社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－SLP－61 (3)
2006／5／11

キーワード主体の頑健な音声インタフェースのための韻律的特徴を用いた発話検証

甲斐 充彦　板倉 雅和[1]

静岡大学工学部　〒432-8561 浜松市城北3-5-1
E-mail: kai@sys.eng.shizuoka.ac.jp, itakura@spa.sys.eng.shizuoka.ac.jp

あらまし 本稿では、音声対話型インタフェースの利用において予想される文法外や冗長な表現、未知語などを含む発話に対して頑健な発話検証を実現することを目的として、キーワードスポッティングと韻律的特徴モデルによる照合を併用したキーワードレベルでの発話検証法を提案する。提案法では、音響・言語的特徴と韻律的特徴の統合による発話検証のための信頼度推定について定式化し、これをキーワードレベルで一般化して推定する手順を述べる。評価実験としては、検証する仮説として発話中に含まれるキーワード数を想定してその判定性能を比較した。結果として、音響信頼度のみによる判定方法と比べて、提案する韻律信頼度を含む統合的な信頼度尺度による判定方法は顕著に判定性能が改善された。
キーワード アクセントモデル、韻律的特徴、HMM、キーワードスポッティング、信頼度

Utterance verification with prosodic feature for keyword-based robust speech interface

Atsuhiko KAI　Masakazu ITAKURA

Faculty of Engineering, Shizuoka University
3-5-1, Johoku, Hamamatsu, Shizuoka, 432-8561, Japan
E-mail: kai@sys.eng.shizuoka.ac.jp, itakura@spa.sys.eng.shizuoka.ac.jp

Abstract This paper propose a keyword-based utterance verification method which can be robust for the utterance of out-of-grammar, out-of-vocabulary, and lengthy expressions, which can be found in the use of working speech interface systems. The proposed method employs a prosodic accent model for keyword-level verification, in addition to the acoustic linguistic features which are traditionally employed for hypothesizing recognition candidates and estimating a confidence measure. This paper presents an experimental result for our proposed utterance verification method, in which the hypothesis for the utterance verification corresponds to the number of keywords in an utterance. As a result, the classification rate of the number of keywords using the proposed integrated confidence measure significantly outperformed one using acoustic-only confidence measure.
Key words Accent model, prosodic feature, HMM, key-word spotting, confidence measure

1 はじめに

近年、音声認識システムの性能は次第に向上してきているが、その応用の一つとして注目される音声対話型のインタフェースにおいては、まだ広く使われるまでには至っていない。その一因として、現状の多くの音声対話型のインタフェースでは、我々人間との音声対話と比べて聞き直しや聞き誤りの修復など不完全な認識を前提とした音声言語の扱いが十分になされていないことが挙げられる。例えば、使用環境や話者、発話内容など様々な変動要因によって認識性能が大きく変化することが予想されても、システム側の扱いとしてはユーザの発話単位で認識結果を棄却するかどうか二値的にしか考えられていないことが多い。これに対して、音声認識システムが出力する認識結果の確からしさ（信頼度）を推定し、その値を単語やキーワードクラス単位での確からしさの情報として、対話システムにおける音声理解や応答候補の生成のために利用する先行研究もある [1, 2, 3]。論文 [2, 3] では、ユーザとシステム間の対話履歴において、ユーザ発話の複数の認識候補の信頼度を利用して単語レベルで“理解スコア”を更新し、対話を通しての頑健な意図理解を試みている。しかし、後者のように認識結果の曖昧さを定量的な情報として与える場合、音声認識システムがどのような単位でどの程度安定してそのような情報を提供できるかが重要な問題となる。

このような背景において、本研究では音声対話型インタフェースでより柔軟な意図理解・応答生成へ

[1]現在、(株) エィ・ダブリュ・ソフトウェアに所属。

の応用を想定して、前述のような認識結果の曖昧さの評価・利用において有用な付加的情報を補強することを考える。特に、機械相手でタスク指向の音声対話型インタフェースを想定して、キーワード主体での発話検証を試みる。機械相手でタスク指向の音声対話型インタフェースを考えたとき、ユーザに対して発話様式に関してあまり制約を与えない場合でも、人間同士での同様なタスクでの発話と比べて言語表現や音響的特徴の違いがみられる。特に、カーナビゲーションシステムへのランドマーク入力タスクを想定してユーザ発話の特徴分析を行なった先行研究 [4] においては、言語表現においては対人間の場合と比べて対機械では簡潔な表現となりやすく、動詞省略表現が多く、一つの発話で多くの情報を伝える、などの特徴が見られている。また、音響・韻律的特徴においても、対機械ではピッチの平均値や最大値が低く、ピッチの変化が小さい、というような特徴がみられる。このことから、対機械のユーザの発話では、対人間の場合と比べて言語特徴、音響・韻律特徴の簡素化を前提とすれば、比較的安定したモデル化や検証が可能になるものと予想される。

そこで、本研究では、対機械のユーザ発話の意図理解において最も影響が大きいことが予想されるキーワード（フレーズ）レベルで、その認識候補の検証のための手掛かりや付加的情報を与えることを目的とする。具体的には、キーワードレベルでの音響・言語的特徴と韻律的特徴を併用することにより、信頼度の推定と検証を行なう方法を提案する。発話検証は 2 段階からなり、文法外の表現や、冗長な表現や未知語などを含む多様なユーザの発話を考慮するため、始めにキーワードスポッティングによって複数候補の仮説を生成する。次に、その仮説候補に対して韻律的特徴のモデルによって照合を行ない、両者の情報を統合した信頼度を推定する。本稿では、特にカーナビゲーションシステムにおけるユーザ主導で対話型の音声入力インタフェース [2] を想定した評価実験による結果を報告する。韻律的特徴の利用に関しては、これまでにも F_0 パターンを用いたアクセント句（キーワードに対応）境界の検出などの先行研究がある [5]。本稿では、アクセント核に注目したモーラ区間単位での韻律特徴の簡易的な統計的モデル化により、キーワードクラス単位での検証の目的において適用が容易な手法について述べる。

2 発話の検証法

2.1 複数候補による信頼度推定

音声認識の問題は、入力音声 $\mathbf{X}$ が与えられたときの事後確率 $P(\mathbf{W}|\mathbf{X})$ を最大化する単語列 $\mathbf{W}$ を見つける問題として、式 (1) のように定式化される。$P(\mathbf{X})$ は $\mathbf{W}$ に対して独立であるので、式 (2) のように等価に表され、一般的な音声認識システムでは音響尤度 $P(\mathbf{X}|\mathbf{W})$ 及び言語確率 $P(\mathbf{W})$ のみで評価がなされる。

$$\hat{W} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{X}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{P(\mathbf{W})P(\mathbf{X}|\mathbf{W})}{P(\mathbf{X})} \quad (1)$$

$$= \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W})P(\mathbf{X}|\mathbf{W}) \quad (2)$$

入力音声に対して尤度の高い順に複数候補 (N-best) の解を求める孤立単語認識の場合、第 i 候補の尤度を $P(\mathbf{X}|w_i)P(w_i)(1 \leq i \leq N)$ とすると、ある単語 w_i の事後確率 $C(w_i)$ は次式 (3) で近似的に与えられ、これを単語 w_i の信頼度と考えることができる。

$$Confidence(w_i) = \frac{P(w_i)P(\mathbf{X}|w_i)}{\sum_{j=1}^{N} P(w_j)P(\mathbf{X}|w_j)} \quad (3)$$

2.2 キーワードレベルでの韻律的特徴

日本語のアクセントには「高」と「低」の 2 レベルがあり、各モーラ（日本語では基本的に 1 モーラは、かな 1 文字にあたる）は 2 レベルのいずれかに対応する。「高」から「低」に変化する位置をアクセント核と言う。韻律的特徴のうち、一般的にキーワード単位ではアクセント核が一つ存在することに注目し、モーラ単位での高低アクセントの特徴をモデル化する。特に発話の基本周波数の特徴量によりモーラ単位の韻律的特徴の統計的モデルを構築するが、モデル化の詳細については 3 章で述べる。

本稿では、キーワードの単位としては複合名詞を含めて考える。複合名詞のアクセントは、その複合の度合いが強いかどうか、何拍語であるか、それがどのような種類の語か、それらアクセントがどうだったか、などによって全体のアクセントが定まる。後述する想定タスクでは、キーワードとしては交通機関の地名や施設名などであり、これらは規則的なアクセントを持っている。そのアクセントは、後部要素（末尾語）によって決まるものが多い（本稿では「県」や「インター」など、キーワードの末尾に付く語を末尾語という）。従って、次の節で述べるキーワード主体の検証法において、必ずしもタスク内の語彙単語全てについて個別のアクセント情報を持っておく必要がなく、検証用モデルの構成は容易である。各キーワードのアクセント規則を表 1 に示す。

2.3 韻律的特徴を含めたキーワードレベルでの検証

本研究では、キーワードスポッティングと韻律的特徴を併用することによって発話検証を行う。キー

表 1: 各キーワードのアクセント規則

都	トーキョート
府	○○○フ
県	○○○ケン
市	○○○シ
区	○○○ク
町	○○○チョー, ○○○マチ
村	○○○ムラ, ○○○ソン
自動車道	○○○ジドーシャドー
高速道路	○○○コードクドーロ
高速	○○○コーソク
道	○○○ドー
インター	○○○インター
インターチェンジ	○○○インターチェンジ
ジャンクション	○○○ジャンクション
駅	○○○エキ
鉄道	○○○テツドー
線	○○○セン
本線	○○○ホンセン
新幹線	○○○シンカンセン

○ 低アクセント
○ 高アクセント
○ アクセント核

ワードスポッティングによる音声認識の N-best 結果である各仮説に対して、音響モデルと韻律パターンモデルを使用し、各仮説の尤度を求める。本研究では、特にアクセント核情報の利用を想定したモデルによって、発話中のキーワードの存在を検証する。

$\mathbf{A}$ を入力音声の音響的特徴系列、$\mathbf{B}$ を入力音声の韻律的特徴系列、$\mathbf{H}_K$ を検証するキーワード（フレーズ）レベルでの仮説とする。ここで、キーワードレベルでの仮説とは、一発話中に含まれるキーワード列（例えば、後述の評価用タスクでは「静岡県」と「浜松インター」）や、キーワードのクラスの列（例えば、「県名」と「インターチェンジ名」）、またはキーワード数、などが考えられる。入力された音声から得られる $\mathbf{A}$ と $\mathbf{B}$ から仮説 $\mathbf{H}_K$ の信頼度を求めるため、次式 (4) のような事後確率を考える。

$$P(\mathbf{H}_K|\mathbf{A},\mathbf{B}) = \frac{P(\mathbf{A},\mathbf{B}|\mathbf{H}_K)P(\mathbf{H}_K)}{P(\mathbf{A},\mathbf{B})} \quad (4)$$

仮説 $\mathbf{H}_K$ は複数の単語列仮説 $\mathbf{W}$ の集合として表されるとすると、

$$P(\mathbf{H}_K|\mathbf{A},\mathbf{B}) \fallingdotseq \frac{\sum_{\mathbf{W}\in\mathbf{H}_K} P(\mathbf{B}|\mathbf{W},\mathbf{A})P(\mathbf{A}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{A},\mathbf{B})} \quad (5)$$

更に、単語列 $\mathbf{W}$ が M 個のモーラの並び $\mathbf{W} = m_1 m_2 \cdots m_M$ として表され、$P(\mathbf{A}|\mathbf{W})$ と $P(\mathbf{B}|\mathbf{W},\mathbf{A})$ はモーラ単位のモデルの連鎖として表現されるとする。また、アクセントパターンはモーラ区間単位で特徴付けられ、各モーラ区間 $(t_1, t_2, \cdots, t_M)$ が音響的特徴 $\mathbf{A}$ にのみ依存するとする仮定すると、

$$P(\mathbf{A}|\mathbf{W}) = P(\mathbf{A}|m_1 m_2 \cdots m_M) \quad (6)$$

$$P(\mathbf{B}|\mathbf{W},\mathbf{A}) \fallingdotseq \sum_{\substack{m_1 m_2 \cdots m_M \\ t_1 t_2 \cdots t_M}} P(\mathbf{B}|m_1 m_2 \cdots m_M, t_1 t_2 \cdots t_M) \cdot P(m_1 m_2 \cdots m_M, t_1 t_2 \cdots t_M|\mathbf{A},\mathbf{W}) \quad (7)$$

ここで、式 (7) は、$\sum$ で仮定するモーラ列や区間候補の集合の範囲を、音響的特徴 $\mathbf{A}$ を用いるキーワードスポッティングによって得られた複数候補の結果に対応するものだけに限定し、近似して求める。

また、アクセント特徴がモーラ単位の高低アクセントの系列としてモデル化され、ある単語列に付与される高低アクセントパターンが $b_1 b_2 \cdots b_M$ のとき、式 (7) の第 1 項の $m_1 m_2 \cdots m_M$ は $b_1 b_2 \cdots b_M$ に置き換えられる。

式 (6),(7) は想定する統計的モデルによって理論上は計算されるが、計算を簡略化するために $\mathbf{A}$,$\mathbf{B}$ の独立性を仮定して、次式のように更に近似して音響信頼度 C_{ac} と韻律信頼度 C_{pros} とに分離して定義する。

$$P(\mathbf{H}_K|\mathbf{A},\mathbf{B}) \fallingdotseq C_{ac}(\mathbf{H}_K) \times C_{pros}(\mathbf{H}_K) \quad (8)$$
$$\equiv Confidence(\mathbf{H}_K) \quad (9)$$

一般的なフレーム特徴系列による統計的なモデルでは、音響モデルの尤度や韻律モデルの尤度のレンジは発話長に依存する。そこで、式 (8) の各項は、次式のように時間（フレーム）長 L で正規化して求める。

$$C_{ac}(\mathbf{H}_K) \equiv \frac{\sum_{\mathbf{W}\in\mathbf{H}_K} \{P(\mathbf{A}|\mathbf{W})\}^{\frac{1}{L}} P(\mathbf{W})}{\sum_{\mathbf{W}} \{P(\mathbf{A}|\mathbf{W})\}^{\frac{1}{L}}} \quad (10)$$

$$C_{pros}(\mathbf{H}_K) \equiv \frac{\sum_{\mathbf{W}\in\mathbf{H}_K} \{P(\mathbf{B}|\mathbf{W})\}^{\frac{1}{L}} P(\mathbf{W})}{\sum_{\mathbf{W}} \{P(\mathbf{B}|\mathbf{W})\}^{\frac{1}{L}}} \quad (11)$$

式 (8) は、更に音響信頼度と韻律信頼度のどちらにどのくらいの重みを置くかということも考えられる。音響信頼度の重みを α、韻律信頼度の重みを β とすると次のように定義される。

$$Confidence(\mathbf{H}_K) \equiv C_{ac}^{\alpha}(\mathbf{H}_K) \times C_{pros}^{\beta}(\mathbf{H}_K) \quad (12)$$

このように定義される信頼度を用いて複数の仮説 $\mathbf{H}_K$ 間で比較をする場合には、式 (12) は次式と等価である。

$$Confidence(\mathbf{H}_K) \equiv C_{ac}(\mathbf{H}_K) \times C_{pros}^{\frac{\beta}{\alpha}}(\mathbf{H}_K) \quad (13)$$

本稿では、後述の評価実験において、仮説 $\mathbf{H}_K$ としては発話中に含まれるキーワード数（1〜3）を仮定しており、キーワード数の判定のためこの式 (13) を用いる。また、式 (10)〜式 (11) において、$P(W)$ は後述の評価実験では全て一様と仮定して無視する。

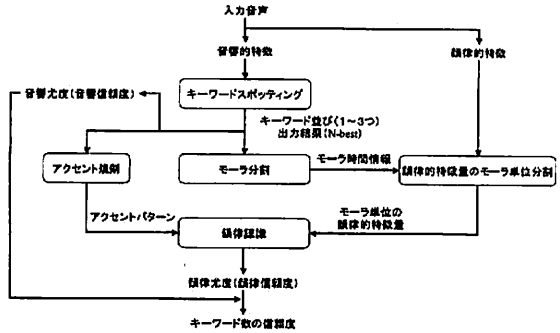

図 1: 発話検証の流れ

2.4 発話検証の流れ

キーワードスポッティングと韻律的特徴を併用した発話検証の手順を図 1 に示す。まず、音響信頼度を求める手順としては、入力音声を音声認識器 SPOJUS[6] に与えてキーワードスポッティングを行い、出力結果として得られるキーワードの並びの尤度から式 (10) を利用して音響信頼度を求める。

次に韻律信頼度を求める手順を説明する。まず、音響信頼度の算出で用いるキーワードスポッティングの結果から、キーワード並びの複数仮説の情報を得る。このキーワード並びの仮説をモーラ単位に分割し、モーラ時間情報を得る。モーラ時間情報を利用して、入力音声から得られた韻律的特徴をモーラ単位に分け、韻律特徴モデルとの照合を行なう韻律認識部に与える。韻律認識部にはキーワード並びに対応するアクセント規則も与える。これらの情報から式 (11) を利用して韻律信頼度を求める。得られた音響信頼度と韻律信頼度から式 (13) を利用して、検証する仮説の信頼度を求める。本稿では検証する仮説はキーワード数がそれぞれ 1,2,3 個のいずれかの場合に対応する。

3 アクセント特徴のモデル化

前述のようにキーワード（キーフレーズ）単位での韻律的特徴、特に高低アクセントによって表されるパターンに関しての照合を可能にするため、モーラ単位でのアクセント特徴のモデル化を行なう。具体的には、高アクセントを H、低アクセントを L、無音区間を P として、先行モーラのアクセントの文脈に依存したモデルを、モーラ単位で構築する。これには 4 状態 2 出力分布の left-to-right 型 HMM を使用する。

使用する音声データの分析条件は表 2 である。特徴量としては、発話内の平均値で正規化された logF0 を用いる。logF0 は時間によって変化するので、1 モーラに対応する logF0 を前半と後半に分け、前半と後半のそれぞれで正規分布を仮定した出力分布を推定する。学習用データとしては、モーラ単位の時間ラベルを用いて分割を行なったあと、HMM の各状態に等フレームで分割割当てし、推定パラメータを固定した。韻律的特徴系列のモデル化の概念図を図 2 に示す。

表 2: 分析条件

サンプリング周波数	16kHz
フレーム幅	25ms
フレーム周期	10ms
特徴パラメータ	logF0 (1 次元)

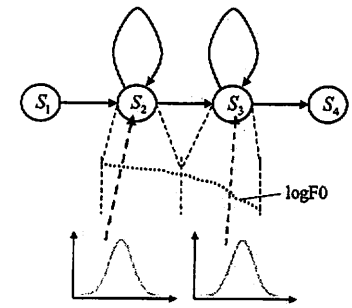

図 2: 統計的韻律モデルの概念図

上述のような統計的モデルで基本周波数の特徴量を扱う場合、一般的な固定次元の特徴ベクトルによる定式化では、無声区間が扱えないという問題がある。本論文では、文献 [7] で提案されている可変次元の多空間上における確率分布に基づいた HMM によるモデル化法を採用する。この方法は、有声区間を 1 次元空間からの出力、無音区間を 0 次元空間からの出力として、それらの混合分布として統計的に有声・無声が混在した韻律的特徴をモデル化するものである。本研究では HTS として提供されているツール [8] を利用した。

モデルの学習に用いる音声データは、後述する評価実験と同様にランドマーク入力タスクに関しての読み上げ音声を用いた。これは男性話者 4 名による発話文の総計 1600 発話である。1 発話には 1～3 キーワードが含まれる。このデータを半分に分け、それぞれ学習用と評価用のデータにした。

4 評価実験

4.1 キーワード抽出実験

文法（CFG）ベースの音声認識器 [6] をもとに、キーワードスポッティングのための言語制約を与えてキーワードスポッティングシステムを実現する。タスクはカーナビゲーションシステムを想定したラ

ンドマーク入力タスクであり、一発話中には最大で3つのキーワードを想定する。例えば、3キーワードを含む発話の文としては「静岡県の東名自動車道の浜松インターだよ」のようなものがある。厳格に定義されたタスクの文法では、音声認識システムの語彙サイズは14430単語、評価セットの単語パープレキシティは1013.9であり、評価データには未知語や文法外の発話が約17.8％含まれる。

本稿でこれ以降に述べる評価実験では、上記の想定するタスクにおいて“キーワード数”を検証する仮説とする。すなわち、ある発話に含まれるキーワード数が1〜3個のいずれかを判定することで評価を行なう。そこで、キーワードスポッティングの仮説としてキーワード数が1から3個までのそれぞれの仮説候補を得るため、それぞれの個数のキーワードを一発話中に許容するキーワードスポッティング用の文法を並列に定義して用いる。また、文法外の発話や冗長な表現、未知語などを含む発話を想定するため、一発話中の複数キーワード間の制約は与えない(キーワード数が2または3の言語制約の場合)。そして、発話の先頭・末尾や、キーワードとキーワードの間、キーワードの末尾語の先行部分(「○○インター」、「△△駅」などの○○や△△の部分)、などには未知語のモデルを許容している。なお、未知語の仮説は任意の音節列としての照合によって得られるものであり[6]、沸き出し誤りを抑えるため未知語候補には一定のペナルティスコアを与えている。

本節では、上記のようなベースとなるキーワードスポッティングシステム構成での評価実験結果を示す。評価方法として、発話されたキーワードの中でどれだけの正解キーワードが抽出できるかを調べた。評価用の音声データとしては、前述のランドマーク入力タスクにおける男性4名の話者による発話のべ800発話(＝200発話×4名)を用いる。この際に、キーワード仮説Kのスコア$S(K)$を式(14)で定義する。認識の途中結果中に含まれるキーワードKに対してそれぞれ$S(K)$を求め、その値がしきい値を超えたかどうかによって、キーワードKを抽出する。

$$S(K)=\frac{(K\text{の尤度})-(K\text{の区間の任意音節列の尤度})}{K\text{の区間長}} \quad (14)$$

キーワードスポッティングを利用した認識において、しきい値を変化させた時の正解キーワード抽出率を図3に示す。ここで、「末尾語なしは不許可」というのは、末尾語が脱落した場合は誤りと見なすということである。例えば「浜松市」というキーワードが発話に含まれる時、「浜松」というキーワードが抽出できていても、正解キーワードが抽出できたことにはならないということである。図3の結果より、キーワード抽出率は4話者のトータルで約93％となっている。

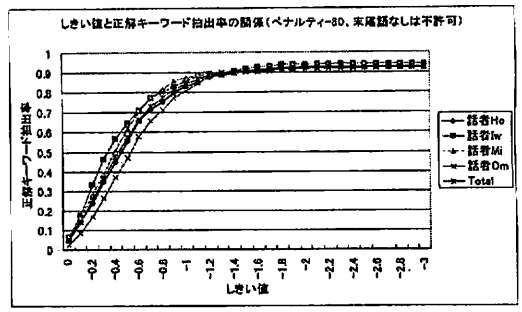

図3: $S(K)$のしきい値と正解キーワード抽出率

4.2 キーワード数の検証

評価にはキーワード数の検証結果の正解率を考える。これは評価データ全体で発話内のキーワード数が正しく判定された発話の割合として定義される。

各信頼度をキーワードスポッティングの各キーワード数の第一位候補のみで計算した。ベースラインとして、音響信頼度(式(10))のみ、韻律信頼度(式(11))のみのキーワード数正解率はそれぞれ66.8％、78.4％であった。

式(13)では、αとβの値を変化させることで、キーワード数の正解率が最大になるように調節できる。評価データを4つに分け、それぞれの評価データ(各データは詳細化・訂正発話50発話ずつ)において、2つの重みの比β/αを変化させた時のキーワード数正解率は図4のようになった。図4より、評価データの違いによらずほぼ$\beta/\alpha=12$付近のパラメータで最良の結果が得られていることが分かる。最も良い条件において、式(13)の信頼度では約90.3％のキーワード数正解率になり、ベースラインでの性能と比較すると、音響信頼度のみによる判定法と比べて約23.5％、韻律信頼度のみによる判定法と比べて約11.9％改善された。

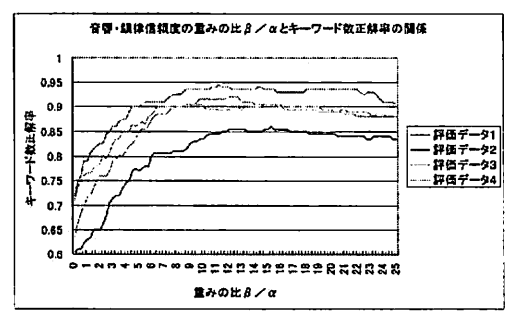

図4: 重みの比β/αとキーワード数正解率の関係

評価データには一発話中に3キーワードを含む例の割合が12％で最も少なかったが、音響信頼度のみ

の判定で3キーワードとしての誤り候補が多くみられた。その詳細を見ると、「(静岡県の東名自動車道の)市」のような末尾語だけの挿入による誤り例が多くを占めており、必ずしもアクセント情報の有用性を示すものとはいえない例が多く含まれた。そこで、検証する仮説として3キーワードの例を除外して、1または2キーワードの評価データ(704発話)を利用して、さらに同様な評価実験を行なった。

その結果、音響信頼度のみと韻律信頼度のみの結果は、それぞれ正解率が88.5 %と82.5 %となり、音響信頼度のほうがやや高い正解率となった。さらに、両者を統合した信頼度($\beta/\alpha = 12$)では95.0 %の正解率となり、韻律的特徴の併用による検証法の効果が示された。内訳をみると、韻律信頼度の併用によって改善される例は、1キーワードの発話に対して音響信頼度のみでは2キーワードとして誤って判定される例がほとんどであった。なお、音響信頼度のみで2キーワードと判定を誤った例では、前述のように「…の市」や「…の駅」のような末尾語のみの挿入例はなく、一般的なキーワード(厳格な文法では文法外となるようなキーワードの組を含む)の挿入誤りとなった例であった。従って、提案する手法は、このような文法外や冗長な表現を想定するような状況において、頑健な発話検証を可能にするものと予想される。

ところで、この評価実験では、冗長な表現部分において一部の未知語や文法外の表現を含む以外では、意味的な文法外(例えばキーワードの並びがタスクの厳格な制約の範囲外)の発話データを全く含んでいない。従って、本研究で想定しているようにそのような発話を含めて評価することで、本稿で提案する韻律信頼度を併用した方法の有効性をさらに示すことができるものと思われる。また、本稿では、検証する仮説としてキーワード数のみに注目しており、キーワードのクラス列やキーワードそのものの検証を含めた総合的な評価を行なっていない。今後は、このような複数のレベルでの検証を含めて、音声対話インタフェースにおける応用においての有用性をさらに調査していく必要がある。

5 まとめ

本研究では一発話に含まれるキーワード単位に注目し、キーワードスポッティングと韻律的特徴を併用した発話検証法を提案した。韻律的特徴を用いた検証では、キーワードクラスに依存する簡単なアクセント規則から、モーラ単位での比較的単純なアクセントモデルをもとにキーワードレベルのモデルを構成し、照合する方法を提案した。評価実験では、文法外や冗長な表現を含むことを想定してキーワードスポッティングを基本とした評価条件において、キーワード数に関しての検証として提案法を適用した。その結果、韻律的特徴を併用する場合において、音響特徴のみによる方法と比べて顕著に高い検証性能が得られた。

なお、本稿で用いた評価用データは、冗長な表現を含む場合を除いて意味的な制約外のような文法外発話の例は少なかったが、そのような例が多い場合には一層有効性が示されるものと期待される。しかし、厳格な文法を用いた場合との比較や、話者やタスクの違いの影響の評価、本稿で定式化したキーワードレベルでの検証法においてキーワード数以外において適用した場合の有効性の評価など、今後さらに検討を進める必要がある。また、今回の実験で用いたアクセントのモデルでは、モーラ区間単位での高低アクセントの違いのみによって比較的単純なモデル化を行なった。従って、更に話者に依存したモデルやキーワードのクラスに特化したモデルなど、モデルの詳細化を検討することも今後の課題である。

謝辞

この研究の一部は中部電力基礎技術研究所の研究助成の支援を受けて実施された。

参考文献

[1] 駒谷, 河原:「音声認識結果の信頼度を用いた効率的な確認・誘導を行なう対話管理」, 情報処理学会論文誌, Vol.43, No.10, pp.3078–3086 (2002.10)

[2] 水谷, 伊藤, 甲斐, 小西, 伊東:「音声認識の信頼度と対話履歴を利用した最尤推定型言語理解」, 情報処理学会研究報告, SLP-45, pp.113-118 (2003.2)

[3] 藤原, 伊藤, 荒木, 甲斐, 小西, 伊東:「認識信頼度と対話履歴を用いた音声言語理解手法」, 電子情報通信学会論文誌, D-II (2006. 予定)

[4] 伊藤, 甲斐, 岩本, 水谷, 由浅, 小西, 伊東:「目的地設定タスクにおける対話状況の違いによる言語・音響的特徴の比較」, 情報処理学会論文誌, Vol.43, No.7, pp.2118–2129 (2002)

[5] 岩野公司 他:「モーラを単位とした基本周波数パターンの確率モデル化とそれによるアクセント句境界の検出」, 情報処理学会論文誌, Vol.40, No.4, pp.1356-1364 (1999.4)

[6] 甲斐, 中川:「冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価」, 電子情報通信学会論文誌, Vol.J80-D-II, No.10, pp.2615–2625, 1997.

[7] 徳田恵一 他:「多空間上の確率分布に基づいたHMM」, 電子情報通信学会論文誌, D- Vol.J83-D- No.7 pp.1579-1589 (2000.7)

[8] HMM-Based Speech Synthesis System (HTS): http://hts.ics.nitech.ac.jp