日英音声翻訳システム ATR MATRIX *

6Q－7

○竹沢寿幸　森元逞　匂坂芳典　ニック・キャンベル　飯田仁
ATR 音声翻訳通信研究所

1　まえがき

自然な話し言葉を対象とした日本語から英語への音声翻訳システム ATR MATRIX (ATR's Multilingual Automatic Translation System for Information Exchange) を構築した．普段我々が使うような自然な音声でもこれを認識し，英語へ翻訳し，合成音声で出力することができる．システム全体はワークステーション1台で動作し，ほぼ実時間で処理を行うことができる．その概要を報告する．

2　システムの概要と特徴

従来の音声翻訳システムへの入力は，文節区切りのようなゆっくり丁寧に発話された文を単位とする音声であった [4]．自然な話し言葉では「あのー」「えーと」といった言葉がはさまれたり，「しょうがないですね」のような話し言葉独特の表現や，「予約お願いします」のような助詞が省略された表現がかなり頻繁に現れる．このような自然で多少くだけた表現があっても認識し，翻訳することができるようになった．また，自然で自発的な発話では二つ以上の文をつないで話すことがあるが，そのような場合は発話分割機能 [7] により1文毎に翻訳を行うことができる．さらに，話し手が男性が女性かを判断してそれに応じた声で音声合成を行うこともできる．

システム構成を図1に示す．音声認識サブシステム，言語翻訳サブシステム，音声合成サブシステムとメインコントローラから構成される．各サブシステムはサテライトコントローラを経由してメインコントローラと接続される．各サブシステムの内容を更新することがあっても，その影響は各サテライトコントローラまででメインコントローラには及ばない．

今回試作したシステムでは，「ホテルの予約」を対象としているが，辞書内容を置き換えることにより，他の場面にも適用することができる．

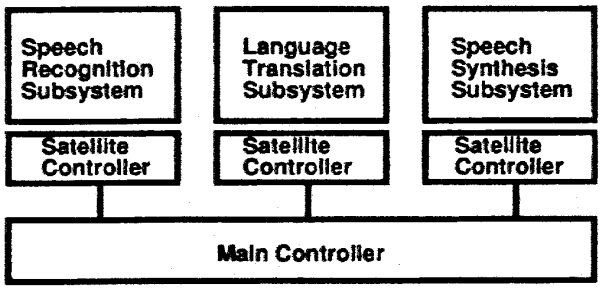

図 1: システム構成

2.1　不特定話者音素環境依存音響モデルと可変長 N-gram 言語モデルを用いた実時間音声認識処理

話者や音素環境の違いにより音声の特徴が大きく異なる．そのため，不特定話者の音素環境依存音響モデルを作成する統計的な手法である ML-SSS 法を提案した [6]．その手法を使って，男性用の不特定話者音素 HMM モデルと女性用の不特定話者音素 HMM モデルを用意した．

また，自然な話し言葉に含まれる多様な言い回しをコンパクトにモデル化できる統計的な手法として可変長 N-gram を提案した [3]．性能を落とさずに音響モデルと言語モデルを利用した認識過程の計算量を削減するために，単語グラフに基づく効率的な単語仮説数削減 [5] を実現している．

さらに，音声の各フレーム (典型的には 10ms) を一つのイベントとし，イベントキューを介して，音声区間検出，前処理，特徴パラメータ抽出，HMM との照合および単語グラフに基づく探索の各モジュールをオンメモリで接続する実装方法としている．各モジュールをパイプライン的に接続する実装方法では原理的に必ず時間遅れが生ずるが，イベントキューに基づく実装方法であれば原理的に時間遅れがない上に，ピッチ抽出等の新たなモジュールの追加が容易であるので拡張性に優れている．

音声認識用辞書の語彙サイズは約 2,000 語とした．人名等の固有名詞を除けば，「ホテルの予約」のように話題と場面を限定すればおおむね十分な規模である．

2.2　音声認識結果を扱うためのロバストな言語翻訳

翻訳では，文の構造を判断するだけでなく，対訳用例を用いることにより，話し言葉に現れる種々の表現を取り扱うことができる [2]．さらに，一部に誤りを含んだ音声認識結果を扱うためのロバストな言語翻訳として部分翻訳機能 [8] を実現している．次の二つのヒューリスティックスを仮定する．

(1) 対訳用例に類似した言語表現は誤認識の度合も少なく，解析結果の信頼性も高い．そこで，文または句の表現と対訳用例との意味的な距離をシソーラスにしたがって計算し，その上限閾値を設定する．

(2) より長い語句の範囲で解析できた結果の方が誤認識の度合も少なく，解析結果の信頼性も高い．そこで，対応する語句を構成する形態素数の下限閾値を設定する．

*A Japanese-to-English Speech Translation System: ATR MATRIX by Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida (ATR Interpreting Telecommunications Research Laboratories, Seika-cho, Kyoto 619-0288, Japan)

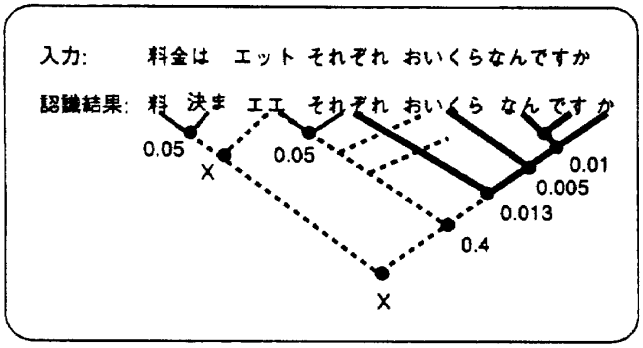

図 2: 部分翻訳の例

図2に部分翻訳の例を示す．例えば，意味的な距離の上限閾値を0.2，形態素数の下限閾値を2とする．「料金は」が「料決ま」と誤認識されており，その構造がそれ以上大きくできないので，棄却される．「エエそれぞれおいくらなんですか」という構造を作ることはできたが，その意味的な距離は0.4となるので，意味的な距離の上限閾値を越えるため，それより一段階小さい「それぞれおいくらなんですか」の部分のみを翻訳対象として選ぶことができる．

なお，言語翻訳部の辞書の語彙サイズは約13,000語である．これは「ホテルの予約」に限定せず，旅行会話に現れる話題と場面を対象としている．音声認識用辞書の語彙はこのサブセットとなっている．

2.3 個性豊かな音声合成

相手言語の合成音声を話し手の音声ないし話し手の声に似た音声で出力できれば，音声翻訳システムを使った個性豊かなコミュニケーションが実現できるだろう．音声認識サブシステムにおいて音声認識結果とともに選ばれた話者モデルの情報を出力することができる．今回は，男性用の不特定話者音素HMMモデルと女性用の不特定話者音素HMMモデルを用意しているので，その情報を活用すれば，入力音声から話し手が男性か女性かを判断することができる．音声合成サブシステム[1]は，選ばれた話者モデルに関する情報をメインコントローラを経由して受け取り，それに応じた声で音声合成を行う．話者モデルの数を増やせば，さらに個性豊かな音声合成を実現できる．

なお，話者モデルの数を増やすことは容易である．認識過程での効率的なビーム探索手法により，不要な話者モデルは発話開始後まもなく枝刈りされてしまうので，話者モデルの数が増えても全体の処理速度が遅くなることはない．

2.4 自然な会話を扱うための技術

自然な会話では，文ごとに区切らず，「ちょっと高いですね。もっと安い部屋は無いですか。」のように二つ以上の文をつないで発話することがある．その場合でも正しく1文ごとに翻訳する必要がある．そのような境界位置にある長さ以上の無音区間(ポーズ)が挿入されることもあるが，そうでないこともある．そこで，境界位置の前2単語と後1単語の合計3単語の範囲の品詞・活用形・活用型を利用する手法を提案した[7]．テキスト入力の書き起こしテキストを用いた予備実験によれば，統計モデルとヒューリスティックスを組み合わせると高い精度で境界を検出できる．

さらに，自然な会話では，「部屋は空いてます?」のように文末を上げることによって疑問を表すことがある．音の高さの変化(韻律)を検出して疑問文かどうかを判断することができるので，その情報を言語翻訳部に渡すことで“Rooms are available.”ではなく“Are rooms available?”のような翻訳を実現することができる．

3 むすび

現在，構築したシステムの性能評価を進めているところである．その結果をもとに，さらに精度向上のための各種改良を進める．今後はさらに，英日方向の同様なシステムを構築し，最終的には日英，英日双方向での会話が可能なシステムを実現する予定である．そして，双方向の会話における発話状況の理解，つまり，音声認識のための次発話予測や相手言語の生成における曖昧性解消の研究を進める．また，多言語間の音声翻訳をめざし，C-STAR IIという国際コンソーシアムを通じて世界各国の研究機関と研究協力を進めており，1999年度に多言語音声翻訳の国際実験を行う予定である．

謝辞

ご支援とご指導をいただいたATR音声翻訳通信研究所山本誠一社長に感謝します．また，システムの実装に協力いただいたベンジャミン・リーブス，西野敦士両氏をはじめとするATR音声翻訳通信研究所の皆様に感謝します．

参考文献

[1] ニック・キャンベル，アラン・ブラック: “CHATR: 自然音声波形接続型任意音声合成システム”，信学技報，SP-96-7, pp 45–52 (1996-05).

[2] 古瀬蔵，美馬秀樹，山本和英，Michael Paul，飯田仁: “多言語話し言葉翻訳に関する変換主導翻訳システムの評価”，言語処理学会第3回年次大会発表論文集，pp. 39–42 (1997-03).

[3] Hirokazu Masataki, and Yoshinori Sagisaka: “Variable-Order N-gram Generation by Word-Class Splitting and Consecutive Word Grouping”, *Proc. of ICASSP'96*, Vol. 1, pp. 188–191 (1996).

[4] 森元逞，田代敏久，竹澤寿幸，永田昌明，谷戸文廣，浦谷則好，鈴木雅実，菊井玄一郎: “音声翻訳実験システム(ASURA)のシステム構成と性能評価”，情報処理学会論文誌，Vol. 37, No. 9, pp. 1726–1735 (1996-09).

[5] 清水徹，山本博史，政瀧浩和，松永昭一，匂坂芳典: “大語い連続音声認識のための単語仮説数削減”，信学論 *D-II*，Vol. J79-D-II, No. 12, pp. 2117–2124 (1996-12).

[6] Harald Singer, and Mari Ostendorf: “Maximum Likelihood Successive State Splitting”, *Proc. of ICASSP'96*, Vol. 2, pp. 601–604 (1996).

[7] 竹澤寿幸，森元逞: “発話単位の分割または接合による言語処理単位への変換”，情報処理学会研究報告，97-SLP-18-4, Vol. 97, No. 101, pp. 19–24 (1997-10).

[8] 脇田由実，河井淳，飯田仁: “意味的類似性を用いた音声認識正解部分の特定法と音声翻訳手法への応用”，人工知能学会研究会資料，SIG-SLUD-9603-2, pp. 7–12 (1997-01).