The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

音声対話システムのためのユーザの発話権維持状態の逐次推定

Incremental estimation of user's turn-taking state for a spoken dialog system

藤江 真也 *1*2 Shinya Fujie
横山 勝矢 *2 Katsuya Yokoyama
小林 哲則 *2 Tetsunori Kobayashi

*1千葉工業大学 Chiba Institute of Technology
*2早稲田大学 Waseda University

Turn-taking state estimation to determine utterance timing of a spoken dialog system is discussed. We propose the recurrent neural network based method to estimate user's turn-taking state incrementally. The proposed method utilizes acoustic feature extracted using a spectrogram autoencoder as well as linguistic feature extracted from a partial speech recognition result using a neural network based language model. The article shows an example of estimation result and discuss the performance of the proposed method.

1. はじめに

音声対話システムの発話タイミング決定のための，ユーザの発話権維持に対する状態を逐次予測する手法を提案する．

従来の音声対話システムの多くは，主に音声認識の前処理として実行される音声区間検出（VAD; Voice Activity Detection）の結果を利用し，音声信号の終了時点をユーザの発話終了として扱っている．更に，多くのシステムはその時点でユーザが発話権をシステムに渡している，あるいは放棄しているとみなし，即座にシステムの発話を開始するような単純な制御をしている．しかしながら，音声区間の途切れ目が発話終了とは限らず，ユーザは発話権を維持してそのまま発話を継続する可能性がある．また，必ずしも音声が途切れていなくともシステムが発話を開始してもよいタイミングが存在するのも事実である．これらに対し，音声区間ごとに発話の継続，終了を推定する手法がある [1, 2]．これらの研究では依然として VAD が検出する音声区間を対象としてその推定を行っている．VAD は，音声信号の終了を正確に検出するために，ハングオーバー処理（一定の長さの無音区間が続くまで音声区間が続いているとする処理）を行う．このため，VAD が検出する音声区間に対して推定を行っている限り，一定の処理遅れが生じることは避けられない．また，音声区間の途中でシステムの発話を開始することは原理的に不可能である．

そこで，逐次的に音声信号を処理し，フレームごとにユーザの発話権に対する状態（維持しようとしているか，譲渡または放棄しようとしているか）を評価することで，システムの発話タイミングを決定しやすくする方法を提案する．従来研究においては，主にシステムの相槌（back-channelling）のタイミング検出のためにこのような処理が行われている [1, 3]．その理由として，相槌生成においてはタイミングの遅れの悪影響が明らかであることが挙げられる．我々は相槌以外のシステム発話に関してもそのタイミングの遅れは好ましくないと考える．

本稿では，単純な方法で発話を認定して行った従来研究 [4] に対して，より厳密な方法で発話権の維持を認定し，予測モデルを構築した結果について報告する．

連絡先: 藤江真也，千葉工業大学 先進工学部 未来ロボティクス学科，shinya.fujie@p.chibakoudai.jp

2. 話者交替タイミングの分析

2.1 分析条件

すでに収集あるいは公開されている人対人の音声対話コーパスに関して，話者交替がどのようなタイミングで行われているかを分析する．

分析対象としては，以下の 4 つのコーパスを選んだ．

- 早稲田コーパス（W）: 早稲田大学と ATR-Trek で独自に収集したもの．レストラン検索サービスを想定した対話．391 対話，計 16 時間 30 分．
- 日本語話し言葉コーパス（CSJ）: 講演者とインタビュワーによる講演に関するインタビュー．58 対話，計 12 時間 10 分．
- RWCP 音声データベース（RWCP）: 自動車購入，海外旅行計画のスタッフと客を想定した対話．96 年度版と 97 年度版両方を対象とした．61 対話，計 8 時間 29 分．
- 重点領域研究「音声対話」対話音声コーパス（PASD）: 様々なタスクの対話で，部分的に人とシステムの対話を含む．62 対話，計 5 時間 8 分．

機械的に話者交替をカウントするために，音声区間を利用する．音声区間情報がコーパスに含まれるもの（CSJ，RWCP）は，それを利用した．その他のコーパスに関しては，公開されているツール（`https://github.com/ASTL-NICT/VAD`）を利用して区間検出を行った．音声区間の間のポーズ長が 700ms 以下の場合は，結合して 1 つの発話区間とした．

以上の処理を施した上で，一方の話者の発話区間の終了タイミングに対して，もう一方の話者の発話区間開始タイミングがどのように分布するかを調査した．

2.2 分析結果

話者交替の発生数は，W で 30,872，CSJ で 8,132，RWCP で 3,493，PASD で 4,131 であった．話者交替タイミングの累積相対度数を図 1 に示す．

タスクの違いによらず，話者交替の約 50%は先行話者の発話終了までに行われており，約 80%は先行話者の発話終了後 600ms まで，約 95%は 1,000ms までに行われていることが分かる．なお，PASD に関してはシステムとの発話が含まれており，システムの発話，あるいはシステムに対する人の発話が人同士の対話に比べて遅れる傾向にあることも分かった．

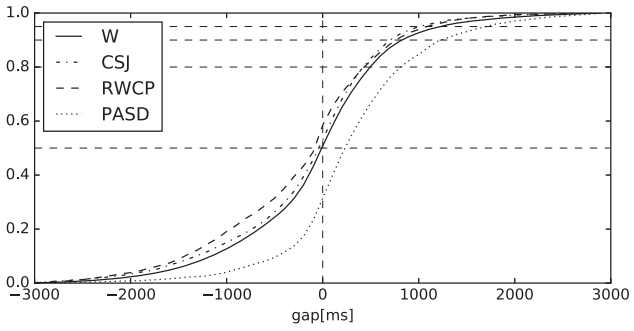

図 1: 話者交替の累積相対度数

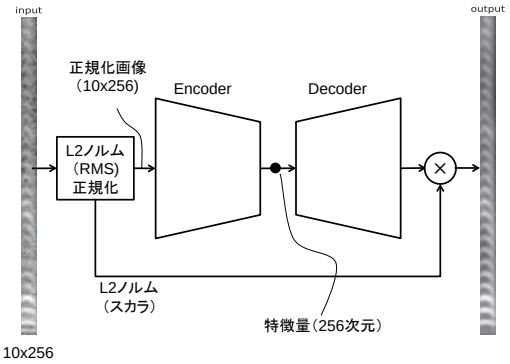

図 2: 音響特徴量抽出器（音声スペクトログラム自己符号化器）

3. 発話権維持の認定

前節の分析では 700ms 以上の無音で区切られた音声区間を発話区間とみなし，その交換を話者交替と考えた．ここでは，より正確に発話権の維持をモデル化するため，対話する二者の音声区間の相互関係をもとに発話権維持を認定する．

まず，音声区間を，それらの間の $\theta_{\text{short\_pause}}$ 以下の長さの無音を無視して 1 つに統合する．これを**発話区間**と呼ぶ．次に，$\theta_{\text{short\_utterance}}$ 以下の短い発話区間は**短い発話**と呼び，単独では発話権を維持する発話ではないと考える．

短い発話でない発話区間（$\theta_{\text{short\_utterance}}$ 以上の長さの発話区間）は，原則的に発話権を維持する発話であり，**発話権維持発話**と呼ぶ．発話権維持発話は，$\theta_{\text{keep\_turn}}$ より短い無音区間を無視して統合される．ここで，$\theta_{\text{keep\_turn}} > \theta_{\text{short\_pause}}$ である．すなわち，一度維持され始めた発話権は，ある程度長い無音区間が挟まれても維持されると考える．

例外的に，$\theta_{\text{short\_utterance}}$ 以上の長さを持つ発話であっても，先行して開始されている相手の発話権維持発話に時間的に内包される（相手の発話権維持発話より遅く始まり，早く終わる）場合は，短い発話とみなす．このルールに従うと，原理的にはどんなに長い発話でも，短い発話と認定される可能性があるが，実際には，本研究で扱うような協調的な対話においては，発話の衝突が起きた場合に発話権を獲得していない話者が喋り続けるということは起こらない．そのため，これを短い発話とみなすことは実用上問題ないと考える．

4. 発話維持状態推定モデル

4.1 問題設定

前節で認定した状態を，フレーム毎に逐次推定する．認定した状態のうち，発話権維持発話と短い発話，さらに発話していない状態の 3 状態は排他的であるため，これらは識別的に推定する．また，単に発話区間か，音声か非音声かについてもそれぞれ別々に同時に推定する．

4.2 特徴量抽出

発話終了の検出などにおいては，ピッチ変動などの韻律情報が重要であることが知られており，従来研究でも広く用いられている．しかしながら，ピッチ抽出は手法により精度が異なり，ある程度の精度を担保するためには複数フレームにより平滑化が必要となる．従って陽にピッチ抽出を行うと，全体の処理遅れが避けられなくなる．一方，音声認識で利用される MFCC などの音響特徴量を用いると，処理遅れは無くなるが，韻律情報が失われるという欠点がある．そこで本研究では，図 2 に示すような狭帯域スペクトログラムを符号化，復号化する

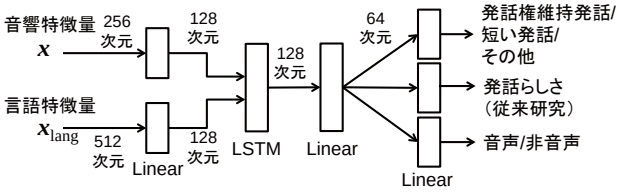

図 3: 発話権状態推定モデル

自己符号化器をニューラルネットワークにより構築し，その中間層の出力を音響特徴量とする．具体的には，10ms 毎に得られる 256 点のパワースペクトルを 10 本分並べたものを入力とし，中間層の出力 256 次元を特徴量とする．

発話権維持状態の推定においては，言語情報も有用であると考えられる．本研究では，Google Cloud Speech API (`https://cloud.google.com/speech/`) のストリーミング音声認識によって逐次得られる部分的な音声認識結果から言語特徴量を抽出する．具体的には，LSTM 言語モデルの中間出力 (512 次元) を用いる．なお，音声認識結果は毎フレーム得られるものではない．音声認識結果が得られないフレームについては，言語特徴量はゼロベクトルとする．

4.3 発話維持状態推定モデル

計算された音響特徴量，言語特徴量を入力とし，発話権維持状態を逐次推定するモデルをニューラルネットワークにより構築する．具体的な構造を図 3 に示す．

発話権維持状態に関しては，発話権維持発話，短い発話，その他（特に発話をしていない状態）を識別的に推定する．また，単に発話をしているかどうか，また，音声か非音声かを推定するネットワークを並列に入れたマルチタスク推定を行う．

5. 推定例

音響特徴量抽出に用いる自己符号化器は，日本語話し言葉コーパスの講演音声 1,000 講演からそれぞれ 10 発話ランダムに抽出した音声で学習した．発話権維持推定モデルは，早稲田コーパスの 100 対話の客役の音声で学習した．$\theta_{\text{short\_pause}}$，$\theta_{\text{short\_utterance}}$ は 1000ms，$\theta_{\text{keep\_turn}}$ は 1500ms とした．

学習に利用していないデータに対する推定例を図 4 に示す．発話権維持状態と短い発話の推定結果を見ると，発話権維持状

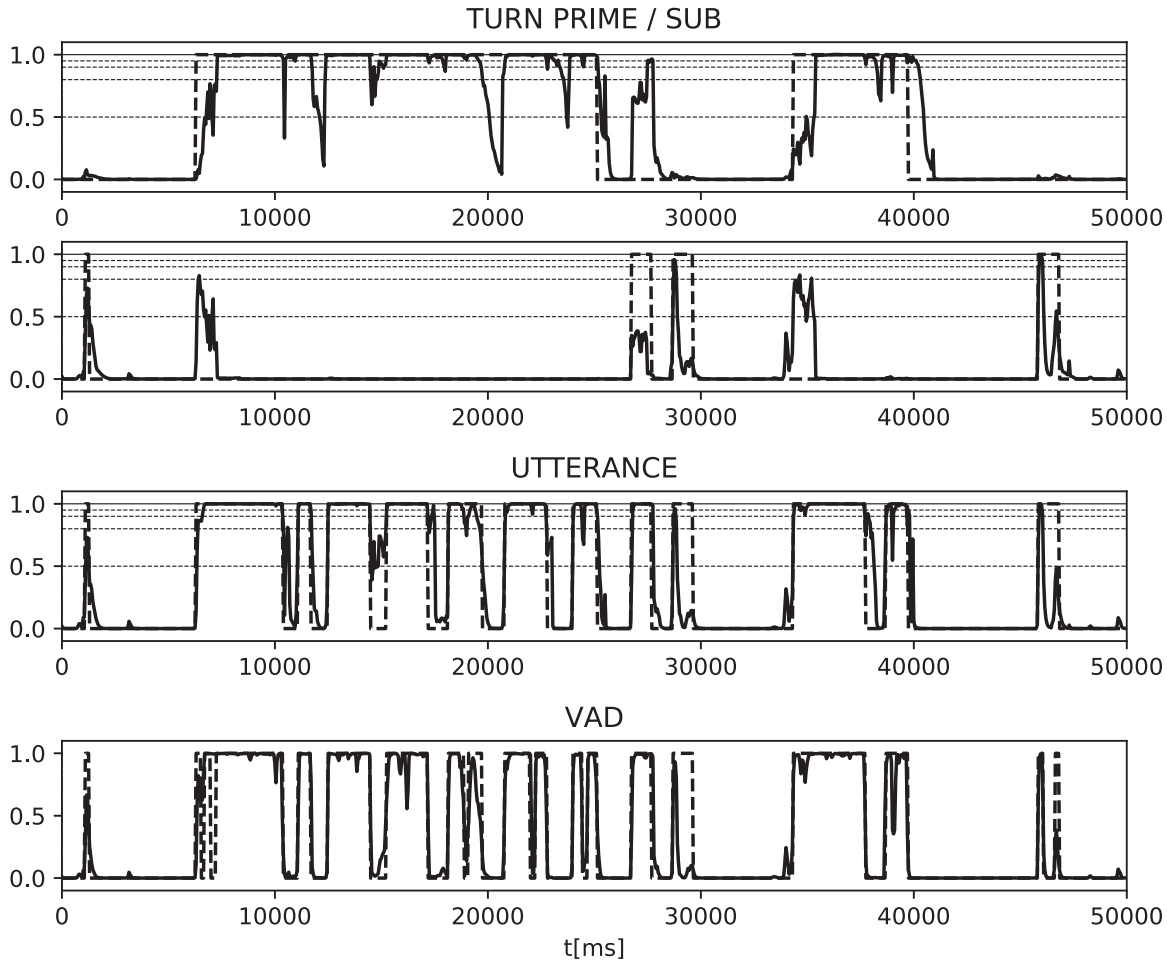

図 4: 予測結果の例．上から発話権維持状態，短い発話状態，発話らしさ [4]，VAD を表す．破線が目標値で，実線が予測結果．

態の始めは短い発話のスコアが高くなり，徐々に発話権維持状態のスコアが高くなっていくという傾向が見られる．これは，音声区間が開始した直後は，短い発話か発話権を維持するかの区別が付きづらいということの現れだと考えられる．また，一度発話権維持状態のスコアが高くなると，ある程度長い無音区間（発話区間や音声／非音声の目標値が 0 になっている区間）があっても，スコアが低くならずに発話権を維持していることを正しく推定できていることが分かる．一方，発話権維持が終了していると判断すべきところでも，しばらくスコアが高いままで，時間とともに徐々にスコアが低くなっている．これは，システムの発話開始の遅延につながるため，素早くスコアが下がるような改善を検討する必要がある．

6. むすび

音声対話システムの低遅延な発話タイミング決定を目指し，ユーザの発話権維持状態の逐次推定モデルを提案，実装した．今後の課題として，本手法で決定したシステム発話のタイミングが，VAD を用いた従来手法に比べて人同士の対話における発話タイミングに近いことを評価することや，実際の音声対話システムに組み込むことが挙げられる．

参考文献

[1] Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanishi, K., and Kawahara, T.: Attentive Listening System with Backchanneling, Response Generation and Flexible Turn-taking, SIGDIAL (2017)

[2] 藤江真也，木下裕介，小林哲則: 発話期待度／意欲度に基づく発話タイミング制御, 人工知能学会研究会資料，SIG-SLUD-B102，pp.43-48 (2011)

[3] Fuji, S., Fukushima, K., and Kobayashi, T.: Backchannel Feedback Generation using Linguistic and Nonlinguistic Information and its Application to Spoken Dialogue System, Proc. Interspeech (2005)

[4] 藤江真也，横山勝矢，小林哲則: 音声対話システムのためのユーザ発話終了タイミングの逐次予測，日本音響学会春季研究発表会講演論文集，2-Q-11 (2018)