E-030

傾聴対話システムのための高齢者音声を用いた発話終了判定
End-of-utterance detection using elderly person voice for an active listening system

伊島 翔大[1)]　藤江 真也[1)]
Shota Ijima　Shinya Fujie

1　はじめに

傾聴対話システムが適切なタイミングで発話を生成するために，話し手の発話終了を判定する手法について検討する. 傾聴対話とは，聞き手が話し手の発話を促進するための対話で，高齢者施設などでは高齢者を話し手として認知症の進行を遅らせる目的で行われている. 専門のボランティアや介護士が聞き手となるが，本研究ではこの代わりを担うロボットを構築することを目的としている. 話を聞く，あるいは対話そのものを楽しむためのロボットはいくつか存在する.

しかしながら，それらのロボットは主にロボットの発話 (質問やリアクション) の内容に注目して開発されている. 一方，特に傾聴対話においては，ユーザである話し手の発話に対して，聞き手であるロボットは適切なタイミングで発話を挟む必要がある. 話し手の方が話す意思を持っているにもかかわらず，ロボットが質問をしたり話題を転換したりすると，話し手の話す意志を削いでしまうことが予想される. 本研究では，話し手の発話音声が途切れたときに，話し手がその発話を続けるかどうかを判別することで，適切なシステムの発話タイミングを決定することを試みる.

我々は，従来から音響情報を使った発話終了判定を行っている [1]. この手法は発話区間の終端における音響特徴の統計量を用いて識別する手法であるが，高齢者を対象とした本研究では，話し手ごとの発話速度の違いの影響により一定区間で求めた統計量では高い性能が得られないことが予想される. そこで，発話区間全体の時系列特徴量を入力としたリカレントニューラルネットワークを用いる手法を提案する. 本稿では高齢者施設で収録した音声と，既存のコーパスに含まれる音声を用いた実験結果について報告する.

2　音声データ

傾聴対話における話し手である高齢者の発話を収録した. 傾聴ボランティアと高齢者が対話している様子を録音したもので，音声区間は発話の区切れ目だと思われるところを手動で区切った. 音声区間ごとに発話が終了であるか，継続であるかを著者の主観でラベル付けした. この結果，5 人分のデータ，継続 179 発話，終了 179 発話が収集された. これらの音声を識別するため，学習データとして既存のコーパスにおける音声を用いることとした. 従来研究 [1] で利用した，千葉大 3 人会話コーパス [2] の音声区間に人手で発話継続/終了をラベル付けしたものを用いる. このコーパスは，千葉大学で収録された，大学生，大学院生，博士研究員を含む同性 3 人からなる友人同士 12 組の雑談を納めたものである. 全 12 セッション，異なる 3 人で会話を行っており，男性 18 人，女性 18 人の合計 36 人で収録されている. また高齢者は発話速度が遅く，千葉大学 3 人会話コーパスのような若者音声では，発話速度の違いにより識別率が低くなることが考えられる．そのため，実際の傾聴対話の音声に近い高齢者の音声コーパスとして，S-JNAS（Senior Japanese Newspaper Article Speech）[3] を用いる．これにより発話速度の違いによる影響が抑えられ識別率の向上につながることが予測される．S-JNAS の音声区間を VAD（Voice Activity Detection）により分割を行い，発話の終端を発話終了とし，その他の区切れ目を発話継続とした．このコーパスは，60 ～ 90 歳の被験者が新聞記事を読み上げた音声を収めたものである．1 人あたり 100 文の新聞記事を読み上げており，男性 151 人，女性 150 人の合計 301 人で収録されている．

1）千葉工業大学 未来ロボティクス学科 藤江真也
shinya.fujie@p.chibakoudai.jp

3　発話継続／終了の判定

3.1　特徴量

識別に利用する特徴量は，ラウドネス，基本周波数，基本周波数の包絡線，有声音確率や，MFCC の 1～15 次元，メル周波数帯域対数パワースペクトルの 1 ～ 8 次元，線スペクトル対の 1～8 次元，それらのΔパラメータ，ΔΔパラメータの，計 105 種類の音響情報である. 年齢や性別による影響を取り除くため，各話者の平均を元の発話から減算をする正規化を各特徴量に施した. 特徴量の抽出には openSMILE [4] を用いた. 特徴量は 10ms ごとに抽出される.

3.2　時系列特徴量による手法

高齢者は発話速度が比較的遅く，人によるばらつきも多いことが考えられる. したがって，従来手法である固定時間長の音声を用いて計算された統計的特徴量は，必ずしも継続/終了の識別に有効なものであるとは限らない. そこで，音声区間全ての特徴量を時系列データとして利用する識別手法を検討する. 識別器として時系列情報を扱うことのできる RNN(Recurrent Neural Network) を利用し，10ms ごとに抽出される音響特徴量を入力した上で，音声区間の最後の特徴量を入力したときの出力によって継続/ 終了を判定する. また通常 LSTM では，時刻 $t-1$ の隠れ層からの出力を時刻 t への隠れ層への入力としているが，時刻 $t+1$ の隠れ層からの出力を時刻 t への入力として用いる，逆方向の LSTM を同時に中間層に用い双方向の LSTM を構成する Bidirectional LSTM を用いて識別を行った．Bidirectional LSTM を用いることで逆方向から系列を入力するため，系列はじめの情報も考慮することができ，性能向上が期待される．

3.3　使用モデル

千葉大学 3 人会話コーパスを用いた学習では，入力を音響特徴量 105 次元ベクトルとし，中間層には共に 32 素子の LSTM，Bidirectional LSTM を使用した，発話継続/終了の 2 値分類を出力とするネットワークである. S-JNAS を用いた学習では，入力を音響特徴量 105 次元ベクトルとし，中間層には共に 16 素子の LSTM，Bidirectional LSTM を使用した，発話継続/終了の 2 値分類を出力とするネットワークである.

表 1　千葉大学 3 人会話コーパスを用いた
傾聴対話における高齢者音声の識別結果

		識別結果	
		終了	継続
正解	終了	144	35
	継続	145	34

(a) LSTM

		識別結果	
		終了	継続
正解	終了	121	58
	継続	118	61

(b) Bidirectional LSTM

表 2　S-JNAS を用いた傾聴対話における
高齢者音声の識別結果

		識別結果	
		終了	継続
正解	終了	100	79
	継続	57	122

(a) LSTM

		識別結果	
		終了	継続
正解	終了	106	73
	継続	52	127

(b) Bidirectional LSTM

4　実験

4.1　若者音声での識別結果

はじめに千葉大学 3 人会話コーパスを用いて傾聴対話における高齢者音声の識別実験を行う．学習させた音声データ数は学習に偏りをなくすため発話継続数，発話終了数ともに 723 発話の合計 1446 発話である．結果を表 1 に示す．

4.2　高齢者音声での識別結果

S-JNAS を用いて傾聴対話における高齢者音声の識別実験を行う．学習させた音声データ数は学習に偏りをなくすため発話継続数，発話終了数ともに 26541 発話の合計 53082 発話とした．結果を表 2 に示す．

5　識別器の比較

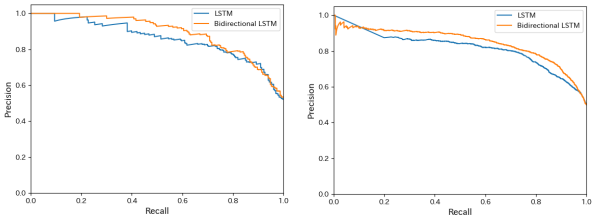
会話への割り込みの問題点として，発話継続である音声を発話終了と間違えることはオーバーラップを生じ，発話終了である音声を発話継続と間違えることは余計なポーズを生じさせる．傾聴対話において実際にロボットが高齢者と対話しているとき，高齢者が発話をしている最中に割り込んで発話を開始することは，高齢者の話す意思を削いでしまう可能性がある，そのようなシステムは実際の傾聴対話と同等の効果が得られるとは言い難い．そのため Precision-Recall 曲線を用い各識別器の性能を比較することで，どちらの識別器がより会話の途中で割り込みをしないか確認する．

5.1　中間層の性能比較

はじめに，LSTM と Bidirectional LSTM の比較を行う．千葉大学 3 人会話コーパスの比較では 1446 発話中 960 発話を学習データとし，残りの 466 発話で話者オープン実験を行った．S-JNAS の比較では 53082 発話中 49088 発話を学習データとし，残りの 3994 発話で話者オープン実験を行った．Precision-Recall 曲線を用いた比較結果図 1 に示す．両データとも，Bidirectional LSTM を用いた識別器の曲線が，LSTM を用いた識別器よりも外側にあることから Bidirectional LSTM を用いることの有用性が確認できた．

5.2　両データでの識別実験の比較

千葉大学 3 人会話コーパスを用いて傾聴対話における高齢者音声を識別する手法と，S-JNAS を用いて傾聴対話における高齢者音声を識別手法の性能の比較を行う．比較結果を図 2 に示す．両データとも LSTM，Bidirectional LSTM での結果は変わらず，Bidirectional LSTM を用いることによる性能の向上は見られなかった．千葉大学 3 人会話コーパスと S-JNAS の性能比較では S-JNAS を用いた識別器の曲線が，概ね千葉大学 3 人会話コーパスを用いた識別器の曲線より外側にあることから，高齢者音声を学習データとして用いることの有用性が確認できた．



(a) 千葉大学 3 人会話コーパス　(b) S-JNAS

図 1　Precision-Recall 曲線を用いた LSTM，
Bidirectional LSTM の性能比較

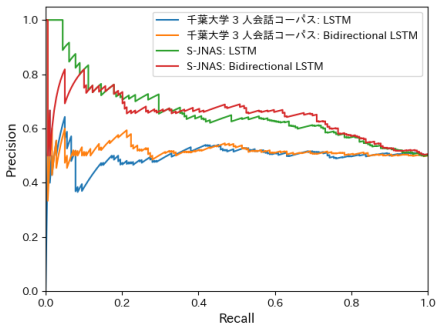


図 2　Precision-Recall 曲線を用いた性能比較

6　おわりに

本研究では傾聴対話システムの発話タイミング決定を目的として，音響情報を用いた発話継続/終了の判定を行う識別器を作成した．LSTM と Bidirectional LSTM を用いた識別器の比較をより，双方向の LSTM である Bidirectional LSTM を用いることが，発話タイミング決定に有用であることが確認された．また，傾聴対話における高齢者音声の識別に対して，高齢者の音声を学習データとして用いる識別器は，若者の音声を学習データとして用いる識別器に比べて性能が高く，その有効性が確認された．今後の課題として，識別率向上，更に，本手法を組み込んだ傾聴対話ロボットの実現が挙げられる．

参考文献

[1] 山崎敦也，藤江真也，“雑談会話における音韻・韻律情報を用いた聞き手の発話タイミングの検出，”日本音響学会春季研究発表会講演論文集，pp. 97-98, March 2016.

[2] Y.DenandM.Enomoto,”Ascientificapproachtoconversational information: Description, analysis, and modeling of human conversation,” In Nishida, T. (Ed.), Conversational informatics: An engineering, pp. 307-330, 2007.

[3] 音声資源コンソーシアム,“新聞記事読み上げ高齢者音声コーパス（S-JNAS）,” http://research.nii.ac.jp/src/S-JNAS.html

[4] F. Eyben, F. Weninger, F. Gross, and B. Schuller. ”Recent developments in opensmile, the munich open-source multimedia feature extractor,” Proc. ACM Int. Conf. Multimedia, pp. 835-838, Oct. 2013.