段階的検索と擬似適合性フィードバックを用いた講演音声ドキュメント検索

南條 浩輝[2,a)]　西尾 友宏[1,b)]　吉見 毅彦[2,c)]

概要：講演や講義などの長い音声ドキュメントの検索のための擬似適合性フィードバック（PRF: Pseudo Relevance Feedback）について述べる．講演音声ドキュメント検索における PRF では，講演の一部を検索対象とした検索（講演パッセージ検索）で初期検索を行い，得られた結果から関連語抽出を行い講演音声を検索対象とした検索をする PRF 手法が効果的であるものの，その効果は十分でない．これに対して，初期検索の検索性能の改善を行ってから関連語抽出を行う PRF 手法を提案する．さらに，講演音声ドキュメント検索の PRF のための関連語抽出方法の提案と種々の PRF による拡張後クエリと拡張前クエリの併用の効果についての調査も行う．CSJ の講演音声を対象とした講演検索および講演パッセージ検索において，平均的な検索精度（11 点平均精度）の向上が確認できた．

キーワード：音声ドキュメント検索，講演音声，擬似適合性フィードバック，クエリ拡張

1. はじめに

音声ドキュメント検索とは，テキストによる書き起こしがなく，ラベルやタグ情報では十分な検索ができない音声データを音声認識によってテキスト化して検索するものである [1]．本研究では，ある程度の長さ（およそ 10 分以上）を持つ講演や講義の音声の自動書き起こし（以下，講演音声ドキュメント）を対象とした擬似適合性フィードバック（PRF: Pseudo Relevace Feedback）を研究する．

PRF とは初期クエリによる音声ドキュメントの検索結果の上位を擬似的に適合ドキュメントとみなして関連語を抽出し，初期クエリに追加することによってクエリ拡張を行う手法である（図 1）．PRF では，各ドキュメントが単一の話題もしくはなるべく少数の話題で構成されていること，および，初期検索の検索性能が低すぎないこと，の 2 点が重要である．

我々は前者の問題に対応する手法をこれまでに提案している．すなわち，講演音声を検索対象とした PRF では，単一の話題もしくはなるべく少数の話題で構成されているとみなすことができる短いパッセージ（講演音声の一部）を検索対象として初期検索を行い，得られた結果から関連語抽出することが効果的であることを明らかにしている [2][3]．この手順は図 2 に示されており，我々は PRFL（PRF for Lectures）とよんでいる．

しかし，PRFL での初期検索は短いパッセージを対象とした検索であるために検索精度が低く，クエリと関連しな

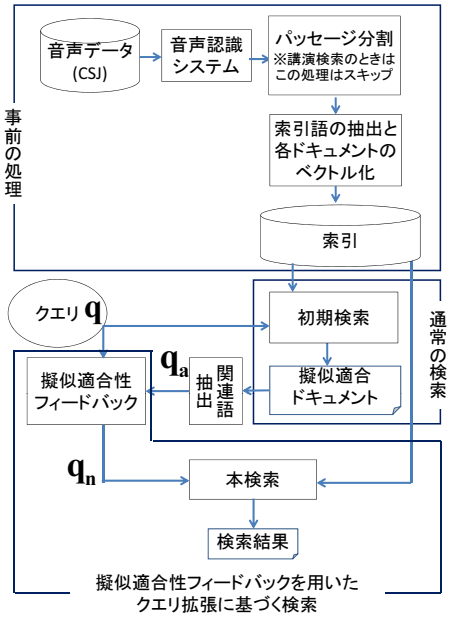

図 1　PRF の概観
Fig. 1　Overview of PRF

[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)] nanjo@nlp.i.ryukoku.ac.jp
[b)] nishio@nlp.i.ryukoku.ac.jp
[c)] yoshimi@nlp.i.ryukoku.ac.jp

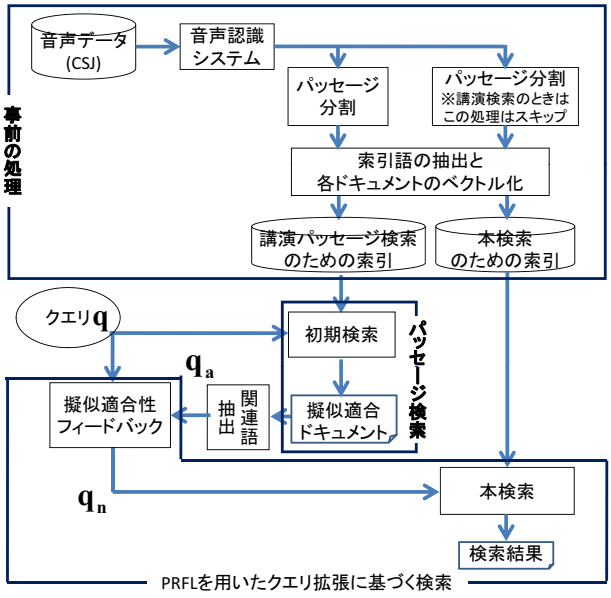

図 2　PRFL の概観
Fig. 2　Overview of PRFL

いドキュメントから関連語抽出が行われる，すなわち PRF が失敗することがある問題を残している．これに対して，本論文では，短いパッセージを対象とした初期検索の検索性能の改善を行い，その結果から関連語を抽出する方法を提案する．具体的には，広域文書と局所文書の類似度統合に基づく講演検索手法 [4] を初期検索（講演パッセージ検索）に適用して初期検索精度を高めて，講演音声ドキュメント検索のための擬似適合性フィードバック（PRFL）を行う手法を提案する．

また，擬似適合性フィードバックにおける関連語抽出の精度向上を目的とした，別のアプローチも提案する．具体的には，関連語抽出において，初期クエリから得られた各擬似適合ドキュメントに共通する話題をクエリに合致する話題とみなしてそこから関連語の抽出を試みる方法，すなわち複数の擬似適合ドキュメントに共通に含まれる語のみを関連語として抽出を行う方法も提案する．

最後に，種々の拡張クエリと拡張前クエリを併用する効果を明らかにする．

2.　講演音声ドキュメント検索

2.1　検索評価に用いるデータ

本研究では，長い音声ドキュメントを対象とした検索実験を行うために，検索対象の音声データとして日本語話し言葉コーパス [5]（CSJ：Corpus of Spontaneous Japanese）の学会講演 987 件と模擬講演 1715 件の合計 2702 件の講演を用いる．

音声ドキュメント検索タスクとして，講演音声ドキュメントそのものを検索単位とするタスクを講演検索タスク，講演音声ドキュメントを分割し，これを検索単位（パッセージ）とするものを講演パッセージ検索タスクがある．本研究では講演パッセージ検索の対象として，秋葉らによるテストコレクションの検索性能の基本評価 [6] でも利用されていた 60 発話単位，30 発話単位，15 発話単位，10 発話単位，および 5 発話単位を採用する．60 発話単位とは単純に講演の先頭から順に 60 発話ごとに区切ったパッセージでありこの各区間が検索対象（パッセージ）となる．30 発話単位と 15 発話単位，10 発話単位，5 発話単位についても 60 発話単位と同様に講演の先頭から順にそれぞれ 30 発話，15 発話，10 発話，5 発話ごとに区切ったものである．

クエリには自然言語文で記述された 125 件のテキスト（NTCIR-9 SpokenDoc の dry run 用クエリ 39 件および formal run 用クエリ 86 件）を用いる．テストコレクションには，適合情報が付与されており，適合度には適合（R）と部分適合（P）がある．本研究では適合ラベル（R）が付与された区間を一部でも含むドキュメントをクエリに対する正解として扱った．

2.2　評価尺度

本研究では，検索結果の上位にどれだけ正解ドキュメントが存在するかを評価することができる 11 点平均精度（11-point Average Precision, “11ptAP” と記す）[7] を用いる．11ptAP は，式（1）に示されるとおり，各検索クエリ Q_k に対して 0.0 から 1.0 まで 0.1 刻みでの各再現率レベル x における補間精度 $IP_{Q_k}(x)$（式（3））を求め，それらの平均 $11ptAP(Q_k)$（式（2））を全検索クエリで平均したものである．

$$\mathrm{11ptAP} = \frac{1}{N}\sum_{k=1}^{N} 11ptAP(Q_k) \tag{1}$$

$$11ptAP(Q_k) = \frac{1}{11}\sum_{i=0}^{10} IP_{Q_k}(\frac{i}{10}) \tag{2}$$

$$IP_{Q_k}(x) = \max_{x \le R_{Q_k}(T)} P_{Q_k}(T) \tag{3}$$

ここで $R_{Q_k}(T)$ と $P_{Q_k}(T)$ は，それぞれクエリ Q_k に対する検索結果の上位 T 番目までの検索結果の再現率と精度である．

3.　検索システム

本論文ではベクトル空間モデル [7] に基づくドキュメント検索システムを用いて，擬似適合性フィードバックの効果を検証する．ベクトル空間モデルは，ドキュメントとクエリをベクトルで表現し，ベクトル間の類似度により検索を実現するモデルである．本論文では，ベクトル間の類似度に SMART[8] を用いる．これはあるクエリ Q とドキュメント $D_i(1 \le i \le N)$ の類似度を，Q と D_i のそれぞれでの索引語 $t_k(1 \le k \le m)$ の正規化出現頻度 q_{t_k} および d_{i,t_k} を用いて，式（4）で与えるものである．

$$\mathrm{SMART}(Q, D_i) = \sum_{k=1}^{m} (q_{t_k} \cdot d_{i,t_k}) \tag{4}$$

ただし，

$$d_{i,t_k} = \begin{cases} \dfrac{1 + \log(\mathrm{tf}_{i,t_k})}{1 + \log(\mathrm{avtf}_i)} \Big/ (1 - \mathrm{slope}) \cdot \mathrm{pivot} + \mathrm{slope} \cdot \mathrm{utf}_i & \text{if } \mathrm{tf}_{i,t_k} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$q_{t_k} = \begin{cases} \dfrac{1 + \log(\mathrm{qtf}_{t_k})}{1 + \log(\mathrm{avqtf})} \log \dfrac{N}{n_{t_k}} & \text{if } \mathrm{qtf}_{t_k} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

ここで，tf_{i,t_k} は D_i 中での t_k の出現数，avtf_i は D_i における単語の出現数の平均を表す．pivot は 1 ドキュメント中の異なり単語数の平均，utf_i は D_i 中の異なり単語数を表す．slope は補間係数（0.2）である．qtf_{t_k} は，Q 中での t_k の出現数，avqtf は Q に含まれる単語の出現数の平均を表す．N は検索対象ドキュメント数を表す．n_{t_k} は，t_k を含むドキュメント数を表す．

本研究では，先行研究 [9] に従って，索引語を動詞と名詞の基本形とする．検索エンジンには汎用連想計算エンジン GETA[*1]を用いる．

これらに基づいて検索システムを構築し，クエリ Q が与えられたとき，全てのドキュメント D_i について Q との類似度 $\mathrm{SMART}(Q, D_i)$ を算出して類似度が 0 より大きいものを高い順に全件出力し，評価を行う．

4. 講演音声ドキュメント検索のための擬似適合性フィードバック

講演検索タスクや講演パッセージ検索のうちでも比較的長いパッセージを対象とする検索タスクでは，中程度の検索精度（11ptAP が 0.3 から 0.5 程度）が得られることがわかっている [6][9]．しかし，このような検索タスクでは，検索で得られる擬似適合ドキュメントは長く複数の話題を含むため，クエリに強く合致する関連語を抽出することが難しく PRF が機能しないことを我々は明らかにしており，講演音声ドキュメント検索のための擬似適合性フィードバック（PRFL: Pseudo Relevance Feedback for Lectures）を提案している [2][3]．この手法の概観は図 2 に示されている．

4.1 PRF の概要

適合性フィードバックとは，得られた検索結果のうち，どのドキュメントが検索意図に適合し，どのドキュメントが適合でないかをユーザが検索システムに入力することにより，クエリベクトル $\mathbf{q}$ を $\mathbf{q_n}$ に修正するものである [7]（式（7））．

$$\mathbf{q_n} = \omega_o \mathbf{q} + \omega_1 \mathbf{d_r} - \omega_2 \mathbf{d}_{\ bfn} \tag{7}$$

このとき，$\mathbf{d_r}$ と $\mathbf{d_n}$ はそれぞれ適合ドキュメントの集合と不適合ドキュメントの集合に含まれる単語の出現頻度を各要素としたクエリベクトルとしたものである．また，ω_0，ω_1，ω_2 は 0 以上の定数である．

ユーザにより適合ドキュメントであるかどうかの判断を必要とせず，ユーザとのインタラクションなしに関連語を抽出し，クエリ拡張を行う手法を擬似適合性フィードバック（PRF: Pseudo Relevance Feedback）という [10]．PRF では，はじめにクエリ Q を用いて検索結果を得る．次に得られた検索結果の上位いくつかを擬似的に適合ドキュメント集合とし，これらのドキュメントから関連語を抽出し，初期クエリに追加することでクエリ拡張を行う．このとき，不適合ドキュメントは用いない．したがって ω_0/ω_1 を β とし $\mathbf{d_r}$ を $\mathbf{q_a}$ とすれば PRF は以下の式（8）で表される．

$$\begin{aligned} \mathbf{q_n} &= \omega_0 \mathbf{q} + \omega_1 \mathbf{d_r} \\ &\approx \frac{\omega_0}{\omega_1} \mathbf{q} + \mathbf{d_r} \\ &\approx \beta \mathbf{q} + \mathbf{q_a} \end{aligned} \tag{8}$$

4.2 PRFL の概要と実験条件

PRFL の概観は図 2 に示されているとおりであり，基本的には PRF（図 1）と同じである．PRF と異なる点は，PRFL では，初期検索があらかじめ短く分割された講演パッセージを対象としたパッセージ検索という点のみである．

PRFL に必要なパラメータとして，擬似適合ドキュメント数，関連語数，重みおよび，初期検索（講演パッセージ検索）のパッセージ単位がある．擬似適合ドキュメント数は，初期クエリで得られた検索結果の上位何件を擬似的に適合ドキュメントとみなすかについてのパラメータである．関連語数は，適合ドキュメントから関連語抽出を行う際に何語抽出するかについてのパラメータである．重みは式（8）における初期クエリに対する重み β である．

本研究では，初期検索のパッセージ単位に 10 発話単位を用いる．それぞれのドキュメント長の平均は 24 秒である[*2]．擬似適合ドキュメント数，関連語数，重みは交差検定によって決定する．

関連語の抽出は単語の出現頻度に基づいて行う．具体的には初期クエリから得られた擬似的な適合ドキュメント集合をひとつのクエリとみなして，式（6）に基づいて各語の正規化出現頻度 q_{t_k} を求め，この値が高いものから順に抽出する．

[*1] 汎用連想検索エンジン GETA(http://geta.ex.nii.ac.jp)

[*2] 単純に全講演の長さの合計を各ドキュメント数で割った平均

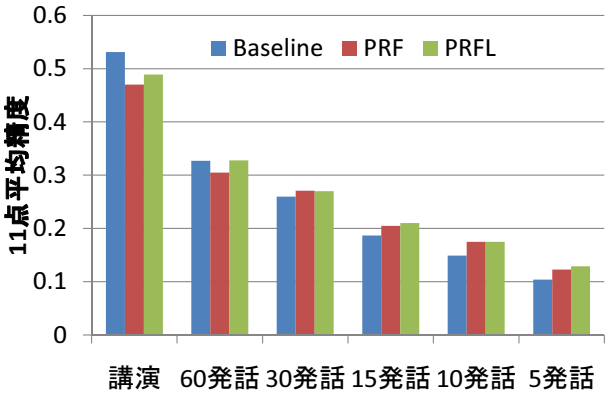

図 3 PRFL の効果（11ptAP の比較）
Fig. 3 Effects of of PRFL (Comparison of 11ptAP)

4.3 PRFL の効果と問題点

講演検索において PRFL を行った結果について述べる．具体的には初期検索（10 発話単位の講演パッセージのパッセージ検索）を行ってクエリを拡張し，2 回目の本検索で講演検索を行った．検索結果を図 3（左端: 講演）に示す．PRFL により PRF よりも高い検索精度が得られるものの，初期クエリを用いたとき（図 3 Baseline）よりも高い検索精度は得られないことがわかる．

次に講演パッセージ検索を行ったときの結果について述べる．具体的には，初期検索（10 発話単位の講演パッセージのパッセージ検索）を行ってクエリを拡張し，2 回目の本検索で講演パッセージ検索（60 発話，30 発話，15 発話，10 発話，5 発話）を行った．PRF と比べて PRFL により高いまたは同等の検索精度を得られることがわかる．

講演音声ドキュメント検索での PRF において，初期検索に講演パッセージ検索（10 発話単位）を行うことが効果的な理由と問題点について考察する．

スライドを利用する講演ではスライド 1 枚につき 1 つの話題について話されていると考えられ，著者らの経験上，スライド 1 枚の説明にかかる時間はおおよそ 30 秒から 1 分程度であることが多く，10 発話単位はこれを擬似的にモデル化していると考えられる．このため，10 発話単位のパッセージは複数話題を含まず，初期検索結果の上位に正解文書があれば，そこから適切な関連語を抽出できる．しかし，10 発話単位講演パッセージの検索精度は 0.149 （図 3 の Baseline，10 発話）と低めであり，初期検索結果の上位に正解文書が存在しない可能性がある．すなわち，PRFL において目的とするドキュメントとは異なるドキュメントから単語を抽出してクエリ拡張が行われてしまい，検索精度が低下するという問題が残っている．

5. 講演音声ドキュメント検索のための類似度統合を用いた擬似適合性フィードバック

前章では，PRFL の効果と問題点について述べた．問題点として具体的に，初期検索（短いパッセージの検索）精度が低いことを述べた．本章では，この問題に対して，広域文書と局所文書の類似度統合に基づく講演検索手法 [4] を初期検索（講演パッセージ検索）に適用して初期検索精度を高めて，講演音声ドキュメント検索のための擬似適合性フィードバックを行う手法を提案する．これを本論文では，講演音声ドキュメント検索のための類似度統合を用いた擬似適合性フィードバック（PRFS: Pseudo Relevance Feedback for lecture with Similarity score integration）と呼ぶ．

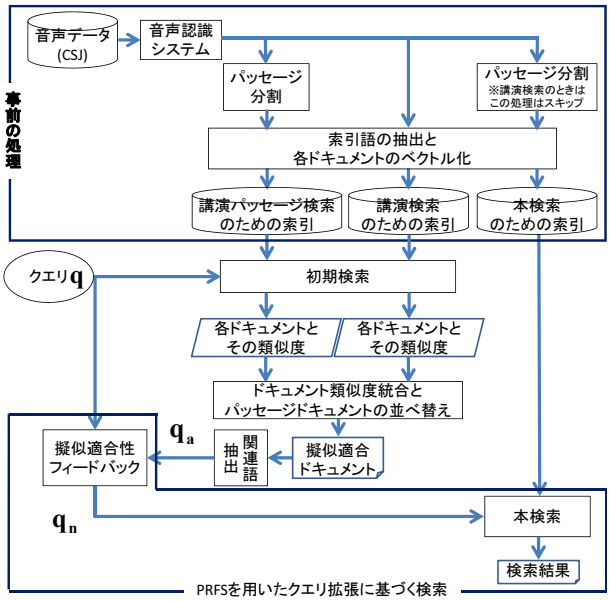

図 4 PRFS を用いた音声ドキュメント検索の概観
Fig. 4 Overview of PRFS

この類似度統合によって検索性能を改善する方法自体は新しくないが，講演音声ドキュメント検索において PRF と組み合わせてその効果を調査して研究はこれまでにみられず，本論文はこの点において新しい．

5.1 PRFS の概要

PRFS の概観を図 4 に示す．基本的には PRF と同じである．PRF と異なる点は，PRFS では，初期検索において，講演検索と講演パッセージ検索のそれぞれを行った上で，類似度を統合することにより講演パッセージの検索結果を調整して初期検索結果とする点である．

本研究では，先行研究 [2][3][4] に従って，講演パッセージ検索には 10 発話単位，類似度統合には対数線形補間（式 9）をそれぞれ用いる．

$$\mathrm{SIM}(Q, D_{i,k}^{k+9}) = \lambda \log \mathrm{SMART}(Q, D_i) + (1-\lambda) \log \mathrm{SMART}(Q, D_{i,k}^{k+9}) \quad (9)$$

ここで，Q は拡張前クエリ，D_i は i 番目の講演ドキュメントを表し，$D_{i,k}^{k+9}$ は i 番目の講演の一部（D_i の k 番目から $k+9$ 番目の発話=10 発話単位のパッセージ）を表す．

λ は統合重みを表す．

すなわち，PRFS では 10 発話単位パッセージの検索結果と初期クエリ Q の類似度に，そのパッセージを含む講演とクエリ Q との類似度を統合して新たな類似度 $\mathrm{SIM}(Q, D_i, {}_k^{k+9})$ を計算し，その統合スコアに基づいて検索結果（10 発話単位）を並べた変えたものを，擬似適合性フィードバックのための初期検索結果として用いる．

PRFS には，以下のパラメータがある．

- 類似度の統合重み λ
- 擬似適合ドキュメント数
- 関連語数
- クエリ拡張時に元のクエリベクトルに付加する重み β

このうち統合重みが PRFS で新しく用いられるパラメータであり，これ以外は PRF で用いられるパラメータである．

本研究では，初めに，統合重みのパラメータについて Leave-one-out の交差検定を行い，次に，PRF に必要な各パラメータについて Leave-one-out の交差検定を行ってそれぞれのパラメータを推定し，その結果を用いた．その際，統合重みは 0.1 から 0.9 まで 0.1 刻みで変化させた．また，PRF に必要な各パラメータにおいては，擬似適合ドキュメント数は 1 件から 5 件までの 1 件刻み，関連語は 10 語から 50 語まで 10 語刻み，β は 1 から 10 まで 1 刻みで変化させて実験を行った．

5.2 PRFS の効果

まず，初期クエリおよび PRF による拡張後クエリ，PRFL による拡張後クエリそれぞれでの講演検索および種々の長さのパッセージ検索（60 発話単位，30 発話単位，15 発話単位，10 発話単位，5 発話単位）の検索結果について述べる．それぞれ 125 件のクエリで検索を行って 11 点平均精度を求め，全クエリでの平均値を求めた．検索結果を表 1（それぞれ Baseline 列，PRF 列，PRFL 列）に示す．講演検索タスクおよび全ての講演パッセージ検索タスクにおいて PRFS によって PRF や PRFL よりも検索精度が向上している．特に，講演検索や 60 発話単位パッセージの検索のような，長い講演音声ドキュメント検索において，効果が見られる．ただし，講演検索タスクでは Baseline よりも検索精度はまだ低い．

これらの結果は長い音声ドキュメント検索のタスクにおいて PRFS が有効であることを示している．

6. 関連語抽出アルゴリズムの変更と擬似適合性フィードバック

これまで，長い音声ドキュメント検索における擬似適合性フィードバックでは，初期検索の検索ドキュメントが長い場合に適切な関連語を抽出できないことを示した．そのために，初期検索では短いパッセージを対象に検索する方法を提案し，その効果を確認した．

表 1 PRFS の効果（11ptAP の比較）
Table 1 Effects of PRFSL (comparison of 11ptAP)

検索単位	Baseline	PRF	PRFL	PRFS
講演単位	0.531	0.470	0.489	0.514
60 発話単位	0.327	0.305	0.328	0.350
30 発話単位	0.260	0.271	0.270	0.289
15 発話単位	0.187	0.205	0.210	0.226
10 発話単位	0.149	0.175	(0.175)	0.184
5 発話単位	0.104	0.123	0.129	0.144

＊（　　）は通常の PRF と同じ

本章では，擬似適合性フィードバックにおける関連語抽出の精度向上を目的とした，別のアプローチを考える．具体的には，複数の擬似適合ドキュメントに共通に含まれる語のみを関連語として抽出を行う方法を提案する．これを，複数の擬似適合ドキュメントに含まれる関連語を用いた擬似適合性フィードバック（PRFC: Pseudo Relevance Feedback using Common terms）とよぶことにする．

6.1 PRFC の概要

従来の擬似適合性フィードバックでは，初期クエリから得られた複数の擬似適合ドキュメント集合をひとつのクエリとみなして，式（6）に基づいて関連語の抽出を行う．このため複数の話題から関連語が抽出され，適切にクエリ拡張が行えない．提案手法は，初期クエリから得られた各擬似適合ドキュメントがそれぞれクエリに合致する話題とその他の話題を含んでいることに着目し，各ドキュメントに共通する話題をクエリに合致する話題とみなしてそこから関連語の抽出を試みるものと言える．

PRFC が PRF と異なる点は，関連語の抽出が初期クエリから得られた擬似的な適合ドキュメントそれぞれから，式（6）に基づいて単語の抽出を行い，抽出された単語群の中から式（10）に基づいて各単語の正規化出現頻度度 q'_{t_k} を求め，この値の高い順に関連語を抽出する点のみである．

$$q'_{t_k} = \begin{cases} \sum_{i=1}^{n} q_{i,t_k} & \text{if } t_k \in \bigcap_{i=1}^{n} W_i \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

ここで，W_i は上位 i 件目の擬似適合ドキュメントに含まれる単語の集合，q_{i,t_k} は上位 i 件目の擬似適合ドキュメントに含まれる単語 t_k の正規化出現頻度（式（6））を表す．

6.2 PRFC の効果

PRFC による拡張後クエリでの講演検索および種々の長さのパッセージ検索（60 発話単位，30 発話単位，15 発話単位，10 発話単位，5 発話単位）の検索結果を評価した．結果（125 件のクエリそれぞれに対する 11 点平均精度の平均値）を表 2（PRFC 列）に示す．通常の PRF と比較して，講演単位，60 発話単位，30 発話単位ではそれぞれ検索精度

表 2 PRFC の効果（11ptAP の比較）
Table 2 Effects of PRFC (Comparison of 11ptAP)

検索単位	Baseline	PRF	PRFC
講演単位	0.531	0.470	0.481
60 発話単位	0.327	0.305	0.335
30 発話単位	0.260	0.271	0.277
15 発話単位	0.187	0.205	0.203
10 発話単位	0.149	0.175	0.166
5 発話単位	0.104	0.123	0.114

が 0.470 から 0.481，0.305 から 0.335，0.271 から 0.277 に向上した．15 発話単位，10 発話単位，5 発話単位ではそれぞれ検索精度が 0.205 から 0.203，0.175 から 0.166，0.123 から 0.114 に低下した．

PRFC は長い講演パッセージ検索において特に有効であることがわかった．この理由として，60 発話単位や 30 発話単位のドキュメントはやや長く[*3]，擬似適合ドキュメントにはクエリに関連する語が十分に含まれており，それぞれの擬似適合ドキュメントに共通して含まれる語を関連語として抽出したときにクエリに関連する語のみが適切に抽出できたと考えられる．15 発話単位，10 発話単位，5 発話単位では，それぞれの擬似適合ドキュメントに含まれる語がそもそも少なく，それぞれの擬似適合ドキュメントに共通して含まれる語を関連語として抽出したときに十分な量の関連語が抽出できなかったと考えられる．

7. 拡張前クエリと種々の PRF での拡張後クエリの併用

擬似適合性フィードバックによるクエリ拡張では，擬似適合ドキュメントから関連語を抽出する．クエリによってはこの擬似適合ドキュメントに真の適合ドキュメントが含まれず，検索精度の低下の原因となる．実際にこれまでに，クエリ拡張によって検索精度が向上するクエリはおよそ 30%から 70%程度であることを確認している [2][3].

この問題に対処する方法として，初期クエリ（拡張前クエリ）での検索結果と拡張後クエリでの検索結果とを併用する方法がある [2][3][11][12][13]．これはクエリ拡張による精度低下を拡張前クエリでの検索結果で補うことを目的とした手法である．

7.1 併用アルゴリズム

ドキュメント D_i と拡張前クエリ Q，通常の PRF による拡張後クエリ Q^{PRF}，PRFC による拡張後クエリ Q^{PRFC}，PRFL による拡張後クエリ Q^{PRFL}，および PRFS による拡張後クエリ Q^{PRFS} 間のそれぞれの類似度 $\mathrm{SMART}(Q, D_i)$，$\mathrm{SAMRT}(Q^{PRF}, D_i)$，$\mathrm{SAMRT}(Q^{PRFC}, D_i)$，$\mathrm{SAMRT}(Q^{PRFL}, D_i)$，

[*3] 単純に全講演の長さの合計をドキュメント数で割ると，それぞれ 1 ドキュメントあたり約 131 秒と約 68 秒

表 3 拡張前クエリと複数の拡張後クエリの併用の効果
Table 3 Effect of combination of original and expanded queries

検索単位	Baseline	Baseline+PRF+PRFC +PRFL+PRFS
講演単位	0.531	**0.551**
60 発話単位	0.327	**0.371**
30 発話単位	0.260	**0.303**
15 発話単位	0.187	**0.231**
10 発話単位	0.149	0.177
5 発話単位	0.104	0.147

太字は Baseline および各 PRF のみの時よりも高い値

$\mathrm{SAMRT}(Q^{PRFS}, D_i)$，を組み合わせて検索を行う．類似度の統合方法には線形対数補間を用いる．具体的には式（11）に基づいて類似度を統合し統合後の類似度 $\mathrm{SIM}(Q, D_i)$ を得る．λ_1，λ_2，λ_3，λ_4 はそれぞれ統合重みを表す．

$$\begin{aligned}\mathrm{SIM}(Q, D_i) = {} & (1-\lambda_1-\lambda_2-\lambda_3-\lambda_4)\log \mathrm{SMART}(Q, D_i) \\ & + \lambda_1 \log \mathrm{SMART}(Q^{PRF}, D_i) \\ & + \lambda_2 \log \mathrm{SMART}(Q^{PRFC}, D_i) \\ & + \lambda_3 \log \mathrm{SMART}(Q^{PRFL}, D_i) \\ & + \lambda_4 \log \mathrm{SMART}(Q^{PRFS}, D_i) \end{aligned} \quad (11)$$

こうして統合された類似度 $\mathrm{SIM}(Q, D_i)$，を用いてドキュメント D_i を類似度の降順に並べ替え最終的な検索結果とする．

7.2 併用の効果

統合重み（式（11）の λ_1，λ_2，λ_3，λ_4）を 0 から 1（ただし $0.1 \leqq \lambda_1+\lambda_2+\lambda_3+\lambda_4 \leqq 0.9$）まで 0.1 刻みで変化させ，125 件のクエリを用いて Leave-one-out の交差検定を行った．

拡張前クエリと複数の拡張後クエリの併用の結果を表 3 に示す．拡張前クエリと複数の拡張後クエリの併用によって検索精度が向上した．特に，講演検索や比較的長い講演パッセージ検索（60 発話単位，30 発話単位）では，拡張前クエリと PRF，PRFC，PRFL，PRFS での拡張後クエリとの併用が効果的であることがわかった．検索対象が短いパッセージ（10 発話や 5 発話）のときに統合の効果が小さかったのは，10 発話単位の検索では PRF と PRFL が同じものであり多様性が減ったことおよび，PRFC が短い講演パッセージ検索ではあまり効果的でないことに起因すると考えられる．

8. まとめ

CSJ の講演を対象とした音声ドキュメント検索において種々の擬似適合性フィードバック（PRF）の研究を行った．

これまでに提案している講演ドキュメント検索のための擬似適合性フィードバック（PRFL），すなわち，初期検索を短いパッセージを対象として行って関連語抽出を行ってPRF を行う方法，の初期検索精度が低いという問題点を明らかにし，この問題に対応した．具体的には，広域文書と局所文書の類似度統合に基づく講演検索手法を初期検索（講演パッセージ検索）に適用して初期検索精度を高めて，擬似適合性フィードバックを行う手法（PRFS）を提案した．さらに，講演音声ドキュメント検索の PRF のための関連語抽出方法（PRFC）の提案と，種々の PRF による拡張後クエリと拡張前クエリの併用を行った．

CSJ の講演音声を対象とした講演検索および講演パッセージ検索において，平均的な検索精度（11 点平均精度）の向上が確認できた．

謝辞

本研究は科研費（24500225）の助成を受けて行われたものである．

参考文献

[1] 相川清明，秋葉友良，伊藤慶明，河原達也，中川聖一，南條浩輝，西崎博光，　胡新輝，松井知子，山下洋一：音声ドキュメント処理ワーキンググループ活動報告，情報処理学会研究報告，2011-SLP-89-4 (2011).

[2] 西尾友宏，南條浩輝，吉見毅彦：講演音声ドキュメント検索のための擬似適合性フィードバック，情報処理学会研究報告，SLP-96-3 (2013).

[3] 西尾友宏，南條浩輝，吉見毅彦：講演音声ドキュメント検索のための擬似適合性フィードバック，情報処理学会論文誌，Vol. 55, No. 5 (2014（採録決定）).

[4] 南條浩輝，弥永裕介，吉見毅彦：広域文書類似度と局所文書類似度を用いた講演音声ドキュメント検索，情報処理学会論文誌，Vol. 53, No. 6, pp. 1654–1662 (2012).

[5] 前川喜久雄：言語研究における自発音声，日本音響学会研究発表会講演論文集（春季），pp. 19–22 (2001).

[6] 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. 501–513 (2009).

[7] 北　研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版株式会社，ISBN4-320-12036-1 (2002).

[8] 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結合処理とその評価，情報処理学会論文誌，Vol. 44, No. SIG 8（TOD 18），pp. 78–91 (2003).

[9] 重安幸治，南條浩輝，吉見毅彦：日本語講演音声ドキュメント検索における索引付けの検討，情報処理学会研究報告，2009-SLP-76-8 (2009).

[10] 真野博子，伊藤秀夫，小川泰嗣：文書検索におけるランキング検索技術，*Recoh technical report(29)*, pp. 21–30 (2003).

[11] Sheldon, D., Shokouhi, M., Szummer, M. and Craswell, N.: LambdaMerge: merging the results of query reformulations., *WSDM'11*, pp. 795–804 (2011).

[12] Aslam, J. A. and Montague, M.: Models for metasearch, *Proc. the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 276–284 (2001).

[13] Montague, M. and Aslam, J. A.: Metasearch consistency, *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 386–387 (2001).