Vol.2014-SLP-102 No.10
2014/7/26

大語彙連続音声認識と音節 N-best 音声認識を用いた Spoken Term Detection の高精度化

長野 徹[1,a)]　倉田 岳人[1]　鈴木 雅之[1]　立花 隆輝[1]　西村 雅史[1,†1,b)]

概要：企業のコールセンターでは，音声通話に含まれる特定のキーワードをチェックするコールモニタリング業務によりコールセンターの品質向上を図っている．一部のコールセンターでは，大語彙連続音声認識技術の利用により日々大量に蓄積される音声データに対するキーワード検索が可能となってきた．ここでは，検索キーワードや業務内容に応じて，再現率を重視したい，適合率を重視したいといった要望がある．本報告では，認識単位の異なる二種類の音声認識システムを用いることで，単にキーワード検出区間を出力するだけではなく各検出区間に対して信頼度のスコアを与え，検索時に再現率・適合率のバランスを調整できるシステムを提案する．提案法では，大語彙連続音声認識を用いて検索キーワード文字列に一致する区間をキーワード検出区間候補として抽出し，それら検出区間に含まれる音節音声認識の N-best 出力と検索キーワード音節列とを比較することで，各検出区間をスコアリングする．実験では認識尤度によるスコアリングを用いた結果との比較を行い，本手法の有効性を示した．

キーワード：音声認識，音声検索語検出，音節音声認識

Improvement of Spoken Term Detection by Combining LVCSR and Syllable-based N-best Speech Recognition Results

Tohru Nagano[1,a)]　Gakuto Kurata[1]　Masayuki Suzuki[1]　Ryuki Tachibana[1]
Masafumi Nishimura[1,†1,b)]

Abstract: In contact centers, it is common to check the call conversations of the call agents with the customers for quality monitoring. Recently, more and more companies have come to use Automatic Speech Recognition (ASR) in call quality monitoring to enable exhaustive search in the calls. Preferences on the search results varies according to the demands; sometimes high recall rates are preferred, and at other times, high precision rates are preferred. Hence, in this paper, we propose a method that not only finds occurrences in the speech data of given search terms, but also gives confidence scores for the found occurrences by combining the recognition results of a word-based Large Vocabulary Continuous Speech Recognition (LVCSR) system and a syllable-based speech recognition system. While the former system is used for finding candidates, the latter system is used for calculating the confidence scores based on the N-best hypotheses. We present experimental results in which the proposed system outperformed a conventional method based on acoustic likelihood in performance.

Keywords: speech recognition, spoken term detection, syllable based speech recognition

1. はじめに

音声認識の高精度化に伴い，様々な場面で音声認識が用いられるようになった．一般的なスマートフォン等に用い

[1] 日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research–Tokyo, IBM Japan Ltd., Toyosu, Koto, 135–8511, Japan
[†1] 現在，国立大学法人 静岡大学 大学院情報学研究科
Presently with Graduate School of Infomatics, Shizuoka University, Hamamatsu, Shizuoka, 432–8011, Japan
[a)] tohru3@jp.ibm.com
[b)] nisimura@inf.shizuoka.ac.jp

られる音声インターフェースとしてだけではなく，企業のバックエンドシステムにおいても音声認識が用いられるようになってきている．例えば，企業内で音声が集約されるコールセンター業務においても音声認識技術が用いられている．コールセンターにおけるコールモニタリング業務では，大量の音声通話の中から特定の単語や不適切な発言等 (以下，単に「キーワード」) をチェックすることで，コールセンターの品質向上やコミュニケータ (オペレータ・販売営業員) の評価を行っている．従来は対象音声データをサンプリングし，エキスパートによる音声聞き取りが中心であったが，近年音声認識システムを用いたコールモニタリングが実用化されており，全通話を対象にモニタリングを行うことができるようになった．音声認識により網羅的に大量の音声からキーワードを検出できるようになった一方，音声認識結果にはある程度の誤りが含まれる．どの程度の誤りが許容されるかは，業務内容によって異なるが，具体的には，検出結果の再現率重視 (音声認識誤りによる過検出を許容するができるだけ漏れなく検出したい)，または適合率重視 (できる限り正確に認識されているもののみを検出したい) といった要求がある．そのため，キーワード検出の性能を調整できる仕組みが望まれている．ただ，大量のデータに対してパラメータを調整しつつ再び音声認識処理を行うことは運用上困難なため，計算量にも考慮した方法が必要である．

「単語を認識単位とする連続音声認識 (以下，単語音声認識)」の結果に対して文字列検索を行うことによりキーワード検出を行う方法を用いる場合，音声認識のパラメータの変更，単語の出現確率を変える等の操作を行えば，ある程度，再現率/適合率の調整が可能であるが，あまり実用的ではない．一方，「音素もしくは音節など単語より短い単位を認識単位とする音声認識 (以下，音節音声認識)」の結果に対してマッチングを行うことによりキーワード検出を行う方法を用いる場合，マッチングの条件を変更することで再現率/適合率の調整が可能である．しかし，単語音声認識を用いる手法と比較すると，言語情報が十分に利用できないため，そもそもキーワード検出性能が低いことが知られている [1]．また，単語・音節音声認識どちらを用いる場合でも，N-best を用いることで適合率は下がるが再現率を高めることができる．一方で，単語音声認識の 1-best のみを用いる場合より，適合率を高めることは難しかった．

そこで本論文では，音声検索アプリケーションの利便性を高めるため，「計算量を大幅に増やさず」「検索性能を落とさず」「適合率を向上させる」ことを目的として，単語音声認識結果と音節 N-best 音声認識結果を組み合わせた音声ドキュメント検索システムを提案する．具体的には，単語音声認識を用いて文字列の一致する区間をキーワード検出区間候補として取り出し，それら検出区間に対応する音節 N-best 音声認識結果と検索キーワード音節列とを比較することで各区間候補に信頼度を与える．この信頼度を利用することでユーザーは適合率の高い音声のみをチェックといった作業が可能になる．

2. 先行研究

音声検索語検出は，一般的に単語を認識単位とする大語彙連続音声認識の書き起こし結果に対して文字列マッチングまたは単語列マッチングを行うが，未知語や認識誤りへの対応として，音素・音節またはそれよりも大きな単位での音声認識結果に対して単位列でのマッチングを行う方法 [2][3] が知られている．これら両者を用いた音声認識結果の利用としては，複数の音声認識システムを用いたシステムコンビネーションによるキーワード検出法 [4][5] などが提案されている．単語音声認識と音節音声認識を，単語音声認識辞書に含まれる語と未知語で切り分けて用いる方法 [1] 等も提案されており，システム全体として高いキーワード検出性能を実現している．また，信頼性尺度という観点からは，単語グラフ事後確率 [6]，N-best[7]，音響尤度，言語尤度を利用した方法など様々な指標が提案され，その有効性が検証されている [8][9]．一方，情報抽出アプリケーションの実用性向上という観点で，単語の文脈一貫性と音響尤度を利用して，信頼度の低い音声ドキュメントを棄却する研究 [10] も行われている．

3. 大語彙連続音声認識と音節 N-best 音声認識を用いたキーワード検索

本論文では辞書に含まれる語のみを検索対象にし，未知語を対象にしない．先行研究をふまえ，本論文でも単語音声認識と音節音声認識を組み合わせて用いるが，未知語に対して音節認識結果を適用するのではなく，辞書に含まれる語に対する検索結果へのフィルタリングに音節音声認識結果を用いる．辞書に含まれる語の検索性能に関しては，一般的に単語音声認識のほうが音節音声認識より優れており，一致区間候補の出力は単語音声認識を用い，区間の評価のみに音節音声認識 N-best を用いる．

3.1 音声認識システム

同一のモデル構築用の音声コーパスから単位の異なる 2 種類の音声認識システムを構築する．

$\mathcal{M}$ 単語を認識単位とした音声認識システム．単語は漢数字を含む漢字と平仮名，片仮名，アルファベットから構成される．
例: あの/大丈夫/ です / よ

$\mathcal{S}$ 音節を認識単位とした音声認識システム．音節は日本語の音節 451 種類から構成される．一般的な日本語音節に対し，長音化した母音，促音を含む音節は別の音節として取り扱う．
例: a / no / da / i / jo: / bu / de / su / yo

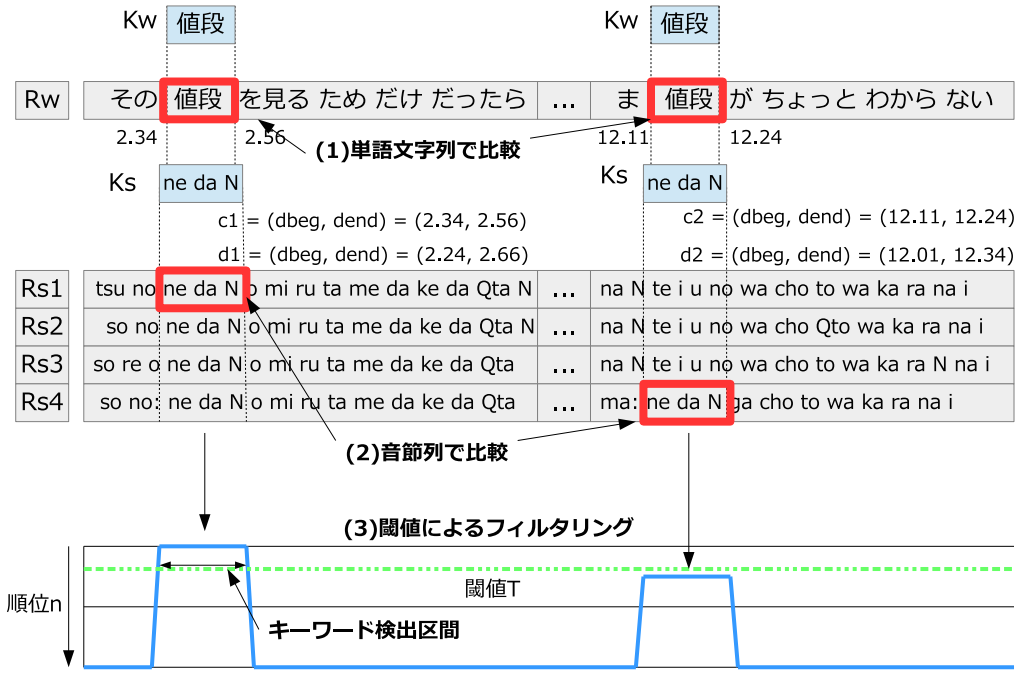

図 1　大語彙連続音声認識と音節 N-best 音声認識を用いたマッチング

検索対象音声を単語単位システム $\mathcal{W}$ および音節単位システム $\mathcal{S}$ を用いて音声認識を行う．システム $\mathcal{W}$ の認識結果は 1-best のみの出力として R_w を，システム $\mathcal{S}$ の認識結果としては N-best: $R_s = R_{s_1} \dots R_{s_N}$ を出力しておく．ここで N-best は検索対象音声から音声認識システムを用いて得られる N 個の仮説の集合である．N は得られる仮説の個数の最大値とし，N-best 出力は認識尤度の降順にソートされているものとする．また，検索キーワードは，文字列 K_w(例:「値段」) で与えられ，辞書または G2P 変換により音節列 K_s(例: ne/da/n) に変換されキーワード検出装置に入力される．各認識結果は，単語アラインメント結果 (開始時間 c_{beg}，終了時間 c_{end}) の列からなり，これらを元に検索用のインデックスを作成しておく．

3.2　キーワード検出手順

検索キーワードを与え，以下の手順で検索にマッチする音声区間を検出する (図 1)．

(1) キーワード文字列 K_w と単語単位システム $\mathcal{W}$ の認識結果 R_w とを比較し，一致する区間を検出する．一致する区間のタイムスタンプは検索用インデックスに含まれる単語アラインメント結果 (c_{beg}, c_{end} の組) から得られる．このタイムスタンプを単語アラインメントの誤差を考慮してわずかな時間 δ だけ広げた値を d_i とし，タイムスタンプ (開始時間 d_{beg}，終了時間 d_{end}) とする．最終的に一致した検出区間候補リスト $D = d_1 \dots d_M (1 \leq i \leq M)$ を得る．

(2) 単語音声認識の検出区間候補 d_i に含まれる音節単位システム $\mathcal{S}$ の認識結果 Rs とキーワード音節列 Ks を比較し，検出区間候補 d_i の信頼度を評価する．
検出区間候補 d_i の評価は音節音声認識 N-best の結果 $R_{s1} \dots R_{sN}$ とキーワード K_s の一致する最小の順位 $n (1 \leq n \leq N)$ を用いる．

(3) 検索時には閾値 T を与え，$n \leq T$ となる検出区間のみを最終的な検出区間とする．

4.　実験

4.1　検索評価用データ

電話録音データを用いて検索評価を行った．各音声ファイルは 8Khz / 16bit sampling で録音され，二話者の音声は予め別チャンネルで保存されているステレオ音声データである．データ量を表 1 に示す．発話時間はパワーヒストグラムを用いた発話区間検出器によって計算した．音声には人手による書き起こしによる正解が付与されており，音声認識に用いられる言語モデルを用いて単語分割されている．

4.2　検索キーワードおよび評価方法

検索キーワードは長さ 1 ～ 10 文字 (2 ～ 11 音節) からなる 38 種類で構成される．各キーワードは予め文字列に対する音節列を与えてある．検索評価用データにはのべ 3248

表 1　検索評価用データ

通話数	100 通話
録音時間	29.97 時間
発話時間	13.57 時間
発話区間数	21853 セグメント
単語数	179K 語

個の検索キーワードが含まれている．また検索キーワードに未知語 (OOV:Out-Of-Vocabulary) は含まない．検索キーワードの例を表 2 に示す．

表 2　検索キーワード例

表記	音節表記
値段	ne da n
東京	to: kyo:
おはようございます	o ha yo: go za i ma su
よろしくお願いします	yo ro shi ku o ne ga i shi ma su

評価は再現率，適合率，およびこれら 2 つを組み合わせた検索性能を示す F 値 (式 1) により行った．

$$\text{再現率} = \frac{\text{検索された正解キーワード数}}{\text{検索評価用データ中のキーワード数}}$$

$$\text{適合率} = \frac{\text{検索された正解キーワード数}}{\text{一致キーワードリストのキーワード数}}$$

$$\text{F 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}} \quad (1)$$

キーワードがマッチしたかどうかの判定は，検索評価用データに付与された書き起こし文字列および，音節列との評価にて行うが，マッチングの判定は発話単位で行う．つまり検出されたキーワードが発話に含まれているかどうかで判定する．1 発話区間の平均の長さは 2.24 秒である．

4.3　音声認識

予め構築された音声認識システムを用いて検索評価用データを単語列および音節列に変換しておく．音声認識モデル構築用データは同様の電話会話を元に生成されており，GMM-HMM を boosted MMI で識別学習した音響モデルである．単語単位音声認識の言語モデルは単語 trigram，音節単位音声認識の言語モデルは音節 trigram を用いて構築してある．単語 1-best による検索評価用データの音声認識率は，文字誤り率で 20.4% である．また音節認識モデルを用いた出力として，N を最大 $N = 1000$ とした N-best 出力を出力しておく．また比較のため単語 N-best についても出力しておく．

4.4　実験結果

比較対象として，単語音声認識 N-best と音節音声認識 N-best 単体での評価を行った．それぞれ，単語音声認識の N 位以内 (N-best) に検索キーワード Kw が含まれている場合，音節音声認識の N 位以内に検索キーワード Ks が含まれている場合，マッチしたと判定した．

$\boldsymbol{W}_{(N)}$　単語音声認識 N-best

$\boldsymbol{S}_{(N)}$　音節音声認識 N-best

$\boldsymbol{C}_{(N)}$　単語･音節音声認識組み合わせ:$\boldsymbol{C}_{(N)} = W_{(1)} \cap S_{(N)}$.
単語認識結果が検索キーワード文字列にマッチし，かつ音節認識結果 N-best が検索キーワード音節列にマッチした場合 (提案手法).

表 3　再現率および適合率

	$N=1$			$N=1000$		
モデル	再現率	適合率	F 値	再現率	適合率	F 値
$\boldsymbol{W}$	0.808	0.770	0.789	0.921	0.330	0.486
$\boldsymbol{S}$	0.626	0.425	0.506	0.810	0.169	0.279
$\boldsymbol{C}$	**0.592**	**0.902**	**0.715**	**0.720**	**0.824**	**0.768**

表 3 に単語音声認識 $\boldsymbol{W}$，音節音声認識 $\boldsymbol{S}$，提案手法 $\boldsymbol{C}$，のそれぞれ 1-best と 1000-best を閾値とした場合の再現率/適合率/F 値を示す．$\boldsymbol{W}$ と $\boldsymbol{S}$ を比較すると，従来研究と同様，単語音声認識の単語検出性能は音節単位での音声認識によるキーワード検出性能に比べ高い．$\boldsymbol{W}$ に関しては $\boldsymbol{W}_{(1000)}$ は $\boldsymbol{W}_{(1)}$ に比べて，再現率は 0.113 ポイント上昇したが，適合率は 0.303 ポイント下がり結果として検索性能 F 値は大きく下がっている．提案法の $N = 1, N = 1000$ における性能を $\boldsymbol{C}$ の行に示す．また，図 2 にそれぞれの検出システムにおいて閾値を $N = 1 \sim 1000$ と段階的に変化させていった結果を示している．$\boldsymbol{W}_{(1)}$ は図中の曲線 W の 1 で示されている点にあたり，$\boldsymbol{C}$ にて $N \to \infty$ とすると $\boldsymbol{W}_{(1)}$ の点と重なる．音節音声認識結果の 1-best を閾値とした $\boldsymbol{C}_{(1)}$ は $\boldsymbol{C}_{(1000)}$ に比べ，F 値が若干下がるが，0.053 ポイントの差にとどまっており，有効なスコアリングが行えていることがわかる．

4.5　音響尤度を用いた手法との比較

一方，音声認識結果の尤もらしさ (信頼度) を測る指標として，音響尤度が幅広く使われている．一般的に音声認識は，単語を w，入力音声を o とすると生成確率最大の w の列 $\hat{w} = \arg\max_{w} P(o|w)P(w)$ を求める問題であり，入力音声に依存する項 $P(o|w)$ を音響的な信頼度を表す音響尤度として利用できる．本実験でも対数音響尤度のフレーム平均 $L(w) = -logP(o|w)/\#offrames$ (対数音響尤度をフレーム数で除したもの) を用いた．前節の N-best の N と同様に，認識区間の音響尤度 $L(w)$ を閾値と比較して，閾値以下のものをマッチしたとする．

表 4 にそれぞれ閾値 $L = 17$ と $L = 25$ のときの検索性能を示してある．また，図 2 に閾値を $L = 17$ から $L = 25$ まで順に変化させていった場合の再現率/適合率の変化を

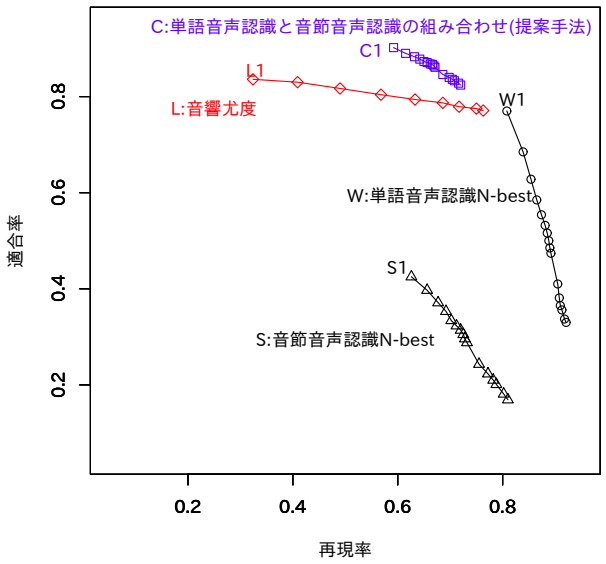

図 2　再現率-適合率曲線

表 4　信頼度として音響尤度を用いた場合の再現率および適合率

	$L=17$			$L=25$		
モデル	再現率	適合率	F 値	再現率	適合率	F 値
WL	0.242	0.838	*0.375*	0.763	0.771	*0.767*

示す．図から読み取れるように閾値を小さくするると再現率が急激に下がるが適合率はそれほど上がらない．この結果から，音響対数尤度は閾値を小さくした際の性能劣化が著しいことがわかる．

4.6　音節音声認識の履歴長

また，さらなる検出性能の向上を目指して，音節音声認識の言語モデル履歴長との関係について調査を行った．一般的に N-gram 言語モデルの履歴長を大きくすると言語モデルの良さを測るパープレキシティは下がるが，効率上の観点から trigram (3-gram) を用いられることが多い．前節でも単語音声認識・音節音声認識共に 3-gram を用いたが，ここでは音節音声認識の履歴長を伸ばすことで，精度の向上を試みた．

表 5 に音節音声認識の履歴長を 3-gram から 4 ～ 7-gram に変化させた際の再現率/適合率を示す．また図 3 に再現率と適合率の関係を示す．音節音声認識単体の性能は 1-best の場合，音節音声認識の履歴長を長くすると再現率/適合率共に改善し，全体の検出性能を示す F 値も向上する．一方，学習コーパス中での 1 単語あたりの音節数は 1.54 だったことから，履歴長の長さという意味では，$1.54 \times 3 = 4.62$-gram がおおよそ単語 3-gram と同じになる．これに近い音節 5-gram ($\boldsymbol{S_{5g}}$) と単語 3-gram ($\boldsymbol{W}$) を比較すると，再現率/適合率ともに単語 3-gram のほうが良かった．

同様に音節 3-gram($\boldsymbol{C_{3g}}$) との組み合わせ結果に比べて，履歴長を長くしたときのほう ($\boldsymbol{C_{4-7g}}$) が 1-best, 1000-best の際の性能は良く，音節音声認識の履歴長を長くすると，音節音声認識単体の性能に加え，単語音声認識と組み合わせた結果も改善した．

表 5　3 ～ 7-gram の音節音声認識モデルを用いた場合の再現率および適合率

	$N=1$			$N=1000$		
モデル	再現率	適合率	F 値	再現率	適合率	F 値
$\boldsymbol{S}_{3g}$	0.626	0.425	0.506	0.810	0.169	0.279
$\boldsymbol{S}_{4g}$	0.730	0.439	0.548	0.872	0.186	0.306
$\boldsymbol{S}_{5g}$	0.781	0.470	0.587	0.894	0.227	0.362
$\boldsymbol{S}_{6g}$	0.811	0.498	0.617	0.906	0.268	0.414
$\boldsymbol{S}_{7g}$	0.837	0.513	0.636	0.911	0.296	0.447
$\boldsymbol{C}_{3g}$	0.592	0.902	0.715	0.720	0.824	0.768
$\boldsymbol{C}_{4g}$	0.688	0.895	0.770	0.768	0.818	0.793
$\boldsymbol{C}_{5g}$	0.727	0.900	**0.804**	0.785	0.826	**0.805**
$\boldsymbol{C}_{6g}$	0.746	0.912	**0.820**	0.792	0.842	**0.816**
$\boldsymbol{C}_{7g}$	0.759	0.923	**0.833**	0.795	0.855	**0.824**

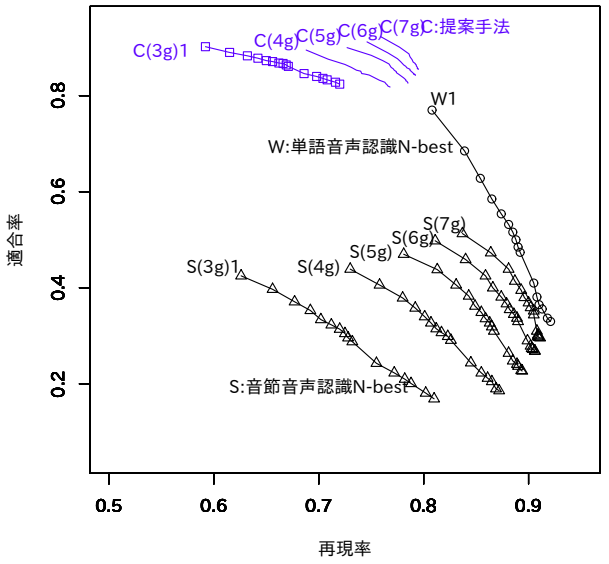

図 3　3 ～ 7-gram の音節音声認識モデルを用いた場合の再現率-適合率曲線

4.7　キーワード音節長と精度との関係

また，モデルの精緻化のために，キーワードの音節長と検索性能との関係を調べた．キーワードの音節長 5 を閾値にし，音節長 $|Ks|$ が $1 \leq |Ks| \leq 4$ のキーワードと $5 \leq |Ks| \leq 11$ のキーワードに分割し，それぞれの検出性能を調べた (表 6，図 4)．モデル $\boldsymbol{W}$ による比較では，わずかに音節長の短いキーワードのほうが F 値が高い．一方，モデル $\boldsymbol{S}$ による比較では，音節列長の長いキーワードのほうが性能が良い．モデル $\boldsymbol{C}$ による比較では，どちらもほ

ぼ同様の傾向を示しているが，単語音声認識 $\boldsymbol{W}$ の性能の良い音節長の短いキーワードのほうが性能が良い．今回検証した範囲では，音節長の短いキーワードのほうが，単語音声認識の性能が良いため，組み合わせの精度も高くなる傾向にあった．

表 6 音節長の違いに対する再現率および適合率

	$N=1$			$N=1000$		
モデル	再現率	適合率	F 値	再現率	適合率	F 値
$\boldsymbol{W}_{(<5)}$	0.814	0.792	**0.803**			
$\boldsymbol{S}_{(<5)}$	0.643	0.371	0.471	0.837	0.144	0.245
$\boldsymbol{C}_{(<5)}$	0.610	0.907	**0.729**	0.736	0.832	**0.781**
$\boldsymbol{W}_{(\geq 5)}$	0.794	0.751	0.772			
$\boldsymbol{S}_{(\geq 5)}$	0.630	0.675	**0.652**	0.770	0.319	**0.451**
$\boldsymbol{C}_{(\geq 5)}$	0.591	0.882	0.708	0.700	0.816	0.754

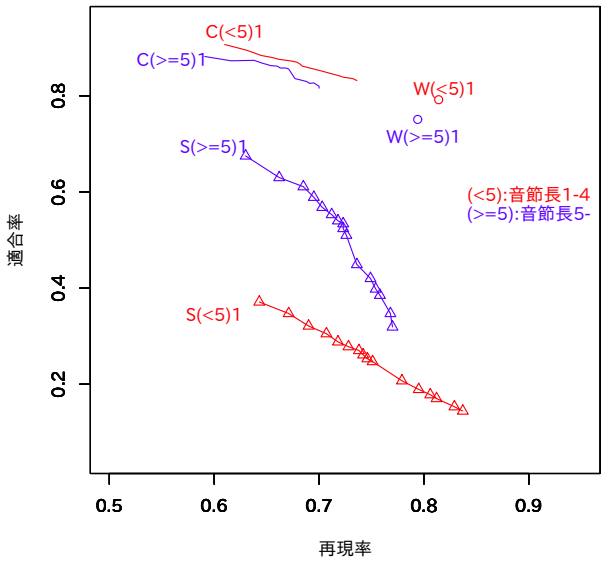

図 4 音節長の違いに対する再現率-適合率曲線

5. 実用的なシステムに向けて

計算量を減らすため，必要な部分のみ信頼度を計算することを考える．信頼度を与えるための音節音声認識 N-best の出力は全ての音声区間について出力しておく必要はなく，単語音声認識でマッチした音声区間のみ出力すればよい．検索キーワード集合が予め既知の場合，インデックスとしては単語音声認識結果のみを用意しておき，単語音声認識結果にマッチした区間のみ音節音声認識 N-best による評価を行うことも実用的な運用として考えられる．全発話区間で単語音声認識と音節 N-best 音声認識を行った場合，$utt(\cdot)$ を検索対象音声データの長さ，r_w, r_s をそれぞれ単語音声認識および音節 N-best 音声認識の単位時間あたりの計算コストとすると，必要な計算量は，$r_w \times utt(\cdot) + r_s \times utt(\cdot)$ であるが，全発話区間で単語音声認識とマッチ区間のみで音節 N-best 音声認識を行った場合，$r_w \times utt(\cdot) + r_s \times utt(match(K_w, R_w))$ となる．

本実験の場合，全音声区間のうち，いずれかの対象キーワードを含む音声区間は 15.6% であった．例えば $r_s = 1.5 \cdot r_w$ (N-best を計算する時間を考慮) とするとマッチ区間のみで音節 N-best 音声認識を行った場合，全区間を対象に単語音声認識と音節 N-best 音声認識を行った場合に比べ，$1.5r_w \times 15.6\% = 0.23r_w$ となり，23% の計算時間の増加にとどまる．さらに，1 つのキーワードを含む音声区間は平均で全発話区間の 0.411% なので，全体の発話量にもよるが，音節 N-best 音声認識を予め行わず，検索時にランタイムで音節 N-best 音声認識および信頼度を計算するといった運用も考えられる．

6. おわりに

単語音声認識の結果に対して音節 N-best 音声認識 の結果を用いてスコアリングすることで，検索性能を劣化させることなく適合率の高い検出結果を得ることができた．一般的な信頼度である音響尤度を用いた結果と比較しても，信頼度として優れているという結果を得た．

参考文献

[1] 西崎博光，中川聖一：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. D-II, 情報・システム, II-パターン処理，Vol. J86-D-II, No. 10, pp. 1369–1381 (2003).
[2] Amir, A., Efrat, A. and Srinivasan, S.: Advances in Phonetic Word Spotting, *Proceedings of the tenth international conference on Information and knowledge management (CIKM '01)*, pp. 580–582 (2001).
[3] 坂本 渚，山本一公，中川聖一：距離付き音節 n グラムインデックスを用いた音声入力による音声ドキュメントの検索語検出法の評価，第 7 回音声ドキュメント処理ワークショップ論文集，pp. 2013–05 (2013).
[4] Mamou, J. et al.: System Combination and Score Normalization for Spoken Term Detection, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, pp. 8272–8276 (2013).
[5] 宇津呂武仁ほか：複数の大語彙連続音声認識モデルの出力の共通部分を用いた高信頼度部分の推定，電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理，Vol. J86-D-II, No. 7, pp. 974–987 (2003).
[6] Wessel, F., Schluter, R., Macherey, K. and Ney, H.: Confidence measures for large vocabulary continuous speech recognition, *IEEE Transaction Speech and Audio Prcesing*, Vol. 9, No. 3, pp. 288–298 (2001).
[7] Rueber, B.: Obtaining confidence measures from sentence probabilities, *Fifth European Conference on Speech Communication and Technology (EUROSPEECH 1997)*, pp. 739–742 (1997).
[8] 中川聖一，堀部千寿：音響尤度と言語尤度を用いた音声認識結果の信頼度の算出，情報処理学会研究報告音声言語情報処理，Vol. 2001, No. 55, pp. 97–92 (2001).
[9] 緒方 淳，有木康雄：音声認識精度向上のための信頼度尺度の比較，情報処理学会研究報告音声言語情報処理，Vol. 2000, No. 119, pp. 113–118 (2000).
[10] 浅見太一ほか：単語の文脈一貫性と音響尤度を用いた音声ドキュメント認識信頼度の性能評価，電子情報通信学会技術研究報告音声，Vol. 110, No. 143, pp. 43–48 (2010).

正誤表（誤：網掛け部）

箇所 1. p.4 右 表 3

表 3　再現率および適合率

モデル	$N=1$ 再現率	$N=1$ 適合率	$N=1$ F 値	$N=1000$ 再現率	$N=1000$ 適合率	$N=1000$ F 値
$\boldsymbol{W}$	0.808	0.770	0.789	0.921	0.330	0.486
$\boldsymbol{S}$	0.626	0.425	0.506	0.810	0.169	0.279
$\boldsymbol{C}$	**0.592**	**0.902**	**0.715**	**0.720**	**0.824**	**0.768**

箇所 2. p.5 右 表 5

表 5　3 ～ 7-gram の音節音声認識モデルを用いた場合の再現率および適合率

モデル	$N=1$ 再現率	$N=1$ 適合率	$N=1$ F 値	$N=1000$ 再現率	$N=1000$ 適合率	$N=1000$ F 値
$\boldsymbol{S}_{3g}$	0.626	0.425	0.506	0.810	0.169	0.279
$\boldsymbol{S}_{4g}$	0.730	0.439	0.548	0.872	0.186	0.306
$\boldsymbol{S}_{5g}$	0.781	0.470	0.587	0.894	0.227	0.362
$\boldsymbol{S}_{6g}$	0.811	0.498	0.617	0.906	0.268	0.414
$\boldsymbol{S}_{7g}$	0.837	0.513	0.636	0.911	0.296	0.447
$\boldsymbol{C}_{3g}$	0.592	0.902	0.715	0.720	0.824	0.768
$\boldsymbol{C}_{4g}$	0.688	0.895	**0.770**	0.768	0.818	**0.793**
$\boldsymbol{C}_{5g}$	0.727	0.900	**0.804**	0.785	0.826	**0.805**
$\boldsymbol{C}_{6g}$	0.746	0.912	**0.820**	0.792	0.842	**0.816**
$\boldsymbol{C}_{7g}$	0.759	0.923	**0.833**	0.795	0.855	**0.824**

箇所 3. p.5 右 図 3

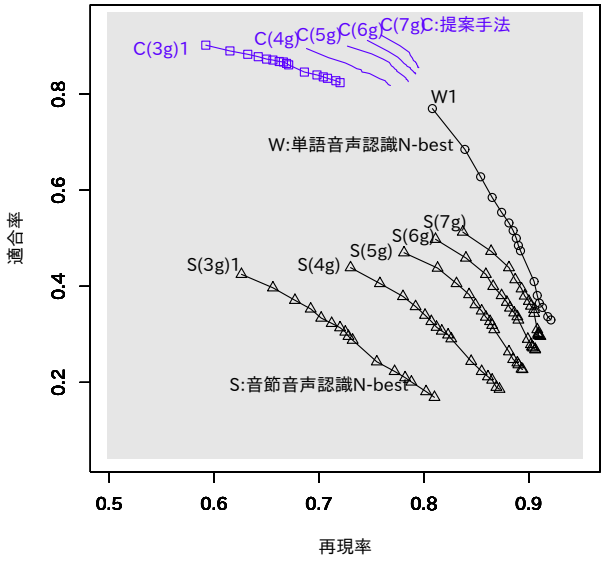

図 3　3 ～ 7-gram の音節音声認識モデルを用いた場合の再現率-適合率曲線

箇所 4. p.6 左 表 6

表 6　音節数の違いに対する再現率および適合率

モデル	$N=1$ 再現率	$N=1$ 適合率	$N=1$ F 値	$N=1000$ 再現率	$N=1000$ 適合率	$N=1000$ F 値
$\boldsymbol{W}_{(<5)}$	0.814	0.792	**0.803**			
$\boldsymbol{S}_{(<5)}$	0.643	0.371	0.471	0.837	0.144	0.245
$\boldsymbol{C}_{(<5)}$	0.610	0.907	**0.729**	0.736	0.832	**0.781**
$\boldsymbol{W}_{(\geq 5)}$	0.794	0.751	0.772			
$\boldsymbol{S}_{(\geq 5)}$	0.630	0.675	**0.652**	0.770	0.319	**0.451**
$\boldsymbol{C}_{(\geq 5)}$	0.591	0.882	0.708	0.700	0.816	0.754

（正）

箇所 1.

表 3　再現率および適合率

モデル	$N=1$ 再現率	$N=1$ 適合率	$N=1$ F 値	$N=1000$ 再現率	$N=1000$ 適合率	$N=1000$ F 値
$\boldsymbol{W}$	0.808	0.770	0.789	0.921	0.330	0.486
$\boldsymbol{S}$	0.618	0.417	0.498	0.809	0.166	0.276
$\boldsymbol{C}$	**0.583**	**0.902**	**0.709**	**0.717**	**0.823**	**0.767**

箇所 2.

表 5　3 ～ 7-gram の音節音声認識モデルを用いた場合の再現率および適合率

モデル	$N=1$ 再現率	$N=1$ 適合率	$N=1$ F 値	$N=1000$ 再現率	$N=1000$ 適合率	$N=1000$ F 値
$\boldsymbol{S}_{3g}$	0.618	0.417	0.498	0.809	0.166	0.276
$\boldsymbol{S}_{4g}$	0.697	0.413	0.519	0.854	0.174	0.289
$\boldsymbol{S}_{5g}$	0.710	0.418	0.526	0.849	0.189	0.309
$\boldsymbol{S}_{6g}$	0.709	0.419	0.527	0.847	0.203	0.328
$\boldsymbol{S}_{7g}$	0.703	0.414	0.521	0.840	0.211	0.337
$\boldsymbol{C}_{3g}$	0.583	0.902	0.709	0.717	0.823	0.767
$\boldsymbol{C}_{4g}$	0.658	0.890	**0.757**	0.757	0.813	**0.784**
$\boldsymbol{C}_{5g}$	0.665	0.885	**0.759**	0.752	0.814	**0.782**
$\boldsymbol{C}_{6g}$	0.667	0.886	**0.761**	0.752	0.816	**0.783**
$\boldsymbol{C}_{7g}$	0.660	0.886	**0.756**	0.748	0.820	**0.782**

箇所 3.

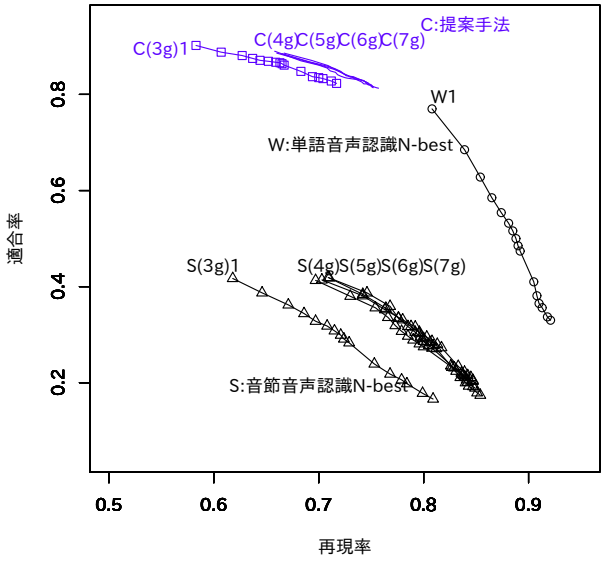

図 3　3 ～ 7-gram の音節音声認識モデルを用いた場合の再現率-適合率曲線

箇所 4.

表 6　音節数の違いに対する再現率および適合率

モデル	$N=1$ 再現率	$N=1$ 適合率	$N=1$ F 値	$N=1000$ 再現率	$N=1000$ 適合率	$N=1000$ F 値
$\boldsymbol{W}_{(<5)}$	0.814	0.792	**0.803**			
$\boldsymbol{S}_{(<5)}$	0.632	0.361	0.460	0.834	0.141	0.241
$\boldsymbol{C}_{(<5)}$	0.597	0.903	**0.719**	0.732	0.830	**0.778**
$\boldsymbol{W}_{(\geq 5)}$	0.794	0.751	0.772			
$\boldsymbol{S}_{(\geq 5)}$	0.630	0.681	**0.654**	0.778	0.322	**0.455**
$\boldsymbol{C}_{(\geq 5)}$	0.590	0.894	0.710	0.706	0.821	0.759