音響的特徴を用いた応答の使い分け・挿入を行う傾聴対話システムの試作

Prototyping of a listening dialogue system that selectively inserts reactive tokens by using acoustic features

福野将人 [1*]　狩野芳伸 [1]

Masato Fukuno[1] Yoshinobu Kano[1]

[1] 静岡大学情報学部
[1]Faculty of Informatics, Shizuoka University

Abstract: In human dialogue, several kinds of reactive tokens are selectively inserted. We developed an active listening dialogue system using acoustic features. This system does not require high performance in speech recognition, can easily be adapted to other languages.

はじめに

人間同士の対話においては、「うん」「はい」などのあいづち(Backchannel)や、相手の語句を反復する復唱(Repetition)、相手の後続発話を補完する共同補完(Collaborative Finish)など、ターンテイキングを伴わない短い応答が自然に使い分けられながら挿入されている[1]。これと同様の機能を傾聴対話システムに持たせようとする研究が行われてきた。

しかし、従来の研究ではユーザ発話に含まれる品詞や係り受け、構文木の深さなどの情報を元にこれを行うものが多く[2][3][4]、リアルタイムに応答を挿入するためには極めて高速に音声認識・形態素解析を行わなければならないと考えられる。現在の音声認識器は認識結果を得られるまでに多少のラグがあり、音声認識誤りも避けられない。

したがって今回試作したシステムでは、ユーザ発話音声の音量・基本周波数といった、リアルタイムに計算可能である音響的特徴のみを用いて応答の使い分け・挿入を行う。このシステムではリアルタイム連続音声認識より簡単な孤立単語認識さえ行えればよく、また特定の言語に根差した特徴量を持たないことから比較的容易に多言語に対応可能である。

応答の分類

どの応答を選択すべきかについて、SVM を用いた多クラス分類を行う。学習・評価用のデータセットについては、日本語用に Chiba3Party、英語用に Santa Barbara Corpus of Spoken American English(SBCSAE)という人間同士の対話の音声および書き起こしテキストを収録したコーパスを用いて自動的に作成した。

*連絡先:　静岡大学情報学部行動情報学科 狩野研究室
〒432-8011 静岡県浜松市中区城北 3-5-1 情報学部 2 号館
mfukuno@kanolab.net

ラベリング

各コーパスの書き起こしテキストを元に、100ms ごとの区間に含まれる発話と、その直後の 100ms 以内にある他の話者の発話から自動でラベルを付与した。ラベルの分類は下記の通りである。

何もしない(0)	以下の全てに当てはまらない場合
あいづち(1)	直後の発話が、Chiba3Party では「応答系・感情表出系感動詞」のアノテーションが振られているもの、SBCSAE では Backchannel と思われる間投詞であるもの。
復唱(2)	直後の発話の語句全体が、区間内の発話に含まれるもの。
共同補完(3)	区間内の発話の語句全体が、直後の発話に含まれるもの。

特徴量

256 フレームずつ分割した各コーパスの音声を元に、自作した音響解析器 nyankyo を用いて音響的特徴を 100ms ごとに計算し、特徴量とした。具体的な特徴量は下記の通りである。但し、SBCSAE では 1 チャンネルに複数人の音声が重なっている箇所があったためその部分はデータセットに含めないことにした。

直前の	特徴	計算
•100ms •500ms •1000ms	•音量 •基本周波数 •フォルマント周波数 (F3 まで) •零交差率	•平均 •中央値 •傾き •分散 •標準偏差

学習

Chiba3Party では 32、SBCSAE では 60 ある対話データのうち 1 つを評価用、残りを学習用のデータとし、ThunderSVM を用いて学習を行った。学習時のパラメータについては c(コスト)と g(ガンマ)のグリッドサーチによる最適化と、各ラベルのデータ数の偏りに応じた重みづけを行い、その他にはデフォルト値を用いた。最適化の際の目的関数にはホールドアウト法で計算される F-Score を用いた。

予測

実環境における予測では、ユーザの発話音声から上記の特徴量を 100ms ごとにリアルタイムで計算し、libsvm で応答の分類を行う。但し予測されるのは選択すべき応答の種類だけであり、実際に何と応答するかは別に決定する必要がある。

応答の生成

孤立単語認識

実環境でシステムが復唱・共同補完を行う際、実際に何と応答するか決めるために必要最低限の単語を認識する必要がある。したがってこの目的で、Julius を用いた孤立単語認識を行う。

日本語についてはＥＤＲ辞書:日本語単語辞書(基本語）から記号等を省いた単語辞書を作成し、Julius の話し言葉モデルキット(ssr-kit)に含まれる音響モデルを用いて孤立単語認識を行うこととした。

しかし英語については音響モデルを準備できなかったことから、今のところ孤立単語認識を行っていない。ただしこの場合でも、応答を使い分けずあいづちだけを挿入することなら可能である。

応答する語句の選択

SVM による予測結果に応じて実際に応答する語句を選択する。

あいづち(1)

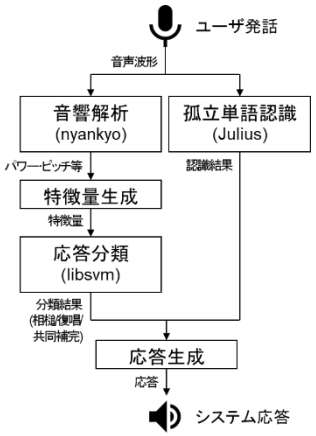
日本語では「うん」、英語では「uh huh」と応答する。

復唱(2)

孤立単語認識で 5 秒以内に認識された単語のうち最新のものを応答とする。5 秒以内に認識された単語が無い場合にはあいづちを返す。

共同補完(3)

孤立単語認識で 5 秒以内に認識された単語全てを半角スペースで区切りながら Google Suggest API のキーワードに入力し、帰ってきた候補のうち最初のものを半角スペースで分割してその最後の文字列を応答とする。5 秒以内に認識された単語が無い場合や API から予測を得られなかった場合にはあいづちを返す。



実環境における動作の概略図

評価

評価用に残しておいたデータを用いて、Chiba3Party では 32 交差検定､SBCSAE では 60 交差検定を行った。結果は以下の通りである。

	Accuracy	Precision	F-Score
Chiba3Party	0.793	0.252	0.339
SBCSAE	0.928	0.250	0.220

※マクロ平均・有効数字 3 桁

但しこの評価は「応答を正しく分類できたか」を示すものであって、実際に復唱・共同補完する際に選択した語句が正しかったかどうかは評価できておらず、また応答の選択に関して完全な正解があるのかは曖昧である。したがって評価手法については現在も検討中である。

結論

ユーザ発話の音響的特徴のみを用いて応答の使い分け・挿入を行う傾聴対話システムを製作することができた。しかしその性能の評価方法についてはさらなる検討の必要がある。

参考文献

[1] P. M. Clancy, S. A. Thompson, R. SuzukiとH. Tao. The conversational use of reactive tokens in English, Japanese, and MandarinJ. Pragmat. 26**(3)**355–387. (1996) .

[2] 竹内真士, 北岡教英, 中川聖一. 韻律・表層的言語情報を発話タイミング制御に用いた　雑談対話システム情報処理学会研究報告音声言語情報処理（SLP）. 2004**(15(2003-SLP-050))**87–92. (2004) .

[3] 中川良太, 中川聖一. 応答タイミングを考慮した音声対話システムとその評価研究報告音声言語情報処理（SLP）. 2009-SLP-77**(22)**1–6. (2009) .

[4] 山口貴史, 吉野幸一郎, 高梨克也, 河原達也. 多様な形態の相槌をうつ音声対話システムのための傾聴対話の分析第77回全国大会講演論文集. 2015**(1)**145–146. (2015) .