イベント継続時間モデルを用いた聞き手反応の検出*

☆森本洋介, 森大毅 (宇都宮大)

1 はじめに

説明場面では話し手は聞き手が出す相槌, 復唱, フィラーを観察し, 発話計画を動的に更新する. たとえば, 聞き手が長い間相槌を打たなければ, 話を理解しているか明示的に確認する, などである. 一方, 音声ガイドシステムは録音された音声を一方的に流すことしかできない. そこで, 本研究ではユーザーの理解状態をモニターし, 必要に応じて柔軟に説明戦略を変更することにより, 効率が高く心的負荷の小さい音声ガイドシステムの実現を目指す. 藤江らは [1], システムとの対話中にユーザーが生成する短い発話の肯定的/否定的態度を認識して, それに応じて発話計画を変更する音声対話システムを提案している. 本研究では, 理解状態を反映する聞き手反応として, フィラーおよび相槌に着目した. 本稿ではこれらの聞き手反応を実時間で検出する方法について述べる.

音響特徴を用いて単純にフレーム単位で分類すると, 現実にはありえないほど短い相槌やフィラーの湧き出しにより, 検出精度が上がらない. 本論文では, この問題に対処するためイベント継続時間モデルを提案し, その有効性について検討する.

2 音声資料と音響特徴量

相槌とフィラー検出のための学習・評価用コーパスとして, 計算尺使い方説明タスクコーパス [2] の 13 セッション分を用いた. 収録時間は約 140 分で説明者 1 名と被説明者 13 名の音声が収録されている. 説明者と被説明者の各イベントの発話数を Table 1 に示す.

音声データから, フレーム幅 25ms または 60ms, フレームシフト 10ms で音響特徴量を抽出した. 音響特徴量は Interspeech 2013 Paralinguistics Challenge のベースライン特徴量 141 次元 [3] を用いた. 特徴量は, ピッチ, パワー,MFCC 12 次元などの LLD や動的特徴量, その前後 4 フレームずつから計算される平均と標準偏差からなる.

3 聞き手反応検出

本稿では, 音響モデルとイベント継続時間モデルを用いて聞き手反応検出を行う. イベントとは, 検出される聞き手反応の 1 つのまとまりを表す. 本節では, 音響モデルのモデル化方法と, イベント継続時間モデルを用いた検出方法について述べる.

Table 1 発話数

	フィラー	相槌	その他
説明者	610	43	3363
被説明者	112	675	1305

3.1 音響モデル

フィラー, 相槌, その他とラベルづけされている音声フレームから抽出した音響特徴量を用いて混合ガウスモデル (GMM) をそれぞれ学習した. F0(基本周波数) については,1 次元 (有声) と 0 次元 (無声) の多空間分布 [4] でモデル化した. 学習後のそれぞれの音響モデルを用いて、確率密度 $p(\boldsymbol{x}|c), c \in \{$ 相槌, フィラー, その他 $\}$ を求める.

3.2 イベント継続時間モデル

フィラー, 相槌, その他の発話の継続時間分布はそれぞれ異なる. 検出されるイベントが本来の継続時間を持って検出されやすくするためにイベント継続時間モデルを用いる. イベント継続時間モデルは, フィラー, 相槌, その他の発話の継続時間のヒストグラムから求めた.

聞き手反応の検出は, 発話ごとのイベントの並びの確率が最大となるように検出される. 発話の総フレーム数を T, 発話の特徴系列を $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$ とする. n 番目のイベントを $S_n = (c_n, d_n)$, イベントの種類を $c_n \in \{$ フィラー, 相槌, その他 $\}$, n 番目のイベントの継続時間を d_n $(\sum_{n=1}^{N} d_n = T)$ とする. イベント継続時間 d の分布 $p(d|c)$ がイベント継続時間モデルである. 各イベントの開始時刻と終了時刻は $f(n) = 1 + \sum_{i=1}^{n-1} d_i$, $t(n) = d_n + \sum_{i=1}^{n-1} d_i$ と書ける. 聞き手反応の検出は, 式 (1) のようにイベント系列 $\boldsymbol{S} = S_1 \cdots S_N$ の事後確率最大化問題として定式化する.

$$
\begin{aligned}
\hat{\boldsymbol{S}} &= \underset{\boldsymbol{S}}{\operatorname{argmax}}\, P(\boldsymbol{S}|\boldsymbol{X}) \\
&= \underset{\boldsymbol{S}}{\operatorname{argmax}}\, P(\boldsymbol{S}) p(\boldsymbol{X}|\boldsymbol{S}) \\
&= \underset{\boldsymbol{S}}{\operatorname{argmax}}\, P((c_n, d_n)|S_1 \cdots S_{N-1}) P(S_1 \cdots S_{N-1}) P(\boldsymbol{X}|\boldsymbol{S}) \\
&= \underset{\boldsymbol{S}}{\operatorname{argmax}} \prod_{n=1}^{N} \{P((c_n, d_n)|(c_{n-1}, d_{n-1})) \prod_{t=f_n}^{t_n} p(\boldsymbol{x}_t|(c_n, d_n))\} \\
&= \underset{\boldsymbol{S}}{\operatorname{argmax}} \prod_{n=1}^{N} \{P(d_n|c_n) P(c_n|c_{n-1}) \prod_{t=f_n}^{t_n} p(\boldsymbol{x}_t|(c_n, d_n))\}
\end{aligned}
\tag{1}
$$

*Response token detection using event duration model. by MORIMOTO, Yosuke, MORI, Hiroki (Utsunomiya University)

Table 2　フィラー検出精度

	再現率	適合率	F 値
継続時間モデル無し	0.905	0.025	0.049
継続時間モデル有り	0.714	0.117	0.201

Table 3　相槌検出精度

	再現率	適合率	F 値
継続時間モデル無し	1.000	0.116	0.208
継続時間モデル有り	0.973	0.468	0.632

4　評価実験

4.1　実験条件

トレーニングデータには説明者 13 セッション, 被説明者 11 セッションのデータを, テストデータには被説明者 2 セッションのデータを用いた. 本稿では発話区間は既知とし, 無音部分以外の検出実験を行った.

音響モデルの混合ガウス分布の混合数は, 実験的に 16 とした.

4.2　評価方法

聞き手反応イベント検出精度は再現率と適合率と F 値で評価する. 再現率は, 正解イベントに対してそのイベントが検出できた割合を表し, 適合率は検出されたイベントが正しく検出できている割合を表す. F 値は, 再現率と適合率の調和平均である. また, 図 1 のように, 検出イベントは正解区間と時間的重なりを持つ 1 つのまとまりとし, 1 つの正解イベントに対して 2 つ以上のイベントが検出された場合, 2 つ目以降の検出イベントは誤検出とみなして適合率を求めた.

4.3　検出精度

イベント継続時間モデルを用いた聞き手反応の検出精度を, イベント継続時間を用いずフレーム単位で認識した場合の検出精度と共に Table 2, 3 に示す. フィラー, 相槌ともに, イベント継続時間を用いることで, 適合率, F 値が上昇していることが見て取れる. イベント継続時間モデルを用いない場合は, 短い継続時間を持つフィラーおよび相槌が多数誤検出されるため適合率が極端に低いが, イベント継続時間モデルを用いることによって, そのような湧き出しが減り, 適合率を改善できた.

4.4　イベント継続時間モデルの効果

Fig. 2 に, 相槌「はい」の音声に対する検出結果の一例を示す. 正解は相槌 1 イベントであるのに対し, 音響モデルのみの検出結果は相槌 4 イベント, フィラー 2 イベントである. 一方, イベント継続時間モデルを用いた検出結果は相槌 1 イベントである. このように, イベント継続時間モデルを用いることでイベントがまとまって検出されている. また, フィラーの誤検出がなくなっていることも見て取れる.

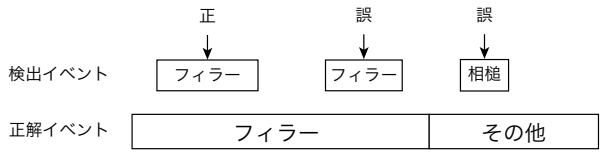

Fig. 1　評価方法

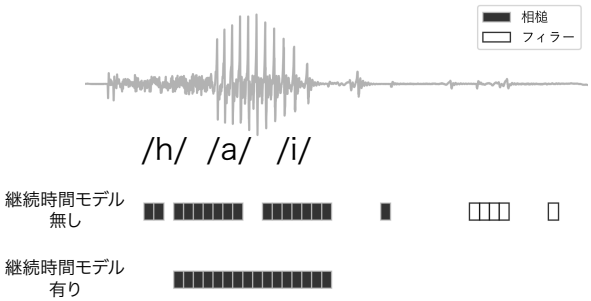

Fig. 2　相槌「はい」の検出結果

イベント継続時間モデルを用いることにより, 誤検出が減り, 聞き手反応検出の精度が向上することが確認できた. しかし適合率の低さからもわかるように, 誤検出がまだ多く見られる. 誤検出を減らすために, 尤度比などの信頼度を併用する必要があると考えられる.

5　おわりに

本稿では, フィラーと相槌検出において, イベント継続時間モデルを用いることで検出精度が向上することを示した. 今後は, リアルタイムで聞き手反応を検出しながら説明するシステムを構築して, 評価実験を行う予定である.

参考文献

[1] 藤江 他, 音講論 (秋), 39–42, 2015.
[2] 藍原 他, HCG シンポジウム論文集, 436–440, 2016.
[3] Shuller *et al.*, in Proc. Interspeech 2013, 148–152, 2013.
[4] 益子 他, 信学論, **J83-D-II**, 1600–1609, 2000.