(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5530729号
(P5530729)
(45)発行日　平成26年6月25日(2014.6.25)　(24)登録日　平成26年4月25日(2014.4.25)

(51)Int.Cl.　F I
G10L 15/10 (2006.01)　G10L 15/10 500T

請求項の数 6　(全 24 頁)

(21)出願番号　特願2010-11175(P2010-11175)
(22)出願日　平成22年1月21日(2010.1.21)
(65)公開番号　特開2010-170137(P2010-170137A)
(43)公開日　平成22年8月5日(2010.8.5)
審査請求日　平成24年11月27日(2012.11.27)
(31)優先権主張番号　61/146,739
(32)優先日　平成21年1月23日(2009.1.23)
(33)優先権主張国　米国(US)

(73)特許権者　000005326
本田技研工業株式会社
東京都港区南青山二丁目１番１号
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100108578
弁理士　高橋　詔男
(74)代理人　100146835
弁理士　佐伯　義文
(74)代理人　100094400
弁理士　鈴木　三義
(74)代理人　100107836
弁理士　西　和哉
(74)代理人　100108453
弁理士　村山　靖彦

最終頁に続く

(54)【発明の名称】音声理解装置

(57)【特許請求の範囲】
【請求項１】
Ｎ個（Ｎは２以上の整数）の言語モデルそれぞれを使用して発話の音声認識を行ない、前記音声認識により得られたＮ個の音声認識結果を出力する音声認識部と、
Ｍ個（Ｍは２以上の整数）の言語理解モデルそれぞれを使用して、前記音声認識部から出力された前記Ｎ個の音声認識結果それぞれの言語理解を行ない、前記言語理解により得られたＮ×Ｍ個の音声理解結果を出力する言語理解部と、
前記言語理解部から出力された前記Ｎ×Ｍ個の音声理解結果であるコンセプトの集合それぞれについて、前記音声理解結果の確からしさを数値化した発話単位信頼度を、前記音声理解結果の特徴を表す値に基づいて算出し、算出された前記発話単位信頼度が最も高い前記音声理解結果を選択する統合部と、
を備えることを特徴とする音声理解装置。
【請求項２】
前記音声理解結果の特徴を表す値は、発話の長さ、前記音声認識を行ったときに得られた音響スコア、前記音声理解結果に含まれるコンセプトの数、前記コンセプトの信頼度、音声理解結果が得られたか否か、及び、前記音声理解結果が肯定発話か否定発話であるかに基づいて得られる値のうち一以上であることを特徴とする請求項１に記載の音声理解装置。
【請求項３】
前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、既知の発話か

ら得られた前記音声理解結果の前記特徴を表す値と、前記音声理解結果が正解であるか否かを表す値とに基づいて、尤度が最大となるように前記特徴を表す値の重みを決定する学習部をさらに備える、

ことを特徴とする請求項１または請求項２に記載の音声理解装置。

【請求項４】

前記学習部は、前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、決定した前記特徴の重みに基づいて他の前記特徴と相関が高い前記特徴を選択し、選択した前記特徴のうち１つを前記発話単位信頼度の算出に用いる、

ことを特徴とする請求項３に記載の音声理解装置。

【請求項５】

前記学習部は、前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、前記発話単位信頼度の算出において所定より影響の小さい前記特徴を選択し、選択した前記特徴を前記発話単位信頼度の算出に用いる前記特徴から除外する、

ことを特徴とする請求項３または請求項４に記載の音声理解装置。

【請求項６】

前記学習部は、前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、前記特徴を表す値を用いたロジスティック回帰式によって前記発話単位信頼度を算出する、

ことを特徴とする請求項３から請求項５のいずれか１項に記載の音声理解装置。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、音声理解装置に関する。

【背景技術】

【０００２】

量的爆発・質的複雑化する情報へのアクセス手段として音声は有望な一手段であり、それを可能にする音声対話システムの開発・運用が行われている。音声対話システムではユーザの発話から得られた意味表現に基づいて応答を生成するため、発話を意味表現に変換する音声理解部が重要である。音声理解は，音声を単語列に変換する音声認識と、単語列を意味表現に変換する言語理解の二つのプロセスからなる。音声認識には音響モデルと言語モデルが必要であるが、音響モデルは音声対話システムのタスクドメインには依存しない。そのため、言語モデルと言語理解モデルを、ドメインごとに必要な対象として考えることができる。

【０００３】

単一の言語モデルと言語理解モデルによる音声理解方式のみを用いる場合では、多様な発話に対して高精度な音声理解を実現することは難しい。これは、発話によって適した言語モデル・言語理解モデルの組み合わせが異なるからである。例えば、音声認識の言語モデルとして文法モデルを用いた場合は、文法内の発話に対して高精度な音声認識が可能となる。しかし、想定外の発話に対して頑健でない。Ｎ－ｇｒａｍモデルは、文法ベースの言語モデルと比較すると、局所的な制約であり、未登録語や認識誤りが生じても回復が容易であるという利点がある。ただし、文全体の制約を表現できないため、一般に想定内の発話に対する性能は文法モデル使用時と比較して低い。言語理解モデルにも、同様に一長一短があるため、正しく理解できる発話を増やすには、複数の言語モデル・言語理解モデルを組み合わせることが有効だと考えられる。

音声理解方式を複数用いると、理解結果が複数得られるため、それら複数の理解結果から最終的な理解結果を求める必要がある。従来は、ＲＯＶＥＲ（Recognizer Output Voting Error Reduction）法のように多数決が用いられることが多かった（例えば、非特許文献１参照）。

【先行技術文献】

【非特許文献】

【０００４】
【非特許文献１】Jonathan G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," Proc. ASRU, pp.347-354, 1997.
【発明の概要】
【発明が解決しようとする課題】
【０００５】
上述したＲＯＶＥＲ法では、複数の音声認識結果や理解結果に対して、重み付き多数決を行い、最終的な結果を得る。多数決では、音声理解性能の高い方式と低い方式とが混在すると、高性能な方式の結果が十分に反映されなくなる場合がある。例えば、多数の音声理解結果が不正解であり、少数の音声理解結果が正解である場合には、正解である音声理解結果が得られる可能性は低い。
【０００６】
本発明は、このような事情を考慮してなされたものであり、発話を高精度に音声理解する音声理解装置を提供することにある。
【課題を解決するための手段】
【０００７】
上記問題を解決するために、請求項１に記載した発明は、Ｎ個（Ｎは２以上の整数）の言語モデルそれぞれを使用して発話の音声認識を行ない、前記音声認識により得られたＮ個の音声認識結果を出力する音声認識部（例えば、実施形態における音声認識部２０）と、Ｍ個（Ｍは２以上の整数）の言語理解モデルそれぞれを使用して、前記音声認識部から出力された前記Ｎ個の音声認識結果それぞれの言語理解を行ない、前記言語理解により得られたＮ×Ｍ個の音声理解結果を出力する言語理解部（例えば、実施形態における言語理解部３０）と、前記言語理解部から出力された前記Ｎ×Ｍ個の前記音声理解結果であるコンセプトの集合それぞれについて、前記音声理解結果の確からしさを数値化した発話単位信頼度を、前記音声理解結果の特徴を表す値に基づいて算出し、算出された前記発話単位信頼度が最も高い前記音声理解結果を選択する統合部（例えば、実施形態における統合部４０）と、を備えることを特徴とする音声理解装置である。
これにより、複数の言語モデルと、複数の言語理解モデルとの全ての組み合わせを用いて発話を音声理解した結果を得、この得られた音声理解結果それぞれについて、音声理解結果の特徴を表す値から、音声理解結果の確からしさを数値として比較可能な発話単位信頼度を算出する。そして、算出した発話単位信頼度を比較し、複数の言語モデルと複数の言語モデルの全ての組み合わせを用いて音声理解した中から最も正解である確率が高い音声理解結果を選択する。
【０００８】
請求項２に記載した発明は、請求項１に記載の音声理解装置であって、前記音声理解結果の特徴を表す値は、発話の長さ、前記音声認識を行ったときに得られた音響スコア、前記音声理解結果に含まれるコンセプトの数、前記コンセプトの信頼度、音声理解結果が得られたか否か、及び、前記音声理解結果が肯定発話か否定発話であるかに基づいて得られる値のうち一以上であることを特徴とする。
これにより、異なる言語モデルや異なる言語理解モデルを用いた場合でも共通して得ることができる特徴量に基づいて発話単位信頼度を算出する。
【０００９】
請求項３に記載した発明は、請求項１または請求項２に記載の音声理解装置であって、前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、既知の発話から得られた前記音声理解結果の前記特徴を表す値と、前記音声理解結果が正解であるか否かを表す値とに基づいて、尤度が最大となるように前記特徴を表す値の重みを決定する学習部（例えば、実施形態における学習部５０）をさらに備える、ことを特徴とする。
これにより、学習データについて得られた特徴を表す値及び言語理解結果に基づいて、発話単位信頼度の算出に用いる各特徴の重みを、言語モデルと言語理解モデルの組み合わ

せに応じて決定する。

【００１０】

請求項４に記載した発明は、請求項３に記載の音声理解装置であって、前記学習部は、前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、決定した前記特徴の重みに基づいて他の前記特徴と相関が高い前記特徴を選択し、選択した前記特徴のうち１つを前記発話単位信頼度の算出に用いる、ことを特徴とする。

これにより、言語モデルと言語理解モデルの組み合わせ毎に、発話単位信頼度の算出に用いる特徴を独立変数とする。

【００１１】

請求項５に記載した発明は、請求項３または請求項４に記載の音声理解装置であって、前記学習部は、前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、前記発話単位信頼度の算出において所定より影響の小さい前記特徴を選択し、選択した前記特徴を前記発話単位信頼度の算出に用いる前記特徴から除外する、ことを特徴とする。

これにより、言語モデルと言語理解モデルの組み合わせに応じて、発話単位信頼度の算出に寄与しない特徴を用いることなく、発話単位信頼度を算出する。

【００１２】

請求項６に記載した発明は、請求項３から請求項５のいずれか１項に記載の音声理解装置であって、前記学習部は、前記Ｎ個の言語モデル及び前記Ｍ個の言語理解モデルの組み合わせ毎に、前記特徴を表す値を用いたロジスティック回帰式によって前記発話単位信頼度を算出する、ことを特徴とする。

これにより、発話単位信頼度を、言語モデル及び言語理解モデルの組み合わせ毎に特徴を重み付けしたロジスティック回帰式により算出し、異なる言語モデル及び言語理解モデルの組み合わせ間で定量的に比較可能な発話単位信頼度を得る。

【発明の効果】

【００１３】

請求項１に記載した発明によれば、複数の言語モデルと複数の言語理解モデルとの全ての組み合わせを用いて発話を音声理解した結果の中から、性能の低いモデルの影響を受けることなく、最も発話単位信頼度が高い結果を選択することができる。よって、言語モデル、言語理解モデルのいずれか片方を複数用いたときより、高精度な音声理解結果を得ることができる。

また、請求項２の発明によれば、異なる言語モデル、異なる言語理解モデルを用いた場合であっても、共通して取得することができる特徴を用いて発話単位信頼度を算出するため、任意の言語モデルや言語理解モデルを実装することができる。

また、請求項３の発明によれば、言語モデルと言語理解モデルの組み合わせ毎に、発話単位信頼度を精度よく算出するための各特徴の重みを決めることができる。

また、請求項４の発明によれば、多重共線性を除去し、発話単位信頼度を精度よく算出することができる。

また、請求項５の発明によれば、発話単位信頼度を、貢献度が低い特徴については用いずに算出することができるため、計算処理の負荷を低くすることができる。

また、請求項６の発明によれば、発話単位信頼度をロジスティック回帰式により算出するため、言語モデルと言語理解モデルのあらゆる組み合わせ間において定量的に比較が可能な発話単位信頼度を精度よく算出することができる。

【図面の簡単な説明】

【００１４】

【図１】本発明の一実施形態による音声理解装置の機能ブロック図である。

【発明を実施するための形態】

【００１５】

以下、図面を参照して本発明の一実施形態を説明する。

【００１６】

[１．本発明の実施形態の概要]

本発明の一実施形態による音声理解装置は、例えば音声対話システムに組み込まれ、複数の言語モデルと複数の言語理解モデルを用いることで、高精度な音声理解を行う。なお、音声認識と言語理解を行うことを音声理解とよび、言語モデルを用いて音声認識した結果を、言語理解モデルを用いて言語理解した結果を音声理解結果とよぶ。ユーザの発話によって適した言語モデルと言語理解モデルの組み合わせは異なることから、単一の音声理解方式で様々な発話に対して高精度な音声理解を実現することは難しい。そこで本実施形態では、まず、複数の言語モデルと言語理解モデルを用いて複数の音声理解結果を得ることで、音声理解結果の候補を得る。次に、得られた複数の音声理解結果に対して、ロジスティック回帰に基づき発話単位信頼度を付与し、その発話単位信頼度が最も高い音声理解結果を選択する。

【００１７】

本実施形態による音声理解装置を用いた音声理解の評価実験では、言語モデルとして、文法モデルとＮ－ｇｒａｍモデルの２種類を用い、言語理解モデルとして、Finite-State Transducer（ＦＳＴ）、Weighted FST（ＷＦＳＴ）、及び、Keyphrase-Extractorの３種類を用いた。この評価実験によれば、本実施形態の音声理解装置によって、言語モデルと言語理解モデルのいずれかを複数用いた場合と比較して、コンセプト理解精度の向上が得られた。また、従来のＲＯＶＥＲ法による音声理解結果の統合と比較し、本実施形態の音声理解装置の有効性が認められた。

【００１８】

本実施形態の音声理解装置では、以下の二つの手法を実装した。

（１）複数の言語モデルと言語理解モデルの使用：Multiple Language models and Multiple Understanding models（ＭＬＭＵ）

（2）音声理解結果の発話単位の信頼度に基づく選択：Confidence-Measure-Based Selection（ＣＭＢＳ）

【００１９】

ＭＬＭＵでは、複数の言語モデルと複数の言語理解モデルを用いて両者のあらゆる組み合わせによる音声理解を行う。これにより、音声認識と言語理解の適した組み合わせによる音声理解結果が得られる。また、後者のＣＭＢＳでは、得られた複数の音声理解結果に対し、ロジスティック回帰により発話単位信頼度を付与し、その発話単位信頼度に基づき適した音声理解結果を選択する。選択時に、音声認識と言語理解結果の特徴を用いることで、誤った音声理解結果が最終結果となることを防ぐ。

以下、２．では、関連研究を詳述し、３．では、本実施形態の音声理解装置において複数出力された音声理解結果から適切な結果を選択する手法について述べる。４．では、音声理解装置の実施例において実装した言語モデルと言語理解モデルについて述べ、５．で評価実験の結果を述べ、６．で本実施形態のまとめを述べる。

【００２０】

［２．関連研究］

これまでも、複数の言語モデルや言語理解モデルを用いた手法が開発されてきた。本実施形態と、従来手法との関係を以下の表１に記す。

【００２１】

【表１】

表1 本実施形態と従来手法との関係

	言語モデル	言語理解モデル	統合手法
単純手法	単一	単一	―
手法1、手法2	複数	―	ROVER法
手法4	複数	単一	決定木による選択
手法3	単一	複数	ROVER法
本実施形態	複数	複数	CMBS

【００２２】

表１に示すように、従来は、音声認識と言語理解とが別々に研究されることが多かった。しかし、音声認識・言語理解をそれぞれ向上させたとしても、それらの組み合わせが適さない場合は、音声理解全体としての性能は向上しない。

【００２３】

なお、従来の手法１については非特許文献１に記載されており、手法２～手法４については、それぞれ以下の文献２～４に記載されている。

文献２（手法２）：H. Schwenk and J.-L. Gauvain, "Combining Multiple Speech Recognizers using Voting and Language Model Information," Proc. ICSLP, pp.915-918, 2000.

文献３（手法３）：S. Hahn, P. Lehnen. and H. Ney, "System Combination for Spoken Language Understanding," Proc.Interspeech, pp.236-239, 2008.

文献４（手法４）：安田宜仁、堂坂浩二、相川清明, ”２つの認識文法を用いた主導権混合型対話制御”, 情報処理学会研究報告, pp.127-132, 2002-SLP-40-22, 2002.

【００２４】

また、発話検証のために複数の言語モデルを用いる手法が開発されている（例えば、文献５、文献６参照）。これらの手法では、音声認識結果の発話検証用として、語彙サイズの大きな言語モデルを用いて音声認識を行い、音響尤度などを比較することで認識結果の信頼性を計る。しかし、これらの方法では、言語理解のために用いる音声認識結果は単一の言語モデルに基づく結果だけである。

【００２５】

文献５：西田昌史、寺師弘将、堀内靖雄、市川　熹, “ユーザの発話の予測に基づく音声対話システム”, 情報処理学会研究報告, pp.307-312, 2004-SLP-12-22, 2004.

文献６：K. Komatani and Y. Fukubayashi and T. Ogata and H. G. Okuno, "Introduning Utterance Verification in Spoken Dialogue System to Improve Dynamic Help Generation for Novice Users, "Proc. 8th SIG-dial Workshop on Discourse and Dialogue, pp.202-205, 2007.

【００２６】

異なる言語モデルを複数用いる研究として、非特許文献１に記載の手法１や文献２に記載の手法２がある。これらの研究では音声認識性能の向上のみが目的であり、言語理解は扱っていない。文献４に記載の手法４では、二つの言語モデルを用いて音声認識を行い、どちらの認識結果を用いるかを識別する決定木を構築している。決定木では、学習時に発話ごとに正解ラベルとして音声理解方式を一意に定める必要がある。複数の音声理解方式が同一の結果を出力する場合が頻繁にあり、正解ラベルを一意に定めることができないという問題がある。

【００２７】

複数の言語理解モデルを用いた研究もされている。文献３に記載の手法３では、ある音

声認識結果に対して、複数の言語理解モデルを用いて言語理解結果を出力し、ＲＯＶＥＲ法を用いて最終的な理解結果を出力している。ただし、音声認識時に使用している言語モデルが単一である。

【００２８】

複数の言語モデルと複数の言語理解モデルを用いて音声理解を行ったときの例を表２に記す。

【００２９】

【表２】

表2 複数の言語モデルと複数の言語理解モデルを用いて音声理解を行ったときの例

U1:六月九日です.	
音声認識結果:	
ー文法	“六月九日です.”
ーN-gram	“六月午後のがです.”
音声理解結果:	
ー文法+FST	“month:6 day:9 type:refer-time”
ーN-gram+WFST	“month:6 type:refer-time”
U2:二十日にお借りします.（下線部は文法外）	
音声認識結果:	
ー文法	“二十日二時ごろです.”
ーN-gram	“二十日に十日二時ます.”
音声理解結果:	
ー文法+FST	“day:20 hour:14 type:refer-time”
ーN-gram+WFST	“day:20 type:refer-time”

【００３０】

表２において、言語モデルと言語理解モデルの組み合わせを、「言語モデル＋言語理解モデル」で表す。音声理解結果は、コンセプトの集合であり、各コンセプトは意味スロットとその値、ならびに、発話タイプからなる。表２においては、「month」、「day」、「hour」が意味スロットであり、その値が意味スロットの後ろの「：」に続けて記述されている。例えば、「month:6」の場合、意味スロット「mouth」の値が「6」であることを示している。また、発話タイプは「type」の後ろの「:」に続けて記述されている。

【００３１】

表２に示すように、発話内容Ｕ１「六月九日です．」は文法に沿った発話であるため、文法モデルを用いて音声認識を行い、ＦＳＴを用いて言語理解を行った結果が正解となりやすい。これに対し、発話内容Ｕ２「二十日にお借りします．」は文法外の発話であるため、局所的な制約であるＮ－ｇｒａｍモデルを用いた方が認識精度は高くなる。さらに、言語理解部でＷＦＳＴを用いることで、言語理解に不要な単語や音声認識時の単語信頼度の低い語を棄却しながら、認識結果をシステムの内部表現、つまり、意味表現であるコンセプト列に変換できる。このように複数の音声理解方式を用いることで、発話内容Ｕ１、Ｕ２の両方の発話に対して正しい音声理解結果を得ることができる。

【００３２】

［３．発話単位信頼度に基づく音声理解結果の選択］

ここでは、まず、本発明の一実施形態による音声理解装置の構成について説明する。図１は、本発明の一実施形態による音声理解装置１の機能ブロック図を示す。

同図において、音声理解装置１は、入力部１０、音声認識部２０、言語理解部３０、統合部４０、及び、学習部５０を備えて構成される。

【００３３】
入力部１０は、発話データの入力を受ける。発話データは、ユーザによる発話の音響データである。入力部１０は、例えば、発話データを有線または無線により接続される他の装置から受信してもよく、コンピュータ読み取り可能な記録媒体から読み出してもよい。
【００３４】
音声認識部２０は、音響モデル記憶部２２、音声認識処理部２４－１～２４－Ｎ（Ｎは２以上の整数）及び発話検証用音声認識処理部２６を備える。
【００３５】
音響モデル記憶部２２は、単語列の音響的な特徴を示す統計的モデルである音響モデルを記憶する。
【００３６】
音声認識処理部２４－ｋ（ｋは１以上Ｎ以下の整数）は、ドメイン依存の言語モデルを記憶する言語モデル記憶部２４１－ｋを備えており、言語モデル記憶部２４１－１～２４１－Ｎに記憶される言語モデルはそれぞれ異なる。言語モデルとは、音響データの音声波形に基づいて得られた単語列の音響スコアや結合確率を得るために用いる規則の集合であり、自然言語に対する統計モデルである。
【００３７】
音声認識処理部２４－ｋは、音響モデル記憶部２２に記憶されている音響モデルと、自身の備える言語モデル記憶部２４１－ｋに記憶されている言語モデルとを用いて、入力部１０に入力された発話データを音声認識し、その結果を出力する。音声認識処理部２４－１～２４－Ｎによる音声認識処理は、ドメイン依存の言語モデルを用いた既存技術の音声認識処理とすることができる。音声認識処理部２４－１～２４－Ｎは、単語列により表される音声認識結果、この音声認識結果の単語列に対する音響スコア及び結合確率を言語理解部３０に出力する。なお、単語列は、１単語からなる場合も含むものとする。
【００３８】
発話検証用音声認識処理部２６は、発話検証用言語モデルを記憶する発話検証用言語モデル記憶部２６１を備えている。発話検証用言語モデルとは、特定のドメインに依存しない大語彙統計モデルを用いた言語モデルである。発話検証用音声認識処理部２６は、音響モデル記憶部２２に記憶されている音響モデルと、発話検証用言語モデル記憶部２６１に記憶されている発話検証用言語モデルを用いて、入力部１０に入力された発話データを音声認識し、その結果を出力する。発話検証用音声認識処理部２６による音声認識処理は、大語彙統計モデルを用いた既存技術の音声認識処理とすることができる。発話検証用音声認識処理部２６は、単語列により表される音声認識結果、この音声認識結果の単語列に対する音響スコア及び結合確率を結合部４０に出力する。
【００３９】
言語理解部３０は、言語理解処理部３２－１～３２－Ｍ（Ｍは２以上の整数）と信頼度算出部３４を備え、言語理解処理部３２－ｊ（ｊは１以上Ｍ以下の整数）は、言語理解モデルを記憶する言語理解モデル記憶部３２１－ｊを備える。言語理解モデルとは、単語列からコンセプトを得るための規則の集合であり、言語理解モデル記憶部３２１－１～３２１－Ｍに記憶される言語理解モデルはそれぞれ異なる。言語理解処理部３２－ｊは、言語理解モデル記憶部３２１－ｊに記憶されている言語理解モデルを用いて、音声認識処理部２４－１～２４－Ｎが出力したＮ個の音声認識結果それぞれを言語理解し、コンセプトの集合である音声理解結果を得る。言語理解処理部３２－１～３２－Ｍによる言語理解処理は、既存技術の言語理解処理とすることができる。
【００４０】
信頼度算出部３４は、言語理解処理部３２－１～３２－Ｍそれぞれが、音声認識処理部２４－１～２４－Ｎから出力されたＮ個の音声認識結果を言語理解することによって得られたＮ×Ｍ個の音声理解結果それぞれについて、所定の規則に従い、各音声理解結果に含まれるコンセプトの信頼度を算出する。このコンセプトの信頼度算出には、既存技術を用いることができる。信頼度算出部３４は、各音声理解結果に併せて、各音声理解結果の特

徴量として、音声理解結果に含まれるコンセプト数や、算出した各コンセプトの信頼度を出力するとともに、音声理解に用いた音声認識結果の音響スコアを出力する。

【００４１】

統合部４０は、発話単位信頼度算出部４２と選択部４４からなる。発話単位信頼度算出部４２は、信頼度算出部３４から出力されたＮ×Ｍ個の各音声理解結果の特徴を表す値、つまり、特徴量から、各音声理解結果の発話単位信頼度を算出する。選択部４４は、発話単位信頼度算出部４２によって算出された発話単位信頼度が最も高い音声理解結果を選択し、その選択した音声理解結果を出力する。音声理解結果は、例えば、図示しない他のアプリケーション実行部に出力してもよく、図示しないディスプレイに表示してもよく、紙などに印刷してもよく、有線または無線により接続される他の装置に送信してもよく、コンピュータ読み取り可能な記録媒体に書き込んでもよい。また、選択部４４は、発話単位信頼度が高い順に所定数、音声理解結果とその発話単位信頼度を出力してもよい。

【００４２】

学習部５０は、学習データを用いて、発話単位信頼度算出部４２が発話単位信頼度の算出に用いる特徴量を選択するとともに、その重みを決定する。

【００４３】

なお、上述の音声理解装置１は、内部にコンピュータシステムを有している。そして、音声理解装置１の音声認識部２０、言語理解部３０、統合部４０、及び、学習部５０の動作の過程は、プログラムの形式でコンピュータ読み取り可能な記録媒体に記憶されており、このプログラムをコンピュータシステムが読み出して実行することによって、上記処理が行われる。ここでいうコンピュータシステムとは、ＣＰＵ及び各種メモリやＯＳ、周辺機器等のハードウェアを含むものである。

【００４４】

また、「コンピュータシステム」は、ＷＷＷシステムを利用している場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。

また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムを送信する場合の通信線のように、短時間の間、動的にプログラムを保持するもの、その場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリのように、一定時間プログラムを保持しているものも含むものとする。また上記プログラムは、前述した機能の一部を実現するためのものであっても良く、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるものであっても良い。

【００４５】

上記構成において、音声認識処理部２４－ｋ（ｋは１以上Ｎ以下の整数）は、音響モデル記憶部２２に記憶されている音響モデルと、自身の備える言語モデル記憶部２４１－ｋに記憶されている言語モデルとを用いて、発話検証用音声認識処理部２６は、音響モデル記憶部２２に記憶されている音響モデルと、発話検証用言語モデル記憶部２６１に記憶されている発話検証用言語モデルとを用いて、入力部１０に入力された発話データを音声認識し、その結果を出力する。言語理解処理部３２－ｊ（ｊは１以上Ｍ以下の整数）は、言語理解モデル記憶部３２１－ｊに記憶されている言語理解モデルを用いて、音声認識処理部２４－１～２４－Ｎが出力したＮ個の音声認識結果それぞれを言語理解し、音声理解結果を得る。

【００４６】

以降、各発話に対して、音声認識処理部２４－１～２４－Ｎによって用いられるＮ個の言語モデルと、言語理解処理部３２－１～３２－Ｍによって用いられるＭ個の言語理解モデルの組み合わせによって出力されたＮ×Ｍ個の各音声理解結果を音声理解結果ｉ（ｉ＝１，…，ｎ；ｎ=Ｎ×Ｍ）と記載する。また、音声理解結果ｉを得るために用いた音声認

識モデルと言語理解モデルの組み合わせを音声理解方式ｉと記載する。つまり、言語理解処理部３２－１が音声認識処理部２４－１～２４－Ｎの音声認識結果を用いたときの音声理解結果がそれぞれ音声理解結果１～Ｎ、言語理解処理部３２－２が音声認識処理部２４－１～２４－Ｎの音声認識結果を用いたときの音声理解結果がそれぞれ音声理解結果（Ｎ＋１）～２Ｎ、…、言語理解処理部３２－Ｍが音声認識処理部２４－１～２４－Ｎの音声認識結果を用いたときの音声理解結果がそれぞれ音声理解結果（Ｎ（Ｍ－１）＋１）～（Ｎ×Ｍ）である。

【００４７】

本実施形態では、音声理解結果である意味表現は、コンセプトの集合であり、コンセプトは、意味スロットとその値の組と、発話タイプとから成る。音声理解結果がコンセプトの集合で示されることについては、例えば、文献７に記載されている。

【００４８】

文献７：J. Glass, j. Polifroni, S. Seneff and V.Zue, "DATA COLLECTION AND PERFORMANCE EVALUATION OF SPOKEN DIALOGUE SYSTEMS: THE MIT EXPERIENCE," Prod.ICSLP, pp.1-4, 2000.

【００４９】

一発話に対する音声理解結果ｉに対し、統合部４０の発話単位信頼度算出部４２は、正解であることの信頼度を表す発話単位信頼度CM_iを付与する。ここで、音声理解結果が正解とは、発話の理解結果が完全に正解、つまり、音声理解結果中に誤ったコンセプトが含まれないことを意味する。

【００５０】

次に、統合部４０の選択部４４は、発話単位信頼度算出部４２によって最も高い発話単位信頼度が付与された音声理解結果を選択し、当該発話に対する最終的な音声理解結果を得て、出力する。つまり、選択結果は$\mathrm{argmax}_i CM_i$が得られた音声理解結果ｉとなる。発話単位信頼度は、音声理解時の特徴に基づくロジスティック回帰式により算出する。ロジスティック回帰式は、学習部５０によって、音声理解方式ｉ毎に以下の式（１）に基づき構築する。

【００５１】

【数１】

$$CM_i = \frac{1}{1+\exp(-(a_{i1}F_{i1}+\ldots+a_{im}F_{im}+b_i))} \qquad (1)$$

【００５２】

学習部５０は、既知の発話データである学習データを用いて上記と同様に得られた音声理解結果に基づき、音声理解方式ｉについて適切な係数（重み）a_{i1}，…，a_{im}と切片b_iを決定する。なお、音声理解ｉに関する独立変数である特徴F_{i1}，F_{i2}，…，F_{im}は、以下の表３に示す特徴である。なお、音声理解方式１～ｎに共通した値となる特徴については、添え字にiを記載していない。

【００５３】

【表３】

表3 音声理解結果iに関する特徴

F_{i1} :音声理解結果iに関する音声認識時の音響スコア
F_{i2} :F_{i1}と発話検証用言語モデル使用時の音響スコアの差
F_{i3} :F_{i1}と発話検証用言語モデル以外の言語モデル使用時の音響スコアとの差
F_4 :発話時間[秒]
F_{i5} :事後確率に基づくコンセプトの信頼度の相加平均
F_{i6} :事後確率に基づくコンセプトの信頼度の音声理解結果i内での最大値
F_{i7} :事後確率に基づくコンセプトの信頼度の音声理解結果i内での最小値
F_8 :F_{i5}の相加平均($\frac{1}{n}\sum_{i}^{n} F_{i5}$)
F_{i9} :F_{i5}の相加平均に対する比(F_{i5}/F_8)
F_{i10}:音声理解結果iに含まれるコンセプト数
F_{11} :音声理解結果1からn内でのコンセプト数の最大値
F_{12} :音声理解結果1からn内でのコンセプト数の最小値
F_{13} :F_{i10}の相加平均($\frac{1}{n}\sum_{i}^{n} F_{i10}$)
F_{i14}:F_{i10}の相加平均に対する比(F_{i10}/F_{13})
F_{i15}:音声理解結果が得られなかったか
F_{i16}:音声理解結果が肯定･否定発話を表すものか

【００５４】

特徴F_{i1}，…，F_{im}それぞれの重みである係数a_{i1}，…，a_{im}と、切片b_iとを決定するために、まず、音声理解装置１の音声認識部２０及び言語理解部３０は、学習データを用いて発話データが入力された場合と同様の音声理解を行い、統合部４０の発話単位信頼度算出部４２は、学習データから得られた各音声理解結果ｉについて、上記の独立変数としての特徴F_{i1}，F_{i2}，…，F_{im}を算出する。そして発話単位信頼度算出部４２によって算出されたそれぞれの独立変数（特徴）の集合に対して、音声理解結果が正解である場合には１を、不正解である場合には０をマニュアルによる入力によって与えてサンプル集合とし、学習部５０は、最尤推定法等によってサンプル集合の対数尤度が最大となるように係数a_{i1}，…，a_{im}と切片b_iを求める。

【００５５】

上記において用いた特徴について述べる。特徴F_{i1}から特徴F_4は、音声認識処理部２４－１～２４－Ｎによる音声認識結果から得られる特徴である。音響スコアは発話時間で正規化する。特徴F_{i1}は、発話単位信頼度算出対象の音声理解結果を得るときに使用した言語モデルに基づく音声認識時の尤度である。特徴F_{i2}と特徴F_{i3}は、音声理解時に用いたモデルとは異なる言語モデル使用時の音響スコアとの比較である。これらの特徴は音声認識結果の信頼性を表す。また特徴F_4は、発話長によって音声認識性能が変化する可能性を考慮して導入した。

【００５６】

例えば、発話単位信頼度算出対象の音声認識結果ｉが、音声認識処理部２４－ｋによる音声認識結果を用いて、言語理解処理部３２－ｊが言語理解処理を行なった結果である場合を仮定する。特徴F_{i1}は、音声認識処理部２４－ｋによる音声認識結果の音響スコアであり、特徴F_{i2}は、音声認識処理部２４－ｋによる音声認識結果の音響スコアから、発話検証用音声認識処理部２６による音声認識結果の音響スコアを減算した値である。特徴F_{i3}は、音声認識処理部２４－ｋによる音声認識結果の音響スコアから、音声認識処理部２４－ｋを除く音声認識処理部２４－１～２４－Ｎの音声認識結果の音響スコアそれ

ぞれを減算した値のうち最も大きい値、つまり、（音声認識処理部２４－ｋによる音声認識結果の音響スコア）－（音声認識処理部２４－１による音声認識結果の音響スコア）、（音声認識処理部２４－ｋによる音声認識結果の音響スコア）－（音声認識処理部２４－２による音声認識結果の音響スコア）、…、（音声認識処理部２４－ｋによる音声認識結果の音響スコア）－（音声認識処理部２４－Ｎによる音声認識結果の音響スコア）のうち最も絶対値が大きい値である。F_4は、音声認識部２０において、入力された発話データから取得する。

【００５７】

特徴F_{i5}から特徴F_{i9}は、言語理解処理部３２－１～３２－Ｍによる音声理解結果の事後確率に基づき算出したコンセプト単位の信頼度に関する特徴である。特徴F_{i5}は、音声理解結果ｉに含まれる全てのコンセプトの信頼度の相加平均である。特徴F_{i6}は、音声理解結果ｉに含まれるコンセプトの信頼度の最大値、特徴F_{i7}は、音声理解結果ｉに含まれるコンセプトの信頼度の最小値である。特徴F_8は、音声理解結果１～ｎについてのF_{i5}の相加平均、特徴F_{i9}は、特徴F_{i5}の特徴F_8に対する比である。

【００５８】

特徴F_{i10}から特徴F_{i14}は、音声理解結果のコンセプト数に関する特徴である。文法外の発話は、発話時間が長くなることがあり、そのような場合、文法モデルに基づく理解結果は正解とならない可能性が高い。特徴F_{i10}は、音声理解結果ｉに含まれるコンセプト数、特徴F_{11}は、各音声理解結果１～ｎに含まれるコンセプト数の最大値、特徴F_{12}は、各音声理解結果１～ｎに含まれるコンセプト数の最小値である。特徴F_{13}は、音声理解結果１～ｎについてのF_{i10}の相加平均、特徴F_{i14}は、特徴F_{i10}の特徴F_{13}に対する比である。

【００５９】

特徴F_{i15}は、音声理解結果が得られた場合、得られなかった場合に応じて２値のうちいずれかの値をとる。特徴F_{i15}により、言語理解処理部３２－１～３２－Ｍにおいて音声認識結果が受理できなかった場合を検出する。言語理解モデルによっては、受理できない音声認識結果が入力されると、音声理解結果は出力されない。そのような場合は、その音声理解結果は正解にならない。

【００６０】

特徴F_{i16}は、音声理解結果が肯定発話であるか、否定発話であるかに応じて２値のうちいずれかの値をとる。特徴F_{i16}は、肯定・否定発話に対しては比較的高精度な音声理解か可能であると考え導入した。具体的には、音声理解結果ｉに含まれるいずれかのコンセプトに、予め指定された肯定、または、否定表現の発話タイプ、あるいは、スロット値が含まれているかに対応して特徴F_{i16}の値を決定することができる。

【００６１】

学習部５０は、ロジスティック回帰式に用いた特徴F_{i1}，F_{i2}，…，F_{im}の特徴量が、平均０、分散１となるように標準化する。また、学習部５０は、特徴F_{i1}，F_{i2}，…，F_{im}から相関が高い特徴を取り除く。本実施形態では、相関係数が０．９以上となる特徴は取り除いた。これは、多重共線性を除去し、学習結果の特徴の係数の絶対値を、有効な特徴順に大きくするためである。

【００６２】

具体的に相関が高い特徴を取り除く処理について説明する。

簡単のため、４つの特徴Ａ，Ｂ，Ｃ，Ｄから相関が高い特徴を取り除く場合を例に説明する。

まず、学習部５０は、特徴Ａ，Ｂ，Ｃ，Ｄのすべての組み合わせの相関係数を算出する。また、学習部５０は、すべての特徴と信頼度の正解値（０または１）との相関係数も算出しておく。これによって、学習部５０は、特徴Ａ，Ｂ，Ｃ，Ｄ及び信頼度の相関係数を要素とする下記の行列を作成する。なお、相関関数が高いと判断する閾値０．９以上の相関係数には「#」を付与している。

【００６３】

【表４】

表4 特徴A～D及び信頼度の正解値の相関係数

	特徴A	特徴B	特徴C	特徴D	信頼度の正解値
特徴A	#1.00	#0.90	-0.16	0.01	0.10
特徴B	#0.90	#1.00	0.39	0.46	0.19
特徴C	-0.16	0.39	#1.00	#0.95	0.20
特徴D	0.01	0.46	#0.95	#1.00	0.17
信頼度の正解値	0.10	0.19	0.20	0.17	#1.00

【００６４】

上記の場合、特徴Ａと特徴Ｂの相関は０．９０、特徴Ｃと特徴Ｄの相関が０．９５であり、相関が高いと判断できる。この場合、相関が高い二つの特徴のうち、信頼度の正解値との相関が低いほうの特徴を削除する。具体的には、特徴Ａと正解信頼度との相関係数は０．１０、特徴Ｂと正解信頼度との相関係数は０．１９のため、正解信頼度との相関がより高いのは特徴Ｂである。よって、特徴Ａを削除し、特徴Ｂを残す。特徴Ｃと特徴Ｄについても同様の操作を行い、特徴Ｄを削除し、特徴Ｃを残す。

上記においては４つの特徴の場合の例を述べたが、特徴が５つ以上であっても、上記と同様に相関係数の行列から相関が高い特徴を検出できる。このように、学習部５０は、特徴F_{i1}，F_{i2}，…，F_{im}と信頼度の正解値との相関係数から上記の同様の行列を作成して、所定の閾値よりも相関が高い二つの特徴を検出し、その二つの特徴のうち、信頼度の正解値との相関係数が最も高い特徴を残し、残りを削除していく。

なお、上記の相関関数の閾値は例であり、ユーザによって設定することが可能である。

【００６５】

さらに、学習部５０は、特徴選択を各音声理解方式ごとに行う。つまり、発話単位信頼度の算出に所定より影響の小さい特徴を選択し、選択した特徴を発話単位信頼度の算出に用いる特徴から除外する。この特徴選択は変数減少法により行う。つまり、最尤推定法等によって決定した係数a_{i1}，…，a_{im}、及び、切片b_iを用いた式（１）から、上記のように相関する特徴を除去した式をフルモデルとして生成する。そして、フルモデルの式から１つずつ特徴を除去していき、発話単位信頼度の精度が所定より低下しない特徴については、発話単位信頼度の算出に用いる特徴からは除外する。

【００６６】

発話単位信頼度算出部４２は、音声理解結果ｉのＣＭｉを算出する場合、最尤推定法等によって決定した係数a_{i1}，…，a_{im}、及び、切片b_iを適用した式（１）から、上記のように選択された特徴の項のみを残した式によって、発話単位信頼度を算出する。この式によって算出された音声理解結果ｉの発話単位信頼度をX_{ie}とする。

【００６７】

音声理解結果の発話単位信頼度の評価尺度は、信頼度の正解値（０または1）との平均誤差ＭＡＥ（Mean Absolute Error）とする。ＭＡＥは、以下の式（２）で求められる。

【００６８】

【数２】

$$MAE=\left(\frac{1}{n}\sum_{i}^{n}\right)|X_{ie}-X_{ia}| \qquad (2)$$

【００６９】

ＭＡＥは、予測値と正解との１発話あたりの誤差の平均を表す。ここで、ｎは全発話数

である。X_{ie}はｉ番目の発話の音声理解結果ｉに対する推定信頼度を表し、X_{ia}は発話単位信頼度の正解値（０または１）を表す。なお、X_{ia}は、人手で与えた。

【００７０】

［４．実施例］

［４．１　実装した言語モデルと言語理解モデル］

本実施形態の言語理解装置１が実現するＭＬＭＵの実施例として、文献８に記載のレンタカー予約システムにおいても用いられている一般的な２種類の言語モデルと、３種類の言語理解モデルを使用できるようにした。

【００７１】

文献８：M. Nakano, Y.Nagano, K. Funakoshi, T.Ito, K. Araki, Y. Hasegawa, and H. Tusujino, "Analysis of User Reactions to Turn-Talking Failures in Spoken Dialogue Systems," Proc. 8th SIGdial Workshop on Discourse and Dialogue, pp,120-123, 2007.

【００７２】

音声認識処理部２４－１～２４－Ｎ（本実施例においてはＮ＝２）にはそれぞれ、以下の言語モデルを用いた。

（１）文法ベース言語モデル（文法モデル）

（２）ドメイン依存統計言語モデル（Ｎ－ｇｒａｍモデル）

【００７３】

レンタカー予約システムにおける文法モデルは、言語理解時に用いるＦＳＴに対応させて人手で記述した。また、Ｎ－ｇｒａｍモデルは、学習データの書き起こしを用いてクラス３－ｇｒａｍを学習し、作成した。語彙サイズは、文法モデルが２７８、Ｎ－ｇｒａｍモデルが３７８である。音声認識器はJulius（vor.4.1.2）を用い、音素毎の音声波形パターンである音響モデルとして、文献９に記載の話者非依存ＰＴＭトライフォンモデルを用いた。文法、Ｎ－ｇｒａｍモデルを用いたときの音声認識時の単語正解精度はそれぞれ、学習データでは６８．１％、８７．５％であり、評価データでは７２．３％、８６．９％であった。

また、音声認識結果を検証するための言語モデル、すなわち、発話検証用音声認識処理部２６において実現する言語モデルとしてドメイン非依存大語彙統計言語モデルを用いた。ドメイン非依存大語彙統計言語モデルは、連続音声認識コンソーシアム配布の、Web文章から学習した単語Ｎ－ｇｒａｍモデルを使用した。語彙サイズは６０，２５０である（文献９参照）。

【００７４】

文献９：T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository," Proc.ICSLP, pp.3069-3072, 2004.

【００７５】

一方、言語理解処理部３２－１～３２－Ｍ（本実施例においてはＭ＝３）にはそれぞれ、以下の３種類の言語理解モデルを用いた。

（１）Finite-State Transducer（ＦＳＴ）

（２）Weighted FST（ＷＦＳＴ）

（３）Keyphrase-Extractor（Extractor）

【００７６】

ＦＳＴは、有限状態オートマンに出力を付与したものであり、入力列に従って状態遷移を行なうことによって、その状態遷移に付与された記号の列を出力する。ＦＳＴによる言語理解では、人手でＦＳＴを作成しておき、それに音声認識結果の単語列を入力することで、言語理解結果を得る。レンタカー予約システムにおいて作成したＦＳＴは、入力可能な単語数は２７８であり、カバレッジは、学習データに対して８１．３％、評価データに対して８６．０％である。入力には音声認識結果の１０－ｂｅｓｔ候補を用い、１０－ｂｅｓｔ候補の１位の候補から順にＦＳＴで受理可能な認識結果を探す。１０－ｂｅｓｔ候

補すべて受理できなかった場合、言語理解結果は出力されない。

【００７７】

ＷＦＳＴによる言語理解は、例えば、文献１０に記載された手法に基づく。ＷＦＳＴでは、ＦＳＴの状態遷移にさらに重みを付加しており、入力列に従った状態遷移に付与された記号の列と、それらの記号に対応した重みの累積を出力する。文献１０に記載のＷＦＳＴでは、音声認識結果をフィラーや単語、コンセプトなどとして抽象化し、これらに対して音素数や音声認識の信頼度を利用した重みを割当てる。ＷＦＳＴの構築には、文献１１に記載のMITToolkitを用いる。ここでは、ＦＳＴにフィラー遷移を付加することで、言語理解に不要な単語を無視する解釈を許容できる。音声認識結果の１０－ｂｅｓｔ候補それぞれをＷＦＳＴによりコンセプト列に変換し、累積重みが最大となるコンセプト列を言語理解結果とする。用いる重み付けの種類は、文献１０に記載されているように、学習データを用いて選択する。ＷＦＳＴによる言語理解では、フィラー遷移の導入により、ＦＳＴでは受理されない音声認識結果に対しても言語理解結果を出力できる。また、音声認識時の単語信頼度を重みに用いるため、音声認識誤りに頑健である。

【００７８】

文献１０：福林雄一朗、駒谷和範、中野幹生、船越孝太郎、辻野広司、尾形哲也、奥乃博，“音声対話システムにおけるラピッドプロトタイピングを指向した言語理解”，情報処理学会論文誌，vol.49，no.8，pp.2762-2772，2008.

文献１１：L. Hetherington，“The MIT Finite-State Transducer Toolkit for Speech and Language Processing，” Proc. ICSLP，pp.2609-2612，2004.

【００７９】

Extractorによる言語理解では、音声認識結果の１位の候補に対して、コンセプトに変換可能な音声認識結果の部分列を単純にコンセプトに変換する。ただし、変換されたコンセプト間に矛盾かおる場合は、矛盾のないコンセプトの組み合わせを、出力コンセプト数が最大となるように出力する。コンセプト間の矛盾は、ＦＳＴを用いて検出した。Extractorによる言語理解は、ＦＳＴでは受理されない音声認識結果に対しても言語理解結果を出力できる。しかし、音声認識結果に誤りが含まれる場合もそのままコンセプト列に変換してしまう。

【００８０】

信頼度算出部３４は、言語理解処理部３２－１～３２－Ｍによって得られた言語理解結果の各コンセプトに対して信頼度を付与する。音声認識結果の１０－ｂｅｓｔ候補を用いて、文献１２の手法に基づき、コンセプトごとに信頼度を計算して用いる。

具体的には、以下のように信頼度を計算する。つまり、各コンセプトに含まれるスロットについてＩＤＦ（inverse document frequency）を算出する。次に、各コンセプトについて、そのコンセプトに含まれるスロットのＩＤＦの和を算出し、算出した和を正規化して信頼度とする。

【００８１】

文献１２：駒谷和範、河原達也，“音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理”，情報処理学会論文誌，vol.43，no.10，pp.3078-3086，2002.

【００８２】

［４．２　ロジスティック回帰式に基づく信頼度の評価］

ロジスティック回帰式に基づき算出した発話単位信頼度の評価を行った。正解理解結果を正しく選択するには、各理解結果に適切な発話単位信頼度が付与されている必要がある。

発話単位信頼度の評価実験に用いる対話データは、被験者３３名に簡単なレンタカーの予約タスクを課し、文献８に記載のレンタカー予約システムと対話をしてもらうことで収集した。結果、４，９８６発話を収集した。収集発話のうち、レンタカー予約システムが検出した発話区間と、人手で付与した発話区間とが一致した４，５１３発話を実験に用いた。これは本実施形態の対象でないＶＡＤ誤りや、タスクに関係のない発話を除くためである。４，５１３発話のうち１６名分２，１９３発話を学習データとし、１７名分２，３

２０発話を評価データとした。学習データを用いて、特徴選択部５０が相関の高い特徴の除去と特徴選択を行った結果、表３に記した１６個の特徴量から、選択された特徴量を表５に示す。

【００８３】

【表５】

表5 選択された特徴量

音声理解方式	選択された特徴
文法+FST	$F_{i1}, F_{i2}, F_{i3}, F_4, F_8, F_{i9}, F_{11}, F_{12}, F_{i15}, F_{i16}$
文法+WFST	$F_{i1}, F_{i2}, F_{i3}, F_4, F_{i7}, F_8, F_{i9}, F_{11}, F_{12}, F_{i15}, F_{i16}$
文法+Extractor	$F_{i1}, F_{i2}, F_{i3}, F_4, F_{i7}, F_8, F_{i9}, F_{11}, F_{12}, F_{i15}, F_{i16}$
N-gram+FST	$F_{i1}, F_{i2}, F_4, F_{i7}, F_8, F_{11}, F_{12}, F_{i15}, F_{i16}$
N-gram+WFST	$F_{i1}, F_{i2}, F_4, F_{i7}, F_8, F_{i9}, F_{11}, F_{12}, F_{i16}$
N-gram+Extractor	$F_{i1}, F_{i2}, F_{i3}, F_4, F_{i7}, F_8, F_{11}, F_{12}, F_{i16}$

【００８４】

表５において、Ｎ－ｇｒａｍ＋Extractorの音声理解方式に関する特徴では、コンセプト数を表す特徴F_{i10}と、コンセプト数の相加平均を表す特徴F_{13}は、コンセプト数の最大値を表す特徴F_{11}と相関が高いため、学習部５０により除かれた。また、事後確率に基づくコンセプトの信頼度の相加平均を表す特徴F_{i5}と、コンセプト信頼度の最大値を表す特徴F_{i6}と、コンセプト信頼度の最小値を表す特徴F_{i7}と相関が高いため、学習部５０により除かれた。さらに、変数減少法による特徴選択の結果、特徴F_{i5}の相加平均に対する比を表す特徴F_{i9}と、特徴F_{i10}の相加平均に対する比を表す特徴F_{i14}、音声理解結果が得られなかったことを表す特徴F_{i15}の三つの特徴が学習部５０により除かれた。このように、学習部５０によって選択された特徴を独立変数とするロジスティック回帰式を用いて、評価データの音声理解結果に対して発話単位信頼度を付与した。

【００８５】

各音声理解方式の結果に対する発話単位信頼度のＭＡＥを表６に示す。

【００８６】

【表６】

表6 ロジスティック回帰式に基づく発話単位信頼度のMAE

音声理解方式	logistic regression	Expect.
文法+FST	0.146	0.333
文法+WFST	0.159	0.331
文法+Extractor	0.147	0.334
N-gram+FST	0.093	0.337
N-gram+WFST	0.146	0.284
N-gram+Extractor	0.135	0.280

【００８７】

表６において、logistic regressionの列に、発話単位信頼度算出部４２がロジスティック回帰式に基づき算出した発話単位信頼度のＭＡＥを示し、Expect.の列にベースラインとして学習データにおける発話単位信頼度の期待値のＭＡＥを示す。ここで、学習データにおける発話単位信頼度の期待値のＭＡＥとは、学習データにおいて各音声理解方式による結果が正解となる割合を推定信頼度としたときの、発話単位信頼度の正解値との誤差を示している。表６において、すべての音声理解方式の結果に対して、本実施例による発話単位信頼度のＭＡＥは、発話単位信頼度の期待値のＭＡＥと比較して小さな値である。つまり、信頼度を予測するモデルとしての性能が高いといえる。これは音響スコアや、事後確率に基づくコンセプトの信頼度など、音声理解結果の精度を表す特徴を用いてロジスティック回帰式を構築した効果である。Ｎ－ｇｒａｍ＋ＦＳＴの理解結果に対する発話単位信頼度のＭＡＥが０．０９３となり最も小さい。これは、Ｎ－ｇｒａｍ＋ＦＳＴにおいて、ＦＳＴでは受理できない音声認識結果が入力され、言語理解結果が出力されなかった場合に低い発話単位信頼度を付与できたからである。

【００８８】

用いた特徴が、発話単位信頼度算出時にどれだけ有効であったかを調べるため、ロジスティック回帰式の係数を調べた。各特徴量の値は標準化されているため、係数の絶対値の大きさを比較することで、特徴の有効性を検証できる。各音声理解方式に対して構築したロジスティック回帰式ごとに、係数の絶対値が大きかった上位５つの特徴と、その係数の値を表７に示す。

【００８９】

【表７】

表7 各回帰式で係数の絶対値が大きかった特徴とその係数の値

文法+FST		N-gram+FST	
F_{i2}	7.37	F_{i15}	-18.08
F_{i15}	-5.51	F_{i2}	4.07
F_{i3}	2.14	F_{11}	-2.31
F_{11}	-1.91	F_{i16}	1.72
F_{8}	1.62	F_{i1}	1.54

文法+WFST		N-gram+WFST	
F_{i2}	6.85	F_{i2}	2.29
F_{i15}	-4.96	F_{i1}	1.93
F_{i3}	1.41	F_{i16}	1.47
F_{11}	-1.38	F_{8}	1.30
F_{8}	1.23	F_{12}	0.73

文法+Extractor		N-gram+Extractor	
F_{i2}	7.47	F_{i2}	2.29
F_{i15}	-5.60	F_{i16}	1.62
F_{i3}	1.96	F_{i1}	1.56
F_{11}	-1.92	F_{i7}	0.98
F_{i1}	1.32	F_{8}	0.93

【００９０】

表７において、全体的に係数の絶対値が大きくなった特徴は、発話検証用言語モデルとの音響尤度差である特徴F_{i2}と、音声理解結果が得られなかったかどうかを表す特徴F_{i15}である。特徴F_{i2}の係数より、音響尤度差が大きいときほど理解結果が正解となりやすいことを示している。特徴F_{i15}の係数が負の大きな値となったのは、音声理解結果が得られないときは、正解とはならなかったためである。他に、文法モデルを使用した音声理解方式では、コンセプト数の最大値を表す特徴F_{11}が有効となり、Ｎ－ｇｒａｍモデルを使用した音声理解方式では、理解結果が肯定・否定発話かどうかを表す特徴F_{i16}が有効であった。

【００９１】

［５．　音声理解実験］

本実施形態の音声理解装置１の実施例によって得られた音声理解結果の評価を行った。評価実験には上述の４．２において述べた４，５１３発話を用いる。学習データ２，１９３発話を用いて、学習部５０により特徴選択とロジスティック回帰式の係数のフィッティングを行い、評価データ２，３２０発話に対して、音声認識処理部２４－１～２４－Ｎ及び言語理解処理部３２－１～３２－Ｍの組み合わせによる音声理解結果に対する発話単位信頼度の付与と、その信頼度に基づく選択を行った。本実施形態では音声理解結果の評価尺度には以下の二つを用いる。

（１）発話完全理解精度

（２）コンセプト理解精度

【００９２】

前者の発話完全理解精度は発話単位の音声理解精度であり、以下の式（３）で求められる。

【００９３】
発話完全理解精度＝（完全正解発話数）／（全発話数）　…（３）
【００９４】
正解理解結果数とは、一発話に含まれるコンセプト列を誤りなく出力できた数である。ここでは、ロジスティック回帰による発話単位信頼度は、音声理解結果が発話単位で完全に正解であるか否かを推定している。本実施形態では、その信頼度が最も高い結果を最終結果として得るため、得られた結果は、発話単位で完全に正解であることが望まれる。発話完全理解精度を用いることで、本実施形態における選択手法が適切に、発話単位で完全に正解の結果を選択できたかを評価する。
【００９５】
後者のコンセプト理解精度とは、コンセプト単位の音声理解精度であり、以下の式（４）で求められる。
【００９６】
コンセプト理解精度
＝１－（誤りコンセプト数／全発話に含まれるコンセプト数）　…（４）
【００９７】
誤りコンセプト数は、置換誤りコンセプト数、削除誤りコンセプト数、挿入誤りコンセプト数の和で求められる。
【００９８】
［５．１　単一方式との比較］
本実施形態と、単一の言語モデル・言語理解モデル使用時の発話完全理解精度を表８に、コンセプト理解精度を表９に示す。表９において、Sub、Del、Insはそれぞれ、置換誤り率、削除誤り率、挿入誤り率を表す。
【００９９】
【表８】

表8 単一の言語モデル･言語理解モデルを用いた方式と本実施例による発話完全理解精度[%]

音声理解方式	発話完全理解精度[%]
文法+FST	79.8
文法+WFST	80.0
文法+Extractor	79.8
N-gram+FST	79.3
N-gram+WFST	84.2
N-gram+Extractor	84.6
MLMU&CMBS	86.8

【０１００】

【表９】

表9 単一の言語モデル・言語理解モデルを用いた方式と本実施例によるコンセプト理解精度[%]

音声理解方式	コンセプト理解精度	Sub	Del	Ins
文法+FST	81.5	11.3	2.6	4.5
文法+WFST	82.0	11.2	3.2	3.6
文法+Extractor	81.5	11.3	2.6	4.5
N-gram+FST	78.7	4.8	13.7	2.8
N-gram+WFST	87.8	6.5	2.9	2.9
N-gram+Extractor	87.7	6.9	2.4	3.0
MLMU&CMBS	89.8	6.3	1.4	2.6

【０１０１】

表８において、Ｎ－ｇｒａｍ＋ＷＦＳＴによる精度が８４．２％、Ｎ－ｇｒａｍ＋Extractorによる精度が８４．６％となり、他の４つの方式と比較して高い値となった。これはＮ－ｇｒａｍモデル使用時の音声認識精度が、文法モデル使用時と比較して高く、より多くの正解コンセプト列を出力できたからである。また、ＷＦＳＴとExtractorによる言語理解では、ＦＳＴでは受理されない音声認識結果に対しても、言語理解結果を出力できたからである。

一方、本実施例の音声理解装置１による発話完全理解精度は８６．８％となった。これは、単一の言語モデル・言語理解モデルを使用したいずれの音声理解方式より高精度である。これは複数の音声理解方式の結果から、本実施形態による選択手法により、適切に正解理解結果を選択できることを示している。

【０１０２】

［５．２　言語モデル・言語理解モデルいずれかを複数を用いた音声理解方式との比較］

本実施形態と、言語モデル・言語理解モデルをいずれか片方だけを複数用いた音声理解方式との比較を行う。それぞれの方式での発話完全理解精度を表１０に、コンセプト理解精度を表１１に示す。

【０１０３】

【表１０】

表10 複数の言語モデル・言語理解モデルを用いたときの発話完全理解精度[%]

使用モデル	CMBS	oracle
LMs+FST	84.4	85.2
LMs+WFST	86.3	88.4
LMs+Extractor	86.4	88.2
文法+LUMs	80.2	80.8
N-gram+LUMs	84.9	85.5
LMs+LUMs (MLMU)	86.8	89.0

【０１０４】

【表１１】

表11 複数の言語モデル・言語理解モデルを用いたときのコンセプト理解精度[%]

使用モデル	CMBS	oracle
LMs+FST	85.4	86.9
LMs+WFST	89.3	91.4
LMs+Extractor	89.2	90.9
文法+LUMs	81.8	82.4
N-gram+LUMs	87.9	89.2
LMs+LUMs(MLMU)	89.8	91.9

【０１０５】

上記の表において、ＬＭｓ、ＬＵＭｓはそれぞれ、言語モデル、言語理解モデルを複数用いることを示す。つまり、ＬＭｓでは、文法モデルとＮ－ｇｒａｍモデルの２種類の言語モデルを使用し、ＬＵＭｓでは、音声認識結果に対して、ＦＳＴ、ＷＦＳＴ及びExtractorの３種類の言語理解モデルを使用した。複数の理解結果の統合手法として、ＣＭＢＳは本実施形態において実現した発話単位信頼度に基づく選択を表し、ｏｒａｃｌｅは、人手による最適な理解結果の選択を表す。人手による選択では、出力された音声理解結果のいずれかを、音声理解精度が最も高くなるように選択した。これは、統合手法の性能の影響を取り除き、複数の言語モデルや言語理解モデルを用いる場合の性能の上限を調べるためである。

【０１０６】

表１０において、言語モデル、言語理解モデルの両方を複数用いて、理解結果を人手によって選択した場合の発話完全理解精度は８９．０％となった。この値は言語モデルと言語理解モデルのいずれかを複数用いた場合より高い精度である。これは、ＭＬＭＵにより言語モデル・言語理解モデルを両方複数用いることで、いずれか片方だけを複数用いる場合より、高精度な音声理解が実現可能であることを示している。

【０１０７】

本実施形態の音声理解装置１の実施例によって、言語モデル、言語理解モデルの両方を複数用いて、ＣＭＢＳにより理解結果を選択した場合の発話完全理解精度は、言語モデルを複数用いた場合と比較してほぼ同等の精度である。この結果は、誤りを全く含まない音声理解結果を得るには、言語モデルを複数用いることが重要であることを示している。音声認識結果に誤りが存在し、正解単語が既に欠落している場合、言語理解部でそれを修復するのは不可能である。複数の言語モデルにより複数の音声認識結果を得ることで、いずれかの音声認識結果に正解が含まれる可能性が増えるため、言語モデルを複数使用したことの方が発話完全理解精度の向上に貢献したと言える。

【０１０８】

［５．３　従来の統合手法との比較］

複数の理解結果を一つの音声理解結果に統合する上で、従来のＲＯＶＥＲ法と、本実施形態の言語理解装置１により実現したＣＭＢＳとを比較する。ＲＯＶＥＲ法は、コンセプト単位の重み付き多数決であり、以下の二つの手順から成る。

【０１０９】

(１)ＤＰマッチングにより、複数の音声理解結果内のコンセプト同士の対応づけを行う。

(２)対応付けられたコンセプトの中に、競合するコンセプトがある場合、スコアに基づき取捨する。アライメント位置ｉにおけるコンセプトcpのスコアは以下の式（５）に基づき算出する。

【０１１０】

Score(cp) =α＊(N(cp,i)/Ns)+(1-α)＊Conf(cp)　…式（５）

【０１１１】

ここでN(cp,i)はアライメント位置iに存在するコンセプトcpの数を表し、Nsは用いた音声理解方式の数、Conf(cp)は、アライメント位置ｉに存在するコンセプトcpの、事後確率に基づくコンセプト信頼度の平均値を表す。αとConf(@）はパラメータであり、学習データを用いて推定する。

【０１１２】

２種類の言語モデルと３種類の言語理解モデルによる６つの理解結果に対し、ＣＭＢＳとＲＯＶＥＲ法、oracleでの統合時の精度を表１２に示す。

【０１１３】

【表１２】

表12 本実施例とROVER法との比較[%]

統合法	発話完全理解精度	コンセプト理解精度
CMBS	86.8	89.8
ROVER法	82.7	85.9
oracle	89.0	91.9

【０１１４】

表１２より、ＲＯＶＥＲ法と比較してＣＭＢＳを実装する本実施形態は、発話完全理解精度、及び、コンセプト理解精度のいずれの尺度でも高い。これは、複数の理解結果に誤った結果が多数ある場合、ＲＯＶＥＲ法では、誤った結果に強く影響された結果を出力してしまうためである。また、本実験において実装したＲＯＶＥＲ法では、事後確率に基づくコンセプト信頼度しか用いておらず、多数の特徴を用いていない。そのため、各アライメント位置ごとのコンセプトのスコアが適切な値とならず、コンセプトの取捨が適切に行われなかったと考えられる。

【０１１５】

音声理解精度の向上に統計的に有意差が見られるのかを調べるため、発話完全理解精度に対して、マクネマー検定を行うとともに、コンセプト理解精度に対して、ウィルコクソンの符号順位検定を行った。マクネマー検定は、対応のとれる二群のカテゴリデータに対し、母比率に差があるかを調べる検定であり、ウィルコクソンの符号順位検定は、対応のとれる二群の間隔尺度・比例尺度のデータに対し、母代表値に差があるかを調べるノンパラメトリック検定である。検定の結果、コンセプト精度に関して、本実施形態と、単一の理解方式で精度が最も高かったＮ－ｇｒａｍ＋ＷＦＳＴや、言語モデル・言語理解モデルいずれか片方だけを複数使用した理解方式で最高性能だったＬＭｓ＋ＷＦＳＴとは有意水準１％で有意差が見られた。しかし、発話完全理解精度に関して、本実施形態と、ＬＭｓ＋ＷＦＳＴ、ＬＭｓ＋Extractorを比較したとき有意差は見られなかった。

【０１１６】

［６．まとめ］

本実施形態では、音声理解の高精度化を目的とし、複数の言語モデルと複数の言語理解モデルを用いた音声理解装置について述べた。評価実験では、言語モデル・言語理解モデルのいずれか片方を複数用いた方式や、ＲＯＶＥＲ法を用いた方式と比較して、本実施形態によるコンセプト理解精度の向上を確認した。

【０１１７】

以上の説明によって、以下がいえる。

（１）言語モデルと言語理解モデルを両方複数用いることの有効性を示した。これまで、言語モデル・言語理解モデルのいずれか片方を複数用いた研究はあったが、どちらも複

数用いるものはなかった。本実施形態では、言語モデルと言語理解モデルを両方複数用いることで、言語モデルまたは言語理解モデルをいずれか複数用いた時より、高精度な音声理解が実現できることを示した。

【０１１８】

（２）複数の理解結果の統合手法として、高精度な音声理解を実現する新しい選択手法を実現した。従来一般的に用いられてきた重み付き多数決では、性能の低い理解方式の結果の影響を受けてしまうという問題があった。本実施形態では、音声理解結果が正解かどうかを予測するロジスティック回帰式を構築し、出力された発話単位信頼度に基づいて選択を行った。これにより、性能の低いモデルの影響を受けることなく、発話ごとに適切な音声理解方式を出力することが可能となった。

【０１１９】

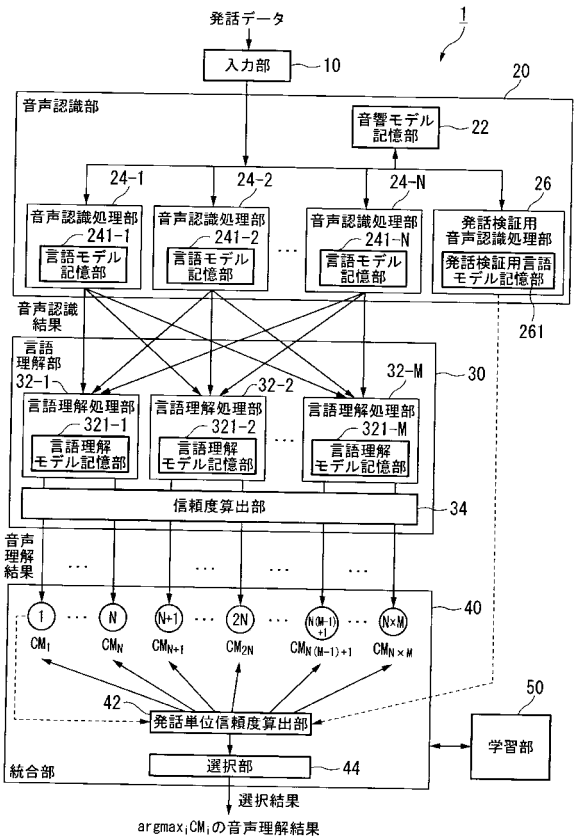
なお、本実施形態では、ロジスティック回帰を発話単位信頼度算出時に用いたが、信頼度を算出する方法は、線形回帰等、様々な手法を用いることもできる。

【符号の説明】

【０１２０】

１…音声理解装置、１０…入力部、２０…音声認識部、２２…音響モデル記憶部、２４－１～２４－Ｎ…音声認識処理部、２４１－１～２４１－Ｎ…言語モデル記憶部、２６…発話検証用音声認識処理部、２６１…発話検証用言語モデル記憶部、３０…言語理解部、３２－１～３２－Ｍ…言語理解処理部、３２１－１～３２１－Ｍ…言語理解モデル記憶部、３４…信頼度算出部、４０…統合部、４２…発話単位信頼度算出部、４４…選択部、５０…学習部

【図１】



フロントページの続き

(72)発明者　中野　幹生
埼玉県和光市本町８－１　株式会社ホンダ・リサーチ・インスティチュート・ジャパン内
(72)発明者　勝丸　真樹
埼玉県和光市本町８－１　株式会社ホンダ・リサーチ・インスティチュート・ジャパン内
(72)発明者　船越　孝太郎
埼玉県和光市本町８－１　株式会社ホンダ・リサーチ・インスティチュート・ジャパン内
(72)発明者　奥乃　博
埼玉県和光市本町８－１　株式会社ホンダ・リサーチ・インスティチュート・ジャパン内

審査官　毛利　太郎

(56)参考文献　特開２００３－２２８３９３（ＪＰ，Ａ）
特開２００７－０４７４８８（ＪＰ，Ａ）
特開平０９－２７４４９８（ＪＰ，Ａ）
特開２００８－２９３０１９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１７／２６