語りに傾聴を示す応答の表出されやすさの推定

伊藤滉一朗 [†1]　村田匡輝 [†2]　大野誠寛 [†3]　松原茂樹 [†4]
[†1] 名古屋大学大学院情報学研究科　[†2] 豊田工業高等専門学校
[†3] 東京電機大学未来科学部　[†4] 名古屋大学情報連携推進本部
ito.koichiro@a.mbox.nagoya-u.ac.jp　murata@toyota-ct.ac.jp
ohno@mail.dendai.ac.jp　matubara@nagoya-u.jp

1　はじめに

語ることは人間の基本的な欲求である．語る行為は，聞き手がいて初めて成立する．日本では，独居高齢者の増加など社会の個人化が進み [1, 2]，聞き手不在の生活シーンが増加している．人が語れる機会を増やすことは現代社会の重要な課題である．これに対し，コミュニケーションロボットやスマートスピーカーなどが語りを聞く役割を担うことが考えられる．これらが聞き手として認められるには，「語りを傾聴していることを話し手に伝達する機能」を備える必要がある．このための明示的な手段は語りに応答することである．以降では，傾聴を示す目的で語りに応答する発話を**傾聴応答**と呼ぶ．

適切なタイミングでの傾聴応答の表出は，話し手の語る意欲を促進する効果が期待できるが，不適切なタイミングでの表出は逆効果になりうる．そのため，傾聴応答の表出では，その表出タイミングが重要となる．本論文では，適切なタイミングでの傾聴応答の自動生成の実現に向けて，あるタイミングが傾聴応答の生成にどの程度適するかを推定する手法を提案する．本研究では，この適切さを表す指標として，傾聴応答の**表出率**を定義し，これを予測する．あるタイミングでの傾聴応答の表出率は，そのタイミングで傾聴応答を表出した聞き手の割合とする．

提案手法は，語りの音響情報とテキスト情報を用いて，傾聴応答の表出率を予測する．具体的には，これらの情報を transformer ベースの手法 [3] でエンコードし，エンコード結果を 1 次元に変換して表出率を算出する．提案手法の予測性能の評価のために，表出率の予測実験を行ったところ，語りの音響情報とテキスト情報の両方を予測に使用する提案手法が有効であることを確認した．

語り	傾聴応答
イタリア旅行をしたことが	はい
一番楽しかったです	はーそうですかー
	素敵ですねー
	ふーん
もう二度と行けないかなと	イタリア旅行
思いながら行ってきましたけど	いえいえそんなー

図 1　語りと傾聴応答の例

2　傾聴応答

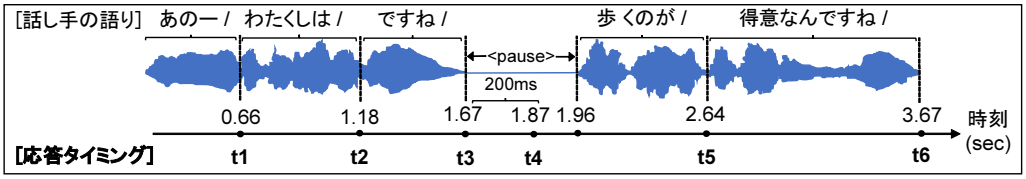
傾聴応答は，話し手の語りに傾聴を伝える応答であり，話し手の語る意欲を促進する効果が期待できる．図 1 に語りと傾聴応答の例を示す．

2.1　傾聴応答の表出率

本研究では，あるタイミングが傾聴応答の生成にどの程度適するかを表す指標として，傾聴応答の表出率を定義し，これを予測する．あるタイミングでの傾聴応答の表出率は，そのタイミングで傾聴応答を表出した聞き手の割合とする．表出率が高いタイミングほど，そのタイミングで応答を表出した聞き手が多いことを意味する．本論文では，傾聴応答の表出率の予測手法を提案する．

2.2　表出率の利用

傾聴応答の生成タイミングに関する従来研究 [4, 5, 6, 7, 8, 9] では，タイミング毎に，傾聴応答の生成に適するか否かの二値分類を経て，生成に適するタイミングを検出してきた．本研究では，従来研究とは異なり，傾聴応答の表出率を予測する．従来研究における，応答生成に適したタイミングか否かの二値分類は，行動の選択といえる．一方，本研究で行う表出率の予測は，あるタイミングでの状況の予測といえる．本研究では，予測された表出率は，適



応答タイミング	使用するテキスト情報（語りのサブワードと応答タイミングを表すタグから構成されるトークン系列）
t3	あ の ー 〈T_B〉 わ た くし は 〈T_B〉 です ね 〈E_B〉
t4	あ の ー 〈T_B〉 わ た くし は 〈T_B〉 です ね 〈T_B〉 〈E_P〉
t5	わ た くし は 〈T_B〉 です ね 〈T_B〉 〈T_P〉 歩 く の が 〈E_B〉

図 2　応答タイミング (上部) とテキスト情報 (下部) の例

表 1　応答タイミングを表現するタグ

< E_B >	文節終端の対象の応答タイミング
< E_P >	ポーズ継続地点の対象の応答タイミング
< T_B >	文節終端の過去の応答タイミング
< T_P >	ポーズ継続地点の過去の応答タイミング

切な応答生成タイミングの判断に利用されることを想定する．表出率を利用した適切な応答生成タイミングの判断方法には，様々な選択肢があると考えられる．最も単純な例としては，表出率がある閾値を超えた際に応答を生成する方法が挙げられる．この方法では，閾値の設定次第で，応答システムの積極性を表現することも可能である．

表出率の予測は，従来の二値の予測に比べ，より細かい粒度での明確な値の予測である．そのため，表出率を予測するシステムは，二値の予測を行う従来のシステムと比べ，移植性が高く，他の要素と組み合わせやすい．例えば，ユーザの個性や，応答システムとユーザの関係性といった，表出率以外の要素と組み合わせて，適切な応答生成タイミングの判断をすることも比較的容易と考えられる．

3　傾聴応答の表出率の予測手法

3.1　表出率の予測タイミング

提案手法では，応答は語りの言語的境界の直後で発話されやすいことを踏まえて，**(a) 語りの文節終端，(b) 語りの文節終端から 200ms ポーズが継続した点**，の二つを傾聴応答の表出率を予測するタイミング（以下，応答タイミング）とする．具体的には，(a) は語りの文節における最終形態素の発話終端時刻である．(b) は，語りにおいてポーズが継続する場合にも応答が発話されやすいことを踏まえた応答タイミングである．図 2 の上部に，応答タイミングの例を示す．「/」は文節終端を表す．t1~t6 の 6 個の応答タイミングのうち，t4 のみ上記の (b) に該当し，それ以外の 5 個の応答タイミングは (a) に該当する．

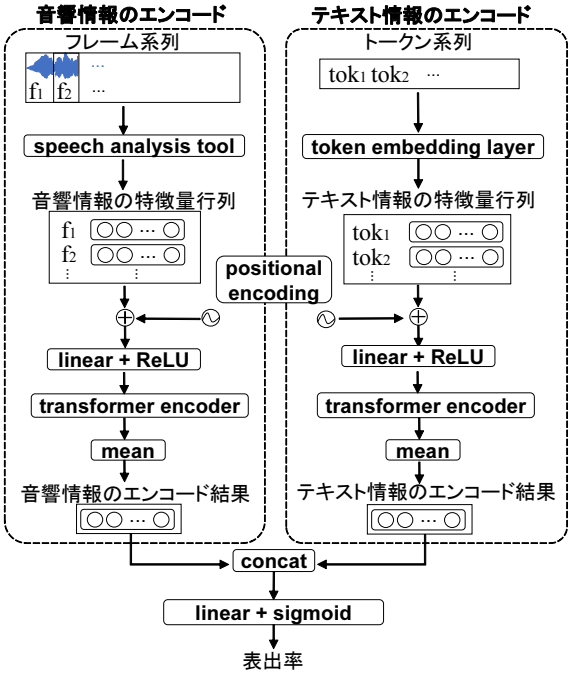


図 3　提案モデルの概略

3.2　使用する特徴量

提案手法は，語りの音響情報とテキスト情報から，表出率を予測する．音響情報としては，対象の応答タイミング直前のフレームの系列を用いる．フレーム単位の特徴量には，MFCC の下位 12 次元とピッチとパワー，及びそれらの Δ と ΔΔ を用いる．

テキスト情報としては，対象の応答タイミング直前のトークンの系列を用いる．トークンは，語りのサブワードと応答タイミングを表現する 4 種のタグ（表 1）から構成される．トークン単位の特徴量には，トークンの埋め込み表現を用いる．テキスト情報の例を図 2 の下部に示す．この例では，テキスト情報として，対象の応答タイミング直前の 3 文節に含まれるトークンの系列を用いている．

応答タイミング	話し手の語り	聞き手の傾聴応答の有無 L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	表出率
t1	あのー											0.0
t2	わたくしは						✓			✓		0.2
t3	ですね	✓								✓		0.2
t4	<pause>							✓	✓		✓	0.3
t5	歩くのが											0.0
t6	得意なんですね	✓		✓		✓				✓	✓	0.5

表 2　傾聴応答の表出率の正解値の算出例

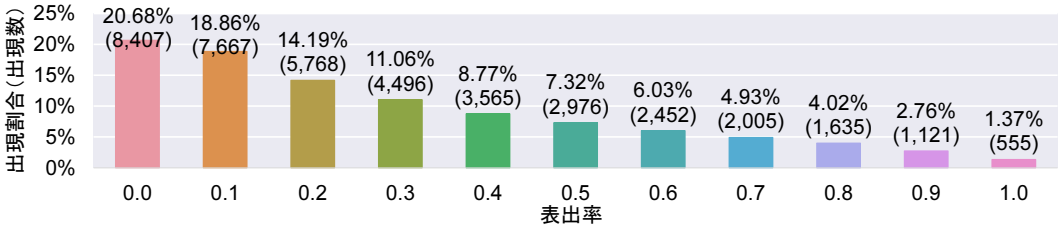


図 4　傾聴応答の表出率の正解値の出現割合

3.3　表出率の予測モデル

提案モデルの概略を図 3 に示す．ただし，$f_i \in [f_1, f_2, \cdots]$ はフレームを，$tok_j \in [tok_1, tok_2, \cdots]$ はトークンを表す．はじめに，音声分析ツールを使用して，対象の応答タイミング直前のフレーム系列から，音響情報の特徴量行列を抽出する．一方，対象の応答タイミング直前のトークン系列をトークン埋め込み層に入力して，テキスト情報の特徴量行列を得る．次に，これら特徴量行列に対して，positional encoding の加算と，線形変換，ReLU 関数を適用したのち，1 層の transformer encoder [3] による変換を行う．その後，変換後の行列に対して，要素ごとに平均値をとり，音響情報とテキスト情報それぞれのエンコード結果であるベクトルを得る．最後に，これら二つのベクトルを連結させたのち，線形変換と sigmoid 関数を適用し，0 以上 1 以下の 1 次元の値に変換することで，表出率を得る．

4　傾聴応答の表出率の予測実験

4.1　実験データ

4.1.1　傾聴応答データ

本実験では，村田らの**傾聴応答データ** [10] を用いる．傾聴応答データにおける傾聴応答は，複数の作業者が録音された語りを聞きながら，それぞれ応答することで収集されており，静的な一つの語りに対して，複数の聞き手の傾聴応答が存在する．録音された語りとして，30 名の高齢者による合計約 8 時間 40 分の語りが収録されているナラティブコーパス JELiCo [11] を使用している．

4.1.2　傾聴応答データにおける表出率

本実験では，傾聴応答データの一部である 10 人の聞き手の傾聴応答 131,616 個を用いた．このデータ中の応答タイミングに対する傾聴応答の表出率の正解値は，そのタイミングで応答した聞き手の割合とした．まず，CaboCha [12] を用いて語りの文節境界を検出したのち，応答タイミングを特定した．次に，応答の発話開始時刻を，語り中の最も近い応答タイミングに対応付けた．最後に，対応付けの結果に基づき，傾聴応答の表出率の正解値を算出した．

表 2 に傾聴応答の表出率の正解値の算出例を示す．この例の語りと応答タイミングは，図 2 のものと同一である．L1 ~ L10 は 10 人の聞き手を表し，✓は聞き手が応答をしたことを意味する．例えば，応答タイミング t2 では，L6 と L9 の 2 人が応答しているので，表出率は 2/10 = 0.2 となる．実験で用いるデータには，40,647 個の応答タイミングが存在した．図 4 に語りに存在する 40,647 個の応答タイミングにおける表出率の正解値の出現割合を示す．

4.2　実験概要

実験データを学習データ，開発データ，テストデータに分割した．それぞれ，23,846 個，8,588 個，8,213 個の応答タイミングが含まれている．

音響の特徴量抽出には Praat [13] を使用した．フレームシフトを 10ms とし，対象の応答タイミングから遡って 25 フレーム（応答タイミングの直前

表 3　各予測手法における評価指標の値

	RMSE											
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	all
random	0.404	0.335	0.287	0.268	0.286	0.334	0.402	0.481	0.566	0.656	0.747	0.378
proposed (P)	0.210	0.160	**0.127**	**0.127**	0.167	0.223	0.290	0.359	0.437	0.505	0.625	0.233
proposed (T)	0.203	0.158	0.133	0.131	**0.164**	0.205	0.262	0.317	0.369	0.431	0.500	0.213
提案手法	**0.174**	**0.155**	0.138	0.147	0.172	**0.204**	**0.247**	**0.288**	**0.323**	**0.383**	**0.450**	**0.199**

250ms）の特徴量を用いた．また，MFCC については，窓幅を 25ms とした．テキスト情報については，対象の応答タイミング直前の 3 文節の語りに含まれるトークン系列を使用した．語りのサブワードへの分割は，MeCab [14] (UniDic Ver. 2.1.2) で形態素解析したのち，subword-nmt [15][1)] により実施した．サブワードの埋め込みの初期値には，日本語 Wikipedia コーパスを用いて GloVe [16] により事前学習した埋め込み表現を利用した．トークンに対する埋め込みの次元数は 300 とした．

モデルの学習では，バッチサイズを 256, 学習率を 2e-4，エポック数を 20 とした．損失関数には MSE (Mean Squared Error) Loss を，最適化手法には Adam [17] を用いた．開発データにおいて損失が最小となったモデルをテストデータに適用して，その性能を評価した．補足として，モデルのハイパーパラメータと実装に用いたライブラリの詳細を，付録 A に記載しておく．

4.3　評価方法

本実験では，テストデータ中の全応答タイミングから算出される RMSE (Root Mean Squared Error) により，表出率の予測性能を評価する．これを RMSE(all) と表記する．さらに，表出率の正解値ごとにも RMSE を算出し，これも評価指標とする．表出率の正解値 c に対する RMSE を RMSE(c) と表記する．補足として，RMSE(all) と RMSE(c) の計算式を付録 B に記載しておく．

提案手法に加え，以下の三つの手法を実装した．

- **random**: 学習データにおける表出率の出現分布に従ってランダムに表出率を予測する手法
- **proposed (P)**: 音響 (Prosody) 情報のみを使用して表出率を予測する手法
- **proposed (T)**: テキスト (Text) 情報のみを使用して表出率を予測する手法

random の評価指標の値は，テストデータに対する 1000 回の試行から得られる評価指標を平均することで算出した．proposed (P) と proposed (T) は，それぞれ，表出率を出力する機構（図 3 の linear + sigmoid）に，音響情報のみを入力する手法と，テキスト情報のみを入力する手法である．

4.4　実験結果

各予測手法に対する評価指標の値を表 3 に示す．提案手法と random を比較すると，提案手法が全ての評価指標で random を上回っており，提案手法がランダムな予測よりも有効であることを確認できた．proposed (P) と proposed (T) についても，同様の結果が得られた．このことは，音響情報とテキスト情報がいずれも表出率の予測に寄与することを示している．一方，提案手法と proposed (P) 及び proposed (T) を比較すると，提案手法が全 12 個の評価指標のうち 9 個で最良の結果を示しており，音響情報とテキスト情報の両方を使用して表出率を予測することの有効性を確認できた．また，分析の結果，提案手法は，正解の表出率が高いほど予測値も高くなる傾向にあることも確認した．付録 C に実験結果の補足を記載しておく．付録の図 5 に提案手法の予測傾向を，図 6 に提案手法による予測例も記載しておく．

5　まとめ

本論文では，傾聴応答の自動生成に向けた，傾聴応答の表出率の予測手法を提案した．提案手法は，語りの音響情報とテキスト情報を transformer ベースの手法でエンコードし，エンコード結果を 1 次元に変換して表出率を算出する．表出率の予測実験によって，音響情報とテキスト情報の両方を使用する提案手法の有効性を確認した．今後は傾聴応答の種類を考慮した表出率の予測に取り組んでいきたい．

謝辞

高齢者のナラティブデータは，奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室から提供いただいた．本研究は，一部，科学研究費補助金 挑戦的研究（萌芽）（No. 18K19811） により実施したものである．

1）　https://github.com/rsennrich/subword-nmt

参考文献

[1]総務省統計局. 平成 27 年国勢調査世帯構造等基本集計結果の概要.

[2]国立社会保障・人口問題研究所. 日本の世帯数の将来推計（全国推計）平成 30 年推計.

[3]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008, 2017.

[4]Hiroaki Noguchi and Yasuharu Den. Prosody-based Detection of the Context of Backchannel Responses. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, pp. 8570–8573, 1998.

[5]Nicola Cathcart, Jean Carletta, and Ewan Klein. A Shallow Model for Backchannel Continuers in Spoken Dialogue. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics (EACL-2003)*, pp. 51–58, 2003.

[6]Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. A Conversation Robot with Back-Channel Feedback Function Based on Linguistic and Nonlinguistic Information. In *Proceedings of the 2nd International Conference on Autonomous Robots and Agents (ICARA-2004)*, pp. 379–384, 2004.

[7]Norihide Kitaoka, Masashi Takeuchi, Ryota Nishimura, and Seiji Nakagawa. Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems. *Journal of Japanese Society for Artificial Intelligence*, Vol. 20, No. 3, pp. 220–228, 2005.

[8]Yuki Kamiya, Tomohiro Ohno, and Shigeki Matsubara. Coherent Back-Channel Feedback Tagging of In-Car Spoken Dialogue Corpus. In *Proceedings of the 11th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL-2010)*, pp. 205–208, 2010.

[9]Takashi Yamaguchi, Koji Inoue, Koichiro Yoshino, Katsuya Takanashi, Nigel G. Ward, and Tatsuya Kawahara. Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents. In *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS-2016)*, pp. 1–12, 2016.

[10]村田匡輝, 大野誠寛, 松原茂樹. 語りの傾聴を話し手に示す応答発話の収集. 電気学会論文誌 C, Vol. 138, No. 5, pp. 637–638, 2018.

[11]Eiji Aramaki. Japanese Elder's Language Index Corpus v2, 2016. https://doi.org/10.6084/m9.figshare.2082706.v1.

[12]Taku Kudo and Yuji Matsumoto. Japanese Dependency Analysis using Cascaded Chunking. In *Proceedings of the 6th Conference on Natural Language Learning (COLING-2002)*, pp. 63–69, 2002.

[13]Paul Boersma and David Weenink. Praat: Doing Phonetics by Computer (version 5.1.05), 2009. http://www.praat.org/.

[14]Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230–237, 2004.

[15]Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016)*, pp. 1715–1725, 2016.

[16]Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)*, pp. 1532–1543, 2014.

[17]Diederik P.Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization, 2014. http://arxiv.org/abs/1412.6980.

A　モデルのハイパーパラメータと実装に用いたライブラリ

3.2 節で述べたとおり，音響の特徴量は 42 次元（MFCC の下位 12 次元とピッチとパワー，及びそれらの Δ と ΔΔ）である．4.2 節で述べたとおり，テキストの特徴量は 300 次元である．表 4 に，proposed (P)，proposed (T) 及び提案手法における transformer encoder [3] の入出力の次元数（d_{model}），multi-head attention のヘッド数（h）を示す．position-wise feedforward network の中間層の次元数（d_{ff}）は，d_{model} の 4 倍とした．これらのハイパーパラメータの値は，開発データを用いて定めた．モデルの実装には，pytorch[2] を用いた．transformer encoder は，pytorch の `nn.TransformerEncoder` と `nn.TransformerEncoderLayer` で実装した．

B　評価指標

RMSE(all) と RMSE(c) は，以下の式によって計算される．ただし，$p(t)$ と $c(t)$ はそれぞれ，応答タイミング t における表出率の予測値と正解値を表し，S_{all} はテストデータにおける全応答タイミングを表し，S_c はテストデータにおける表出率の正解値が c であるような応答タイミングの集合を表す．

$$\text{RMSE(all)} = \sqrt{\frac{1}{|S_{\text{all}}|}\sum_{t\in S_{\text{all}}}(p(t)-c(t))^2} \qquad \text{RMSE}(c) = \sqrt{\frac{1}{|S_c|}\sum_{t\in S_c}(p(t)-c(t))^2}$$

C　実験結果の補足

図 5 に，提案手法による傾聴応答の表出率の予測傾向を示す．この図は，表出率の正解値が高いほど予測値も高くなる傾向を示している．この結果から，提案手法は，傾聴応答の表出率の高さに対応した応答タイミングの特徴をうまく捉えられているといえる．最後に，提案手法による表出率の予測例を図 6 に示す．

表 4　各予測手法における transformer encoder のハイパーパラメータの値

	音響側		テキスト側	
	d_{model}	h	d_{model}	h
poposed (P)	512	16		
poposed (T)			256	16
提案手法	128	16	512	4

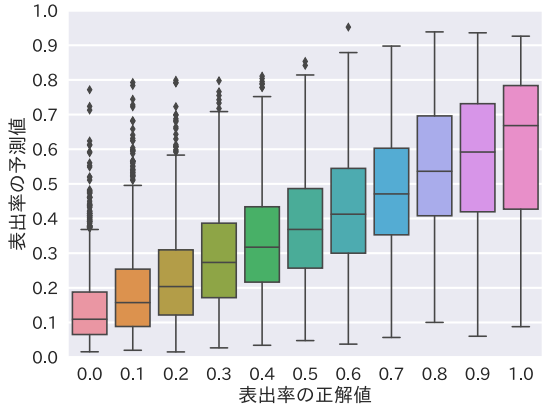


図 5　提案手法による傾聴応答の表出率の予測傾向

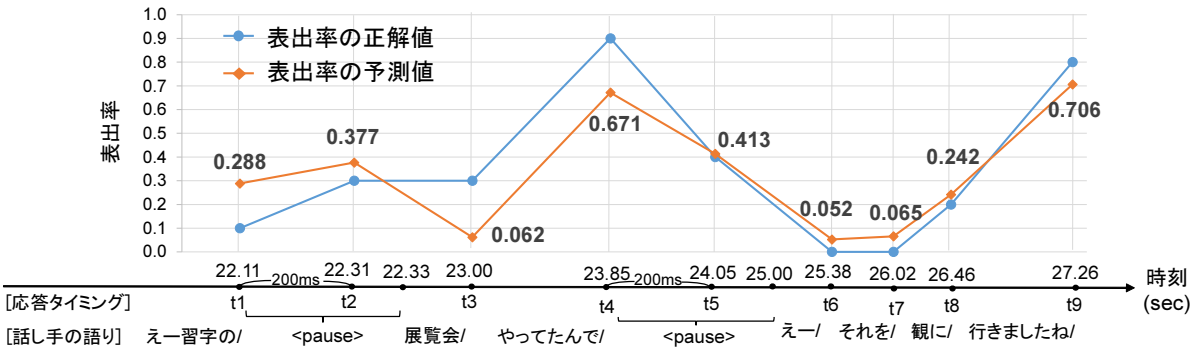


図 6　提案手法による傾聴応答の表出率の予測例

2）https://pytorch.org/docs/stable/index.html