ハンズフリーの音声インタフェースを実現する自由音声対話技術

EUROPA: A Framework for Building Hands-Free Spoken Dialogue Interfaces

笹島　宗彦　SASAJIMA Munehiko
屋野　武秀　YANO Takehide

より自然な話しことばによる計算機との対話を目標として，新音声対話システム構築方式を開発した。話しことばに対応するために，音声認識方式としてキーワード スポッティングを採用し，多様な言いまわしを反映させる文法記述を可能とした。また，音声認識結果に対して高速解析が可能な構文解析方式を開発し，実用的な数の文パターンを，ほぼ実時間で処理できる見通しが得られた。この方式をベースとするカーナビゲーション音声対話システムは，対象語彙（ごい）数300から700，対象文パターン200万以上を受理できる。すべてのプロセスは，ノートパソコン（PC）1台で動作し，利用者の発話終了から，おおむね2秒以内に応答が可能である。

We have developed a new framework for building spoken dialogue systems, called EUROPA (Environment for utterance recognizable packages). To cope with spoken language, the voice recognition module of EUROPA performs keyword spotting. The grammar rules for EUROPA allow various ways of representing users' intentions. Furthermore, we have developed BTH, a new method for identifying input sentences from a very-large-scale keyword lattice within a short time.

We have applied EUROPA to build MINOS (Mobile interactive navigation speech system), a prototype system for car navigation, which can accept more than 2 million sentence patterns and answer within 2 seconds in most cases.

1 まえがき

計算機性能の飛躍的な向上に伴い，従来は実現困難であった音声やジェスチャなどのマルチメディア情報を利用した計算機への入力も可能となりつつある。特に，音声対話機能は，手を使わずに操作できるハンズフリーや，目で見なくても操作できるアイフリーの特性を備えており，利用者が“ながら（何かほかのことをしながら）”利用する状況，例えば，カーナビゲーション（以下，カーナビと略記）システムなどへの適用が望まれている。

実環境に音声インタフェースを適用する場合の課題として，当社は次の三つを考えている。

一つ目は話しことばへの対応である。ここで話しことばとは，われわれが文法規則などを意識せずにふだん使っていることばや文を意味する。話しことばには，「えーと」や「あのー」など不要語の挿入や，言いよどみ，助詞の言い間違いなどの文法規則で想定していない現象が起きる。車の運転など緊張している状態では，これらの現象は更に起きやすい。

二つ目はインタフェースの即応性である。インタフェースの快適性を考えると，入力から応答までの時間は短いほうが良い。特に，カーナビなど数秒遅れると応答文がまったく無意味になるようなアプリケーションシステム（以下，アプリケーションと略記）の場合は，即応性が更に強く要求される。

最後は，個別システムの構築容易性である。機器に応じて，どのような音声対話機能を搭載すべきかが異なる。機器に合わせて柔軟に機能変更できる枠組みが必要である。

以上の課題を解決するために，当社は，音声対話フレームワークEUROPA（Environment for Utterance RecOgnizable PAckages）を開発し，カーナビ音声対話システムMINOS（Mobile Interactive NavigatiOn Speech-system）の試作に適用した。ここでは，EUROPAの設計方針及びシステム構成と，MINOSについて述べる。

2 EUROPAシステムの概要

EUROPAでは，**図1**のように音声認識，構文解析，問題解決の順に処理を進め，利用者に応答を返す。音声認識は利用者の音声入力を認識し，EUROPA内部で利用するためのシンボルに変換する。構文解析では，あいまい性を持つ音声認識結果からEUROPAで受理可能な入力文候補群を抽出する。問題解決では，各文候補が表す質問に対する回答を知識ベースを参照しながら検索し，その結果から応答文を生成する。知識ベースは対象タスクに関連する情報を蓄積した意味ネットワークである。最終的に生成された文章は音声合成され，利用者に提示される。

2.1 音声認識部 ― 話しことばへの対応

話しことばに対応することを念頭におくと，音声認識部についても話しことば特有の現象による影響を軽減するように考慮する必要がある。話しことばには，不要語や言いよどみなどが現われるが，これらの現象をシンボルに変換し

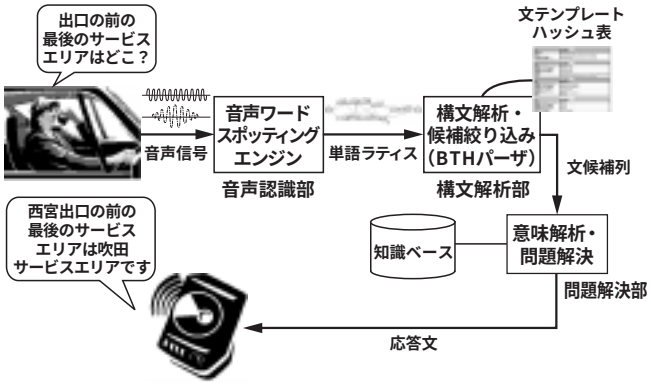

図1．EUROPA のフレームワーク　構文解析部及び問題解決部を高速化することによって，全体の高速化が実現されている。
Configuration of EUROPA system

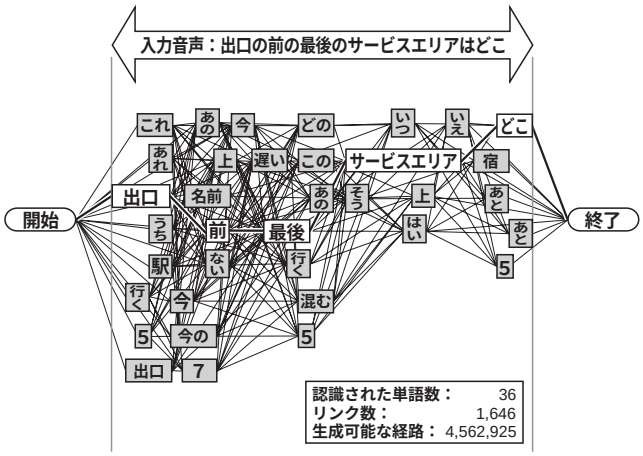

図2．キーワード ラティスの例　認識する単語が増えると，キーワード ラティスの規模が爆発的に大きくなる。
Example of keyword lattice

ても入力理解には直接必要のない情報となる。処理量を抑えるためにも，できるだけ早い段階でこのような必要でない情報は無視できることが望ましい。

そこで，当社は，音声認識部にキーワード スポッティング方式を採用した。この方式は，認識対象語彙（キーワード）に該当する音声信号区間を入力音声信号から検出し，その区間と対応するキーワードの組を出力する音声認識方式である。システムは，抽出されたキーワードを時間的な条件に基づいて並べたキーワード系列を文として受理し，意味解析を行う。これにより，話しことば入力に特有の不要語や言いよどみ現象による影響を軽減する音声認識が可能となる。

2.2　構文解析部 ― 即応性の確保

構文解析部は，音声認識結果からシステムで受理可能な文候補群を選択する役割を持つ。上述のように，EUROPA では音声認識部にキーワード スポッティングを採用している。音声認識部は，入力された音声信号のどこがキーワードの区切りなのか一意には判断できないので，あいまい性を残したまま認識結果を出力する。具体的には，音声信号に対して仮定できるキーワードの区切り方すべてを対応させたものを認識結果として出力する。例えば，音声信号から“ DEGUCHI ”という音の並びが得られた場合，その音声信号に対してキーワード“ 出口（DEGUCHI）”のほかに，自宅を表すキーワード“ うち（UCHI）”を対応させる。更に，似ている音声信号と対応するキーワードも認識結果として出力する。例えば，“ DEGUCHI ”の“ ぐ（GU）”という音は数字の“ 5（GO）”に似ているので，キーワード“ 5 ”も出力される。

音声認識結果は，音声信号から抽出されたキーワードを一つのノードとしたグラフ構造として扱われる。このグラフ構造をキーワード ラティス（以下，ラティスと略記）と呼ぶ。**図2**は，「出口の前の最後のサービスエリアはどこ」という発話の認識結果に対応するラティスの例である。左右方向は時間軸を表し，各ノードはその出現時刻に合わせて配置され，その幅はキーワードが占める時間を表している。

構文解析部は，ラティスから文として受理可能なキーワードの並びを抽出する。一般に，大きなタスクを音声対話システムで扱おうとすると音声認識の対象語彙が増加するため，認識結果のあいまい性が増加する。それにより，ラティスから抽出可能なキーワードの並びの数は爆発的に増加し，構文解析部の負荷が増大する。例えば，図2の場合，生成可能な経路の数，すなわち構文解析部が検証しなければならないキーワードの並びの数は約450万通りとなる。インタフェースの即時性を保つためには，多くのキーワードを含む大規模なラティスについても，高速に解析する構文解析部が必要である。

この問題に対処するため，当社は文テンプレートハッシュ方式（BTH：Bun（meaning ″sentence ″in Japanese）Template Hash）を開発した。BTHは，キーワード系列に展開することなくラティスを解析することにより，受理可能なキーワード系列の集合を高速に抽出することが可能である。BTHによる構文解析の最大計算量は，ノード間の辺の数に比例する回数の集合演算となる。ラティスにおける辺の数は，ラティスを文系列に展開した場合のキーワード系列数と比較して圧倒的に少ないため，従来方式と比較してより高速に受理可能な文を検索することができる。

2.3　問題解決部 ― 個別システム構築容易性の実現

問題解決部では，構文解析部からの出力であるシステムで受理可能なキーワード系列（文候補）の集合を入力として，その個々の文候補に対応する問題解決を実施し，問題解決結果に基づき応答文章を生成する。

各文候補の解析は次のように行われる。文候補集合から一つずつ候補が取り出され，意図変換器によってユーザーの意図を表現するデータ構造（ユーザー意図表現）に変換さ

れる。各ユーザー意図表現に対する問題解決が実行され，意味的な観点から各文候補に対してスコアが与えられる。最後に，音声信号としての観点と意味的な観点から，もっともスコアが高かったものについてユーザーに対する応答文を生成する。応答文は音声合成モジュールに入力され音声で出力される。

当社が開発したMINOSの問題解決部は，図3のような構成となっている。問題解決部のモジュールやデータ群は，対象領域（ドメイン）独立のものとドメイン依存のものに分かれている。前者は，ユーザー発話に対応する地点を地理データベースから検索するエンジン（場所解決器）など，いわゆるアプリケーション特有の問題解決モジュールと，応答発話パターンの辞書（応答生成テンプレート）など，アプリケーションに特有なデータで構成される。後者は知識ベースへのアクセス管理部（知識ベースマネージャ）など汎用のデータマネジメントモジュールが主となって構成される。

これらのモジュールやデータを利用して対話システム全体を制御する手続きは，その対話システムの応答生成方針やアプリケーションに依存する。当社は，対話システムの動作を記述するためのスクリプト言語USHI（Unification-based Script Handling Instruction set），及びUSHIインタプリタ（スクリプト言語を解釈して実行するモジュール）を開発した。対話システムの開発者は，ドメイン独立／依存のモジュールを呼び出し，制御するコードをUSHIスクリプトにより記述し，システムがそれを解釈実行することでシステム全体の動作を制御できる。構築するシステムに応じた機能の取捨選択が可能であり，この点で個別システムの構築容易性を高めている。

2.4　カーナビ音声対話システム MINOS

EUROPAをベースに，MINOSを試作した（図4）。MINOSは，対象語彙数300から700，対象文パタン200万以上

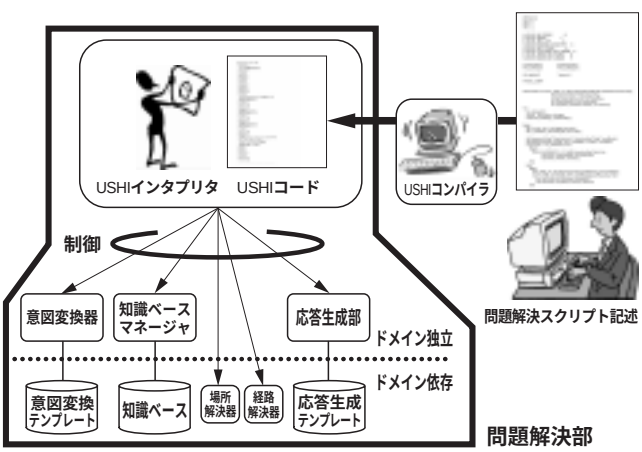

図3．問題解決部の構成例　EUROPAの枠組みをカーナビ音声対話システムの問題解決部に適用した例である。
Example of problem-solving module: car navigation system

図4．MINOS全景　MINOSは，ノートPC1台で動作する。
External view of MINOS

を受理し，音声認識―意図理解―応答生成のすべてのプロセスをノートPC1台上で動作させる。展示会会場程度の騒音下であっても，利用者の発話終了からおおむね2秒以内に返答することが可能である。

MINOSは，出発地から目的地までの経路上にある道路施設について，その地点や距離情報又は所要時間などの質問に答えることができる。また，「そこまで何分？」，「その手前にコンビニある？」のように，システムが応答したものを指示代名詞で参照することも可能である。

3　あとがき

ハンズフリー インタフェースである音声対話を実環境に適用する際の課題と，その解決策として当社が提案する音声対話システム構築フレームワークEUROPAについて述べた。今後は実環境への対応とともに，組込み機器などへの適用を目指した対話システムの高速化と，コンパクト化を更に進める所存である。

謝　辞

この研究の立案及び遂行を通じて，奈良先端科学技術大学院大学 情報科学研究科 河野恭之助教授より多大なる御指導をいただきました。ここに深く感謝の意を表します。

笹島　宗彦　SASAJIMA Munehiko, D.Eng.
研究開発センター　マルチメディアラボラトリー，工博。
音声対話の研究・開発に従事。人工知能学会，情報処理学会会員。
Multimedia Lab.

屋野　武秀　YANO Takehide
研究開発センター　マルチメディアラボラトリー。
音声対話の研究・開発に従事。電子情報通信学会会員。
Multimedia Lab.