4Q-2

音声認識と言語理解を動的に選択する音声理解フレームワーク

勝丸 真樹 † 中野 幹生 ‡ 駒谷 和範 † 成松 宏美 § 船越 孝太郎 ‡
辻野 広司 ‡ 高橋 徹 † 尾形 哲也 † 奥乃 博 †

† 京都大学大学院 情報学研究科 知能情報学専攻 ‡(株) ホンダ・リサーチ・インスティチュート・ジャパン
§ 津田塾大学 学芸学部 情報数理科学科

1. はじめに

音声対話システムで用いる言語モデルや言語理解方式には一長一短があるため，非常に大量の訓練データがある場合を除いて単一手法で高い性能を出すことは難しい．これに対して，複数の音声理解方式を用いることが有効だと考えられる．各理解方式がうまく理解できる発話は異なるので，正しい理解結果が含まれる可能性が高くなるからである．これまでに，複数の言語モデルや言語理解方式を使う方式として，ROVER 法 [1] や複数の言語理解方式の統合 [2] が提案されてきた．これらはいずれも言語モデルと言語理解方式のどちらかだけを複数用いる方式であり，さまざまな開発の制約下で，十分な性能が得られるとは考えにくい．

本稿では，システム開発時の制約の下でできるだけ高精度な音声理解を行うためのフレームワーク MLMU (Multiple Language models and Multiple Understanding models) を報告する．MLMU では，複数の言語モデルと複数の言語理解方式を用いてユーザ発話を理解し，複数の理解結果を出力する．MLMU の 1 実装として，言語モデルを 2 種類，言語理解方式を 3 種類を用いて，それらの任意の組合せを用いて音声理解を行い，複数の結果から発話ごとに適した理解結果を動的に選択する音声理解システムを構築した．

2. 複数の言語モデル・言語理解方式を用いる発話理解フレームワーク MLMU

我々が開発した MLMU は，言語モデルと言語理解方式の組み合わせである音声理解方式を複数用いることができるフレームワークである (図 1)．システム開発者は，タスクドメインごとに利用可能な言語モデルと言語理解方式を列挙することによって，複数の音声理解方式に基づく発話理解を行える．

複数の音声理解方式を用いることで，単一の音声理解方式よりも高精度な音声理解が可能となる．たとえば，文法でカバーできる発話に対しては，文法と Finite States Transducer (FST) の組み合わせによる音声理解の性能が高い場合が多いものの，必ずしもカバー率と予測性能の高い文法が構築できるとは限らない．文法でカバーできない発話に対しては，統計的言語モデルとキーフレーズスポッティング (スポッティング) や Weighted FST (WFST) が有効な場合が多い．複数の理解方式が必要な発話例を図 2 に記す．U1 は文法に沿った発話であるため，文法で音声認識し FST で言語理解した結果が正解となりやすい．これに対し，U2 は文法外の発話であるため，単語間の制約が弱い統計的言語モデルを用いた認識の方が正解が多くなる．さらに言語理解部でスポッティングを

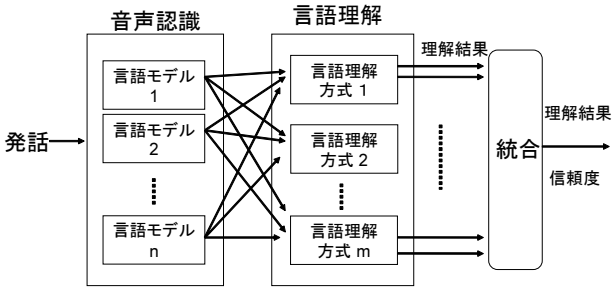

図 1: MLMU における音声理解の流れ

U1: 六月九日です．	
文法 + FST	認識結果 : 六月九日です.
	理解結果 : month:6, day:9
統計 + Spotting	認識結果 : 六月午後のがです.
	理解結果 : month:6
U2: 九日と十日 です．(下線部は文法外)	
文法 + FST	認識結果 : 九日土曜からです.
	理解結果 : start-day:9, day-of-week:Sat
統計 + Spotting	認識結果 : 九日とも十日です.
	理解結果 : start-day:9, end-day:10

図 2: 複数の音声理解方式が必要となる例

用いることで，認識結果が理解用文法に沿っていなくとも正しい言語理解結果に変換できる．このように，音声理解方式ごとに得意とする発話が異なるので，複数の音声理解を適用すれば，理解結果の中に正解が含まれる可能性が高まる．つまり，得られた理解結果の中から適切な理解結果の選択法や統合法の開発が，理解精度の向上の鍵である．また，複数の音声理解結果から，特定の音声理解結果に絞る必要はない．複数の候補を保持しておき，ユーザとの対話を通して最終的な理解結果を同定することも有益である．さらに，理解結果の一致度などを調べたり，ランク付けや多数決を行うことで信頼度の算出ができる．言語理解に信頼度を付与することで，信頼度に基づく効率的な対話管理 [3] が可能となる．

3. 複数の音声理解方式に基づく音声理解の実装とその評価

我々は MLMU をマルチドメイン対話システム構築ツールキット RIME-TK (Robot Intelligence based on Multiple Experts Tool Kit) [4] 上に実装した．システム開発者があらかじめ言語モデルと言語理解方式を用意することで，それぞれの組み合わせを簡便に指定できる．MLMU の有効性を示すため，レンタカー予約システムと人間の対話コーパス [5] でのユーザ発話の音声理解実験を行った．

3.1 言語モデル・言語理解の種類

本実験で言語モデルは，文法ベース言語モデルとドメイン依存統計言語モデルの 2 種類を用いた．文法ベース言語モデルで用いる文法は，言語理解で用いる意味文法と同じものである．ドメイン依存統計言語モデルは，言語

Speech Understanding Framework that Uses Multiple Language Models and Language Understanding Methods: *Masaki Katsumaru (Kyoto Univ.), Mikio Nakano (HRI Japan), Kazunori Komatani (Kyoto Univ.), Hiromi Narimatsu (Tsuda College), Kotaro Funakoshi, Hiroshi Tsujino (HRI Japan), Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

表 1: 音声認識時の特徴

S1:	検証用言語モデル使用時の音響スコア
S2:	文法ベース言語モデル使用時の音響スコア
S3:	統計言語モデル使用時の音響スコア
S4:	検証用言語モデル使用時と文法ベース言語モデル使用時との音響尤度差
S5:	検証用言語モデル使用時と統計言語モデル使用時との音響尤度差
S6:	文法ベース言語モデル使用時と統計言語モデル使用時との音響尤度差
S7:	発話時間 [秒]

表 2: 言語理解結果から得られる特徴

L1〜L6:	音声理解結果 1〜6 の事後確率に基づくコンセプトの信頼度の相加平均
L7:	L1 から L6 の相加平均
L8 〜 L13:	信頼度の相加平均の比 (= L1, ... L6 / L7)
L14 〜 L19:	音声理解結果 1 から 6 のコンセプト数
L20:	L14 から L19 の相加平均
L21 〜 L26:	コンセプト数の比 (= L14, ... L19 / L20)

理解用文法から自動生成した 10,000 文から単語 N-gram を学習させた．語彙サイズは文法ベース，ドメイン内統計言語モデルいずれも 257 である．これらに加えて，検証用として語彙サイズ 60,250 のドメイン非依存大語彙統計言語モデル [6] を用いた．

言語理解方式は，FST，WFST，スポッティングの 3 種類である．FST による言語理解では，言語理解用文法から FST を生成し，それに音声認識結果を与えることで言語理解結果であるコンセプト列を得る．WFST による言語理解は福林らの手法に基づく [7]．WFST のパラメータは評価用データとは異なる 1 名の 105 発話から推定した．スポッティングによる言語理解では，音声認識結果においてコンセプトに変換できる単語のまとまりを，全てコンセプトに変換する．

3.2 音声認識と言語理解における特徴に基づく理解結果の選択

複数出力される理解結果から最適な理解結果を選択するため，音声認識時と言語理解結果から得られる特徴を入力とした識別器を構築する．音声認識時に得られる特徴を表 1 に示す．音響スコアは発話時間で正規化した．特徴 S1 から S7 で示した音響尤度や発話長を考慮して，それぞれの音声認識結果の信頼性を検証する．言語理解結果から得られる特徴を表 2 に示す．L1 から L13 では理解結果の事後確率に基づく信頼度 [3] と，信頼度の理解結果間の関係を考慮した．L14 から L20 はコンセプト数に関する特徴である．コンセプトの個数に応じて理解方式の性能が変化する可能性も考慮した．L21 から L26 は理解結果間のコンセプト数の比を表す．この値が大きい場合は，他の理解結果よりコンセプトを多く出力しており，挿入誤りとなっている可能性が大きい．これらの特徴により，適切な理解結果を選択する決定木を構築する．決定木の構築には C5.0 [8] を用いた．

3.3 評価対象発話データ

評価データには，レンタカー予約システムと人間の対話データ (22 名× 8 対話) 中のユーザ発話 3,086 発話を用いた．音声認識器は Julius (ver. 4.0.2) を用い，音響モデルは話者非依存 PTM トライフォンモデルである [6]．文法ベース，統計ベースの言語モデルを用いたときの音声認識精度はそれぞれ，68.7%と 76.2%である．理解結果選択の決定木構築のため，学習データとして発話ごとに音声理解方式の正解ラベルを付与した．正解ラベルは，6 つの音声理解方式のいずれかあるいは，棄却かであり，発話ごとに Concept Error Rate (CER) が最も低くなる音声理解方式を付与した．CER は (システムが誤ったコンセプト数) / (発話に含まれるコンセプト数) で計算する．棄却のラベルは，挿入誤りが多く，6 つすべての音声理解で CER が 1 を越える場合に付与した．CER が最も低くなるラベルが複数存在する場合，全学習データに対する精度が最も良い理解方式を正解ラベルとした．

表 3: 各音声理解方式ごとの CER

音声理解方式 (言語モデル + 言語理解方式)	CER[%]
(1) 文法 + FST	32.3
(2) 文法 + WFST	34.2
(3) 文法 + Spotting	32.3
(4) 統計 + FST	36.8
(5) 統計 + WFST	31.6
(6) 統計 + Spotting	27.9
(1) 〜 (6), 棄却から選択 (本実装)	26.0

3.4 実験結果と考察

10-fold クロスバリデーションで発話データから決定木の構築と理解結果の選択を行なった．単一の言語モデル・言語理解方式を用いたときの CER と，本実装での CER を表 3 に示す．(1) から (6) の理解方式の中で最も精度が良かったのは (6) 統計+スポッティングである．WFST による言語理解の精度が，単純なスポッティングより低いのは，WFST のパラメータが不適切であったことが一因と考えられる．本実装での音声理解精度は，(6) の精度より 1.9 ポイント高い．これは複数の言語モデルと言語理解方式を考慮した結果である．6 つの理解結果と棄却から人手で正しい理解結果を選んだ場合の CER は 16.5%となった．この値は本実装より 9.5 ポイント高く，複数の言語モデルと言語理解方式を用いた場合の性能の上限と考えられる．本実装の性能の向上には，まず特徴量の検証が必要である．

4. おわりに

本研究では，複数の言語モデルと言語理解方式を用いた音声理解の枠組みとその 1 実装を述べた．本枠組みの有効性の検証のためには，他ドメインでの実験や，言語モデルの性能を変化させた実験が必須である．また，今回は理解精度のみを評価基準としたが，効率的な対話管理のためには理解結果への適切な信頼度付与は今後の重要な課題である．

謝辞 本研究の一部は，科研費，GCOE の支援を受けた．

参考文献

[1] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *ASRU-1997*, pp. 347–354.

[2] S. Hahn *et al.* System Combination for Spoken Language Understanding. In *Interspeech-2008*, pp. 236–239.

[3] 駒谷他. 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. 情処学論, Vol. 43, No. 10, pp. 3078–3086, 2002.

[4] M. Nakano *et al.* A framework for building conversational agents based on a multi-expert model. In *SIGdial-2008*, pp. 88–91.

[5] M. Nakano *et al*. Analysis of user reactions to turn-taking failures in spoken dialogue systems. In *SIGdial-2007*, pp. 120–123.

[6] T. Kawahara *et al.* Recent progress of open-source LVCSR Engine Julius and Japanese model repository. In *ICSLP-2004*, pp. 3069–3072.

[7] 福林他. 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. 情処学論, Vol. 49, No. 8, pp. 2762–2772, 2008.

[8] J. R. Quinlan. C5.0. http://www.rulequest.com/see5-info.html.