて詳細に述べる．

教師なし推定 1 教師なし推定 1 では，単語重要度推定のための学習データには以下のものを用いた．

- 検索要求のテンプレートと特徴語から生成した擬似検索要求の集合（$\boldsymbol{W}_{p_1}$）
- 擬似検索要求に対しランダムに誤りを発生させて生成した擬似音声認識結果の集合（$\boldsymbol{W}'_{p_1}$）
- 擬似検索要求 $\boldsymbol{W}_{p_1}$ の各検索結果から生成した擬似正解データの集合（$\boldsymbol{C}_{\boldsymbol{W}_{p_1}}$）

擬似検索要求の集合 $\boldsymbol{W}_{p_1}$ はテンプレートと特徴語を用いて 10000 件作成した．検索要求のテンプレートは，NTCIR-3 WEB 検索タスク [24] の音声入力検索課題 47 件を元に作成した．具体的には，検索要求中の一部を特徴語を当てはめるためのブランクに置き換えることで作成した．擬似検索要求生成のための特徴語は，3.2 節で述べた音声認識システムの単語辞書中の単語（本実験では名詞）とした．これは，音声入力型情報検索システムでは，音声認識辞書中の単語のみを検索に用いる語とするので，それらの重要度が推定されれば十分と考えるためである．実際に検索要求からテンプレートを作成し，擬似検索要求を生成した例を図 2 に示す．

擬似音声認識結果の集合 $\boldsymbol{W}'_{p_1}$ として，ランダムに単語を挿入，削除，置換することで，擬似検索要求 $\boldsymbol{W}_{p_1}$ の各文ごとに 5 種類の擬似音声認識結果を生成し，これを用いた（合計 50000 文）．挿入および置換する単語リストは 3.2 節で述べた音声認識システムの単語辞書に登録されている全ての単語とした．擬似音声認識結果は 50000 文の WER の平均が 50%となるように生成した．この手法で実際に生成された擬似検索要求と擬似音声認識結果の例を図 3 に示す．

検索の正解集合 $\boldsymbol{C}_{\boldsymbol{W}_{p_1}}$ には，擬似検索要求 $\boldsymbol{W}_{p_1}$ の各検索課題について検索結果の上位 20 件を求め，それを用いた．

擬似検索要求 1:	課税 に は どんな もの が ある か
擬似音声認識結果 1:	完全 に は あっ どん な もの が ある か
擬似検索要求 2:	病人 は どこ で 見 られる か
擬似音声認識結果 2:	病院 は 合っ て どこ で 見 られる か

図 4 教師なし推定 2 の擬似検索要求と擬似音声認識結果の例

教師なし推定 2 教師なし推定 2 では，単語重要度推定のための学習データには以下のものを用いた．

- 検索要求のテンプレートと特徴語から生成した擬似検索要求の集合（$\boldsymbol{W}_{p_2}$）
- 擬似検索要求に対し音声認識シミュレートを行い誤りを発生させて生成した擬似音声認識結果の集合（$\boldsymbol{W}'_{p_2}$）
- 擬似検索要求 $\boldsymbol{W}_{p_2}$ の各検索結果から生成した擬似正解データの集合（$\boldsymbol{C}_{\boldsymbol{W}_{p_2}}$）

擬似検索要求の集合 $\boldsymbol{W}_{p_2}$ はテンプレートと特徴語を用いて 20000 件作成した．検索要求のテンプレートは，NTCIR-3 WEB 検索タスク [24] の音声入力検索課題 47 件と NTCIR-9 SpokenDoc[25] formal run 用の検索課題 86 件の計 133 件を元に作成した．テンプレートの生成手法は教師なし推定 1 と同様である．擬似検索要求生成のための特徴語は，3.2 節で述べた音声認識システムの単語辞書中の単語（本実験では名詞）とした．

擬似音声認識結果の集合 $\boldsymbol{W}'_{p_2}$ として，統計的機械翻訳の枠組みに基づきサブワード単位で音声認識結果をシミュレート [23] することで，擬似検索要求 $\boldsymbol{W}_{p_2}$ の各文ごとに擬似音声認識結果を生成し，これを用いた（合計 74325 文）．音声認識シミュレートのためのシステムは，古谷ら [19] のものと同一とした．この手法で実際に生成された擬似検索要求と擬似音声認識結果の例を図 4 に示す．

検索の正解集合 $\boldsymbol{C}_{\boldsymbol{W}_{p_2}}$ には，擬似検索要求 $\boldsymbol{W}_{p_2}$ の各検索課題について検索結果の上位 20 件を求め，それを用いた．

単語重要度推定の評価は，評価用の検索要求（擬似音声認識誤りを含む）に対して推定された単語重要度に基づく WWER を計算し，その擬似音声認識誤りによる検索精度の低下率（IRDR）との相関を求めることで行った．高い正の相関が得られるほど，WWER による IRDR の予測精度が高いことを示しているため，適切な単語重要度が得られているといえる．評価データ用の検索要求には NTCIR-9 SpokenDoc[25] に含まれている検索要求合計 125 件とその正解集合を用いた．これらの検索要求に対して，ランダムに単語の削除，挿入，置換を行い，擬似的に音声認識誤りを含む検索要求を作成した．挿入および置換する単語リスト

表 1　WWER と IRDR の相関

単語重要度の決定方法	相関係数
一様	0.53
品詞	0.53
教師なし推定 1	**0.59****
教師なし推定 2	**0.56****

一様：　全ての単語重要度が 1
品詞：　名詞と動詞に 10，機能語に 1

は 3.2 節で述べた音声認識システムの単語辞書から作成した．その際，WER がおよそ 10%，20%，30%，...，100%となるようにし，各 WER ごとにそれぞれ 100 件作成した．この手法を用いて，評価データとして 125000 件のペアを準備した．このペアを用いて 3.3 節のシステムで検索を行って IRDR を計算し，IRDR > 0 のペアを評価データとした．なお，IRDR の最大値は 1 となるため，WWER と IRDR の相関を求める際に，WWER > 1 の場合は WWER $= 1$ として計算した．また，比較のために，ベースラインとして以下の 2 種類の単語重要度を設定した．

- 全ての単語に 1（“一様” と表記，このとき WWER $=$ WER）
- 内容語（名詞と動詞）に 10，機能語（それ以外）に 1（“品詞” と表記）

3.5.4　単語重要度自動推定の実験結果

教師なし推定の実験結果を表 1 に示す．表中の太字はベースラインの相関係数より高い相関係数を表し，** は提案手法とベースラインの相関係数の間で Meng-Rosenthal-Rubin Method [26] を用いて検定を行い，有意水準 1%で有意差があったものを表す．

一様な単語重要度を用いた場合の WWER（WER），品詞ごとに一様な単語重要度を用いた WWER と IRDR の相関係数はどちらも 0.53 であった．これに対して，教師なし推定 1 で決定した単語重要度を用いた WWER と IRDR の相関係数は 0.59 と高かった．また，ベースラインの相関係数と比較して，有意水準 1%で有意差があった．教師なし推定 2 で決定した単語重要度を用いた WWER と IRDR の相関係数も 0.56 と高く，ベースラインの相関係数と比較して，有意水準 1%で有意差があった．このことは，一様な単語重要度および品詞ごとに一様な単語重要度よりも，教師なし推定を行うことで得られた単語重要度から，IRDR をより近似する WWER を得ることができることを示している．実際にそれぞれの教師なし推定により得られた単語重要度の例を図 5 に示す．文書特定能力が高い単語である “講演” や “音声” には，他の単語よりも相対的に高い重要度が設定されていることがわかる．これらのことは，推定した単語重要度を用いて WWER を定義し，それを最小化する音声認識を行うことが，情報検索の精度向上につながりやすいことを示す．

	講演	音声	の	特徴	について	...
教師なし推定 1	3.25	1.95	1.00	1.00	1.00	
教師なし推定 2	5.89	5.29	1.32	1.00	1.00	

図 5　それぞれの教師なし推定で得られた単語重要度の例

表 2　ベースライン音声検索システムの結果

音声認識手法	11 点平均精度
事後確率最大化	0.358
WER 最小化	0.358
WWER-1 最小化	0.358
WWER-2 最小化	0.358

WWER-1：　教師なし推定 1 の単語重要度に基づく WWER
WWER-2：　教師なし推定 2 の単語重要度に基づく WWER

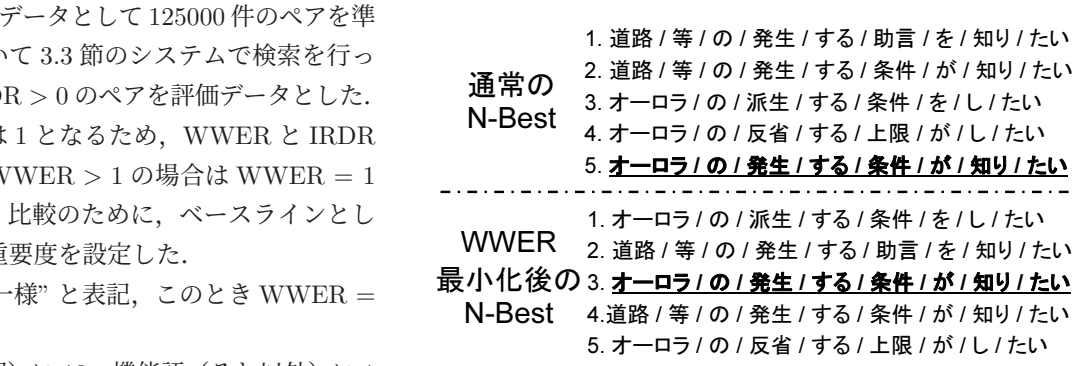

図 6　事後確率最大化音声認識と WWER 最小化音声認識の N-Best の比較

3.6　ベースライン音声検索システムでの実験結果

ベースラインの音声検索システムの評価結果を表 2 に示す．事後確率最大化音声認識を用いた場合の 11 点平均精度は 0.358 となった．WER 最小化音声認識を用いた場合，WER は 0.266 から 0.264 となり誤りが削減された．WWER-1（教師なし推定 1 の単語重要度）を最小化する WWER 最小化音声認識を用いた場合，WWER は 0.316 から 0.313 となり誤りが削減された．WWER-2（教師なし推定 2 の単語重要度）を用いた場合も同様に誤りが削減された．しかし，いずれの音声認識を用いた場合でも 11 点平均精度は 0.358 となり，本タスクではベースラインとの差が見られなかった．このことは，ベイズリスク最小化音声認識を行うことで重要語の音声認識誤りの削減は可能であるものの，実際に検索性能の向上に結びつきにくいタスクが存在することを示している．

4.　N-Best リストからの検索要求生成と評価

4.1　N-Best リストを用いるメリット

事後確率最大化音声認識および WWER 最小化音声認識を行った際の N-Best リストの比較を図 6 に示す．これは，説明のために作成した例であり，実例ではない．この例では，通常の音声認識結果の 5 番目に適切と考えられる仮説（図中の下線を付与している仮説）が出現している．一方，WWER 最小化音声認識の結果には，適切な仮説は 3 番目

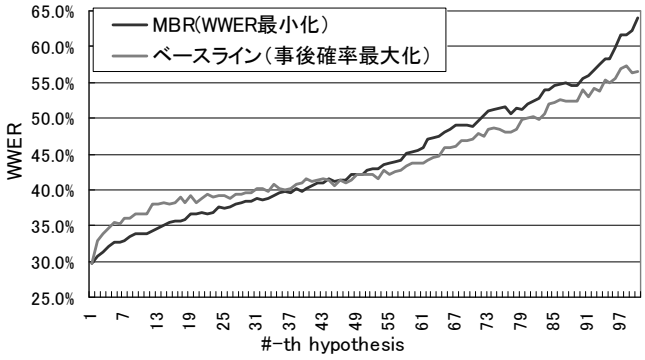

図 7 事後確率最大化音声認識と WWER 最小化音声認識の N-Best の音声認識率の比較

に出力されている．このように，WWER 音声認識を行えば，検索の観点で適切な仮説が 1-Best に出現しない場合であっても，上位の仮説として出現しやすくなると考えられる．

実際に，3 章で述べた音声認識システムを用いて，事後確率最大化音声認識および MBR 音声認識（WWER 最小化音声認識）を行い，これにより生成した N-Best リストの評価を行った．具体的には，各音声認識の結果得られた N-Best リストの 1-Best の仮説の認識率（WWER）の平均，2-Best の仮説の認識率の平均，...，100-Best の仮説の認識率の平均を求めた．その結果を図 7 に示す．1-Best での認識率は同等であるものの，2-Best 以降の結果では，上位の仮説の認識率は MBR 音声認識を行った場合に認識率が高くなる（WWER が低くなる）傾向が見られている．このことから，MBR 音声認識を行うことで N-Best リスト中の上位に認識率の高い仮説を得られやすくなる，すなわち上位の質の高い N-Best リストを生成できることがわかる．このことは，検索要求を 1-Best から生成することと比較して，N-Best 中の上位の仮説から生成することにより，適切な検索要求を生成できる可能性を示している．

4.2 N-Best リストを用いた検索要求生成方法

4.2.1 仮説の順位に基づく重みを与える仮説統合

音声認識結果の N-Best リストは仮説の信頼度に基づきランク付けされている．このため，上位の仮説に含まれる単語は信頼度が高く，正解単語である可能性が高いと考えられる．このことから，式（3）における索引語 t の出現数 qtf_t を N-Best リストから生成する際に，順位に基づいた重みを与えて各仮説を併用することを提案する．順位に基づく重みの与え方として，以下の 3 つを提案する．

- 順位の重みを与えない方法（以下，“N-Best（重みなし）”と記述）．
- 順位の逆数を重みとして与える方法（以下，“N-Best（線形）”と記述）．
- 順位の対数の逆数を重みとして与える方法（以下，“N-Best（対数）”と記述）．

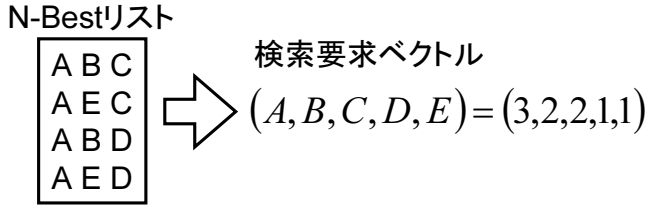

図 8 仮説の順位を基づく重み付きの仮説統合の例（N-Best（対数））

N-Best（重みなし），N-Best（線形），N-Best（対数）はそれぞれ式（7），式（8），式（9）で qtf_t を計算する．

$$\mathrm{qtf}_t = \sum_{n=1}^{N} \mathrm{qtf}_{t,n} \tag{7}$$

$$\mathrm{qtf}_t = \sum_{n=1}^{N} \frac{\mathrm{qtf}_{t,n}}{n} \tag{8}$$

$$\mathrm{qtf}_t = \sum_{n=1}^{N} \frac{\mathrm{qtf}_{t,n}}{\log_2(n+1)} \tag{9}$$

ここで，$\mathrm{qtf}_{t,n}$ は t 番目の索引語が N-Best リストの n 番目の仮説に出現した回数を表し，N は N-Best の候補数を表す．$\mathrm{qtf}_{t,n}$ が小数部を持った場合，最も近い整数に切り上げを行う．なお，$N = 1$ の場合には，1-Best の仮説の単語数に基づき検索要求を生成した結果に一致する．N-Best（対数）に基づいた仮説統合の例を図 8 に示す．図 8 の “A”，“B”，“C”，“D”，“E” はそれぞれ単語を表す．“B” と “E” に着目すると，“B” が初めて出現した仮説は 1 番目の仮説であり，検索要求のベクトルの要素は 2 となる．一方，“E” が初めて出現した仮説は 2 番目の仮説であり，検索要求のベクトルの要素は 1 となる．例に示すとおり，上位の仮説に含まれる索引語の出現数が高い値となる．

4.2.2 仮説のアライメントに基づく仮説統合

本研究で用いる検索要求ベクトルは索引語の出現数に基づいているため，音声認識では単語単位での認識精度が重要となる．そこで，N-Best リストから WTN（Word Transition Network）を生成し，検索要求を生成することを提案する．具体的には，ROVER 法 [27] と同様の手法に基づき，N-Best リストを WTN に変換する．N-Best リストから WTN を生成する例を図 9 に示す．図 9 の “A”，“B”，“C”，“D”，“E” はそれぞれ単語を表し，コロンの後ろの数値は各単語のクラスタ内でのスコアを表す（例えば，“B:0.4” は，単語 “B” のスコアが 0.4 であることを表す）．具体的な手法を以下に示す．

(1) 音声認識の仮説 $W_n\,(n = 1 \ldots N)$ から N-Best リストを生成する．
(2) W_1 を WTN_1 とする．
(3) $n = 2 \ldots N$ において，DP マッチングを用いて WTN_{n-1} と W_n のアライメントを求め，WTN_n を生成する．

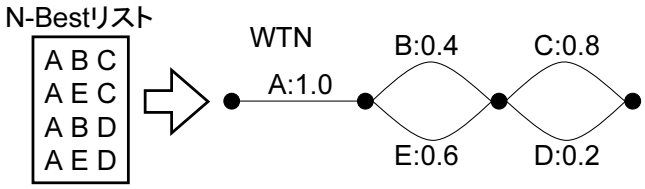

図 9 WTN を用いた仮説統合の例

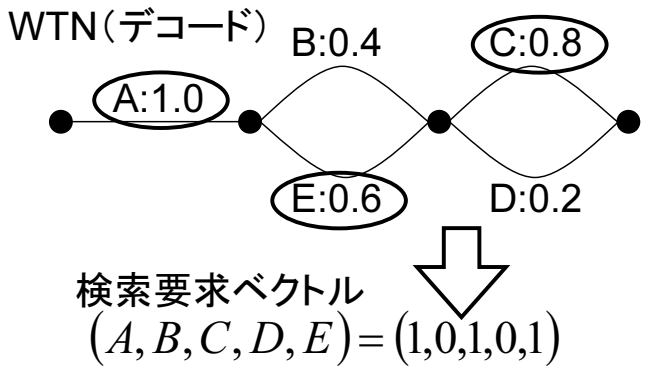

図 10 WTN のデコードに基づく検索要求生成の例

(4) WTN_N の各クラスタごとに，各索引語 t のスコア $S_{i,t}$ および NULL 遷移のスコア $S_{i,NULL}$ を計算する．このとき，クラスタ内に索引語以外の単語が含まれていた場合は NULL 遷移として扱う．

WTN_N の i 番目のクラスタ内の索引語 t のスコア $S_{i,t}$ は式（10）に基づき計算する．

$$S_{i,t} = \frac{CM_{i,t}^{\gamma_1} \cdot CNT_{i,t}^{\gamma_2}}{\sum_t CM_{i,t}^{\gamma_1} \cdot CNT_{i,t}^{\gamma_2}} \tag{10}$$

ここで，$CM_{i,t}$ は i 番目のクラスタの索引語 t の単語の音声認識時の信頼度を表す．$CNT_{i,t}$ は i 番目のクラスタの索引語 t の出現数を表す．γ_1 と γ_2 はそれぞれ $CM_{i,t}$ と $CNT_{i,t}$ の重みパラメータである．

WTN_N からの検索要求の生成手法として，以下の 3 つを提案する．

- 各クラスタの最もスコアの高い単語から検索要求を生成する（以下，“WTN（デコード）” と記述）．
- 各単語ごとにクラスタ間のスコアに応じた重みを与えて検索要求を生成する（以下，“WTN（スコア）” と記述）．
- 各クラスタの最もスコアの高い単語とのスコアの比によって WTN の枝刈りを行う（以下，“WTN（枝刈り）” と記述）．

それぞれについて詳細に述べる．

WTN（デコード）

WTN（デコード）は，WTN_N の各クラスタごとに最もスコアが高い単語を選択する手法である．具体的には，式（3）における索引語 t の出現数 qtf_t を式（11）に基づき計算する．

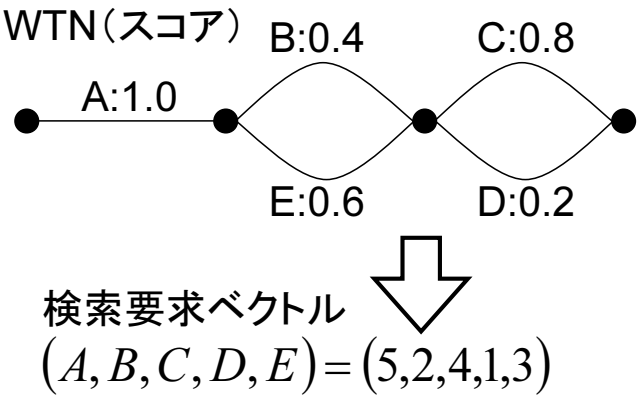

図 11 WTN のスコアに基づく重みを用いた検索要求生成の例（式（12）の $K = 5$）

$$\mathrm{qtf}_t = \sum_{i=1}^{M} IsMax_{i,t} \tag{11}$$

$$IsMax_{i,t} = \begin{cases} 1 & \text{if } S_{i,t} = \max_t S_{i,t} \\ 0 & \text{otherwise} \end{cases}$$

ここで，$S_{i,t}$ は WTN_N の i 番目のクラスタの索引語 t のスコア（式（10））を表し，M は WTN_N のクラスタの総数を表す．$IsMax_{i,t}$ は，$S_{i,t}$ が i 番目のクラスタの中で最もスコアが高い場合に 1，それ以外の場合は 0 となる．WTN（デコード）の例を図 10 に示す．この例では，WTN は “A E C” とデコードされ，これに基づき検索要求が生成される．

WTN（スコア）

WTN（スコア）は，WTN_N の各クラスタごとに単語のスコアに応じて索引語の出現数を決定する手法である．具体的には，式（3）における索引語 t の出現数 qtf_t を式（12）に基づき計算する．

$$\mathrm{qtf}_t = \sum_{i=1}^{M} K \cdot S_{i,t} \tag{12}$$

ここで，$S_{i,t}$ は WTN_N の i 番目のクラスタの索引語 t のスコア（式（10））を表し，M は WTN_N のクラスタの総数を表す．K は索引語の出現数を整数化するパラメータである．qtf_t が小数部を持った場合，小数第 1 位で四捨五入を行う．WTN（スコア）の例を図 11 に示す．2 番目のクラスタに着目すると，“B” と “E” のスコアはそれぞれ 0.4 と 0.6 である．$K = 5$ とすると，“B” の出現数は 2，“E” の出現数は 3 となる．

WTN（枝刈り）

WTN（枝刈り）は，WTN_N の各クラスタごとに単語のスコアに応じて索引語の出現数を決定する手法である．具体的には，式（3）における索引語 t の出現数 qtf_t を式（13）に基づき計算する．

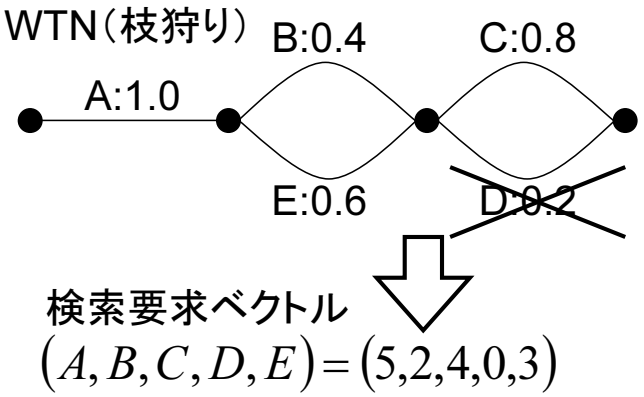

図 12 WTN の枝刈りを用いた検索要求生成の例（式（13）の $K = 5$, $\alpha = 3$）

$$\mathrm{qtf}_t = \sum_{i=1}^{M} Score_{i,t} \tag{13}$$

$$Score_{i,t} = \begin{cases} K \cdot S_{i,t} & \text{if } \dfrac{\max_t S_{i,t}}{S_{i,t}} \leq \alpha \\ 0 & \text{otherwise} \end{cases}$$

ここで，$S_{i,t}$ は WTN_N の i 番目のクラスタの索引語 t のスコア（式（10））を表し，M は WTN_N のクラスタの総数を表す．K は索引語の出現数を整数化するパラメータである．$Score_{i,t}$ は $S_{i,t}$ と i 番目のクラスタ内で最も高いスコアとの比がしきい値 α 以下のときは $K \cdot S_{i,t}$，それ以外のときは 0 となる．qtf_t が小数部を持った場合，小数第 1 位で四捨五入を行う．なお，しきい値 α を無限大とすると，式（13）は式（12）に一致する．WTN（枝刈り）の例を図 12 に示す．図 12 の 3 番目のクラスタに着目すると，このクラスタ内で最も高いスコアを持つ索引語は “C” であり，しきい値 $\alpha = 3$ すると，“C” と “D” のスコアの比は 4 ($> \alpha$) であるため “D” は検索要求に含まれない．

4.3 N-Best リストを用いた情報検索の評価

4.3.1 実験条件

N-Best リストから検索要求を生成する各手法の比較の実験を行った．検索タスク，音声認識システムの構成，情報検索システムの構成，評価尺度は 3 章で述べたベースライン音声検索システムと同様とした．検索要求生成手法として，1-Best（従来法：3 章の検索要求生成と同様），N-Best（重みなし），N-Best（線形），N-Best（対数），WTN（デコード），WTN（スコア），WTN（枝刈り）を用いた．検索要求および WTN 生成のための N-Best リストは各音声認識結果の上位 5 件（N = 5）を用いた．式（12）と式（13）の K は，WTN 生成に用いる N-Best と同一（$K = \mathrm{N} = 5$）とした．式（10）の γ_1 と γ_2，および式（13）の α は，検索要求 39 種類を 1 種類ずつに分割し，39 分割交差検定を行うことで決定した．

表 3 N-Best リストからの検索要求生成手法の比較

検索要求生成手法	音声認識手法			
	事後確率最大化	MBR		
		WER	重要度 1	重要度 2
1-Best	0.358	0.358	0.358	0.358
N-Best（重みなし）	0.351	0.352	0.356	0.355
N-Best（線形）	0.354	0.354	0.356	0.356
N-Best（対数）	0.353	0.355	0.357	0.356
WTN（デコード）	0.343	**0.360**	**0.359**	**0.361**
WTN（スコア）	0.352	0.356	**0.360**	**0.360**
WTN（枝刈り）	0.347	0.357	**0.362**	**0.361**

重要度 1： 教師なし推定 1 の単語重要度に基づく WWER
重要度 2： 教師なし推定 2 の単語重要度に基づく WWER

4.3.2 実験結果

N-Best リストを用いて検索要求を生成する情報検索の実験結果を表 3 に示す．なお，1-Best の結果は表 2 と同一である．

はじめに事後確率最大化音声認識の結果から検索要求を生成した結果について述べる．いずれの検索要求生成手法を用いた場合も，1-Best よりも検索精度が低下した．これは，ベイズリスク最小化音声認識を行わずに N-Best リストを用いて検索要求を生成しても，検索性能の向上を得ることは難しいことを示している．

次に WER 最小化音声認識の結果から検索要求を生成した結果について述べる．N-Best（重みなし），N-Best（線形），N-Best（対数），WTN（スコア），WTN（枝刈り）の生成手法を用いたときには，1-Best よりも検索精度が低下した．一方，WTN（デコード）の生成手法を用いたときには検索精度は 0.360 となり，1-Best よりも高い検索精度が得られた．

次に重要度 1（教師なし推定 1 で決定した単語重要度に基づく WWER）を用いた WWER 最小化音声認識の結果から検索要求を生成した結果について述べる．N-Best（重みなし），N-Best（線形），N-Best（対数）の生成手法を用いたときには，1-Best よりも検索精度が低下した．一方，WTN（デコード），WTN（スコア），WTN（枝刈り）の生成手法を用いたときには，検索精度はそれぞれ 0.359，0.360，0.362 となり，1-Best よりも高い検索精度が得られた．

また，重要度 2（教師なし推定 2 で決定した単語重要度に基づく WWER）を用いた WWER 最小化音声認識の結果から検索要求を生成したときも同様であり，WTN（デコード），WTN（スコア），WTN（枝刈り）の生成手法を用いたときに 1-Best よりも高い検索精度が得られた．

MBR 音声認識を行った上で WTN を構築し，検索要求を生成することにより，検索精度の向上が得られた．特に，重要度 1 に基づく MBR 音声認識と WTN（枝刈り）の組み合わせによって，本研究で最大の検索精度が得られた．WTN からの検索要求の生成は，N-Best リスト中に含まれ

ている索引語のうち，信頼度の低いものを検索要求に含めない，もしくは検索要求中での出現頻度を小さくする手法であり，このような検索要求生成法が有効であることを示している．

次に，各音声認識手法について比較を行う．N-Best（対数）に着目すると，事後確率最大化音声認識，WER 最小化音声認識，WWER 最小化音声認識（重要度 1）の結果はそれぞれ 0.353，0.355，0.357 であり，WWER 最小化音声認識を用いた際に最大の検索精度が得られている．N-Best（重みなし），N-Best（線形），WTN（デコード），WTN（スコア），WTN（枝刈り）においても，事後確率最大化音声認識よりも WER 最小化音声認識，WWER 最小化音声認識の結果で高い検索精度が得られている．また，WWER 最小化音声認識を行ったときは WER 最小化音声認識を行ったときよりも高い，もしくは同等の検索精度が得られている．この結果は，WWER 最小化音声認識を行うことで N-Best リストの上位候補の質が向上し，適切な検索要求を生成することができるようになることを示している．

WWER 最小化音声認識を行って質の高い N-Best リストを生成してから WTN を構築した後に，WTN から信頼度を考慮して検索要求を生成する効果を示した．

5. おわりに

音声入力型情報検索のための音声認識手法と検索要求生成手法について検討を行った．具体的には，ベイズリスク最小化音声認識を行い，その結果得られた N-Best リストを用いて検索要求を生成する手法を提案した．実験の結果，WWER 最小化音声認識を行って重要単語の誤りが少ない N-Best リストを生成し，その N-Best リストの上位を用いて WTN を構築して信頼度を考慮することで，適切な検索要求が生成できることを示した．

謝辞　本研究は科研費基盤研究（B）（21300066）の助成を受けたものである．

参考文献

[1] 翠　輝久，河原達也：限定されたドメインにおける質問応答機能を備えた文書検索・提示型対話システム，情報処理学会研究報告，第 62 回音声言語情報処理研究会，pp. 69–74 (2006).

[2] 桐山伸也，広瀬啓吉，峯松信明：話題知識を導入した文献検索音声対話システム，電子情報通信学会論文誌，Vol. J85-D-II, No. 5, pp. 863–876 (2002).

[3] 大淵康成，神田直之：音声検索実用化の現状と課題，情報処理学会研究報告，第 88 回音声言語情報処理研究会 5, pp. 1–4 (2011).

[4] Hori, C., Hori, T., Isozaki, H., Maeda, E., Katagiri, S. and Furui, S.: Deriving Disambiguous Queries in a Spoken Interactive ODQA System, *Proc. IEEE-ICASSP*, pp. 624–627 (2003).

[5] 翠　輝久，駒谷和範，清田陽司，河原達也：音声対話によるソフトウェアサポートのための効率的な確認戦略，電子情報通信学会論文誌，Vol. J88-DII, No. 3, pp. 499–508 (2005).

[6] 勝丸真樹，中野幹夫，駒谷和範，成松宏美，船越孝太郎，辻野広司，高橋　徹，緒方哲也：複数の言語モデル・言語理解方式を用いた音声理解の高精度化，情報処理学会研究報告，第 75 回音声言語情報処理研究会，pp. 45–50 (2009).

[7] Matsushita, M., Nishizaki, H., Utsuro, T. and Nakagawa, S.: Improving Keyword Recognition of Spoken Queries by Combining Multiple Speech Recognizer's Output for Speech-driven WEB Retrieval, *IEICE TRANS. & SYST.*, Vol. E88-D, No. 3, pp. 472–480 (2005).

[8] 南條浩輝，河原達也，七里　崇：音声理解を指向したベイズリスク最小化枠組みに基づく音声認識，電子情報通信学会論文誌，Vol. J91-D, No. 5, pp. 1314–1324 (2008).

[9] 松尾宏規，西田昌史，古谷　遼，南條浩輝，山本誠一：単語の重要度を考慮したベイズリスク最小化音声認識を用いた音声入力型情報検索システムの評価，日本音響学会講演論文集，秋季研究発表会，pp. 201–202 (2011).

[10] Goel, V., Byrne, W. and Khudanpur, S.: LVCSR rescoring with modified loss functions: A decision theoretic perspective, *Proc. IEEE-ICASSP*, Vol. 1, pp. 425–428 (1998).

[11] Stolcke, A., König, Y. and Weintraub, M.: Explicit word error minimization in N-best list rescoring, *Proc. EUROSPEECH*, pp. 163–166 (1997).

[12] 河原達也，李　晃伸：連続音声認識ソフトウエア Julius，人工知能学会誌，Vol. 20, No. 1, pp. 41–49 (2005).

[13] 古谷　遼，南條浩輝：音声認識エンジン Julius への重要単語誤り最小化機能の実装，日本音響学会講演論文集，秋季研究発表会，pp. 153–154 (2011).

[14] 志々見亮，西田昌史，南條浩輝，山本誠一：音声入力型情報検索に対する単語信頼度によるリスコアリングを適用したベイズリスク最小化音声認識，日本音響学会講演論文集，秋季研究発表会，pp. 205–206 (2012).

[15] Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y. and Itou, K.: Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data, *IPSJ-journal*, Vol. 50, No. 2, pp. 82–94 (2009).

[16] 七里　崇，重安幸治，南條浩輝，吉見毅彦：音声クエリによる講演音声ドキュメント検索の基礎的評価，第 4 回音声ドキュメント処理ワークショップ，No. 16 (2010).

[17] 河原達也，武田一哉，伊藤克亘，李　晃伸，鹿野清宏，山田　篤：連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要，情報処理学会研究報告，第 49 回音声言語情報処理研究会，pp. 325–330 (2003).

[18] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *Proc. ISCA & IEEE-SSPR*, pp. 7–12 (2003).

[19] 古谷　遼，七里　崇，南條浩輝，松尾宏規，西田昌史，山本誠一：ベイズリスク最小化に基づく音声入力型情報検索のための単語重要度の自動推定，第 6 回音声ドキュメント処理ワークショップ，No. 13 (2012).

[20] 高野明彦，西岡真吾，今一　修，岩山　真，丹羽芳樹，久光　徹，藤尾正和，徳永健伸，奥村　学，望月　源，野本忠司：汎用連想計算エンジンの開発と大規模文書分析への応用 (2002).http://geta.ex.nii.ac.jp/pdf/itx002.pdf.

[21] 小作浩美，内山将夫，井佐原均，河野恭之，木戸出正継：WWW 検索における複数検索結果の統合処理とその評価，情報処理学会論文誌，Vol. 44, No. SIG 8(TOD 18), pp. 78–91 (2003).

[22] 北　研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版.ISBN 4-320-12036-1.

[23] 七里　崇，南條浩輝，吉見毅彦：音声入力型情報検索における単語重要度推定のための統計的機械翻訳を用いた

音声認識シミュレート，日本音響学会講演論文集，春季研究発表会 ，pp. 98–99 (2010).
[24] 江口浩二，大山敬三，石田栄美，神門紀子，栗山和子：NTCIR-3 WEB: Web 検索のための評価ワークショップ，*NII Journal*, No. 6, pp. 31–56 (2003).
[25] Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T. and Matsui, T.: Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop, *NTCIR*, 9, pp. 223–235 (2011).
[26] Meng, X. L., Rosenthal, R. and Rubin, D. B.: Comparing correlated correlation coefficients, *Psychological Bulletin*, Vol. 111, No. 1, pp. 172–175 (1992).
[27] Fiscus, J.: A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), *Proc. IEEE-ASRU*, pp. 347–354 (1997).