情報処理学会論文誌 Vol.56 No.8 1646–1656 (Aug. 2015)

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)]

受付日 2014年11月22日, 採録日 2015年5月9日

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含む発言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．一方，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．このような現場では，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．本論文では，認識単位の異なる 2 種類の音声認識システムを用いて，各検出区間に対して信頼度を与え，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キーワード文字列に一致する区間を検出し，それら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検出区間に信頼度を与える．実験では，大語彙連続音声認識システム単体および音節音声認識システム単体のみを用いた場合との比較を行った．その結果，2 種類の音声認識システムの組合せにより，誤認識音声の検出を 42.1%～76.7%減らすことができ，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)]

Received: November 22, 2014, Accepted: May 9, 2015

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 vary according to the searched keywords and the purpose of the search; 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. The experimental results show that the proposed method reduce 42.1%–76.7% false-positive error by combining LVCSR and the syllable-based speech recognition system.

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

1. はじめに

音声認識精度の向上にともない，様々な場面で音声認識

[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., Koto, Tokyo 135–8511, Japan
[†1] 現在，国立大学法人静岡大学大学院総合科学技術研究科
Presently with Graduate School of Integrated Science and Technology, Shizuoka University, Hamamatsu, Shizuoka, 432–8011, Japan
[a)] tohru3@jp.ibm.com
[b)] nisimura@inf.shizuoka.ac.jp

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られる音声インタ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．たとえば，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，全通話のモニタリングができるようになった．音声認識によって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，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．特に，大量の検出結果に対して人手にて聴取作業を行うような場合，正確に認識されている可能性の高い音声から聴取できると作業効率が高まる．そのため，主に適合率を高める方向で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．一般的には音声認識パラメータの変更，単語の出現確率を変える等の操作を行った後，再度音声認識を行えば，ある程度，再現率・適合率の調整が可能であるが，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．

キーワード検出では，一般的に単語を認識単位とする大語彙連続音声認識（以下，単語音声認識）の書き起こし結果に対して文字列での比較または単語列での比較を行うが，未知語（辞書に含まない語）や認識誤りへの対応として，サブワードや音素・音節またはそれに準じる単位での音声認識（以下，音節音声認識）の結果に対して単位列での比較を行う方法 [1], [2], [3] が知られている．また，単語単位の音声認識において直接的に信頼度の算出が困難な場合，音素や音節を認識単位とした認識システムを用いて信頼度を推定する方法 [4] も提案されており，未知語の棄却に良い性能を示している．音節音声認識は単語音声認識を用いる手法と比較すると，言語情報を利用していないためキーワード検出性能が低い [5] が，各種認識誤りに対応した距離を定義することで再現率・適合率の調整が可能である．距離の定義としては，編集距離 [1]・Word Confusion Network（WCN）[2] により距離を定義する方法，音素ラティスから生成される *tri*-gram インデックスと編集距離を組み合わせた方法 [3] 等が知られている．これら両方の利点・欠点を考慮して，「単語音声認識システム」と「未知語を扱うことのできる単語・音節音声認識システム」を併用したシステム [5] も提案されており，単語音声認識と音節音声認識の併用が未知語・既知語を含めた検出精度の向上に役立つことが分かっている．システム統合という観点からは，信頼度の高い認識区間を推定するために複数の音声認識システムの認識結果単語列の論理積部分を抽出する方法 [6] や，音声認識精度を向上させるためのシステム統合手法（ROVER 法）[7]，各システムから得られるスコアの正規化方法 [8] 等について研究が行われており，それぞれ単体のシステムを用いた場合に比べ良い性能を示している．ただし，5〜数十種類のパラメータの異なる音声認識器を用いており，計算コストが高い．

本論文では，音声検索語検出システムの利便性を高めるため，「計算量を大幅に増やさない」「作業効率を向上させる」「高い適合率に調整が可能である」ことを目的として，単語音声認識と音節音声認識を組み合わせたシステムを提案する．提案法では音節音声認識システムを単語音声認識システムにより検出された区間の信頼度の計算に用いることで，単語音声認識単体ではカバーできなかった適合率を優先させたキーワード検出を実現する．具体的には，(1) 単語音声認識結果を用いてキーワード文字列の一致する区間を検出し，(2) それら検出区間に対応する音節音声認識結果とキーワード音節列を用いて信頼度を計算する．従来，音節音声認識は未知語への対処として用いられることが多かったが，既知語（辞書に含まれる語）に信頼度を与えるシステムとして用いる．信頼度の与え方として，単語グラフ事後確率 [9]，N-best [10]，音響尤度，言語尤度を利用した方法等，様々な指標が提案され，その有効性が検証されている [11], [12] が，中でも単語グラフを用いた単語事後確率による信頼度付与および N-best を用いた単語事後確率を求める方法の性能が高いとされている [13], [14]．本研究においても性能が高く，単語グラフを直接扱うより計算コストの低い，N-best を用いた音節事後確率による方法を用いて信頼度の計算を行った．また計算量の観点から，計算量の少ない，N-best の N 位以内に含まれるかどうか（ランキング）を信頼度とした指標についても検討を行った．実験ではまず，ベースシステムとして単語音声認識 1-best を用いた場合と，単語音声認識 1-best に音節音声認識 N-best を用いて信頼度計算を行った場合とを比較し，2 つの音声認識器の組合せが有効であることを示す．次に，ベースシステムとして単語音声認識 N-best を用いて信頼度計算を行った場合と，このベースシステムの結果に対してさらに音節音声認識 N-best を用いて信頼度計算を行った場合とを比較し，ベースのシステムによらず 2 つの音声認識器の組合せが有効であることを示す．さらに単語音声認識 N-best から計算される信頼度と音節音声認識 N-best から計算される信頼度を組み合わせることにより，

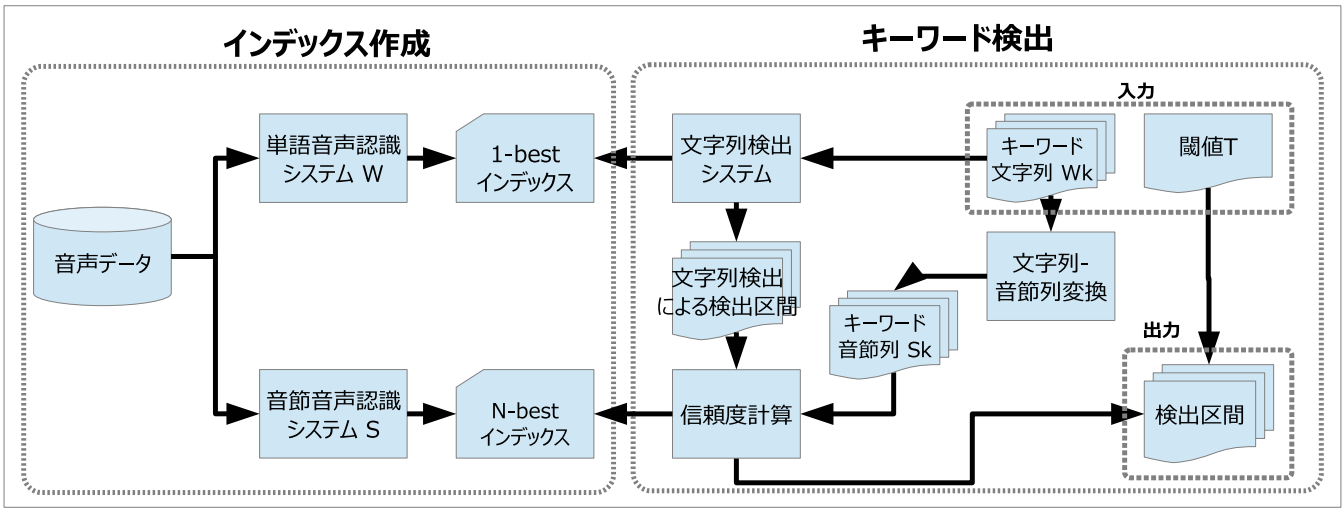


図 1 大語彙連続音声認識と音節 N-best 音声認識を用いたキーワード検出
Fig. 1 Term Detection using LVCSR and N-best syllable-based speech recognition results.

作業効率の良い信頼度を付与できることを示す．有効性の評価は，各システムで得られる適合率の範囲と，実際のコールモニタリング作業を模した作業効率（信頼度上位から音声を聴取した際に誤った音声が含まれる割合）により行った．実験の結果，組合せにより適合率の上限は，それぞれのベースシステムに比べて 0.248，0.035 ポイント向上し，また誤検出の割合を再現率 0.5 のポイントにおいてそれぞれ 76.7%，42.1%減らすことができた．

2. 大語彙連続音声認識と音節音声認識を併用したキーワード検出

音声検索語検出システムでは単なる音声の書き起こしと異なり，未知語や認識誤り対策のため音節音声認識が併用されることが多い [5], [15], [16]．本研究では，このように 2 種類の音声認識システムが利用可能な状況で，適合率を高めることのできるシステムについて検討を行う．音節音声認識は未知語検出に対して適用するのではなく，既知語に対する検出区間への信頼度付与に音節音声認識結果を用いるため，本論文では既知語のみを検出対象にし未知語を対象としない．システムの概要を図 1 に示す．システムは大きく分けて，検出対象データに対して音声認識を行い各キーワードが含まれているかを示すインデックスを付与するインデックス作成部と入力されたキーワードを元に検出対象データからキーワードの検出を行うキーワード検出部から構成されている．

2.1 インデックス作成

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．音声コーパスには音声に対応する単語列が付与されており，辞書と音声コーパスを用いて音声認識モデルを構築できる．また，この音声コーパスと辞書を用いて，音声コーパスに対して音節列を付与し音節単位の音声認識モデルを構築できる．音声認識モデルの音響モデル・言語モデルをそれぞれのシステムで個別に構築することもできるが，本研究では同一の音響モデルを用い，言語モデルのみを単位の異なる 2 種類の言語コーパスから構築する．下記に 2 種類の音声認識システムを示す．

$\mathcal{W}$ 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：右 / クリック/ ですね

$\mathcal{S}$ 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 “Q” を含む音節は別の音節として取り扱う．
例：mi / gi / ku / ri / Qku / de / su / ne

検出対象の音声を単語単位の音声認識システム $\mathcal{W}$ および音節単位の音声認識システム $\mathcal{S}$ 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 のみの出力として単語列 w^1 を，システム $\mathcal{S}$ は N-best の認識結果 $s^1 \dots s^N$ を出力する．ここで N-best は検出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出力されているものとする．また，キーワードは，文字列 w_K（例：「値段」）で与えられ，辞書またはテキスト音素列変換 [17], [18] により音節列 s_K（例：ne / da / n）に変換され信頼度計算装置に入力される．本実験では辞書を用いて変換を行ってある．各認識結果は，単語アラインメント結果 $d = \langle$開始時間 t_{beg}, 終了時間 $t_{end}\rangle$ の列からなり，これらを元に検出用のインデックスを作成する．

2.2 キーワード検出

次に以下の手順でキーワードに一致する音声区間を検出

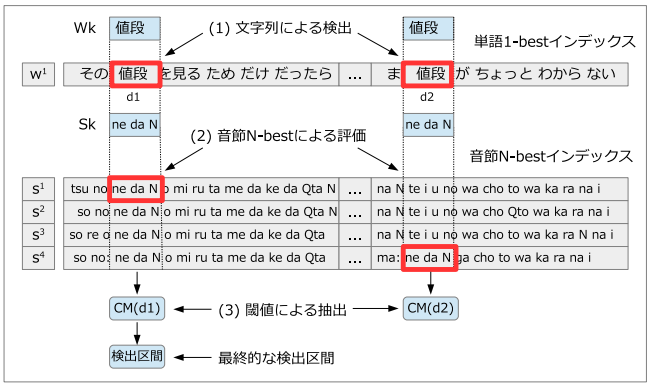


図 2 キーワード検出
Fig. 2 Procedure of keyword dection.

する（**図 2**）．キーワードは，文字列 w_K および音節列 s_K で与えられるものとする．

(1) **文字列による検出**
キーワード文字列 w_K と単語音声認識システム $\mathcal{W}$ の認識結果 w^1 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出用インデックスに含まれる単語アラインメント結果 $d = \langle t_{beg}, t_{end} \rangle$ から得られる[*1]．最終的に M^{w_K} 個の一致した検出区間リスト $D(w_K) = d_1^{w_K} \ldots d_{M^{w_K}}^{w_K}$ を得る．

(2) **音節 *N*-best による信頼度評価**
単語音声認識の検出区間 $d_i^{w_K} (1 \leq i \leq M^{w_K})$ に含まれる音節音声認識システム $\mathcal{S}$ の認識結果 $s^1 \cdots s^N$ とキーワード音節列 s_K を用いて検出区間 $d_i^{w_K}$ の信頼度 $CM(s_K, d_i^{w_K})$ を求める．検出区間の信頼度の計算は 2.3 節に記す 2 種類の信頼度を用いて行う．

(3) **閾値による抽出**
検出時には閾値 T を与え，信頼度 $CM(s_K, d_i^{w_K}) \geq T$ となる検出区間のみを最終的な検出区間とする．

2.3 信頼度

信頼度として N-best を用いた 2 種類の尺度について検討を行う．以下に，本研究で用いる 2 種類の尺度について示す．

ランキング

計算量の少ない N-best の評価方法として N-best の N を変えたときに N-best に含まれるかどうかを判断する．1-best から順にキーワードとの比較を行うだけなので，最も検出時間が短くなると期待できる．検出区間 d_i の評価は音節音声認識 N-best の結果 $s^1 \ldots s^N$ とキーワード s_K の一致する最小の順位 $n (1 \leq n \leq N)$ を用いる（式 (1)）．$\mathcal{L}_{d_i}^{s^n}$ は $d_i^{w_K} (1 \leq i \leq M^{w_K})$ 中に含まれる N-best の n 番目の部分音節列集合を表し，δ は一致するときに 1，それ以外のときに 0 を返すクロネッカーのデルタである．

$$CM_{Rank}(s_K, d_i^{w_K}) = 1 - \frac{min_{n=1}^{N}(n \cdot \delta(s_K, \mathcal{L}_{d_i^{w_K}}^{s^n})) - 1}{N} \quad (1)$$

事後確率

N-best を用いた事後確率の値を信頼度とする．あるキーワード s_K が検出区間 $d_i^{w_K} (1 \leq i \leq M^{w_K})$ に出現すると仮定したときの文の事後確率を，区間 $d_i^{w_K}$ に s_K を含む文の生成確率を仮説集合全体 N-best の生成確率の和で除したものとして定義する（式 (2)）．x は入力音声を表し，$p(s)$ は音節を単位とした言語モデルによる尤度，$p(x|s)$ は音響モデルによる尤度を表す．N-best 全体から事後確率が計算されるので，ランキングを用いた信頼度より計算量が多い．

$$CM_{Post}(s_K, d_i^{w_K}) = \frac{\sum_{n=1}^{N} p(x|s^n) \cdot p(s^n) \cdot \delta(s_K, \mathcal{L}_{d_i^{w_K}}^{s^n})}{\sum_{n=1}^{N} p(x|s^n) \cdot p(s^n)} \quad (2)$$

3. 実験

単語音声認識システムおよび音節音声認識システムを構築し，実データを用いて，単語音声認識システムの結果に対し音節音声認識システムの認識結果を用いて信頼度を与えることの有効性を検証する．

3.1 検出評価用データ

電話録音データを用いて検出評価を行った．各音声ファイルは 8 kHz/16 bit サンプリングで録音され，二話者の音声はあらかじ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（**表 1**）．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．

表 1 検出評価用データ
Table 1 Test data.

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	21,853 セグメント
単語数	179 K 語

3.2 キーワードおよび評価方法

キーワードは未知語を含まず，長さ 1〜11 文字（1〜12 音節）からなる 40 語で構成される．キーワードの長さごとの語数は長さ 1 から 11 まで順に $(2, 9, 5, 4, 6, 6, 1, 1, 2, 3, 1)$ である．各キーワードはあらかじめ文字列に対する音節列を与えてある．キーワードの例を**表 2** に示す．検出評価用データにはのべ 3,248 個のキーワードが含まれている．

最終的な評価は作業効率や検出に要する検出時間を含め

[*1] タイムスタンプの開始時間と終了時間は単語アラインメントの誤差を考慮してわずかな時間 Δ だけ広げてある．Δ の幅は検出評価用データにおける 1 音節あたりの継続長が 0.15 秒であったことから 2 音節分の 0.3 秒とした．

表 2 キーワード例
Table 2 Example of keywords.

表記 w_K	音節表記 s_K
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*

て行うが，本章では基本的な検出性能である再現率，適合率，およびこれら 2 つを組み合わせた検出性能を示す F 値（式 (3)）により評価を行う．キーワードが正しく検出されたかどうかの判定は発話単位で行う．検出評価用データの各発話ごとに書き起こし文字列および音節列が付与されており，それを元にキーワードが含まれているかどうかをあらかじめ各発話に対して正解ラベルを付与しておき，システムが検出した区間がラベル付けられた発話区間に含まれていれば正しく検出されたものとする．検出評価用データでの 1 発話区間の平均の長さは 2.24 秒であり，1 発話区間に同一キーワードが複数回含まれる可能性は小さいことから発話ごとに検出の判定を行うこととした．

$$\text{再現率} = \frac{\text{正しく検出された発話数}}{\text{検出評価用データ中の正解ラベル付き発話数}}$$

$$\text{適合率} = \frac{\text{正しく検出された発話数}}{\text{システムにより検出された発話数}}$$

$$\text{F 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}} \tag{3}$$

3.3 音声認識

音声認識システムを用いて検出評価用データを単語列および音節列に変換しておく．音声認識モデルは 150 時間の電話会話の録音音声を元に作成されており，GMM-HMM を boosted MMI で識別学習した音響モデルである．単語音声認識システムの言語モデルは単語 *tri*-gram，音節音声認識システムの言語モデルは音節 5-gram を用いて構築してある．単語 1-best による検出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，信頼度計算に必要な N-best 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．一発話あたりの N-best の出力数は，単語音声認識では (最小値, 最大値, 平均) = (1, 877, 91.6) 個，また音節音声認識では (最小値, 最大値, 平均) = (1, 495, 36.6) 個であった．**表 3** に音声「右クリックですね」を入力としたときの単語音声認識 N-best 出力，また**表 4** に音節音声認識出力の結果を示す．誤りなく認識されているのはそれぞれ w^1 および s^4 となる．

3.4 ランキングを信頼度として用いた実験

2 種類の信頼度 $CM_{Rank}(s_K, d^{w_K})$，$CM_{Post}(s_K, d^{w_K})$ のうち，まずランキング $CM_{Rank}(s_K, d^{w_K})$ を用いた場合の提案手法の評価を行う．キーワード文字列 w_K が単語音声認識の 1-best である $\boldsymbol{W}_{(1)}^{Rank}$ に一致した区間に対して，音節音声認識 N-best によって計算される $\boldsymbol{S}_{(T)}^{Rank}$ で信頼度付与した結果を示す．比較として，単語音声認識 N-best と音節音声認識 N-best 単体での評価も行う．単語音声認識 N-best と音節音声認識 N-best に関してはそれぞれ $CM_{Rank}(w_K, d^{w_K})$, $CM_{Rank}(s_K, d^{s_K})$ による評価を行い，単語音声認識システム単体でどのような再現率・適合率の傾向があるか調べた．式 (1) と同様に，

$$CM_{Rank}(w_K, d_i^{w_K}) = 1 - \frac{min_{n=1}^{N}(n \cdot \delta(w_K, \mathcal{L}_{d_i^{w_K}}^{w^n})) - 1}{N}$$

および

$$CM_{Rank}(s_K, d_i^{s_K}) = 1 - \frac{min_{n=1}^{N}(n \cdot \delta(s_K, \mathcal{L}_{d_i^{s_K}}^{s^n})) - 1}{N}$$

と定義される．$d_i^{w_K}, d_i^{s_K}$ はそれぞれ s_K を音節 N-best 中に含む区間，w_K を単語 N-best 中に含む区間である．それぞれの方法で得られる検出区間の集合を以下のように示す．

$\boldsymbol{W}_{(T)}^{Rank}$ 単語音声認識 N-best
全検出評価用データに対する単語音声認識 N-best 出力に対し，閾値を n 位に対応する $T = 1 - (n-1)/N$ とした場合に $CM_{Rank}(w_K, d^{w_K}) \geq T$ となる音声区間の集合

$\boldsymbol{S}_{(T)}^{Rank}$ 音節音声認識 N-best
全検出評価用データに対する音節音声認識 N-best 出力に対し，閾値を n 位に対応する $T = 1 - (n-1)/N$ とした場合に $CM_{Rank}(s_K, d^{s_K}) \geq T$ となる音声区間の集合

$\boldsymbol{C}_{(T)}^{Rank}$ 単語・音節音声認識組合せ（提案手法）
$\boldsymbol{C}_{(T)}^{Rank} = \boldsymbol{W}_{(1)}^{Rank} \cap \boldsymbol{S}_{(T)}^{Rank}$．単語認識結果がキーワード 1-best 文字列に一致し，かつ音節音声認識 N-best

表 3 単語 N-best 出力の例
Table 3 Example of word N-best output.

n	認識単語列 w^n	単語列対数認識尤度
1	右 クリック ですね	−81.1235
2	右 クリック です よね	−81.3125
3	右 クリック です よ ね	−81.6375
4	右 クリック でしょう ね	−81.8335
5	右 クリック です よ	−81.9705

表 4 音節 N-best 出力の例
Table 4 Example of syllable N-best output.

n	認識単語列 s^n	音節列対数認識尤度
1	*mi gi ku ri Qku su su me*	−80.6225
2	*mi gi ku ri Qku o su su me*	−80.7070
3	*mi gi ku ri Qku su de*	−80.7140
4	*mi gi ku ri Qku de su ne*	−80.7615
5	*mi gi ku ri Qku de su ne:*	−81.4520

表 5 $CM_{Rank}(s_K, d^{w_K})$ を用いた $\boldsymbol{C}^{Rank}$ の再現率・適合率
Table 5 Recall and Precision of $\boldsymbol{C}^{Rank}$ using $CM_{Rank}(s_K, d^{w_K})$.

	$T=1$			$T=0$		
モデル	再現率	適合率	F 値	再現率	適合率	F 値
$\boldsymbol{W}^{Rank}$	**0.746**	0.637	0.688	*0.900*	*0.236*	*0.374*
$\boldsymbol{S}^{Rank}$	0.710	0.418	0.526	0.849	0.189	0.313
$\boldsymbol{C}^{Rank}$	0.634	**0.827**	**0.717**	**0.706**	**0.735**	**0.720**

出力に対し，閾値を n 位に対応する $T = 1-(n-1)/N$ とした場合に $CM_{Rank}(s_K, d^{w_K}) \geq T$ となる音声区間の集合

表 5 に単語音声認識 $\boldsymbol{W}^{Rank}$，音節音声認識 $\boldsymbol{S}^{Rank}$ および提案手法 $\boldsymbol{C}^{Rank}$ において閾値を最大・最小とした場合の再現率・適合率・F 値を示す．$\boldsymbol{W}^{Rank}_{(1)}$ と $\boldsymbol{S}^{Rank}_{(1)}$ の点がそれぞれ単語音声認識 1-best 出力および音節音声認識 1-best 出力に対応する再現率・適合率である．$\boldsymbol{W}^{Rank}_{(1)}$ と $\boldsymbol{S}^{Rank}_{(1)}$ の点を比較すると音節音声認識の F 値 0.526 に対し単語音声認識の F 値 0.688 と 0.162 ポイント高くなっており，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．$\boldsymbol{W}^{Rank}$ に注目すると，$\boldsymbol{W}^{Rank}_{(0)}$ は $\boldsymbol{W}^{Rank}_{(1)}$ に比べて，再現率は 0.154 ポイント上昇したが適合率は 0.401 ポイント下がり，結果として F 値は 0.314 ポイント下がっている．同様に音節音声認識の F 値も閾値の変化 $T = 1 \rightharpoonup 0$ にしたがい，0.213 ポイント下がる．

図 3 に各モデルにおいて閾値を $T = 1 \rightharpoonup 0$ と段階的に変化させていった結果を示している．なお，図中の $\boldsymbol{W}^{Rank}$ において $T = 1$ 以外の点も図中にプロットしてあるが，$\boldsymbol{C}^{Rank}$ を求める際には 1-best にあたる $\boldsymbol{W}^{Rank}_{(1)}$ 以外の点は用いていない．音節音声認識結果の 1-best を閾値とした $\boldsymbol{C}^{Rank}_{(1)}$ は $\boldsymbol{W}^{Rank}_{(1)}$ に比べ F 値が向上し (0.688 → 0.717)，適合率の上限を 0.190 ポイント向上させることができている．

3.5 事後確率を信頼度として用いた実験

次にランキング $CM_{Rank}(s_K, d^{w_K})$ の代わりに音節列事後確率 $CM_{Post}(s_K, d^{w_K})$ を信頼度として定義し，同様に実験を行った．キーワード文字列 w_K が単語音声認識の 1-best である $\boldsymbol{W}^{Rank}_{(1)}$ に対して一致した音声区間に対して，音節 N-best を用いた事後確率で信頼度付与した結果を示す．音節の事後確率 $\boldsymbol{S}^{Post}_{(T)}$，および $\boldsymbol{W}^{Rank}_{(1)}$ と $\boldsymbol{S}^{Post}_{(T)}$ から計算される検出区間の集合を以下のように示す．音節モデルの事後確率による信頼度は

$$CM_{Post}(s_K, d_i^{s_K}) = \frac{\sum_{n=1}^{N} p(x|s^n) \cdot p(s^n) \cdot \delta(s_K, \mathcal{L}_{d_i^{s_K}}^{s^n})}{\sum_{n=1}^{N} p(x|s^n) \cdot p(s^n)}$$

として定義される．

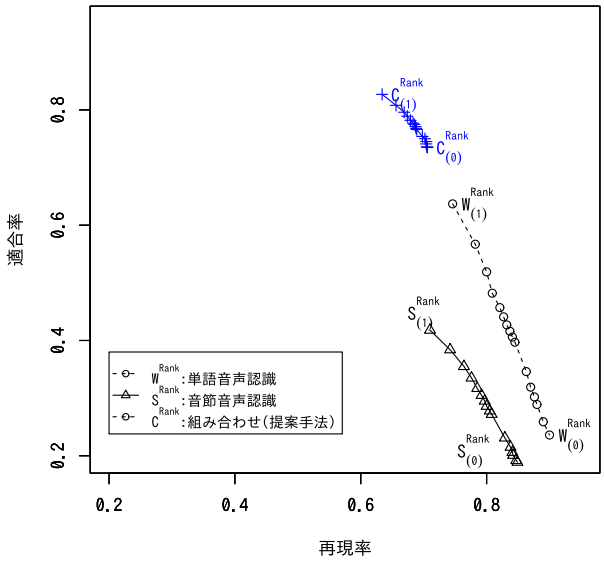


図 3 $\boldsymbol{C}^{Rank}$ の再現率–適合率曲線
Fig. 3 Recall–Precision curve of $\boldsymbol{C}^{Rank}$.

表 6 $CM_{Post}(s_K, d^{w_K})$ を用いた $\boldsymbol{C}^{Post}$ の再現率・適合率
Table 6 Recall and Precision of $\boldsymbol{C}^{Post}$ using $CM_{Post}(s_K, d^{w_K})$.

	$T=1(0.94)$			$T=0$		
モデル	再現率	適合率	F 値	再現率	適合率	F 値
$\boldsymbol{W}^{Rank}$	**0.746**	0.637	**0.688**	*0.900*	*0.236*	*0.374*
$\boldsymbol{S}^{Post}$	0.433	0.494	0.461	**0.849**	0.189	0.309
$\boldsymbol{C}^{Post}$	0.402	**0.885**	0.553	0.706	**0.735**	**0.720**
$\boldsymbol{S}^{Post}_{(0.94)}$	0.516	0.507	0.511			
$\boldsymbol{C}^{Post}_{(0.94)}$	0.480	**0.888**	0.623			

$\boldsymbol{S}^{Post}_{(T)}$ 音節音声認識 N-best
全検出評価用データに対する各音節音声認識 N-best 出力に対し，音節列事後確率が $CM_{Post}(s_K, d^{s_K}) \geq T$ となる音声区間の集合．

$\boldsymbol{C}^{Post}_{(T)}$ 単語・音節音声認識組合せ（提案手法）
$\boldsymbol{C}^{Post}_{(T)} = \boldsymbol{W}^{Rank}_{(1)} \cap \boldsymbol{S}^{Post}_{(T)}$．単語認識結果がキーワード 1-best 文字列に一致し，かつ音節認識 N-best 出力から得られる音節列事後確 $CM_{Post}(s_K, d^{w_K}) \geq T$ となる音声区間の集合．

信頼度として単語事後確率 $CM_{Post}(s_K, d^{w_K})$ を用いた場合の音節音声認識 $\boldsymbol{S}^{Post}$ および組合せ提案手法 $\boldsymbol{C}^{Post}$ のそれぞれの閾値を $T = 1, 0$ としたときの再現率・適合率・F 値を**表 6** の $\boldsymbol{S}^{Post}$，$\boldsymbol{C}^{Post}$ 行に示す．$\boldsymbol{C}^{Post}$ の適合率の上限は $\boldsymbol{W}^{Rank}$ に比べ 0.248 ポイント高く，改善している．また，図 3 と同様に**図 4** に閾値を $T = 1 \rightharpoonup 0$ と段階的に変化させていった結果を示す．閾値 T が 1 に近づくとともに適合率は若干下がる傾向にあり，適合率が最高になったのは閾値が $T = 0.94$ の点で適合率は 0.888 であり，この点での再現率・適合率・F 値を表 6 の下 2 行に示す．ランキング $CM_{Rank}(s_K, d^{w_K})$ を信頼度とした $\boldsymbol{C}^{Rank}_{(1)}$ とを比較すると $\boldsymbol{C}^{Post}_{(0.94)}$ のほうが 0.061 ポイント最高適合率が高

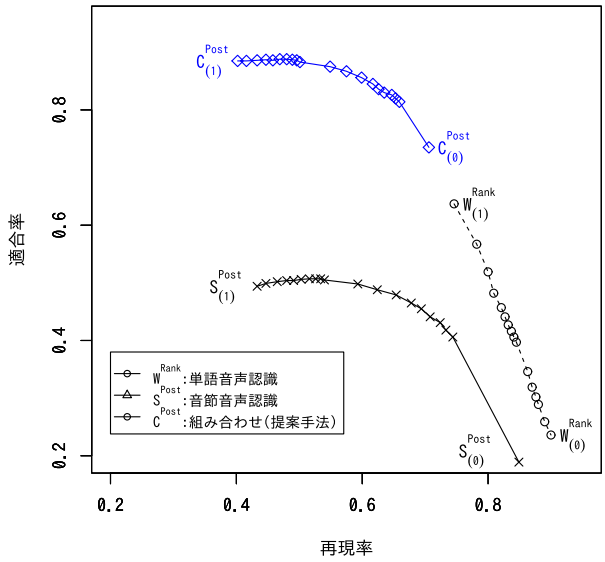


図 4 $CM_{Post}(s_K, d^{w_K})$ を用いた場合の再現率–適合率曲線
Fig. 4 Recall–Precision curve of $CM_{Post}(s_K, d^{w_K})$.

かった．

$\boldsymbol{C}^{Rank}$，$\boldsymbol{C}^{Post}$ どちらの信頼度を用いた場合においても，適合率の上限を高めるための調整方法として有効に働いていることが分かる．いずれも単語 1-best の結果に対して音節音声認識 N-best を使って信頼度付与した結果であり，検出対象音声に対し単語 1-best と音節 N-best の認識結果を得ておけば，キーワード検出時には通常の文字列検出を行い，キーワードが検出された音声区間に対してのみ信頼度評価を行うことで，適合率の高い音声区間のみを得ることができる．

3.6 単語音声認識に対して事後確率を信頼度として用いた実験

単語 1-best の結果に対して音節音声認識 N-best を $CM_{Rank}(s_K, d^{w_K})$ および $CM_{Post}(s_K, d^{w_K})$ により計算される信頼度により信頼度付与できることが分かったが，単語認識 N-best の結果から直接 $CM_{Post}(s_K, d^{w_K})$ を用いて信頼度の評価を行った．音節モデルの事後確率による信頼度は

$$CM_{Post}(w_K, d_i^{w_K}) = \frac{\sum_{n=1}^{N} p(x|s^n)\cdot p(s^n)\cdot \delta(w_K, \mathcal{L}_{d_i^{w_K}}^{w^n})}{\sum_{n=1}^{N} p(x|w^n)\cdot p(w^n)}$$

として定義される．

単語音声認識 N-best の結果に対して直接単語事後確率を用いた結果 $\boldsymbol{W}^{Post}$ と，提案手法である組合せ $\boldsymbol{W}^{Post}$ の事後確率の最も高い $\boldsymbol{W}^{Post}_{(1)}$ に対して音節 N-best を用いた事後確率 $\boldsymbol{S}^{Post}$ を用いて信頼度付与した結果から計算される検出区間の集合を以下のように示す．

$\boldsymbol{W}^{Post}_{(T)}$ 単語音声認識 N-best
全検出評価用データに対する各音節音声認識 N-best 出力に対し，単語事後確率が $CM_{Post}(w_K, d^{w_K}) \geq T$ となる音声区間の集合

$\boldsymbol{G}^{Post}_{(T)}$ 単語・音節音声認識組合せ（提案手法）
$\boldsymbol{G}^{Post}_{(T)} = \boldsymbol{W}^{Post}_{(1)} \cap \boldsymbol{S}^{Post}_{(T)}$．単語認識結果がキーワード W_K が単語事後確率が最大となる単語列 $\boldsymbol{W}^{Post}_{(1)}$ に一致し，かつ音節認識結果の事後確率 $CM_{Post}(s_K, d^{w_K}) \geq T$ となる音声区間の集合．

表 7 $\boldsymbol{W}^{Post}$ モデルおよび $\boldsymbol{G}^{Post}$ モデルの再現率および適合率
Table 7 Recall and Precision of $\boldsymbol{W}^{Post}$and $\boldsymbol{G}^{Post}$.

	$T=1$			$T=0$		
モデル	再現率	適合率	F 値	再現率	適合率	F 値
$\boldsymbol{W}^{Post}$	0.346	0.924	0.503	**0.900**	0.236	0.374
$\boldsymbol{S}^{Post}$	**0.433**	0.494	0.461	0.849	0.189	0.309
$\boldsymbol{G}^{Post}$	0.272	**0.959**	**0.423**	0.331	**0.943**	**0.490**

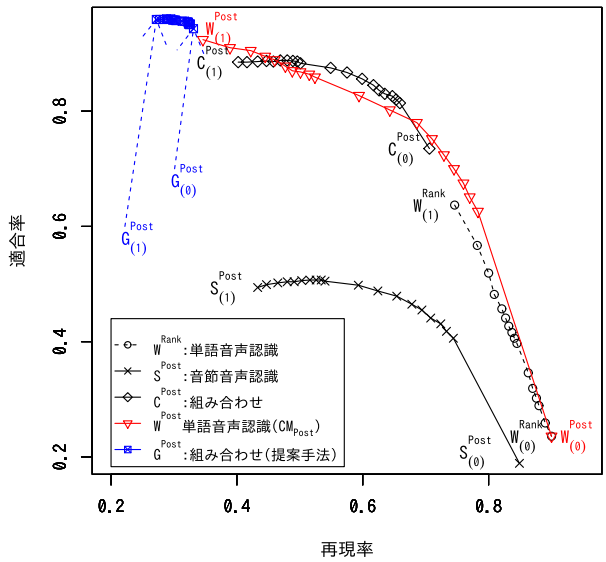


図 5 $\boldsymbol{W}^{Post}$ モデルおよび $\boldsymbol{G}^{Post}$ モデルの再現率–適合率曲線
Fig. 5 Recall – Precision curves of $\boldsymbol{W}^{Post}$and $\boldsymbol{G}^{Post}$.

表 **7** に閾値を最大・最小とした際の性能を示し，図 **5** に閾値を $T=1 \rightharpoonup 0$ と段階的に変化させていった結果を示す．表 7 の $\boldsymbol{W}^{Post}$ に着目すると閾値が $T=1$ とした場合 $\boldsymbol{W}^{Post}_{(1)}$ の適合率は $\boldsymbol{C}^{Post}_{(1)}$ の適合率より高く，0.924 であった．さらに $\boldsymbol{W}^{Post}_{(1)}$ で検出された音声区間において $\boldsymbol{S}^{Post}$ を用いて信頼度付与を行ったところ，$\boldsymbol{W}^{Post}_{(1)}$ の最大適合率 0.924 から 0.959 に 0.035 ポイント向上し，46%の適合率向上を示している．さらに，$\boldsymbol{W}^{Post}$ モデルと $\boldsymbol{C}^{Post}$ モデルのそれぞれの閾値を独立に変えて組み合わせた場合の F 値の最も高い点をつないだ結果を図 **6** に示し $\boldsymbol{G}^{PostMax}$ とする．これは本提案手法にて得られる最も高い性能を示す．

また，キーワードの文字列長の違いにより再現率・適合率に違いがあるかを調べるため，40 語のキーワードを文字列長が 4 文字以下のキーワード 20 語と 5 文字以上のキーワード 20 語に分け，それぞれの再現率・適合率・F 値を調べた．表 **8** に結果を示す．$\boldsymbol{W}^{Post}$ モデルに着目すると，文字列長の長いキーワードの F 値が高いが $T=0$ におけ

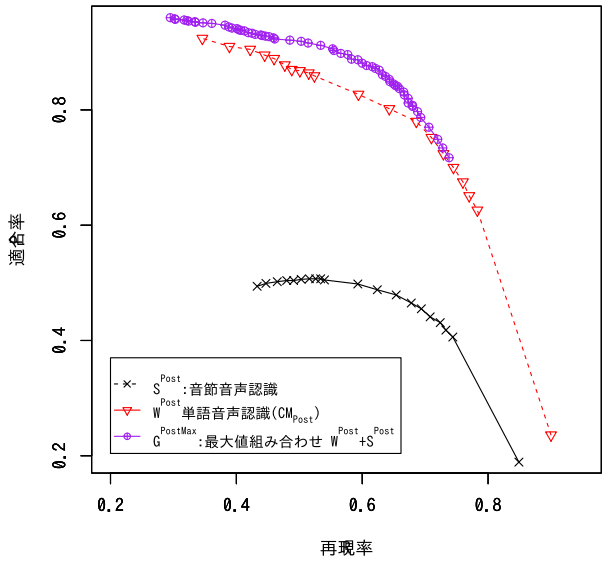


図 6 $\boldsymbol{W}_{(T)}^{Post}$ と $\boldsymbol{S}_{(T)}^{Post}$ を用いて得られる最も F 値の高い再現率–適合率曲線

Fig. 6 Recall – Precision curve that has highest F-values using $\boldsymbol{W}_{(T)}^{Post}$ and $\boldsymbol{S}_{(T)}^{Post}$.

表 8 単語の文字列長の違いによる再現率および適合率

Table 8 Recall and Precision of long keyword and short keyword.

モデル	$T=1$ 再現率	$T=1$ 適合率	$T=1$ F 値	$T=0$ 再現率	$T=0$ 適合率	$T=0$ F 値
	文字列長 4 文字以下					
$\boldsymbol{W}^{Post}$	0.336	0.921	0.493	**0.910**	0.218	0.352
$\boldsymbol{S}^{Post}$	**0.465**	0.423	0.443	**0.856**	0.161	0.271
$\boldsymbol{G}^{Post}$	**0.282**	**0.961**	**0.436**	0.325	0.942	0.483
	文字列長 5 文字以上					
$\boldsymbol{W}^{Post}$	**0.365**	**0.930**	**0.524**	0.879	**0.284**	**0.430**
$\boldsymbol{S}^{Post}$	0.369	**0.844**	**0.513**	0.836	**0.294**	**0.436**
$\boldsymbol{G}^{Post}$	0.250	0.954	0.396	**0.343**	**0.944**	**0.503**

る再現率は文字列長の短いキーワードの方が高い．また $\boldsymbol{S}^{Post}$ に関しては $T=1$，$T=0$ のどちらにおいても文字列長の短いキーワードが，再現率は高いが適合率は大幅に低い結果となっている．文字列の短いキーワードは音節長も短く，同じ音節列を持つ別の語（文字列の異なる語）に一致しやすく，適合率を下げる要因になっている．$\boldsymbol{G}^{Post}$ モデルでは，$T=0$ における F 値は文字列長の長いキーワードのほうが高いが，これは $\boldsymbol{W}^{Post}$ の $T=1$ の値に依存して高くなっている．また $\boldsymbol{G}^{Post}$ モデルでの $T=1$ の F 値は文字列長の短いキーワードのほうが高く，$\boldsymbol{G}^{Post}$ モデルで用いている $\boldsymbol{S}^{Post}$ モデルの再現率の高さが正しい信頼度付与に寄与しているものと考えられる．

4. 考察

大規模なコールセンター [19] の例では，通話時間が約 10 分程度の問合せが月に 35 万件あり，平日のみの受電と仮定すると 1 日あたり約 17,000 件．約 3,000 時間の音声がコールモニタリングの対象となる．これら 1 日数千時間の音声に対して音声認識処理を日々行うためには，100 台規模の大きな計算資源が必要となる．本システムもこのような規模のコールセンターデータを対象としており，実験の結果をふまえ，本研究の目的とする「計算量を大幅に増やさない」，「作業効率を向上させる」，「高い適合率に調整が可能である」という観点から考察を行う．

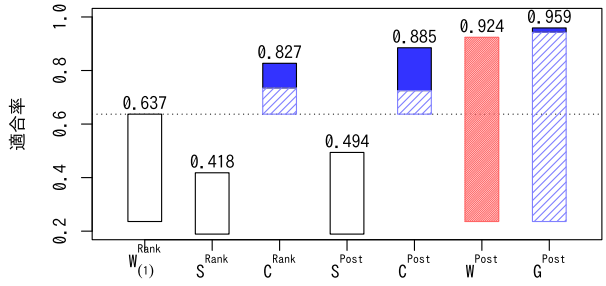


図 7 適合率の範囲

Fig. 7 Range of precision.

4.1 適合率の範囲と作業効率

実験の結果，音節音声認識による信頼度の付与を行うことで検出の適合率を高めることができた．すでに認識された結果を用いるので，再認識処理することなく検出時に閾値を指定することにより適合率の高い区間の音声をチェックできる．前章の実験で得られた適合率の区間を図 7 にまとめる．横軸に平行な 適合率 = 0.637 の破線は単語 1-best で得られる適合率を示す．$\boldsymbol{C}^{Rank}$，$\boldsymbol{C}^{Post}$，$\boldsymbol{G}^{Post}$ はそれぞれベースシステムとの組合せとして用いられるので，それぞれのベースモデルの適合率の下端から $T=0$ における適合率の区間を斜線の区間にて示す．単語 1-best である $\boldsymbol{W}_{(1)}^{Rank}$ に対して，$\boldsymbol{C}^{Rank}$ においては 0.190，$\boldsymbol{C}^{Post}$ においては 0.248 ポイント適合率を向上させることができている．同様に，$\boldsymbol{W}^{Post}$ に比べて $\boldsymbol{G}^{Post}$ では 0.035 ポイント適合率を向上させることができている．

また，最終的な検出結果の確認（キーワードが実際に音声区間と一致しているか）として人手による聴取作業を前提としている．したがって，できるだけ誤検出が少なくかつ全体を網羅できると効率的に作業が行える．実際のコールモニタリング作業を模した作業効率により評価を行った．再現率・適合率の結果を用いて，図 8 に信頼度上位から音声を聴取した際にどれだけ誤った音声が含まれるかを示した．横軸に検出区間数，縦軸にその検出区間に含まれる誤り数を示す．単語事後確率 $\boldsymbol{W}^{Post}$ および音節事後確率 $\boldsymbol{S}^{Post}$ の結果をベースにした $\boldsymbol{G}^{PostMax}$ の検出誤りが全区間において最も少ないことが分かる．

正解であるのべ 3,148 個のキーワードの 25%，50%を網羅する再現率 0.250，0.500 のポイントにおいて誤検出数を比較した結果を表 9 に示す．再現率 0.500 の場合，単語

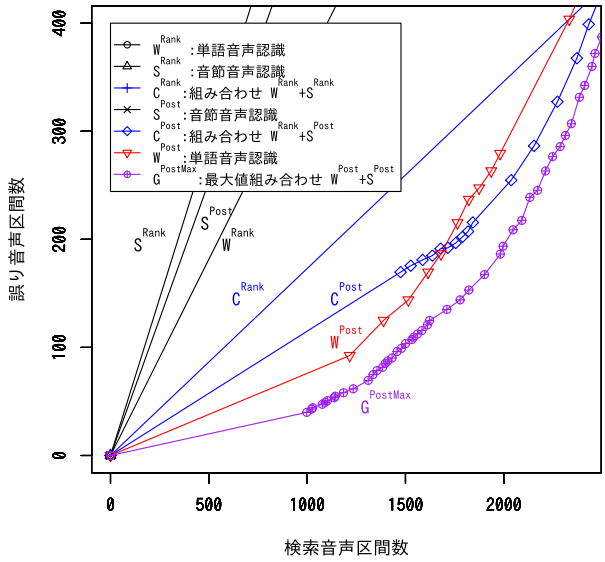


図 8 検出区間に対する誤り数
Fig. 8 Number of false-alarm to matched result.

表 9 再現率 0.250, 0.500 のポイントにおける誤り検出数
Table 9 Number of false-alarm at 0.250 and 0.500 recall rate.

	再現率 0.250%（正解数 812）		再現率 0.500%（正解数 1,624）	
モデル	検出数	誤検出数	検出数	誤検出数
$\boldsymbol{W}^{Rank}_{(1)}$	1,274	462	2,550	926
$\boldsymbol{C}^{Rank}$	970	**158**	1,940	**316**
$\boldsymbol{C}^{Post}$	917	**105**	1,839	**215**
$\boldsymbol{W}^{Post}$	879	**67**	1,871	247
$\boldsymbol{G}^{PostMax}$	846	**34**	1,767	**143**

1-best である $\boldsymbol{W}^{Rank}_{(1)}$ と提案法を比較すると，$\boldsymbol{W}^{Rank}$ の場合 2,550 個の検出に対して 1,624 個の正解と 926 個の誤りを含むのに対し，ランキングを信頼度とした $\boldsymbol{C}^{Rank}$ の場合 316 個，事後確率を信頼度とした $\boldsymbol{C}^{Post}$ の場合 215 個の誤り数となり，それぞれ 65.8%，76.7%の誤検出を減らすことができている．また単語 N-best を用いた $\boldsymbol{W}^{Rank}$ と提案法である $\boldsymbol{G}^{PostMax}$ を比較した場合，それぞれ 247 個の誤りと 143 個の誤りを含み 42.1%の誤検出を減らしている．再現率 0.250 の場合では $\boldsymbol{G}^{PostMax}$ の誤りが最も少なく次いで $\boldsymbol{W}^{Post}$ の誤り数が次に少ない．この場合でも提案法 $\boldsymbol{G}^{PostMax}$ は単語 N-best のみを用いた $\boldsymbol{W}^{Post}$ に比べて誤りを 49.2%減らすことができている．どちらの場合においても本手法にて大幅に作業効率が削減できていることが分かる．

4.2 計算時間

単語音声認識と音節音声認識の組合せにより，未知語や認識誤りへの対応だけでなく辞書語の検出性能向上にも寄与できることが分かったが，前述したように大量の音声を限られた時間で処理するためには計算量に対して考慮する必要がある．コールモニタリングを目的とした場合，一般的な検索エンジンのように自由なキーワードを検出時に入力するのではなく，あらかじめ設定しておいたキーワードに対して検出を行う場合が多い．キーワードが既知であるため，$\boldsymbol{C}^{Rank}$，$\boldsymbol{C}^{Post}$ ではあらかじめ $\boldsymbol{W}^{Rank}_{(1)}$ により絞り込まれた検出区間のみに対して $CM_{Post}(s_K, d^{w_K})$ の計算（音節音声認識および音節 N-best の出力）を行えばよい．また $\boldsymbol{G}^{Post}$ では単語 N-best に含まれる検出区間のみに対して $CM_{Post}(s_K, d^{w_K})$ の計算を行えばよい．CPU が Intel Xeon 3.16 G Hz，メモリ 4 GB の計算機を用いて実際に計測を行った結果，システムで用いた音声認識器では，単語音声認識 1-best 出力に要する時間を r とすると単語 N-best 出力に $1.37\,r$，音節音声認識 1-best に $1.77\,r$，音節 N-best 出力に $5.08\,r$ 要していた．なお，文字列検出および信頼度計算のステップは音声認識時間に比べると小さいため，これらの数字には文字列検出および信頼度計算の時間も含まれるものとする．**表 10** に，インデックス作成およびキーワード検出に要する実行時間をモデルごとに示す．表 10 中の a は $\boldsymbol{W}^{Rank}_{(1)}$ により絞り込まれた検出区間の全検出対象データに対する割合，b は単語 N-best にキーワードが含まれる検出区間の全検出対象データに対する割合である．本実験で用いた検出評価用データおよび検出キーワードの場合 40 語の合計で $a = 0.159$，$b = 0.566$ であった．この場合の実行時間を条件 1 の列に記す．一方，実際のコールセンターのモニタリングでは，検出対象の音声が非常に大量ではあるものの，検出キーワードは低頻度で出現する語が用いられる．このような場合 a，b はともに 0 に近くなり，信頼度付与のためのインデックス作成に要する時間は無視できる．$a = 0$，$b = 0$ とした場合の実行時間を条件 2 の列に示す．

表 10 モデルの違いに対する計算時間
Table 10 Computation time of each model.

モデル	実行時間	条件 1	条件 2
$\boldsymbol{W}^{Rank}_{(1)}$	r	r	r
$\boldsymbol{S}^{Rank}$, $\boldsymbol{S}^{Post}$	$1.77\,r$	$1.77\,r$	$1.77\,r$
$\boldsymbol{C}^{Rank}$, $\boldsymbol{C}^{Post}$	$r + a \times 5.08\,r$	$2.81\,r$	$1.00\,r$
$\boldsymbol{W}^{Post}$	$1.37\,r$	$1.37\,r$	$1.37\,r$
$\boldsymbol{G}^{Post}$	$1.37\,r + b \times 5.08\,r$	$4.24\,r$	$1.37\,r$

条件によって計算時間は大きく異なるが，比較的頻度の高いキーワードを検出対象とした場合，全検出対象データに対して単語 N-best を計算する $\boldsymbol{W}^{Post}$ が検出効率および計算時間の面から有利である．一方，コールモニタリング業務のような大量のデータから低頻度のキーワードを検出する場合は，計算時間にほとんど影響をおよぼさず，作業効率の良い $\boldsymbol{C}^{Post}$ が有利となる．また $\boldsymbol{G}^{Post}$ は計算資源に余裕があり同様に低頻度のキーワードを検出する場合は人手による聴取作業を最小限にすることができる．$\boldsymbol{S}^{Rank}$，$\boldsymbol{S}^{Post}$ 単体では検出効率および実行時間の面からはあまり

実用的ではなかった．

5. おわりに

今回我々は，「計算量を大幅に増やさない」「作業効率を向上させる」「高い適合率に調整が可能である」という観点で単語音声認識結果と音節音声認識結果を組み合わせたシステムを提案した．本論文では，既知語の適合率改善のために音節音声認識の結果を用い，ランキング，事後確率による信頼度のいずれにおいても適合率の高い音声区間の検出を試みた．この信頼度を利用することでユーザは適合率の高い音声のみを検出漏れをできるだけ少なくチェックといった作業が可能になる．有効性の評価は，各システムで得られる適合率の範囲と，実際のコールモニタリング作業を模した作業効率（信頼度上位から音声を聴取した際に誤った音声が含まれる割合）により行った．信頼度の計算方法として，ランキングと単語事後確率とを比較すると，事後確率のほうが適合率の上限が高いが，ランキングの場合，必要な適合率の幅に応じて計算量を減らすことができる．出力を 10-best に限定して適合率の幅を減らすと，事後確率に基づく場合の 0.32 倍程度の計算量でキーワードの検出ができる．ベースシステムとして単語音声認識 1-best を用いた場合と，単語音声認識 1-best に音節音声認識 N-best を用いて信頼度計算を行った場合とを比較すると，適合率の上限は最大 0.248 ポイント向上し，作業効率の観点からは誤検出の割合を 76.7%減らすことができた．また，ベースシステムとして単語認識 N-best を用いて信頼度計算を行った結果に対してさらに音節音声認識 N-best を用いて信頼度計算を行うことで，適合率の上限は 0.035 ポイント向上し，作業効率の観点からは誤検出の割合を 42.1%減らすことができた．一方，ベースシステムにかかわらず音節認識 N-best を用いて信頼度計算を行う手法は信頼度計算にコストがかかるが，信頼度計算のコストはベースシステムにより検出されるキーワードの頻度に比例するため，一般的なコールモニタリングで用いられる低頻度のキーワードを対象にした場合，全体の計算コストにほとんど影響を与えることなく，高い作業効率を実現することができる．

今後は，音声認識を用いた情報抽出アプリケーションの実用性向上という観点で，本研究の目的であった適合率の高い効率的なキーワードの検出を含め，アプリケーションの効率化に取り組んでいきたい．

参考文献

[1] Amir, A., Efrat, A. and Srinivasan, S.: Advances in Phonetic Word Spotting, *Proc. 10th International Conference on Information and Knowledge Management* (*CIKM '01*), pp.580–582 (2001).

[2] Mamou, J. and Ramabhadran, B.: Vocabulary Independent Spoken Term Detection, *The 30th Annual International ACM SIGIR Conference* (*SIGIR 2007*), pp.615–622 (2007).

[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] Liu, P., Tian, Y., Zhou, J. and Soong, F.: Background Model Based Posterior Probability for Measuring Confidence, *The 9th European Conference on Speech Communication and Technology* (*INTERSPEECH2005*), pp.1465–1468 (2005).

[5] 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).

[6] 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).

[7] Fiscus, J.: A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), *1997 IEEE Workshop on Automatic Speech Recognition and Understanding* (*ASRU 1997*), pp.347–354 (1997).

[8] 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).

[9] Wessel, F., Schluter, R., Macherey, K. and Ney, H.: Confidence measures for large vocabulary continuous speech recognition, *IEEE Trans. Speech and Audio Prcesing*, Vol.9, No.3, pp.288–298 (2001).

[10] Rueber, B.: Obtaining confidence measures from sentence probabilities, *5th European Conference on Speech Communication and Technology* (*EUROSPEECH 1997*), pp.739–742 (1997).

[11] 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).

[12] 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).

[13] Jiang, H.: Confidence measures for speech recognition: A survey, *Speech Communications*, Vol.45, No.4, pp.455–470 (2005).

[14] Roark, B., Saraclar, M. and Collins, M.: Corrective language modeling for large vocabulary ASR with the perceptron algorithm, *2004 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP2004*), pp.749–752 (2004).

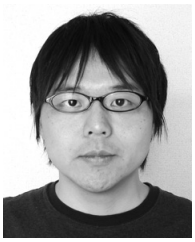
[15] Vergyri, D., Shafran, I., Stolcke, A., Gadde, R.R., Akbacak, M., Roark, B. and Wang, W.: The SRI/OGI 2006 Spoken Term Detection System, *The 8th Conference in the Annual Series of INTERSPEECH* (*INTERSPEECH2007*), pp.2393–2396 (2007).

[16] Akbacak, M., Vergyri, D. and Stolcke, A.: Open-vocabulary spoken term detection using graphone-based hybrid recognition systems, *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP2008*), pp.5240–5243 (2008).

[17] Chen, S.: Conditional and Joint Models for Grapheme-to-Phoneme Conversion, *2003 European Conference on Speech Communication and Technology* (*EUROSPEECH2003*), pp.2033–2036 (2003).

[18] Hahn, S., Lehnen, P., Wiesler, S., Schluter, R. and Ney, H.: Improving LVCSR with Hidden Conditional Ran-

dom Fields for Grapheme-to-Phoneme Conversion, *The 14th Interspeech Conference* (*INTERSPEECH2013*), pp.495–499 (2013).
[19] 厚生労働省：(30) コールセンター事業（年金電話相談事業）(2014)，入手先 ⟨http://www.mof.go.jp/budget/topics/budget_execution_audit/fy2014/sy2607/2607d.html⟩.



長野 徹 （正会員）

1998 年筑波大学大学院工学研究科修士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，自然言語処理および音声言語処理の研究に従事．日本音響学会会員．

倉田 岳人 （正会員）

2004 年東京大学大学院情報理工学系研究科修士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，音声認識および自然言語処理に関する研究に従事．2013 年東京大学大学院情報理工学系研究科より博士号取得．日本音響学会会員．博士（情報理工学）．

鈴木 雅之

2013 年東京大学大学院工学系研究科博士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，音声認識および自然言語処理に関する研究に従事．IEEE，ICSA，電子情報通信学会，日本音響学会各会員．博士（工学）．

立花 隆輝

1998 年東京大学大学院工学系研究科修士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，音楽電子透かしおよび音声合成・音声認識の研究に従事．2007 年大阪大学大学院工学研究科博士課程修了．電子情報通信学会，日本音響学会各会員．博士（工学）．

西村 雅史

1983 年大阪大学大学院基礎工学研究科博士前期課程修了．同年日本アイ・ビー・エム（株）入社．同社東京基礎研究所にて，音声認識等の音声言語情報処理の研究に従事．2014 年静岡大学大学院情報学研究科教授．IEEE，電子情報通信学会，日本音響学会，人工知能学会各会員．博士（工学）．