1C-01

応答義務推定の補助としての繰り返し発話検出手法の比較検討

川井 雄太† 藤田 寛泰‡ 谷川 晃大† 山下 峻†† 船越 孝太郎‡‡
(株)Nextremer† 高知工科大学 情報学群‡ 北海道大学 情報科学研究科††
(株)ホンダ・リサーチ・インスティチュート・ジャパン‡‡

1. はじめに

公共の場における人間-ロボット間の対話において，ロボットが複数ユーザと対話する状況が存在する．このとき，ロボットはユーザの発話だとしても，ユーザ同士の対話や独り言に対して応答すべきでない．そこで，ロボットが応答すべきユーザ発話を推定する応答義務推定技術が提案されている[1]．しかしユーザの振る舞いが人それぞれである中，応答義務推定の精度を100%に近づけることは困難である．一方，システムの誤認識に対し，ユーザは同じ内容の言い直し(繰り返し)で対処することが多い[2]．発話の繰り返しを高精度に検出できれば，応答義務推定の精度はそのままでも，インタラクション全体の質を向上できる見込みがある．本研究では応答義務推定における偽陰性の誤推定に起因する繰り返しを対象とする．具体的には，システムが応答すべき発話を無視する場合である．この場合，同一発話の単純な繰り返しがなされることが多い．従って，本研究では単純繰り返し発話を対象に検討を行う．

2. 関連研究

Kitaoka ら(2003)はカーナビの地名入力タスクにおける訂正発話を検出する[2]．第一の特徴量として10次元LPCメルケプストラム系列間のDPパスによる，照合開始位置から照合終端位置までの最小累積距離，第二の特徴量として音声認識候補集合間の重なり度を用いる．言い直し確率はこれら2つを説明変数とするロジスティック関数に従うと仮定して訂正発話を検出する．またLevitan(2014)らは音声検索クエリの再試行検出課題を扱う[3]．Levitanらが使用した特徴量は3カテゴリに分類される．類似性カテゴリに属する特徴量は単語，文字レベルでの2クエリ間の編集距離の未加工値，正規化値の双方，2クエリ間の共通単語数，最長共通単語列長(相対・絶対)である．正確性カテゴリの特徴量として，ユーザがシステムの提示結果と対話を行ったか否かのブール値を用いる．認識性カテゴリの特徴量として，クエリの長さや代替発音の量を用い，誤認識の可能性が高い音声の特性をモデル化する．これら3カテゴリの特徴量をロジスティック回帰の特徴量とし，音声クエリを *NO RETRY*, *REPETITION*, *REPHRASE*, *SEARCH RETRY*, *OTHER* の5つの再試行タイプに分類する．本研究ではこのうち *NO RETRY*(繰り返しなし)と *REPETITION*(同一音声の繰り返し)の識別を扱う．

3. 提案手法

本研究では，Kitaokaらの特徴量をベースとし，複数の機械学習アルゴリズムで繰り返し発話検出精度の比較を行なう．Levitan らからは編集距離を導入する．また，MFCC系列間の連続DPに基づく特徴量と，音声のフレーム長に基づく特徴量を新たに提案する．

3.1. MFCC系列間の連続DPに基づく特徴量

HTK 3.4.1(http://htk.eng.cam.ac.uk/)を用いて12次元MFCCを抽出する．特徴抽出パラメータは全てデフォルトに従う．ここではMFCC系列間のDPマッチングにおける最短経路に基づき特徴量を提案する．経路の移動方向は，縦，横，斜めの3方向がある．これら3方向の連続移動数を使って特徴量を計算する．連続移動数についての例を以下の図1に示す．

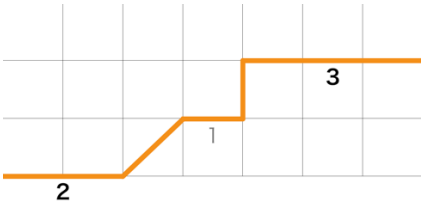

図1. 横方向連続移動数

連続移動数とは同一方向に2フレーム以上連続して移動したときの移動数を示す．図1で横方向連続移動数としてカウントされるのは2と3である．総連続移動数は各方向の連続移動数の総和．最大連続移動数は連続移動数の最大値，平均連続移動

Comparative Study of Repeated Utterances Detection Method for Estimating Response Obligation
Yuta KAWAI†, Hiroyasu FUJITA‡, Akihiro TANIKAWA†, Shun YAMASHITA†† and Kotaro FUNAKOSHI‡‡
† Nextremer Co., Ltd.
‡ Kochi University of technology
†† Hokkaido University
‡‡ Honda Research Institute Japan Co., Ltd.

数は連続移動数の平均値を表す．3 方向に対し総連続移動数，縦・横の 2 方向に対し最大連続移動数，平均連続移動数を計算し，特徴量として用いる．また，6 区間の連続移動数ヒストグラムも特徴量として用いる．直前発話と現発話のフレーム長の差も特徴量とする．

3.2. 音声の N-best 認識候補に基づく特徴量

本研究では N-best 認識候補集合間の最小編集距離を特徴量とする．これは事前実験において，重なり度よりも 5-10 ポイント高い精度を示したためである．式(1)に最小編集距離の定義を示す．

$$\min_{1\le i,j\le N} ed(n_i, m_j) \qquad (1)$$

n_iは直前発話nのi番目の認識候補，m_jは現発話mのj番目の認識候補，$\mathrm{ed}(n_i, m_j)$はn_iとm_jの編集距離，Nは候補解数を表す．ここでは$\mathrm{N} = 10$とする．

4. 実験

本研究では合成音声を学習に使う．評価は合成音声，人間音声の双方で行なう．音声認識には Julius 4.3.1(http://julius.osdn.jp/) を用いる．Julius の音響モデルには DNN-HMM を用いる．実験は同一単語による単純な繰り返しを扱う．

4.1. 実験方法

合成音声は Mac OS X 10.10.5 の say コマンドで作成する．人間音声は男性話者 2 名で録音する．フレーム長の差のみで識別可能になることを避けるため，単語は 4 モーラで統一する．合成音声は，話者“Otoya”，44 単語，5 発音，5 話速で計 1100 作成する．発音は単語モーラ間への記号挿入や，表記の変更により変化させる．また，実際の使用環境に近づけるため，音声の先頭・末尾にランダムなフレーム追加・削除処理を施し，最後にホワイトノイズを重畳する．人間音声は各話者 26 単語，5 発音(普通・普通・速く・遅く・強く)で 130 発話を adintool で録音する．

異なる発話のペアをランダムに生成し，単語が同一の場合は正解ラベルを 1，異なる場合は 0 としてペアに与える．レコードはペアの組み合わせ順列により 26400 作成する．正例負例の比率は 1:1 である．

テスト用データセットも学習用データセットと同様の方法で，500 レコードずつ作成する．

Weka 3.8.0(http://www.cs.waikato.ac.nz/ml/weka/) を使い，RandomForest(RF)，ロジスティック回帰(LR)，サポートベクターマシン(SVM)で学習を行う．人間音声の正解率は 2 名の平均値で評価する．

4.2. 実験結果

実験結果を表 1 に示す．表中(a)は MFCC 系列間の連続 DP に基づく提案特徴量．(b)は式(1)による最小編集距離．(c)は DP パスにおける最小累積距離．(d)は重なり度を表す．なお，(c)+(d)は先行研究[2]で Kitaoka らが使用した特徴量である．

表 1. 実験結果

特徴量	合成音声			人間音声		
	RF	LR	SVM	RF	LR	SVM
(a)	82.8	81.0	82.0	78.1	82.5	81.6
(b)	72.8	68.7	72.8	80.8	80.3	81.3
(c)	67.9	75.4	75.4	66.9	71.4	72.1
(d)	67.0	67.0	67.0	74.9	75.3	75.3
(c)+(d)	69.7	76.4	75.8	79.6	82.7	82.3
(a)+(c)+(d)	85.2	**85.0**	84.6	78.6	78.4	85.5
(b)+(c)+(d)	73.5	77.8	78.0	79.1	84.5	84.0
(a)+(b)+(c)+(d)	**85.6**	84.8	**84.8**	**88.7**	**88.9**	**88.6**

(b)+(c)+(d) は合成音声でのみ一貫して(a)+(c)+(d)より悪い結果となった．これは合成音声に施した先頭・末尾のフレーム処理に起因すると考えられる．発話の一部欠損，もしくは余分なノイズ区間の追加により音声認識の精度が低下し，語彙を正しく推定できないからである．一方，(a)+(c)+(d)ではフレーム処理に関わらず，合成音声で(c)+(d)よりも精度向上が見られる．このことから，音声データの欠損時においても有効な特徴量が提案できたことが示唆される．

5. おわりに

本研究では応答義務推定の補助を目的とし，繰り返し発話の検出を行った．提案特徴量を先行研究の特徴量に加えることにより，精度が向上する可能性が示唆される．今後は今回導入できなかった Levitan らの提案する他の特徴量についても実験する．また，実際に対話システムに組み込み，応答義務推定精度の向上に関する実験を行なう．

参考文献

[1] Sugiyama, et al.; “Estimating Response Obligation in Multi-Party Human-Robot Dialogues”, In Proc. Humanoids, 2015.
[2] Kitaoka, et al.; “Detection and Recognition of Correction Utterance in Spontaneously Spoken Dialog”, In Proc. INTERSPEECH, 2003.
[3] Levitan, et al.; “Detecting Retries of Voice Search Queries”, In Proc. ACL, 2014.