Vol.2017-SLP-118 No.9
2017/10/13

音声認識技術を用いたコンタクトセンターオペレータ支援

長野 徹[1,a)]　壁谷 佳典[2]　岡原 勇郎[2]　吉田 一星[3]　倉田 岳人[1]　立花 隆輝[1]

概要：コンタクトセンターにおける音声認識と検索を用いたオペレータ支援システムについて紹介する．企業の販売チャネルが対面からインターネットへと移行するにつれ，コンタクトセンターに寄せられる問い合わせは多岐にわたるようになってきた．迅速に顧客への応答を行うため，マニュアルや FAQ をデータベースとした検索機能の充実が図られているが，さらなる顧客応答時間の短縮・自動化を目指し，音声認識技術と検索機能とを組み合わせた音声によるリアルタイム検索システムを構築した．

キーワード：音声認識, 類似文検索, コンタクトセンター, 支援システム

Agent Supporting System using Speech Recognition in Contact Center

Tohru Nagano[1,a)]　Yoshinori Kabeya[2]　Isao Okahara[2]　Issei Yoshida[3]　Gakuto Kurata[1]
Ryuki Tachibana[1]

Abstract: We developed an agent support system using speech recognition and similarity search. As a customer contact point is changing from face-to-face to the internet in a decade, contact center is becoming more and more important to the customers. The contact center is tacking to organize existing manuals, faqs, and related data to answer to the customers' inqueries rapidly and to improve quality of responses. In order to shorten the response time and reply more accurately, we made the system that automatically reocognize speech conversations between contact centers' agent and customer, and search similar texts from the database in real time.

Keywords: Speech recognition, Similarity search, Contact center, Support system

1. はじめに

ディープラーニングを用いた近年の音声認識の精度向上は著しく，Switchboard データを対象にした音声認識誤り率 (単語誤り率) は 2014 年には 10.4%[1] であったが，2015 年には 8.0%[2]，2016 年には 6.9%[3]，今年 3 月には 5.5%[4][5] そして 8 月には 5.1%[6] [7] の音声認識誤り率を達成している．人間の聞き取り能力についても様々な分析が行われているが [4][8]，Switchboard タスクに限定するとほぼ人間の聞き取り能力に達したと言ってもよい．一方，他人同士の電話会話を模した Switchboard データとは異なり，同じ電話会話でもトピックを限定しない，家族間のよりくだけた音声を多く含む CallHome データでは 10.1%[4] の音声認識誤り率と，一般的な状況下において人間の聞き取り能力に達しているとまでは言えない状況にある．電話会話の音声認識に加え，多人数発話に対する音源分離 [9]，高雑音下での音声認識 [10]，Far Filed での音声認識 [11]，回線・圧縮歪みによる帯域の欠損した音声への対応，非定常ノイズに対する処理，等，多種多様な音声を「人間と同等に」認識できるようになるにはまだ少し時間がかかるが，ビジネスへの応用は急速に進みつつある．

音声認識の応用として，車載機器におけるコマンド音声認識を代表とする，キーボード以外の代替入力インターフェースとしての役割がよく知られており，現在も Siri, Cortana, Alexa といった音声アシスタント機能のフロント

[1] 日本アイ・ビー・エム（株）東京基礎研究所
IBM Reseach - Tokyo
[2] 同社 グローバル・ビジネス・サービス
Global Business Services, IBM Japan
[3] 同社 ソフトウェア開発研究所
Software Development Laboratory, IBM Japan
[a)] tohru3@jp.ibm.com

エンドインターフェースとして進化し続けている．一方，音声認識には，人手での処理がコスト的に適さない大量の音声データを処理するための手段としても用いられており，例えば筆者らは，2009 年頃からコンタクトセンターの不適切な発話を検出するためのコンタクトセンターモニタリングに音声認識技術の活用を行っている．コンタクトセンターに蓄積される大量の音声データを人間の聴取作業によりチェックするには，録音された音声と同じ時間を必要とするが，99.9% 以上の発話時間は問題のない発話であり，コスト面から人手での聴取は適さない．大量の録音音声をバッチ作業にて音声認識し，認識結果から業務上不適切な発話を検出することで，人手による検出精度には及ばないが網羅的かつ低コストでのモニタリング作業が可能となった．

そして，さらなる音声認識の応用として，2014 年からコンタクトセンターにおいてリアルタイムに音声認識し，リアルタイムにオペレータを支援するオペレータ支援システムの研究・開発を行っている．これは従来の IVR（Interactive Voice Response）のように顧客が主体となってシステムに話しかけるのではなく，顧客とオペレータは通常どおり対話を行い，その内容を横で聞いてオペレータの顧客対応を支援するコンタクトセンターのスーパーバイザの役割を果たすように設計されている．このシステムの利用により，顧客に対し素早くかつ正確に応答することが可能となり，顧客応対の質の向上が期待できる．本論文ではこのコンタクトセンターにおけるオペレータ支援システムについて説明する．

2. コンタクトセンターでのオペレータ支援

電話応対時間の短縮は，コンタクトセンターにおける最も基本的な課題の一つである．商品の機能，サービスの内容が複雑化するにつれ，コンタクトセンターに寄せられる問い合わせも複雑化し，オペレータに求められる能力も高いものとなってきた．また個々のオペレータの能力も熟練度により差があるため，長時間の電話応対になるケースも多い．一般的にオペレータが顧客応対を行う際，

- 対話を通じて問題を定義
- 想定された問題を質問応答データベースにて検索
- もっとも一致する応答を顧客に案内

というサイクルを，問題が解決するまで繰り返すが，経験の浅いオペレータは，問題定義にも検索にも時間を要する．そこで，本システムでは，この問題定義と検索の時間を短縮するため，オペレータと顧客との対話をリアルタイムに音声認識し，その認識結果を用いてデータベースの検索を行う．オペレータは検索された結果からもっとも一致すると思われる応答を顧客に案内するだけでよい．

正しい応答をオペレータに提示し，オペレータ支援の効率を高めるためには

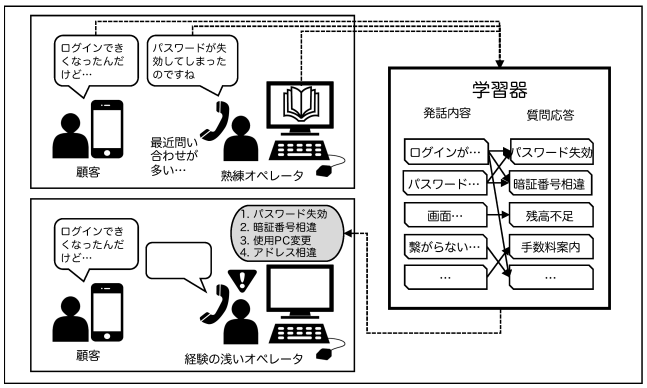

図 1 知識ベースを介したオペレータの相互支援

- 認識精度の高い音声認識
- 音声認識結果に対して正しい応答を返す類似文検索

が必要となるが，これらコンポーネントを人手でチューニングするのはコストがかかるため，ほとんど人手のかからない方法でこれらコンポーネントのチューニングを行う．音声認識精度の向上のためには，音声認識結果に対してデータセレクションを行い，得られた音声認識結果のテキストを音声認識の言語モデル用コーパスとして用いる．コーパスは Watson 音声認識の言語モデルカスタマイズの仕組みを通じて言語モデルに追加される．類似文検索に対しては，オペレータの簡単な操作により正解ラベルの付与を行うことで教師データを取得し，モデルの改善に用いることで精度を高めている．

コンタクトセンター全体で考えると，熟練したオペレータの音声認識結果・正解ラベルがシステムに反映されることにより，システム利用者全員に対して精度の高い応答結果を返すことができるようになる．熟練したオペレータがスキルの足りないオペレータを直接補助するのはコストがかかるが，システムを通してスキルの高いオペレータの能力を低コストで共有できることが期待できる（図 1）．

3. オペレータ支援システム

3.1 システム概要

システムの概要を図 2 に示す．オペレータ支援システムの入力はオペレータと顧客との音声データであり，

(1) オペレータが顧客からの問い合わせを受電すると同時に音声データがストリームとして音声認識エンジンに入力される．
(2) 入力された音声は音声認識エンジンによりテキスト化され，類似文検索システムへの入力となる．
(3) 類似文検索システムは，入力テキストを質問文として，その質問に最も類似した質問応答を関連度の高い順に出力する．

そして，オペレータは，複数の質問応答の中から最も適切

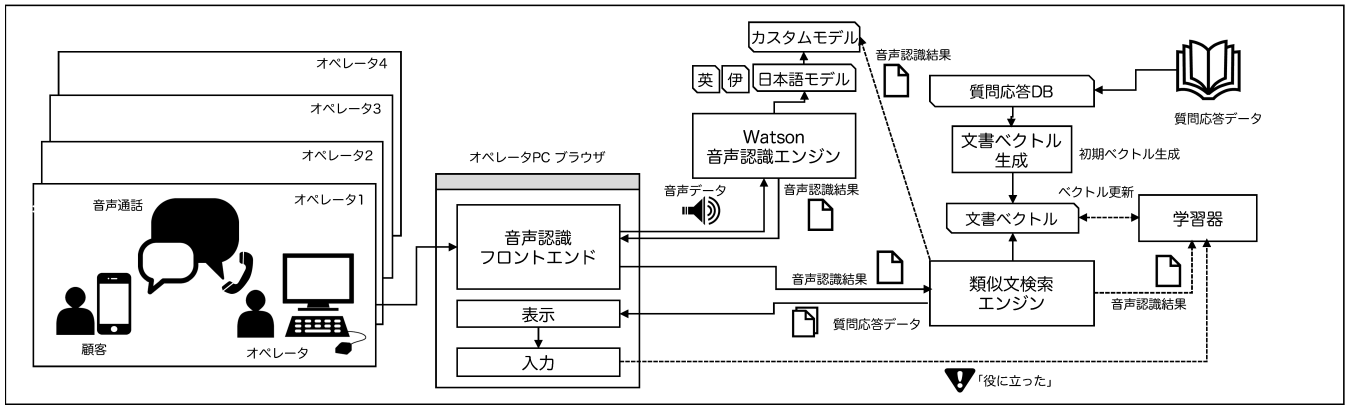

図 2 システム概要

と思われる応答文を参照し，顧客へ適切な返答を返す．この一連の流れを通話が終了するまで繰り返す．

通話が終了すると，コールログの記述など通常の業務と同時に，オペレータがシステム利用中に参照した質問応答に対して「役に立った」というフラグを立てることができる．この「役に立った」フラグを入力テキストである音声認識結果に対する正解ラベルとみなして用いて類似文検索システムのパラメータを学習する．また，音声認識結果自体も音声認識モデルの改善に用いられる．以下では (2) 音声認識と (3) 類似文検索の動作について解説する．

3.2 音声認識

3.2.1 Watson 音声認識

本システムでは Watson 音声認識システム [12][*1] を用いて音声認識を行う．Watson 音声認識はクラウド上のサービスとして REST(`https`) や Websocket(`wss`) プロトコルを通じて提供される．例えば日本語の狭帯域用音声認識モデルは，以下のような簡単な REST API を通じてアクセスすることができる．

```
https://stream.watsonplatform.net/
speech-to-text/api/v1/
recognize?model=ja-JP_NarrowbandModel
```

Watson 音声認識システムには，

- 発話の終了を待つことなくストリーミングで音声を受理し，リアルタイムに音声認識結果を出力できる
- 音声認識言語モデルのカスタマイズを行い，認識結果を改善できる

という特徴があり，本システムのような特定のトピックに特化したリアルタイムアプリケーションに適している．

3.2.2 音声入力フロントエンド

オペレータと顧客との会話は，受話器に接続された録音装置または電話交換機（PBX）に接続された通話録音装置（ロガー）を通じて取得される．音声データは顧客側は 8KHz の帯域，オペレータ側は 8KHz または 16KHz の帯域を持った 2 チャンネル音声である．音声入力フロントエンド（図 3）は VAD（Voice Activity Detector）を通して入力された音声を Watson 音声認識にストリームとして送信，非同期に音声認識結果を受け取る．音声認識結果は発話ごとにまとめられ，サーバー上の類似検索エンジンに送信される．

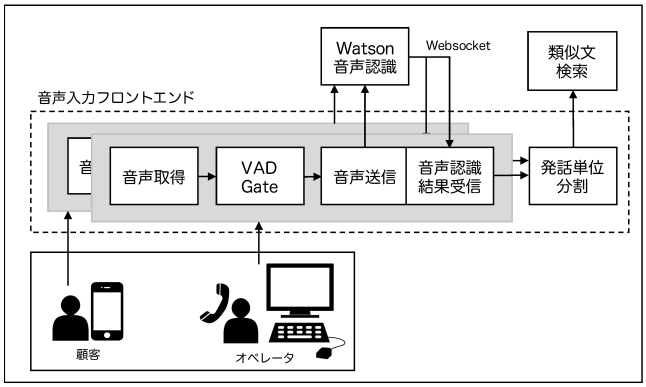

図 3 音声入力フロントエンドの動作

3.2.3 言語モデルカスタマイズ

音声認識の結果を改善するために Watson 音声認識では言語モデルのカスタマイズ機能が提供されている．対象分野に関する単語（表記と読みの組）とコーパスとを与えることで，分野適応が可能となる．カスタマイズを用いた音声認識は以下のようにカスタマイズの利用毎に発行される customization_id を URI に追加するだけで利用できる．

```
https://stream.watsonplatform.net/
speech-to-text/api/v1/
recognize?model=ja-JP_NarrowbandModel
&customization_id=03dedede-1234-abcd
```

例えば，実業務で生じる音声データに対し，1,470 文 21,000 単語を含む関連する日本語テキストを言語モデルカスタマ

[*1] なお，本稿執筆時におけるサービスの仕様に沿って説明するが，サービスの仕様は随時変わる可能性があるため適宜 API 仕様を確認されたい．また URL の一部は省略してある．

イズに利用したところ，22.2%の音声認識単語誤り率が，カスタマイズモデルを用いることで 18.7% へと 15.7 ポイント誤りが削減された．

3.3 類似文検索

3.3.1 類似文検索システム

質問応答データベースには「よくある質問（FAQ）」などが登録されている．本システムでは，顧客に対して直接応答を返すのではなく，オペレータを支援するのが目的である．膨大な質問応答データベースから顧客の質問に最も近い質問を返すこと（図 4）で，オペレータはマッチした質問応答を参照し，ユーザーに応答部分を踏まえた返答を行う．質問応答データベースは，質問部と応答部が対となった質問応答の集合からなる．音声認識結果を入力として，データベースに含まれる質問部（応答部を含めてもよい）と類似度を計算し，類似度の高い質問応答から N-best を出力する．

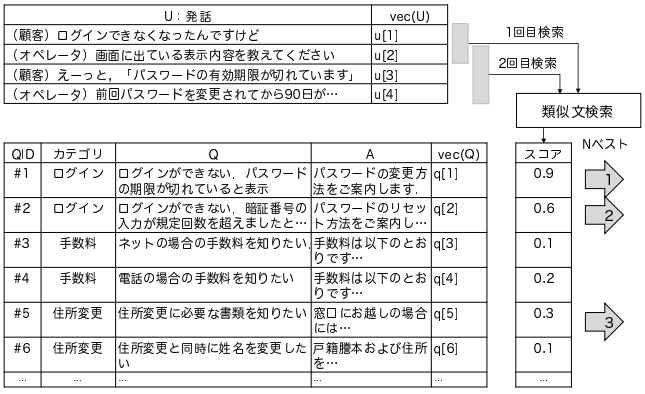

図 4 類似文書検索を用いた質問応答

3.3.2 学習

入力テキスト（音声認識結果）を単語ベクトルとして表現し，同じく単語ベクトルとして表現された質問部との類似度を計算する．初期状態での文書類似度は必ずしもオペレータの期待したものではなく，またシステムが長期間使われる間，新たな質問応答の変更・追加が発生する．これらデータベースの変更に対し人手で単語ベクトルの重みの調整を行うことは効率が悪い．そこで，オペレータ自身の最小限の作業で単語ベクトルの重みを更新できる以下のような教師あり学習の枠組みを，システム内に導入している（図 5）．学習のための正解ラベルはオペレータが 1 件の顧客応対終了時に付与する「役に立った」ラベルが利用される．

ランキング学習 「役に立った」ラベルから，クエリと適合した文書を正例，それ以外の文書（全てまたはランダムに選択された複数の文書）を負例として各単語の重みを学習する．

カテゴリ学習 参照回数の少ないロングテールの質問応答は，そもそも「役に立った」ラベルが付与されることが期待できない．一方，同じカテゴリの文書では同じ単語の重みはある程度共有できると考えられ，個別の文書の重みの学習ではなく，同じカテゴリに属する文書は同様に正例とみなして学習する．

クリック学習 「役に立った」ラベルから，クエリと適合した文書のペアの重みを更新する．

例えば，実業務で生じる音声データに対し，学習データを用いて類似文書検索モデルを改善した．正解が検索結果の上位 N 件に含まれたかを N 位再現率で評価を行ったところ，N=1,5,10 において，初期状態で 16%,44%,57% であった再現率がこれらの学習を行うことで 46%,80%,88%と，すべての N に対して 30%から 40%近く改善した．

図 5 類似文検索の学習フェーズ

4. おわりに

本稿では音声認識を用いたコンタクトセンターにおけるオペレータ支援技術を紹介した．本システムを用いることで，オペレータに音声認識の存在を意識させることなく，オペレータをサポートすることが可能になり，顧客応対の質の向上が期待できる．

音声認識や類似文検索の精度は日々向上しているが，現状では，どの分野においても汎用的に高い精度を達成しているシステムはなく，分野適応が重要な技術となっている．本システムでは，オペレータと顧客との対話自体・オペレータによるフィードバックにより，フィードバックしたユーザーだけではなく，システム利用者全員にメリットがあるシステムとなっており，「育てて」いくことで，常に最新のデータに対応できるようになっている．

これら学習されたコンポーネントは，オペレータ支援だけでなく，音声による質問応答システムや，テキストベースのチャット型質問応答システムにも利用できる．一方，様々なコンポーネントの組み合わせでは，各コンポーネントの誤りが組み合わされることによる限界が予想され，今後は End-to-End で問題解決をする仕組みが重要になると考える．

参考文献

[1] Soltau, H., Saon, G. and T.N., S.: Joint training of convolutional and non-convolutional neural networks, *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2014)*, pp. 5609–5613 (2014).
[2] Saon, G., Kuo, H., Rennie, S. and Picheny, M.: The IBM 2015 English Conversational Telephone Speech Recognition System, *arXiv preprint*, p. arXiv:1505.05899 (2015).
[3] Saon, G., Rennie, S. and Kuo, H.: The IBM 2016 English Conversational Telephone Speech Recognition System, *arXiv preprint*, p. arXiv:1604.08242 (2016).
[4] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L. amd Roomi, B. and Hall, P.: English Conversational Telephone Speech Recognition by Humans and Machines, *arXiv preprint*, p. arXiv:1703.02136 (2017).
[5] Kurata, G., Sethy, A., Ramabhadran, B. and Saon, G.: Empirical Exploration of Novel Architectures and Objectives for Language Models, *The 18th conference in the annual series of INTERSPEECH (INTERSPEECH2017)*, pp. 279–283 (2017).
[6] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X. and Stolcke, A.: The Microsoft 2017 Conversational Speech Recognition System, Technical report (2017).
[7] Kurata, G., Ramabhadran, B., Saon, G. and Sethy, A.: Language Modeling with Highway LSTM, *Proc. ASRU* (2017 (to appear)).
[8] Stolcke, A. and Droppo, J.: Comparing Human and Machine Errors in Conversational Speech Transcription, *The 18th conference in the annual series of INTERSPEECH (INTERSPEECH2017)*, pp. 137–141 (2017).
[9] Hershey, J., Chen, Z., Le Roux, J. and Watanabe, S.: Deep Clustering: Discriminative Embeddings for Segmentation and Separation, *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2016)*, pp. 31–35 (2016).
[10] Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W., Espi, M., Higuchi, T., Araki, S. and Nakatani, T.: The NTT CHiME-3 System: Advances in Speech Enhancement and Recognition for Mobile Multi-Microphone Devices, *IEEE Automatic Speech Recognition and Understanding Workshop 2015 (ASRU2015)*, pp. – (2015).
[11] Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T. and Bacchiani, M.: Generation of Simulated Utterances in Virtual Rooms to Train Deep Neural Networks for Far-field Speech Recognition in Google Home, *The 18th conference in the annual series of INTERSPEECH (INTERSPEECH2017)*, pp. 379–383 (2017).
[12] IBM: Speech to Text 音声認識 — Watson Developer Cloud, IBM（オンライン），入手先 ⟨https://www.ibm.com/watson/jp-ja/developercloud/speech-to-text.html⟩（参照 2017-09-20）.