音声対話の要約技術

Spoken Dialogue Summarization

議事録作成支援に向けて

会議などの自然発話においては「言語的曖昧性」を含む話し言葉から要旨を把握するのは容易ではない。本研究では、議事録作成の前段階として、このような音声対話をテキスト化し、発話内容の要点をまとめる要約技術について検討を行った。さらに、提案手法を模擬会議データで評価した結果について述べる。

執筆者
先端技術応用研究所
情報技術グループ
瀬川　修

1 背景と目的

これまで会議など音声対話の要約の研究が多数試みられているが、自然発話の認識の難しさに加え、曖昧な話し言葉から要旨を把握するのは容易ではない。また、音声対話においては発話分割と併せ話者識別という困難な問題が存在する。

最近では深層ニューラルネットワークに基づく音声認識技術の発展によって、音声対話の認識性能が向上し、対話内容の要約が技術的なスコープに入ってきた。そこで、本研究では議事録の自動生成に向けた音声対話の要約技術について検討を行い、模擬会議データによる提案手法の評価を行った。

2 音声対話の要約技術

音声対話の要約の実現に向けては、主な要素技術として、発話分割と話者識別、音声認識、および要約などが挙げられる。各要素技術と技術課題を第１表に示す。

第１表　音声要約技術の課題

要素技術	課題
発話分割と話者識別	・複数話者の識別 ・発話のオーバーラップ
音声認識	・自然発話の認識 ・未知語の学習
要約	・話し言葉からの要旨把握 ・要約文の生成

3 アルゴリズム検討

以下では、各要素技術の概要と課題を解決するためのアルゴリズムについて述べる。

(1) 発話分割と話者識別

長時間の音声に対し発話分割と話者識別（誰がいつ喋ったか）を行う手法のことを総称して話者ダイアライゼーション（Speaker Diarization）と呼ぶ。当該分野では、2019年頃からニューラルネットワークに基づくEnd-to-End方式が盛んに検討されている。本研究では、Transformerに基づくEnd-to-End方式をベースとして、話者数既知のマルチチャネル録音に対し、マイクを装着した「主話者」の音声を選択的に判別するダイアライゼーション方式を考案した（特許出願中）。

マルチチャネル録音では、話者別に口元で録音したとしても周囲の話者の声がオーバーラップして混入する。そこで、提案手法では、各チャネルの話者数が１名という条件の下で、モノラル混合音声の各フレームにつき主話者の発話区間かノイズ区間かの２値判別を行う。具体的には、入力音声の全体構造から話者区間を推定するglobalネットワークと、入力音声の局所的構造から話者区間を推定するlocalネットワークの２つ推論結果の統合によって、チャネルごとの主話者区間を推定する。そして、各チャネルの推定結果（話者ラベル）を統合することによって最終結果を得る。

(2) 音声認識

2015年頃から系列変換モデルとAttentionに基づくEnd-to-End音声認識手法の検討が盛んに行われるようになった。本研究では、TransformerとCTCを併用したEncoder-Decoderによるアルゴリズムを用いた。認識の最小単位（トークン）は日本語キャラクタである。

End-to-End音声認識手法では、モデル学習に音声と書き起こしテキストのペアが必要であり、語彙の追加にコストを要していた。そこで、我々はニューラル音声合成(Text-to-Speech :TTS) を用いて、テキストデータから対応する音声データを自動生成する「データ拡張」の枠組みを検討した。

(3) 要約

要約には、発話の重要度を評価して抜粋する「抽出型要約」と、対話全体を勘案して要約文を生成する「生成型要約」がある。最近では、大規模言語モデル（Large Language Model: LLM）を利用して生成型要約を行う方式が盛んに検討されるようになり、長期文脈を考慮した要旨の把握が可能になりつつある。本稿では、Transformerに基づく自己回帰型のLLM（オープンソース）を用いて生成型要約を行う方式の初期検討を行った。

ここで、オープンソースのLLMを利用するメリットであるが、分野やタスクに特化した独自の学習（ファインチューニング）が可能なこと。また、クローズドな環境で運用の内製化が可能などの点が挙げられる。

4 評価

(1) 評価データ

日本語に関しては、研究開発に利用可能な十分な質量の音声対話と要約の評価データが整備されておらず、今回独自に模擬会議コーパスの収録を行った。収録にあたっては、マルチチャネル録音装置と話者数分のピンマイクを用意し、各話者の胸元にマイクを装着して話者に対応する音声チャネルを記録した。評価に用いた模擬会議データの概要を第２表に示す。

第２表　模擬会議データの概要

議題	時間長 (min)	収録場所	話者数
1.情報セキュリティ	58	会議室A	5
2.画像アノテーション業務説明	61	会議室B	4
3.音声書き起こし業務説明	64	会議室C	4
4.コロナ&テレワーク	70	会議室C	4
5.音声書き起こし業務進捗確認	33	会議室C	3
6.他の業務の作業付与	14	会議室C	4
7.安全衛生、コンプライアンス	64	会議室A	3

(2) 発話分割と話者識別

モデル学習には日本語講演音声コーパスCSJに含まれる3,212講演からランダムに選択した2話者の発話を組み合わせて作成したシミュレーション対話を使用した。また、評価データは第2表のコーパスから4セッション(2,3,4,6)を用いた。

実験では評価指標として検出誤り率DERを用い、30.8%という性能が得られた。パワーが大きく周波数帯域が類似した話者はラベルの間違いが発生することがあるため、このようなエラーの低減は今後の課題である。

(3) 音声認識

音声認識の基本性能、および前述のTTSによるデータ拡張の有効性を評価するため、第2表のコーパスを用いた実験を行った。模擬会議コーパスのうち4セッション(5,663発話)を学習データ、3セッションを評価データ(4,616発話)として用いた。

実験では、以下の3つの条件設定で性能比較を行った。「CSJ-Transformer」は日本語講演音声CSJで事前学習されたベースラインモデル。「CSJ-Transformer +FT」は模擬会議コーパスでファインチューニングしたモデル。「CSJ-Transformer +FT +TTS」は、人間系で作成した未知語を含む414の例文をTTSでデータ拡張してファインチューニングしたモデルである。学習時のミニバッチにおけるリアルデータと拡張データの割合は2:1とした。

評価結果を第3表に示す（評価指標は文字誤り率CER)。同表から、ベースラインに対し模擬会議音声でファインチューニングを行った効果が確認できる。また、TTSデータ拡張によって未知語が学習され、CERが低減していることがわかる。データ拡張の有無によって改善した発話の例を第４表に示す（太字が例文で学習された未知語)。

第３表　音声認識性能評価

Model	CER (%)
CSJ-Transformer	33.9
CSJ-Transformer +FT	17.6
CSJ-Transformer +FT +TTS	**16.8**

第４表　TTSデータ拡張による未知語学習の効果

データ拡張なし	データ拡張あり
あと一違うい知かなもうそっかあそれが出た事あって	あと**非通知**違う**非通知**からもうっかりそれが出たことあって
社員者もだから	**社員証**もだから

(4) 要約

今回用いたLLMはオープンソースとして公開されているLlama2の日本語モデル（ELYZA-Llama-2-instruct、パラメータ数70億、130億）である。試行評価では、別途収録した模擬会議（議題「議事録の要約学習データ作成業務説明」、話者4名、時間長21min）の音声認識結果（人手修正あり）を用いた。要約で用いたプロンプトと要約例を第５表に示す。

同表より、モデルのパラメータ数に着目すると70億と比較して、130億のほうが具体的で要旨を捉えた生成結果になった。また、プロンプトに条件を与えることにより、口語調の対話からフォーマルな様式の要約文が生成できることがわかった。なお、ファインチューニングの効果については、執筆時点で会議要約の学習データが十分用意できなかったため、今後の検討課題である。

第５表　会議音声要約のプロンプトと要約例
議題「議事録の要約学習データ作成業務説明」

プロンプト: あなたはプロの編集者です。これから文章を渡すので、以下の条件に従ってその要約を作成してください。
条件
・重要なキーワードを取りこぼさない。
・「あー」や「えーと」など、言い淀み（間投詞、フィラー）を取り除く。
・文章の意味を変更しない。
・架空の表現や言葉を使用しない。

要約例: ELYZA-Llama-2-7b-instruct（パラメータ数 70億）
学習データの作成のために、議事録の要約を行う必要がある。議事録は、日時、場所、出席者、発言内容などの項目から構成されている。また、議事進行の方の話の要点を書き、それに対して質問をしている姿を想像できるような感じでまとめる。

要約例: ELYZA-Llama -2-13b-instruct（パラメータ数 130億）
業務の一つとして、議事録の要約作成がある。議事録は厚生労働省のワクチン打ち合わせの会議のものを例に、分割して作成することが推奨される。司会者が議事内容をまとめている箇所を抽出し、要約とする方針で作業を行う。今回は、地震津波に関する会議の例を20点ほど頂いており、要約は手間がかかるが、量は多くないため、1枚程度にまとめるのが適切である。

5 まとめ

本研究では議事録作成の前段階として、会議における音声対話の要約技術の検討を行い、実現性を確認した。（本研究は㈱Human Dawaware Lab.との共同研究として実施した）