論文 / 著書情報
Article / Book Information

論題(和文)	連続音素認識を用いた単語認識誤りに頑健な講演音声検索
Title(English)	
著者(和文)	岩田 憲治, 篠田 浩一, 古井貞熙
Authors(English)	Kenji Iwata, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2008年春季講演論文集, Vol. , No. 2-10-20, pp. 113-116
Citation(English)	, Vol. , No. 2-10-20, pp. 113-116
発行日 / Pub. date	2008, 3

連続音素認識を用いた単語認識誤りに頑健な講演音声検索*

◎岩田憲治, 篠田浩一, 古井貞熙 (東工大)

1 はじめに

本研究では講演音声の検索をする際に発生する，単語認識誤りによる検出不能の問題を解決するために音素認識を利用する．音素認識は西崎らの研究 [1] に挙げられるように主に未知語の認識に用いられる場合が多かったが，未知語だけでなく多少の認識誤りを許容することで，単語認識では検出が不可能な区間を検出することが可能であると考えられる．また，音素認識と単語認識の検索結果を融合することで，単語認識のみを用いた検索結果よりも精度の良い結果が得られることが期待できる．

音素認識を利用した音声検索を行っている研究として Yu らの研究 [2] がある．そこでは，音素認識結果からキーワードの検索を行うために，音素認識結果をグラフで表現し，その音素グラフとキーワードの音素列との DP マッチングを行っている [3]．この手法の欠点として，不必要な区間を検出する湧きだし誤りが非常に多くなってしまうことが挙げられる．そこで本研究では，湧きだし誤りを抑える音素認識検索手法を提案し，音素検索結果と単語検索結果を融合することで検索性能の向上を目指す．

2 音素認識を用いたキーワード検索

ここでは音素認識結果からキーワードの検索を行う従来手法 [3] の詳細を述べる．この手法は音素グラフのあるフレーム区間とキーワードの音素列を DP マッチングすることにより，そのフレーム区間におけるキーワードの信頼度を求めるものである．

まず音素グラフのアークの集合 L のうち，1 つのアークを $l=(b,e,p,s)$ と表す．p は認識された音素，b はそのアークの開始フレーム，e は終了フレーム，そして s は認識器によって得られた対数の底が 10 の音響対数尤度である．次に以下の関数を定義する．

$$V(b,e,p)=\begin{cases} s & \text{if } (b,e,p,s)\in L \\ -\infty & \text{otherwise} \end{cases} \tag{1}$$

この関数を利用して音素グラフとキーワードの DP マッチングを行う．まず $C(i,t)$ という関数を用意する．これは，キーワードの i 番目までの音素と音素グラフの DP マッチングを行った結果，フレーム t で終了したときの最大のスコアを返す関数である．この関数を以下のような再帰的な表現で計算することにより，フレームベースの DP マッチングを行う．

$$\forall t\ C(0,t) = 0 \tag{2}$$

$$C(i,e) = \max_b \begin{cases} C(i-1,b)+V(b,e,p_i) \\ C(i-1,b)+(e-b)P_s \\ \quad +\max_z V(b,e,z) \\ C(i,b)+(e-b)P_i \\ C(i-1,e)+P_d \end{cases} \tag{3}$$

このとき P_s，P_i，P_d はそれぞれ置換ペナルティ，挿入ペナルティ，削除ペナルティである．キーワードの音素数が N のとき，あるフレーム t において $C(N,t)$ を求め，またそのスコアが最大となった際使用した $C(0,s)$ のフレーム s を記憶しておくことで，開始フレーム s，終了フレーム t の候補が得られる．

次に，得られた候補の信頼度を求める．開始フレーム b，終了フレーム e，キーワードの音素数 N である検索候補 I の信頼度 P_I は

$$P_I=\frac{10^{C(N,e)}}{\mathrm{ML}(b,e)} \tag{4}$$

で求めることができる．このとき $\mathrm{ML}(b,e)$ は音素グラフ全体での最尤パスのうちフレーム b から e までの尤度を返す関数である．この信頼度により候補の優劣をつけ，検索結果を信頼度の大きいものから順に出力する．

3 湧きだし誤りを抑える音素認識検索手法

3.1 KLD による置換ペナルティの付与

従来手法では，置換ペナルティにおいて置換する音素組を制限して計算量と誤りを減らしていたが，置換する音素組それぞれに対して異なるペナルティを与えることは行っていなかった．本研究では音素間 KLD(Kullback-Leibler Divergence)[4] を利用し，置換音素組によって置換ペナルティを変更する．2 つの GMMs，$\tilde{s}$ があったとき，s から $\tilde{s}$ への KLD は

$$D_{\mathrm{KL}}(s||\tilde{s}) \approx \frac{1}{2N}\sum_{m=1}^{M}\omega_m\sum_{k=1}^{2N}\log\frac{p(\boldsymbol{o}_{m,k}|s)}{p(\boldsymbol{o}_{m,k}|\tilde{s})} \tag{5}$$

に近似される．ただし N は音響特徴量の次元数，M は混合数，ω_m は GMM における m 番目の Gaussian kernel の混合重み，$\boldsymbol{o}_{m,k}(1\le k\le 2N)$ は m 番目の Gaussian kernel の k 番目の sigma point とす

*Robust lecture retrieval against word recognition errors using continuous phoneme recognition
By Kenji Iwata, Koichi Shinoda and Sadaoki Furui (Tokyo Institute of Technology)

る．sigma point には，$o_{m,k} = \mu_m + \sqrt{N\lambda_{m,k}}u_{m,k}$，$o_{m,k+N} = \mu_m - \sqrt{N\lambda_{m,k}}u_{m,k}(1 \leq k \leq N)$ を選んだ．ここで μ_m は m 番目の Gaussian kernel の平均ベクトル，$\lambda_{m,k}$，$u_{m,k}$ は m 番目の Gaussian kernel の共分散対角行列における k 番目の固有値と固有ベクトルである．KLD は非対称な距離である．

式 (5) により全ての GMM 間の KLD を算出し，動的計画法を用いて 2 つの音素間の KLD を近似的に求める．この音素間 KLD を置換ペナルティに乗ずることで，音響的特徴が類似している音素組はペナルティが軽く，類似していない音素組はペナルティが重く付与される．

3.2 アーク同士の繋がりを考慮した検索手法

従来手法のアルゴリズムはフレームベースの DP マッチングのため，DP マッチングを行う際にアークの終端と他のアークの始端のフレームが一致していれば繋げて先に進めていた．本研究では，アーク同士が 1 つのノードで繋がっているものと区別するため，アーク同士が 1 つのノードで繋がっていない候補にペナルティを与える．

Insertion が起きた際もこの点を考慮したアルゴリズムにする．計算量削減のために Insertion は 1 本のアークの時間内で起こると仮定し，Insertion が起きた区間の前後のアーク同士が 1 本のアークで繋がっていなければ，それは 1 本のパスとして繋がっていないと判断しペナルティを与えることとした．

この手法を従来のアルゴリズムに追加する際，今までのフレームベースでの DP マッチングではアーク同士が繋がっているかを判断するのが困難である．そこでノードベースの DP マッチングを行うようにアルゴリズムを変更する．このアルゴリズムに，置換する音素組によって置換ペナルティの大きさを変更する機能を加えることも可能である．

3.3 フレーム区間の類似した候補の削減

ここまで述べてきたアルゴリズムは，非常に多くの候補を検出してしまう．そこで，類似したフレーム区間を持つ候補は一番高い信頼度を持つ候補だけ残し，他は削除することで候補の削減を行う．このとき候補同士が似たフレーム区間を持っているかを測る指標が必要であるが，それは 2 つの候補 A，B が

$$\frac{\min(e_A, e_B) - \max(b_A, b_B)}{\max(e_A, e_B) - \min(b_A, b_B)} > \tau_d \qquad (6)$$

という条件を満たすとき，A と B は似たフレーム区間を持つ候補であると定義する．b_A，e_A が候補 A の開始フレームと終了フレーム，b_B，e_B が候補 B の開始フレームと終了フレームである．τ_d は閾値で，その値を小さくすることで多くの候補が削減される．

4 単語認識を利用した検索結果との融合

4.1 単語認識結果からの検索手法

単語認識を用いた検索では，まず音声に対し 1-best の単語認識を行う．このときアラインメントも同時に行う．検索の際には，1-best の認識結果にキーワードが存在した場合，アラインメントにより得られたフレーム区間が検出区間となる候補が得られる．

候補の信頼度の計算にはフレーム平均尤度を用いる．得られた候補のうち最尤のフレーム平均尤度を $l_{\max}$，信頼度を測る対象となる候補 I のフレーム平均尤度を l_I とすると，その候補の信頼度 W_I は

$$W_I = \frac{l_I}{l_{\max}} \qquad (7)$$

で得ることができる．

4.2 単語，音素の結果を融合する手法

単語認識，音素認識を用いた検索により得られたそれぞれの信頼度を用いて結果の融合を行う．単語検索で得られた候補 I と音素検索で得られた候補 J が

$$\frac{\min(e_I, e_J) - \max(b_I, b_J)}{\max(e_I, e_J) - \min(b_I, b_J)} > \tau_m \qquad (8)$$

という条件を満たすとき，今までの候補 I，J を削除し融合候補 K を生成する．融合候補 K の検出区間は I と J の区間の平均をとる．信頼度 M_K は

$$M_K = {W_I}^{\lambda} \cdot {P_J}^{(1-\lambda)} \qquad (9)$$

で与えられる．τ_m は閾値で，値を小さくすると融合する候補の組が多くなる．λ は融合重みであり，$0 \leq \lambda \leq 1$ を満たす．融合する相手が見つからなかった場合は，その相手の信頼度を非常に小さい値 D として上式にて融合し，候補を生成した．

5 実験

5.1 実験条件

講演音声データベースとして，CSJ(Corpus of Spontaneous Japanese)[5] の学会講演音声と模擬講演音声を使用した．講演数は 2701，総講演時間は約 530 時間である．このデータを話者性がオープン，かつ同じ種類の講演が均等に分かれるように 2 つに分割し，交差検定 (cross-validation) を行った．音響モデルは triphone，言語モデルは音素認識では単純な文法，単語認識では trigram を使用した．認識デコーダは音素認識は HTK，単語認識は Julius を使用した．

キーワードの決定には tf-idf を利用した．tf-idf の計算をデータベース内の全単語について行い，その tf-idf の値が大きかった単語をキーワードとした．た

だし結果の信頼性を高めるため，データベース全体で100回以上存在している単語のみを対象とし，また助詞，助動詞，フィラー，感動詞といった実際にキーワードとして用いられる可能性の低い単語も対象外とした．

キーワードの正解発声区間はmonophoneによる強制切り出しにより作成した．評価に使用する候補群は交差検定により認識，検索を行い得られた候補のうち，評価セットの講演の検索結果である候補を選択し作成した．評価の際は候補のフレーム区間と正解発声区間が一部でも合致していれば検索成功とした．評価の尺度にはMAPとPrecision，Recallを用いる．MAPとはMean Average Precisionの略で，各キーワードにおいてAverage Precisionを求め，その平均をとることで計算される．Average Precisionは検索された各正解の順位における精度の和を全正解数で割ったものであり，式で表すと

$$\mathrm{AP}_i = \frac{1}{N_i} \sum_{k=1}^{M_i} \frac{k}{r_{i,k}} \tag{10}$$

となる．ここでN_iは単語iの正解数，M_iは正解検出数，$r_{i,k}$は単語iのk個目の正解が出現した順位である．PrecisionとRecallの評価の際には，信頼度が閾値τ_fより大きい候補を使用した．τ_fを調節することで評価に使用する候補の数が増減し，PrecisionとRecallの値が変化する．

5.2 実験結果

5.2.1 音素認識を用いた検索結果

比較的小規模な評価データを用いて，提案した音素検索手法の評価を行った．評価セットはCSJのテストセット31講演で，キーワードはtf-idf上位100単語である．そして置換ペナルティを置換音素組によって変化させる手法(KLD)，アーク同士が繋がっていない候補にペナルティを与える手法(ArcPenalty)，そのどちらも使用する手法(KLD+Arc)，どちらも使用しない手法(Base)の4つの手法の比較を行った．どの手法においてもフレーム区間が類似した候補は削除している．PrecisionとRecall評価のための閾値を$\tau_f = 10^{-16}$とした．表1に結果を示す．この結果からArcPenaltyを使用することで性能の大幅な向上が見られるが，KLDを使用することで若干性能が低下してしまっている．この原因として，生成した音素グラフの規模が十分以上に大きく，正解音素がほとんど出現していたため，KLDが効果のある事例が少なかったと考えられる．ArcPenaltyはCorrectのみでマッチングする順位が高い候補のうち，誤りであったものの順位を効果的に落とすことができたため，性能向上につながったと考えられる．また単語認識を用いた検索結果と比較すると，Recallは高く単語認識では検出できないような候補を音素認識で検出できているが，Precisionが非常に低くなってしまっており，湧きだし誤りを抑えるには更なる改善の余地があるといえる．

Table 1　音素認識を用いた検索結果

手法	MAP	Precision(%)	Recall(%)
Base	0.176	0.81	66.9
KLD	0.174	0.81	66.8
ArcPenalty	0.365	1.11	73.8
KLD+Arc	0.364	1.11	73.8
単語認識	0.546	82.5	66.1

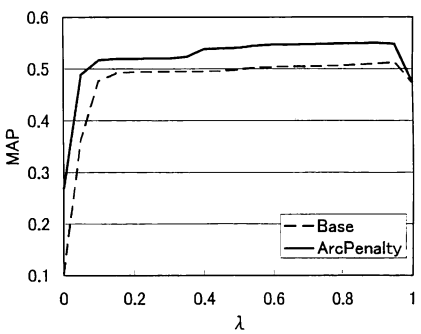

Fig. 1　音素検索と単語検索の融合結果

5.2.2 音素検索と単語検索の融合結果

音素検索結果と単語検索結果を融合した結果の評価を行う．評価セットはCSJ全体の2701講演で，キーワードはtf-idf上位1000単語である．融合する際に使用するパラメータの値は$D = 10^{-1000}$，$\tau_m = 0.3$とした．図1に融合重みλを変化させたときのMAPの変化を示す．点線が音素検索手法にBaseを使用したもの，実線がArcPenaltyを使用したものである．結果から，音素検索結果と単語検索結果を融合することによりMAPの向上が見られる．最もMAPが高いのはArcPenaltyを用いた音素検索結果を$\lambda = 0.9$で融合したときで，単語検索結果の0.475より高い0.550となった．

最もMAPが高くなった条件で，閾値τ_fを変化させて評価する候補数を変化させたときのPrecision-Recall曲線を描いたのが図2である．実線が単語検索結果のみ，点線が音素検索結果と単語検索結果を融合したものである．結果から，50%程度のRecallの場合は融合した方がPrecisionが高くなっている．これは融合したことにより，単語検索結果の候補の順位を効果的に入れ換えることができたためだと考えられる．

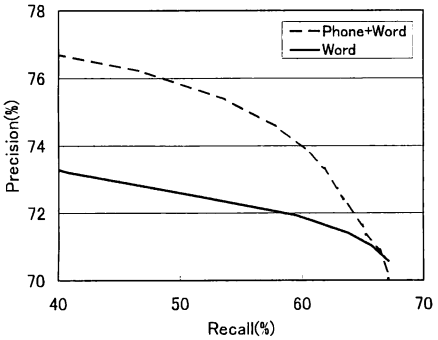

Fig. 2　融合した検索結果の Precision-Recall 曲線

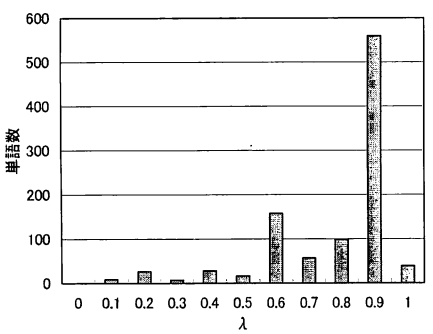

Fig. 3　融合重み λ の値と単語数の関係

5.2.3　単語ごとの評価

融合した結果について単語ごとに評価を行う．単語ごとに MAP が最大となる融合重み λ を求める．λ は 0 から 1 まで 0.1 ずつ変化させて検証を行った．それぞれの λ を使用する単語数を示したのが図 3 である．大半の単語は単語検索結果を重視すると良い結果が得られるが，音素検索結果を単語検索結果と同等またはそれ以上に重視すると良い単語も存在する．融合重みを単語ごとに振り分けると効果があることが分かる．単語ごとに最適な λ を使用したところ，MAP は若干ではあるが 0.555 に向上した．

次に単語ごとに最適な融合重みにした上で，閾値 τ_f を単語ごとに変化させて F 値が最大となる値を求めた．それぞれの閾値の変域を使用する単語数を示したのが図 4 である．横軸は τ_f を λ で割り，底が 10 の対数をとったものとなっている．融合する際相手が存在しない音素検索候補の信頼度には $D^{\lambda} = 10^{-1000\lambda}$ を乗じているので，相手が存在しない音素検索候補を使用すると結果が良くなる単語は，横軸の計算を上で述べたようにすることによりほぼ-901 以下の値となる閾値を持つ．結果から F 値が最大となるのは，閾値を大きくすることで上位の候補のみを使用したときか，小さくすることで音素検索結果にのみ出てくる候補も利用したときのどちらかに分かれたといえる．単語ごとに最適な λ と τ_f を使用したところ，F 値は単語検索結果の 68.8 より若干高い 69.1 となった．

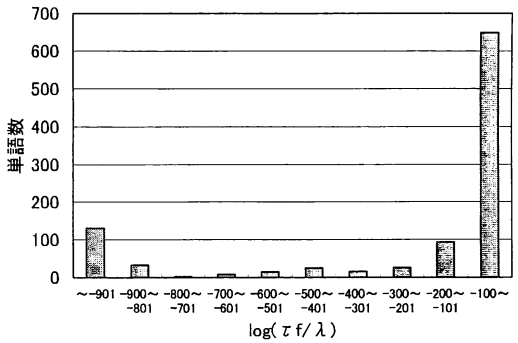

Fig. 4　信頼度の閾値 τ_f と単語数の関係

6　おわりに

本研究では認識誤りをある程度補償可能な音素グラフ形式を利用し，キーワードの音素列との DP マッチングにより検索を行った．その際湧きだし誤りを削減するため，グラフのアークの繋がりを考慮してマッチングを行うなどの手法を提案した．そして単語検索結果との融合を行い，Recall を 50%に固定した場合，単語検索結果に比べ 3.2 ポイント Precision を向上することができた．また単語ごとの最適な融合重みや閾値について考察を行った．

今後の課題としてはさらなる性能の向上のため，音素言語モデルを利用した検索を行うことが考えられる．またアルゴリズムの速度向上のため，グラフ表現をコンパクトにまとめる Confusion Network[6] を用いる手法を検討する必要がある．単語ごとに最適な融合重みや閾値を自動的に推定する手法が実現が可能であるかについても検討していきたい．

参考文献

[1] 西崎 他, “音声認識誤りと未知語に頑健な音声文書検索手法”, 電子情報通信学会論文誌, J86-D-II, pp.1369-1381, 2003.

[2] P.Yu *et al.*, “Vocabulary-Independet Indexing of Spontaneous Speech,” IEEE Transactions on Speech and Audio Processing, pp.635-643, 2005.

[3] D.A.James, S.J.Young, “A Fast Lattice-Based Approach to Vocabulary Independent Wordspotting,” ICASSP, pp.377-380, 1994.

[4] P.Liu *et al.*, “Divergence-based similarity measure for spoken document retrieval”, ICASSP, pp. IV-89-IV-92, 2007.

[5] K.Maekawa *et al.*, “Spontaneous Speech Corpus of Japanese,” LREC, pp.947-952, 2000.

[6] L.Mangu *et al.*, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” Computer Speech and Language, 14, pp.373-400, 2000.