(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2017-102247
(P2017-102247A)
(43)公開日　平成29年6月8日(2017.6.8)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 15/10	(2006.01)	G10L	15/10	5００T
G10L 15/22	(2006.01)	G10L	15/22	３００U
G10L 13/00	(2006.01)	G10L	13/00	１００M
G10L 15/02	(2006.01)	G10L	15/02	３００K
G10L 15/197	(2013.01)	G10L	15/197	

審査請求 未請求　請求項の数 7　ＯＬ　（全 11 頁）

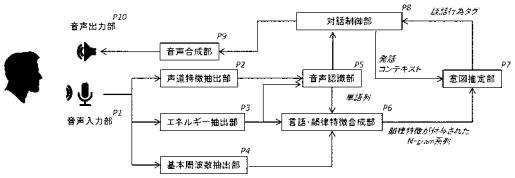
(21)出願番号　特願2015-234835 (P2015-234835)
(22)出願日　平成27年12月1日 (2015.12.1)

（出願人による申告）平成２７年度国立研究開発法人科学技術振興機構研究成果展開事業　戦略的イノベーション創出推進プログラム「高齢者の記憶と認知機能低下に対する生活支援ロボットシステムの開発」委託研究、産業技術力強化法第１９条の適用を受ける特許出願

(71)出願人　301021533
国立研究開発法人産業技術総合研究所
東京都千代田区霞が関１－３－１
(72)発明者　佐土原　健
茨城県つくば市東１－１－１　国立研究開発法人産業技術総合研究所つくばセンター内

(54)【発明の名称】音声対話システム、音声対話制御法およびプログラム

(57)【要約】　（修正有）
【課題】談話行為識別精度の向上を図る。
【解決手段】音声入力部Ｐ１と、声道特徴系列を計算する声道特徴抽出部Ｐ２と、エネルギー系列を計算するエネルギー抽出部Ｐ３と、基本周波数系列を計算する基本周波数抽出部Ｐ４と、声道特徴系列とエネルギー系列に基づいてタイムスタンプ付きの単語列を出力する音声認識部Ｐ５と、タイムスタンプ付きの単語列とタイムスタンプ付きの前記エネルギー系列と前記基本周波数系列から韻律特徴が付加された拡張単語Ｎ－ｇｒａｍを生成する言語・韻律特徴合成部Ｐ６と、拡張単語Ｎ－ｇｒａｍと該発話のコンテキスト情報から談話行為タグを推定する意図推定部Ｐ７と、談話行為タグが表す発話意図と対話文脈を考慮してシステム発話を生成する対話制御部Ｐ８と、該生成されたシステム発話を音声信号に変換する音声合成部Ｐ９と、該変換された音声信号を音声出力装置で再生する音声出力部Ｐ１０とからなる。
【選択図】図１



【特許請求の範囲】
【請求項１】
人との音声対話インタフェースを含む音声対話システムであって、
音声入力装置からの音声入力を処理して音響信号に変換する音声入力部(Ｐ１)と、
その音響信号を処理して声道特徴系列を計算する声道特徴抽出部(Ｐ２)と、
その音響信号を処理してエネルギー系列を計算するエネルギー抽出部(Ｐ３)と、
その音響信号を処理して基本周波数系列を計算する基本周波数抽出部(Ｐ４)と、
該計算された声道特徴系列とエネルギー系列に基づいてタイムスタンプ付きの単語列を出力する音声認識部(Ｐ５)と、
該出力されたタイムスタンプ付きの単語列とタイムスタンプ付きの前記エネルギー系列と前記基本周波数系列から韻律特徴が付加された拡張単語Ｎ－ｇｒａｍを生成する言語・韻律特徴合成部(Ｐ６)と、
該生成された拡張単語Ｎ－ｇｒａｍと該発話のコンテキスト情報から該発話の談話行為タグを推定する意図推定部(Ｐ７)と、
前記該発話のコンテキスト情報を提供し該推定された談話行為タグが表す発話意図と対話文脈を考慮してシステム発話を生成する対話制御部(Ｐ８)と、
該生成されたシステム発話を音声信号に変換する音声合成部(Ｐ９)と、
該変換された音声信号を音声出力装置で再生する音声出力部(Ｐ１０)とからなることを特徴とする音声対話システム。
【請求項２】
言語・韻律特徴合成部(Ｐ６)において、前記拡張単語Ｎ－ｇｒａｍに付加された韻律特徴は離散化された韻律特徴であることを特徴とする請求項１に記載する音声対話システム。
【請求項３】
意図推定部(Ｐ７)において、前記該発話の談話行為タグを推定は、次の談話行為タグを表す確率変数Ｉの事後確率の最大化により行うことを特徴とする請求項２に記載の音声対話システム。
ただし、Ａは音響信号、A_Sはその声道成分、Ｗは単語列、W^fは前記離散化された韻律特徴が付与された拡張単語列とする。
【数１２】

$$P(A|I) \approx \sum_W P(A_S|W)P(W^f|I) \qquad (12)$$

【請求項４】
言語・韻律特徴合成部(Ｐ６)において、前記離散化を、Ｗに対応するフレーム列において、当該基本周波数の変化量が平均＋標準偏差よりも大きい場合は＋を、変化量が平均－標準偏差よりも小さい場合は－を、変化量が平均±標準偏差の範囲内であれば０を付与する３値の離散化であることを特徴とする請求項３に記載の音声対話システム。
【請求項５】
意図推定部(Ｐ７)において、前記尤度Ｐ(Ｗ^f｜Ｉ)は、次式により計算されることを特徴とする請求項４に記載の音声対話システム。
ただしW_i^fはＷに含まれるＮ－ｇｒａｍ、W_1，…，W_mに前記離散韻律特徴を付与した、拡張Ｎ－ｇｒａｍ列W^f_1，…，W^f_mの１つとする。
【数１３】

$$P(W^f|I) = \prod_{i=1}^{m} P(W_i^f|I) \qquad (13)$$

【請求項６】

意図推定部(Ｐ７)において、前記尤度Ｐ(Ｗ$_{i}^{f}$｜Ｉ)は、次式により計算されることを特徴とする請求項５に記載の音声対話システム。

ただし、ｊは自然数、ｊ＞０の場合は、基本周波数の勾配、強勢等、ｆで用いられている離散韻律特徴が付与されたＮ－ｇｒａｍを表し、特に、ｊ＝０の場合は、韻律特徴を用いないＮ－ｇｒａｍを表す。

【数１４】

$$P\left(W_i^f \middle| I\right) = \sum_{j=0}^{J} \lambda_j P(W_i^{f_j} | I), \quad \sum_{j=0}^{J} \lambda_j = 1 \qquad (14)$$

【請求項７】

人との音声対話インタフェースを含む音声対話プログラムであって、請求項１乃至請求項５のいずれか１項に記載される音声対話システムの各処理を実行することを特徴とする音声対話プログラム、および当該プログラムを記憶したプログラム媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、音声対話を利用してユーザと情報のやりとりを行う、音声対話システム、音声対話制御手法、プログラムに関する。

【背景技術】

【０００２】

従来、機器やシステムのインタフェースとして、ユーザが日常的に用いているコミュニケーション手段である音声対話を用いたインタフェースが利用されてきた。

音声対話インタフェースにおいては、ユーザの発話からユーザの意図を抽出し適切な応答を生成する対話制御技術が必要になる。

【０００３】

ユーザの意図の中でも、対話文脈における発話の機能、例えば、相手の発言に対して、肯定しているのか、否定しているのか、あるいは疑問を発しているか等を判定する問題は、談話行為識別と呼ばれている。

例えば、非特許文献１では、発話意図を表す談話行為タグを予測するための確率モデルを、発話に談話行為タグが付与された音声対話データから、統計的に機械学習させ、学習されたモデルを用いて、音声対話中の発話に談話行為タグを付与する技術が開示されている。

【０００４】

ところで、これまでの発話意図推定技術においては、音声に含まれる言語特徴が意図を推定するための重要な手がかりとして用いられてきた。

また、表情や韻律(音声のリズム、抑揚、速度、強勢など)等の非言語特徴も、発話者の意図を反映することが良く知られている。

【０００５】

例えば、特許文献１では、表情と韻律を用いたユーザの感情の推定技術が開示されている他、特許文献２では、システムの間違いをユーザが指摘した箇所を同定する技術が開示されている。

また、非特許文献１および非特許文献２では、韻律に関する発話の統計量を言語情報に付加して、談話行為識別精度を向上させる技術が開示されている。

【０００６】

さらに、特許文献３および非特許文献３においては、ピッチの変化等、韻律特徴の軌跡を離散符号化した系列を言語特徴とともに利用して談話行為識別精度を向上させる技術が開示されている。

いずれも、言語特徴に付加する形で、言語特徴とは独立な、発話文の音響的な特徴として韻律特徴が用いられている。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】特開２００６－３１３２８７号公報

【特許文献２】特開２０１３－２０５８４２号公報

【特許文献３】米国特許出願公開第２００６／０１２２８３４号明細書

【非特許文献】

【０００８】

【非特許文献１】A. Stolcke et al. : “Dialogue act modeling for automatic tagging and recognition of conversational speech”, Computational Linguistics, Vol.26, No.3, pp.339-373, 2000.

【非特許文献２】E. Shriberg et al. : “Can prosody aid the automatic classification of dialog acts in conversational speech?”, Language and speech, Vol.41, No.3-4, pp.443-492,1998.

【非特許文献３】V.K.R. Sridhar et al. : “Combining lexical, syntactic and prosodic cues for improved online dialog act tagging”, Computer Speech and Language, Vol.23, No.4, pp.407-422, 2009.

【非特許文献４】A. Black et al. : “Predicting the intonation of discourse segments from examples in dialogue speech. ”, Computing Prosody, pp.117-128, 1997.

【非特許文献５】K.Sadohara et al. : “Sub-lexical dialogue act classification in a spoken dialogue system support for the elderly with cognitive disabilities”. In Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies. pp.93-98, 2013.

【非特許文献６】鹿野清宏他：“音声認識システム”, オーム社, ２００１.

【非特許文献７】河原達也, 荒木雅弘：“音声対話システム”, オーム社, ２００６.

【非特許文献８】P.Taylor: “The tilt intonation model”,In Proceedings of the International Conference on Spoken Language Processing, Vol.4,pp.1383-1386,1998.

【非特許文献９】S.Ananthakrishnan: “Categorical prosody models for spoken language applications”,PhD Thesis,University of Southern California,2008.

【発明の概要】

【発明が解決しようとする課題】

【０００９】

しかしながら、談話行為識別精度の向上のために用いられる、韻律特徴の従来の利用法は以下の理由で十分ではない。

まず、韻律特徴を発話文の特徴として抽出するためには、ユーザの一連の発話の中から、各発話文を正しく抜き出す必要がある。

しかし、文法や話し方のスタイルに制約のない自由発話においては、そもそも文の境界は不明瞭になりがちであり、しかも音声認識の間違いを含んだ発話を文の単位に正しく分節することは一般に難しい。

【００１０】

例えば、「もう一回言ってくれない」は文の最後にかけてピッチを上げながら話せば依頼を意味するが、十分な長さの無音区間を置かずに直後に「もう一回」と念を押した場合、この発話のピッチは文末にかけて上昇するとは限らない。

もしも、両者を別個の文として分離できない場合は、依頼の発話であっても文末のピッチ上昇は観測できなくなってしまう。

【００１１】

また、談話行為識別において、ある種の談話行為タグは、特定のフレーズに特定の韻律の変化を伴うことがしばしば観察される。

例えば、システムの発話をもう一度聞きたいという意図に対応する「言い直し要求」では、「言ってくれる？」、「言ってください？」のような特定のフレーズと同時にピッチの上昇が観察される場合が多い。
その場合、特徴の共起関係をより確実な識別の手がかりとすることで、より頑健かつ高精度な識別が期待できる。
【００１２】
ところが、韻律特徴を言語特徴とは独立に発話の特徴としてモデル化する従来の技術では、言語特徴と韻律特徴の共起関係を捉えることができない。
このような問題点に鑑み、本発明は、言語特徴と韻律特徴の相関を直接モデル化し、もって談話行為識別精度の向上を図る技術を提案する。
【課題を解決するための手段】
【００１３】
まず、入力されたユーザの音声から、音声認識に用いられる声道特徴量系列を抽出すると同時に、エネルギーや基本周波数など韻律に関する特徴量の時系列を抽出する。
【００１４】
次に、得られた音声認識結果と韻律特徴時系列から、談話行為識別に用いる特徴系列を合成する。
【００１５】
この合成特徴は、音声認識によって得られた単語Ｎ－ｇｒａｍに対して、その時区間に対応する離散的な韻律特徴を付与して得られる拡張単語Ｎ－ｇｒａｍとなっている。
【００１６】
こうして得られた拡張単語Ｎ－ｇｒａｍを入力として、言語特徴を利用した談話行為識別技術を適用し、識別モデルの学習やタグの予測を行う。
【発明の効果】
【００１７】
このように、言語特徴と韻律特徴を合成した特徴を用いることで、両者の相関を考慮した談話行為識別が可能になるだけでなく、韻律特徴が、発話文ではなく、単語Ｎ－ｇｒａｍに付与されていることで、発話文の正確な分節が得られない場合でも、韻律特徴を効果的に用いた談話行為識別が可能になる。
【図面の簡単な説明】
【００１８】
【図１】音声対話システム装置構成をあらわす図である。
【図２】実験結果(談話行為タグ識別精度)をあらわす図である。
【図３】実験結果(情報要求の適合率)をあらわす図である。
【図４】実験結果(言い直し要求の適合率)をあらわす図である。
【発明を実施するための形態】
【００１９】
次に、図１を参照して、本発明の音声対話装置の全体構成例を説明する。
【実施例１】
【００２０】
音声入力部(Ｐ１)において、発話はマイクロホン等を用いてアナログ信号として取得された後、ただちにデジタル信号に変換される。
【００２１】
声道特徴抽出部(Ｐ２)において、音声認識において用いられる、Mel Frequency Cepstral Coefficient(ＭＦＣＣ)等の声道特徴量が計算される。
【００２２】
また、エネルギー抽出部(Ｐ３)において、フレーム毎のエネルギーが計算され、声道特徴系列と併せてエネルギー特徴系列が音声認識部(Ｐ５)に送られ、音声認識が行われ単語列に変換される。
この時、各単語の発話における時区間を表すタイムスタンプを同時に計算しておく。

また、ここで音声認識部の出力を単語列としているが、正確には当該音声認識システム(Ｐ５)の使用する辞書に登録されている認識ユニットの列を意味しており、言語学的単語の列に限定されるものではない。

日本語のように単語に分かち書きされない言語の場合には、形態素解析のエラー等により、認識結果が正しく単語に分かち書きされない場合もあり、そのような場合には、音素やモーラのようなサブワードを認識ユニットとして用いて、サブワードユニットのＮ－ｇｒａｍを使って談話行為識別を行った方が良い場合もある。

また、識別に特徴的なフレーズ(単語列)がある場合は、フレーズを認識ユニットとして用いた方が良い場合もある。

本明細書では、典型的な認識ユニットである単語を用いて説明を行うが、認識ユニットは言語学的単語に限定されるわけではなく、単語Ｎ－ｇｒａｍは、認識ユニットのＮ－ｇｒａｍと読み替えることができる。

【００２３】

声道特徴やエネルギーと並行して、基本周波数抽出部(Ｐ４)では、フレーム毎に基本周波数が計算される。

【００２４】

次に、音声認識部(Ｐ５)で計算されたタイムスタンプ付きの単語列、およびタイムスタンプが付与されたエネルギー系列と基本周波数系列が言語・韻律特徴合成部(Ｐ６)に送られ、韻律特徴が付加された拡張単語Ｎ－ｇｒａｍが生成される。

【００２５】

この拡張単語Ｎ－ｇｒａｍと、対話制御部(Ｐ８)が提供する発話のコンテキスト情報から、意図推定部(Ｐ７)において、発話の談話行為タグが推定される。

【００２６】

引き続いて、対話制御部(Ｐ８)では、談話行為タグが表す発話意図と対話文脈を考慮して、システム発話が生成され、引き続く音声合成部(Ｐ９)で音声信号に変換された後、スピーカー等の音声出力部(Ｐ１０)を通して音声が再生される。

【００２７】

ここで、音声入力部(Ｐ１)、声道特徴抽出部(Ｐ２)、エネルギー抽出部(Ｐ３)、基本周波数抽出部(Ｐ４)、音声認識部(Ｐ５)、対話制御部(Ｐ８)、音声合成部(Ｐ９)、音声出力部(Ｐ１０)には公知の技術を用いることができる(非特許文献６、非特許文献７)。

【００２８】

以下では、言語・韻律特徴合成部(Ｐ６)および意図推定部(Ｐ７)についてのみ詳細に説明する。

【００２９】

本発明の１つの実施形態で用いられる、談話行為識別のための基本的な原理は、談話行為タグを表す確率変数Ｉの事後確率の最大化である。

【００３０】

【数１】

$$I^* = \underset{I}{\operatorname{argmax}} P(I|A) = \underset{I}{\operatorname{argmax}} P(A|I)P(I) \qquad (1)$$

ここで、Ａは音響信号を表す。

【００３１】

音響信号Ａが、声道成分A_sと韻律成分A_pに分離できると仮定すると、Ａの尤度は以下のように書ける。

【００３２】

【数２】

$$P(A|I) = \sum_W P(AW|I) = \sum_w P\left(A_s A_p \middle| WI\right) P(W|I) \qquad (2)$$

ここでＷは単語列を表す。

【００３３】

声道成分と韻律成分が条件付き独立であると仮定すると、

【００３４】

【数３】

$$P(A|I) \approx \sum_W P(A_s|WI) P\left(A_p \middle| WI\right) P(W|I) \qquad (3)$$

【００３５】

さらに、声道成分は、単語列のみに依存すると仮定すると、

【００３６】

【数４】

$$P(A|I) \approx \sum_W P(A_s|W) P\left(A_p \middle| WI\right) P(W|I) \qquad (4)$$

と書ける。

【００３７】

非特許文献１では、ここからさらに、韻律成分は単語列に依存しないと仮定し、

【００３８】

【数５】

$$P(A|I) \approx \sum_W P(A_s|W) P\left(A_p \middle| I\right) P(W|I) \qquad (5)$$

というモデル化を行う。

【００３９】

非特許文献２および非特許文献３においても、モデル化手法は異なるが、基本的に、韻律成分が単語列に依存せず、発話意図のみに依存するとしてモデル化を行っている。

本発明では、このような仮定(非特許文献１乃至非特許文献３)を置かず、数式(４)を、

【００４０】

【数６】

$$P(A|I) \approx \sum_W P(A_s|W) P\left(A_p \middle| WI\right) P(W|I) = \sum_W P(A_s|W) P\left(A_p W \middle| I\right) \qquad (6)$$

と、単語列Ｗと韻律特徴A_pを同時にモデル化する。

【００４１】

そのために、離散化された韻律特徴が付与された拡張単語列W^fを導入し、

【００４２】

【数７】

$$P(A|I) \approx \sum_W P(A_s|W)P\left(W^f\middle|I\right) \qquad (7)$$

とモデル化する。

【００４３】

離散化された韻律特徴ｆとしては、例えば、Ｗに対応するフレーム列において、基本周波数の変化量が平均＋標準偏差よりも大きい場合は＋を、変化量が平均－標準偏差よりも小さい場合は－を、あるいは変化量が平均±標準偏差の範囲内であれば０を付与する３値の離散化を用いることができる。

【００４４】

あるいは、強勢であれば、エネルギーが平均＋標準偏差よりも大きい場合はｓ＋を、平均－標準偏差よりも小さい場合はｓ－を、あるいは平均±標準偏差の範囲内であればｓ０を付与する３値に離散化を用いることができる。

【００４５】

例えば、「もう１回言って」という単語列に、強勢とピッチの上昇が観察されれば、「もう１回言って$^{+}$，$^{s+}$」と単語列が拡張されることになる。

もちろん、離散化のやり方は、これ以外の方法を考えることもでき、離散化の粒度も３値に限るものではない。

【００４６】

このような韻律特徴の離散化は、非特許文献８記載の、ｔｉｌｔ特徴と基本的な考え方は同じであり、このような特徴を非特許文献４では音声合成に、非特許文献９では音声認識に用いているが、本発明では談話行為識別に用いる。

【００４７】

また、非特許文献３では、韻律特徴量の軌跡を離散符号化し、符号のＮ－ｇｒａｍを談話行為識別に利用しているが、単語列との相関は考慮されていない。

【００４８】

ところで、このような韻律特徴の離散化の際には、話者間の変動や発話環境の変動に対処するため、話者毎また発話環境毎に特徴量の正規化を行うことが望ましい。

例えば、同一話者の直近複数の発話を用いて、平均値や標準偏差を計算することができる。

【００４９】

以上述べたような離散化を適用することで、言語・韻律特徴合成部(Ｐ６)において、音声認識により得られた単語列Ｗ＝w_1，…，w_nは離散韻律特徴が付与されたＮ－ｇｒａｍ列に拡張される。

【００５０】

その過程をより詳細に述べる。まず、ＷからＮ－ｇｒａｍ(Ｎ≧１)を抽出する。
例えば、Ｎ＝２であれば、

【００５１】

【数８】

$$\langle s\rangle w_1,\ w_1w_2,\ w_2w_3,\cdots,w_{n-1}w_n,\ w_n\langle/s\rangle \qquad (8)$$

が抽出される。

ここで、＜ｓ＞，＜／ｓ＞はそれぞれ文頭、文末を表す記号である。

このとき、単語の一種として、短い無音区間を表すw_i＝＜ｓｐ＞を含めれば、別種の韻律特徴を拡張Ｎ－ｇｒａｍの中に取り込むことができる。

【００５２】

次に、各単語に付与されたタイムスタンプに基づいて、各Ｎ－ｇｒａｍ毎に、対応する時区間におけるエネルギー系列と基本周波数系列の部分区間を抽出し韻律特徴を計算する。

その際、欠損値があれば線形補完等で補い、当該時区間の平均や変化量等の統計量を計算し、前述の正規化を施した後に離散化し各Ｎ－ｇｒａｍに付与される。

【００５３】

次に、意図推定部(Ｐ７)を説明する。

音声信号Ａが所与のとき、数式１を最大化する談話行為タグＩを計算する。

【００５４】

事前確率Ｐ(Ｉ)は、訓練データから予め計算した値を利用することができる。

このとき、発話のコンテキストを用いることでタグの識別精度が向上することが広く知られている。

例えば、非特許文献５では、直前のシステム発話の談話行為タグＣが所与であることを利用して、Ｐ(Ｉ)の代わりにＰ(Ｉ｜Ｃ)を用いることでタグの識別精度を向上させている。

【００５５】

数式１の尤度Ｐ(Ａ｜Ｉ)の計算には数式７を用いる。

数式７においてＰ(A_s｜Ｗ)は、音声認識で用いている音響モデルから計算される音響尤度を用いることができる。

つまり、音声認識部から出力される尤度上位ｎ個の単語列を正しい認識の候補と考える場合、各候補W_i(１≦ｉ≦ｎ)の音響尤度Ｐ(A_s｜W_i)を重みとするＰ(W_i^f｜Ｉ)の重みづけ和としてＰ(Ａ｜Ｉ)が計算される。

【００５６】

この重みづけ和を計算する際、一般に音響尤度は非常に小さな値になるので、桁落ちを防ぐために、非特許文献５では、音響尤度を尤度の最大値Ｍで正規化して用いており、本発明でも有効な計算方法である。

【００５７】

【数９】

$$P(A_s|W)/M,\ M=\max_i P(A_s|W_i) \qquad (9)$$

【００５８】

尤度Ｐ(W^f｜Ｉ)は、Ｎ－ｇｒａｍでモデル化される。つまり、Ｗに含まれるＮ－ｇｒａｍ、W_1，…，W_mに対して離散韻律特徴を付与した、拡張Ｎ－ｇｒａｍ列$W^f{}_1$，…，$W^f{}_m$が条件付き独立であると仮定し、

【００５９】

【数１０】

$$P(W^f|I)=\prod_{i=1}^{m}P(W_i^f|I) \qquad (10)$$

のように計算する。

ここで用いる、拡張Ｎ－ｇｒａｍの尤度Ｐ($W^f{}_i$｜Ｉ)は、予め訓練データから推定しておく。

【００６０】

ただし、離散韻律特徴で拡張されているので、訓練データが十分に多くない場合は、必ずしも拡張Ｎ－ｇｒａｍ　W_i^fが訓練データに含まれない場合が想定される。

そのような場合は、以下のような平滑化を行うことが望ましい。

【００６１】

【数１１】

$$P(W_i^f|I) = \sum_{j=0}^{J} \lambda_j P(W_i^{f_j}|I), \quad \sum_{j=0}^{J} \lambda_j = 1 \qquad (11)$$

ここで、ｊは自然数であり、ｆ$_j$(ｊ＞０)は、基本周波数の勾配、強勢等、ｆで用いられている離散韻律特徴である。

【００６２】

特に、ｊ＝０は、離散韻律特徴が付与されていないＮ－ｇｒａｍの生起確率であり、これ自身、ゼロ頻度問題に対処するために、Ｇｏｏｄ－Ｔｕｒｉｎｇ法など公知の方法(非特許文献６)を用いて平滑化されているものとする。

【００６３】

図２は、基本周波数(Ｆ０)の変化量を韻律特徴として用いた場合の、本発明による談話行為タグ識別率の向上を示している。

【００６４】

図２で「Ｆ０なし」として示されているのは、言語特徴のみを用いて識別した場合の識別精度を示している。

また、「Ｆ０変化(発話単位)」として示されているものは、一つの発話における基本周波数の勾配を３値に離散化した韻律特徴を、言語特徴とは独立に用いた場合の識別精度を示している。「Ｆ０変化(２ｇｒａｍ)」として示されているのは、単語２－ｇｒａｍ毎に、３値の離散韻律特徴を付与した拡張２－ｇｒａｍ特徴を用いた場合の識別精度を示している。

【００６５】

この結果から分かるように、韻律特徴を言語特徴と独立に付与しても、必ずしも識別精度は向上しない一方で、言語特徴と韻律特徴の相関を考慮する本発明によれば、識別率がおよそ１％向上していることが分かる。

【００６６】

図３と図４は、韻律特徴が寄与すると予想される「情報要求」と「言い直し要求」の２つの談話行為タグの適合率、すなわち、それぞれのタグを予測した発話の中で、正しい予測の割合を示している。

タグ個別にみると、「情報要求」でおよそ５％、「言い直し要求」で２％適合率が向上していることが分かる。

【００６７】

本発明のシステムは、マイクロホンとスピーカーとパーソナルコンピュータを用い、図１に示した各処理部を実行するプログラムをＣおよびＰｅｒｌ言語で作成し、実行して確認した。

作成したプログラムは、上で述べたように、汎用計算機を用いた汎用的なプログラムであってもよいし、各種音声対話システム・装置・機器にのみ適合する固有のプログラムであってもよい。

また、プログラムは、内蔵式、埋め込み式(Imbedded)、読み込み式、ダウンロード方式、分散型、あるいはクラウドコンピューティングであってもよい。

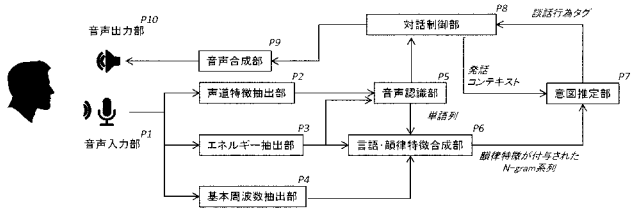
【００６８】

音声入力に使用したマイクロホンは、機器の一部として備わるマイクロホンであってもよいし、その設置場所は近接地・遠隔地を問わず、音声入力装置であれば足りる。

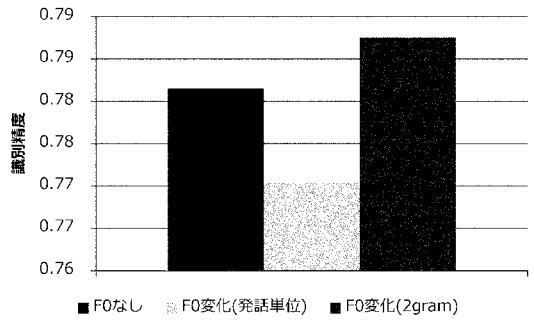
音声出力に使用したスピーカーは、機器の一部として備わるスピーカーや、イヤホンであってもよいし、その設置場所は近接地・遠隔地を問わず、音声出力装置であれば足りる。

音声入力信号はアナログ音響信号だけでなく、本発明の内部処理に適してデジタル化された音響信号のいずれであってもよい。

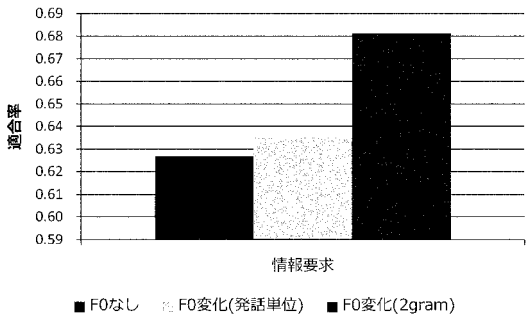
【図１】


音声出力部
P10
音声合成部
P9
対話制御部
P8
談話行為タグ
声道特徴抽出部
P2
音声認識部
P5
発話
コンテキスト
意図推定部
P7
単語列
音声入力部
P1
エネルギー抽出部
P3
言語・韻律特徴合成部
P6
韻律特徴が付与された
N-gram系列
基本周波数抽出部
P4

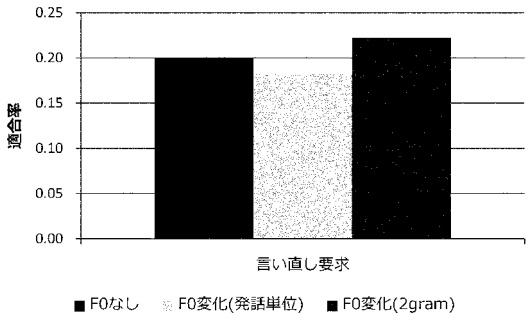
【図２】


0.79
0.79
0.78
0.78
0.77
0.77
0.76
識別精度
■F0なし
F0変化(発話単位)
■F0変化(2gram)

【図３】


0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.60
0.59
適合率
情報要求
■F0なし
F0変化(発話単位)
■F0変化(2gram)

【図４】


0.25
0.20
0.15
0.10
0.05
0.00
適合率
言い直し要求
■F0なし
F0変化(発話単位)
■F0変化(2gram)