Vol.2013-SLP-98 No.6
2013/10/25

クラウド型音声認識サービスの車載機利用のための音声処理技術開発

本間 健[1,a)]　額賀 信尾[1]　大淵 康成[2]

概要：近年、スマートフォン向けのクラウド型音声認識サービスが、多数登場している。これらのサービスは、従来の組み込み型音声認識システムに比べ、幅広い単語や文を認識することができる。そのため、車載機の音声インタフェースに利用することにより、ユーザの利便性を向上できると考えられる。しかし、多くのクラウド型音声認識サービスは、スマートフォン向けに開発されているため、走行雑音がある車内環境では、性能を発揮できない場合がある。われわれは、クラウド型音声認識サービスを車載機で利用することを目的として、技術開発を行った。第 1 に、走行雑音がある環境で頑健に動作する発話区間検出モジュールを開発した。第 2 に、クラウド型音声認識サービスの前段で使用する雑音抑圧モジュールを開発した。開発したモジュールの性能評価を行った結果、走行雑音環境下で高い効果が得られることを確認した。

1. はじめに

1.1 背景

音声インタフェースは、手や目を拘束しないため、自動車の運転中でも安全に操作できるインタフェースとして、多くの車載機に搭載されてきた。従来の車載機の音声インタフェースでは、組み込み型計算機の限られたリソース上で音声認識を行わなければならないため、受理できる単語のバリエーションや、音声で検索できる地名のバリエーションに制限があった。そのため、ユーザは、あらかじめ発話できる単語や言い回しを覚えておく必要があった。

一方、近年、スマートフォン向けのクラウド型音声認識サービスが、多数登場している。これらのサービスは、組み込み型の音声認識システムに比べて、認識できる単語や言い回しが多く、認識率も実用レベルに達している。そのため、車載機の音声インタフェースにクラウド型音声認識サービスを導入することは、発話できる単語や言い回しを広げるための有力な 1 手段となる。われわれは、クラウド型音声認識サービスを車載機に適用することで、音声インタフェースの利便性を向上させる検討を行ってきた。

クラウド型音声認識サービスを車載機で利用する際の課題として、走行雑音への頑健性の向上が挙げられる。スマートフォン向けのクラウド型音声認識サービスの多くは、スマートフォンの通常用途で最適となるように開発されている。すなわち、通常の生活環境で、スマートフォンに近い位置から発話すれば、性能を発揮することができる。一方、自動車内の環境では、マイクは運転者から離れた位置に設置されており、かつ、エンジン音やロードノイズといった走行雑音が存在する。そのため、認識対象となる音声には走行雑音が混入し、発話区間検出の困難性が増し、音声認識率の低下が発生する。これらの走行雑音による性能低下に対処する技術が必要である。

走行雑音に対する音声認識の性能向上の研究は、多数の研究者が行ってきた。従来研究の多くは、音声認識技術と雑音対応技術は不可分のものとして扱われてきた。たとえば、耐雑音型の音響モデル [1] のように、音声認識器の内部に雑音対応の技術を入れ込む多くのアプローチが、認識性能を大きく向上させることが知られている。一方、車載サービス事業者がクラウド型音声認識サービスを利用する場合、クラウド型音声認識サービスはブラックボックスであるため、音声認識器内部の改造によって雑音対応をすることはできない。また、クラウド型音声認識サービスは複数存在し、それぞれ備える機能 (言語、語彙範囲など) が異なる。そのため、車載サービス事業者がユーザに対して幅広いサービスを提供するためには、複数の音声認識サービスを使い分けられることが望ましい。そのため、走行雑音対応をする際には、さまざまなクラウド型音声認識サービスに適用できる、汎用的な手法であることが求められる。

[1] (株) 日立製作所 中央研究所
Central Research Laboratory, Hitachi, Ltd., Tokyo 185–8601, Japan
[2] クラリオン (株)
Clarion Co.,Ltd., Saitama 330–0081, Japan
[a)] takeshi.homma.ps@hitachi.com

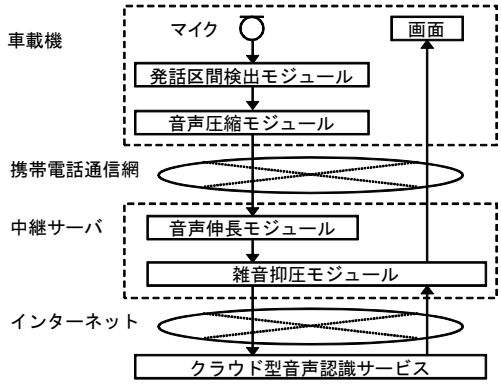

図 1 システム構成
Fig. 1 System diagram.

1.2 本研究の目的

以上の背景から、われわれは、クラウド型音声認識サービスを車載機に適用するにあたり、音声認識サービスに依らずに使用できる走行雑音対応のための技術を開発することを目的として、研究を行った。第 1 に、走行雑音環境下でも頑健に発話区間を検出できる発話区間検出モジュールを開発した。第 2 に、走行雑音による認識率低下を防ぐための雑音抑圧モジュールを開発した。

本稿の構成は以下のとおりである。2 章では、クラウド型音声認識サービスに走行雑音対応技術を適用するためのシステム構成を示す。3 章では、発話区間検出モジュールおよび雑音抑圧モジュールの手法および性能評価結果を示す。最後に、4 章にて、まとめと今後の予定を示す。

2. システム構成の検討

クラウド型音声認識サービスに走行雑音対応技術を適用するにあたり、システム構成の検討を行った。図 1 に、検討結果となるシステム構成を示す。

システム構成のうち、クラウド型音声認識サービスは、外部の事業者が提供する前提である。そのため、検討対象となる範囲は、ユーザの音声をクラウド型音声認識サービスへ届けるまでの範囲、および、クラウド型音声認識サービスの認識結果をユーザに提示するまでの範囲となる。

車載機では、ユーザが発話した音声がマイクで集音されたのち、発話区間検出モジュールへ入力される。発話区間検出モジュールは、音声から、ユーザが発話している時間区間だけを抽出し、後続の処理へ渡す。

発話音声のデータは、携帯電話の通信網を通じて、中継サーバに送信される。ユーザが認識結果を得るまでの応答時間を早めるうえで、データ転送速度が遅い携帯電話網による通信量を極力低減する必要がある。そのため、音声データは、送信のまえに、車載機にて圧縮される。

中継サーバは、われわれが独自に開発したものである。この中継サーバには、雑音抑圧モジュールがあり、車載機から来た音声に対する雑音抑圧処理が行われる。

音声認識サービスを図 1 のようなサーバと端末 (車載機) の構成で実現する場合、各モジュールを端末に置くか、サーバに置くかを選択する余地が出てくる。われわれは、雑音抑圧モジュールはサーバに置き、発話区間検出モジュールは車載機に置く構成とした。

雑音抑圧モジュールをサーバに置いた理由は、次のとおりである。第 1 に、雑音抑圧技術は、いまだに新しい手法が開発されている分野の技術である。最新の技術をユーザへ提供するためには、ソフトの更新が容易で、かつ最新のアルゴリズムを駆動する上で計算機リソースが豊富であるサーバのほうが都合がよいからである。第 2 に、クラウド型音声認識サービスは複数存在するため、それぞれのサービスに相性がよい雑音抑圧方式を実装するためには、サーバ側にソフトがあるほうが対応しやすいからである。

一方、発話区間検出モジュールを車載機に置いた理由は、ユーザの発話からいち早く発話開始/発話終了を判断し、即座にユーザに結果を伝える必要があるからである。発話区間検出においても、サーバ上に置くことで、豊富な計算機リソースを利用した最新のアルゴリズムを使用することができる。しかし、携帯電話通信網の通信遅延により、リアルタイムに発話開始/発話終了をユーザに伝えることは困難になる。そのため、たとえばユーザは発話を終えているのに、画面には「お話しください」といった発話を促す表示が出続けるといったことが起こり、ユーザの使い勝手を落とすことになる。そのため、本研究では、発話検出の結果を即座にユーザに伝えられることを重視し、車載機に置く構成とした。

3. 走行雑音対応技術

3.1 発話区間検出モジュール

3.1.1 手法

発話区間検出モジュールに求められる要件は、走行雑音が混入している音声波形から、ユーザが発話している時間区間 (発話区間) だけを抽出することである。発話区間を抽出する基本的な手法は、音声の短時間ごとのパワーを計算し、このパワーが定めた閾値より大きい場合、発話区間だと判断する手法である。この手法は、音声の背景にある雑音が小さい場合には、うまく動作する。しかし、自動車のような背景雑音が大きい環境では、発話区間と無発話区間のパワーの差が小さいため、適切なパワーの閾値を事前に設定することが困難となり、発話区間の検出誤りが高頻度で発生する。

われわれは、この問題に対処するため、音声に対して、背景雑音の抑圧処理を行った後に、パワーに基づいて発話区間を判定する手法を取った [2]。図 2 に、発話区間検出モジュールの処理手順を示す。発話された音声波形は、短時間に分割され、スペクトルに変換された後、Minima-Controlled Recursive Averaging(MCRA) アルゴリズムにより、背景雑

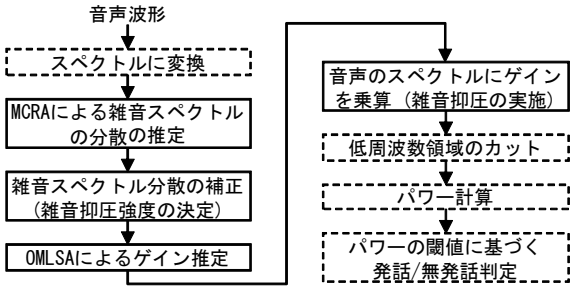

図 2 発話区間検出モジュールの処理手順
Fig. 2 Process of voice activity detection module.

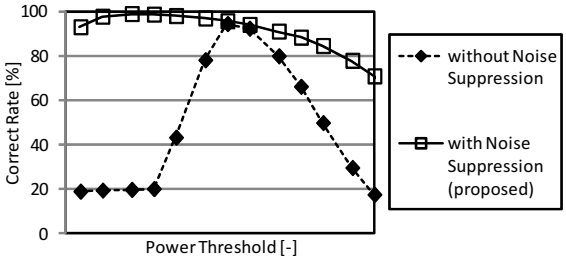

図 3 発話区間検出の正解率
Fig. 3 Correct rate of voice activity detection.

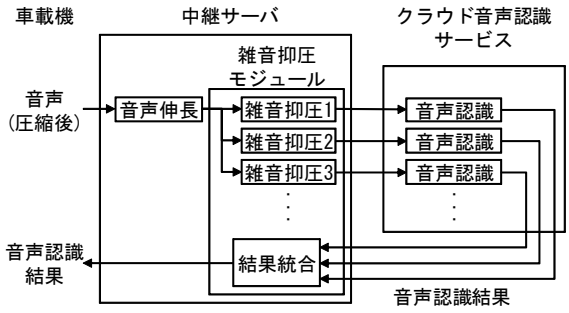

図 4 雑音抑圧モジュールの構成
Fig. 4 Block diagram of noise suppression module.

音の特徴量 (スペクトルの分散) が推定される。つぎに、推定された背景雑音の特徴量をもとに、Optimally-Modified Log Spectral Amplitude(OMLSA) アルゴリズム [3] により、音声中のユーザ発話音声の比率をゲインの形で推定する。つぎに、求められたゲインを音声のスペクトルに乗算することで、背景雑音が抑圧された音声のスペクトルを得る。最後に、走行雑音のみが存在する低周波数領域をカットしたのち、パワーを計算し、このパワーが閾値を上回った場合に、発話区間であると判定する。

OMLSA に基づく雑音抑圧では、一般に、推定された背景雑音の特徴量に対して、定数を乗算したのち、音声のゲイン推定が行われる。この定数が大きいほど、雑音抑圧効果が高くなるが、雑音抑圧後の音声に含まれるひずみも多くなる。音声強調を目的とした雑音抑圧の場合、音声のひずみを小さくするため、この定数に 1 未満の値を用いることが多い。一方、本研究のように発話区間検出の前処理に用いる場合、音声のひずみは問題にならず、むしろ強めに雑音抑圧をかけることによって、発話区間と無発話区間のパワーの差を大きくすることが重要である。この理由から、本研究では、推定された背景雑音の特徴量に対して、1 以上の定数を乗算した。

3.1.2 性能評価

最初に、時速 100 km/h で高速道路を走行し、自動車内に取り付けたマイクにより、走行雑音を収録した。収録時のマイクの位置は、サンバイザーおよび車載機筐体内部とした。そして、収録した背景雑音を、静音環境で収録した施設名を発話した約 2,300 個の音声データに重畳した。雑音重畳時の SNR には、別途、高速道路走行中に既定の文言を発話する実験を行い、それぞれのマイクの収録音声から求めた SNR を用いた。

つぎに、開発した発話区間検出モジュールによる発話区間検出の正解率を求めた。ここでは、正解の発話開始/終了時刻と、発話区間検出モジュールが出力した発話開始/終了時刻とが、ともに 0.5 s 以下の誤差であった場合に、正解とみなした。また、比較のために、雑音抑圧を行わずに、図 2 の破線で示した処理のみで発話区間を検出する手法 (従来手法) の正解率も算出し、開発した手法 (提案手法) との比較を行った。

図 3 に、発話区間検出の正解率の関係を示す。横軸は、発話区間と判定するパワーの閾値である。縦軸は、各マイクの音声に対する発話区間の正解率を平均した値である。雑音抑圧を行う提案手法では、発話区間とみなすパワーの閾値を最適に設定した場合、正解率が 98.8 %となり、従来手法を上回った。また、従来手法では、パワーの閾値によって正解率が大きく変動するが、提案手法では、パワーの閾値の広い範囲で、90 %以上の正解率が得られた。このことから、提案手法は、走行環境でも高い発話区間検出性能を示すことを確認できた。

3.2 雑音抑圧モジュール

3.2.1 手法

1 章でも述べたとおり、車載サービス事業者にとって、クラウド型音声認識サービスで使われている音声認識器の中身は、ブラックボックスである。そのため、音声認識率を高めるために前段で雑音抑圧を行う場合においても、音声認識器に最適な 1 個の雑音抑圧手法を事前に知ることは、実質不可能である。そこで、本研究では、雑音抑圧の手法を 1 個に固定するアプローチは取らなかった。その代わりとして、複数の雑音抑圧手法で生成した複数の音声データをクラウド型音声認識サービスに送信し、得られた複数の認識結果を統合する形を取った。

雑音抑圧モジュールの構成を図 4 に示す。中継サーバへ送られた圧縮済みの音声データは、伸長されたのち、複数の異なる雑音抑圧アルゴリズムによって処理される。それ

ぞれの雑音抑圧後の音声は、同時並行で、クラウド型音声認識サービスへ送信され、それぞれの音声認識結果が得られる。中継サーバでは、得られた複数の音声認識結果を統合し、採用する音声認識結果を決定し、車載機へ送信する。音声認識結果の統合方法の詳細は割愛するが、たとえば、それぞれの音声認識結果の信頼度の高い順から N-best を構成する方法などを取ることができる。

雑音抑圧のアルゴリズムには、双方向型 OMLSA(BOMLSA)[4] を用いた。異なる雑音抑圧音声を得るために、雑音抑圧の強度 (背景雑音の特徴量に乗算する定数) をさまざまに変えた雑音抑圧器を用意した。

BOMLSA は、通常の OMLSA[3] と比べて、高い雑音抑圧性能を持つ。一方、雑音特徴の推定に、時間逆方法のスペクトル推定を行う必要があるため、発話区間検出が終了した後でなければ雑音抑圧処理を開始することができない。そのため、仮に車載機側で BOMLSA を実装する場合、(a) 発話区間検出が完了するまで全音声データをメモリに保持する必要があるため、メモリ消費量を増加させる点、(b) 発話区間終了後に雑音抑圧を行うまでサーバへの音声送信を開始できないため、応答時間が長くなる点、といったデメリットが発生する。しかし、サーバに BOMLSA を実装する場合、(a) のメモリ消費量については、PC 上の実装になるため、問題にはならない。また、(b) の応答時間についても、車載機からの音声送信が発話区間検出の終了を待たずに開始でき、かつ発話区間検出終了後の処理 (雑音抑圧処理、認識処理) は、すべて有線のネットワークで接続された PC で行われるため、短時間で完了する。そのため、応答時間の増加を抑えたまま、高い雑音抑圧性能を実現できる。

3.2.2 性能評価

3.1.2 節の評価で用いた音声のうち、サンバイザーマイクの走行雑音を重畳した施設名発話音声を用い、提案した手法の効果を検証した。雑音抑圧および認識結果統合の手法は、以下の 4 条件とした。

(a) 通常の OMLSA / 4 パターンの雑音抑圧音声＋認識結果統合
(b) BOMLSA / 1 パターンの雑音抑圧音声 (認識結果統合なし)
(c) BOMLSA / 4 パターンの雑音抑圧音声＋認識結果統合
(d) 雑音抑圧なし

それぞれの条件において、あるクラウド型音声認識サービスを使用して音声認識実験を行った。ある音声に対する認識結果の 5-best 候補のなかに、正解が含まれなかった場合を認識誤りと定義した。さらに、各条件での認識誤りの数が、雑音抑圧なしの条件と比べてどの程度減少したか (誤り削減率) を算出した。

各条件における誤り削減率を図 5 に示す。(a)(b) の条件では、誤り削減率は約 35 %であった。これに対して、(c) の BOMLSA/結果統合の条件では、誤り削減率は 58 %となり、もっとも大きな誤り削減率を得た。このことより、提案する手法を用いて、走行雑音環境下における認識性能を大幅に改善できることを確認できた。

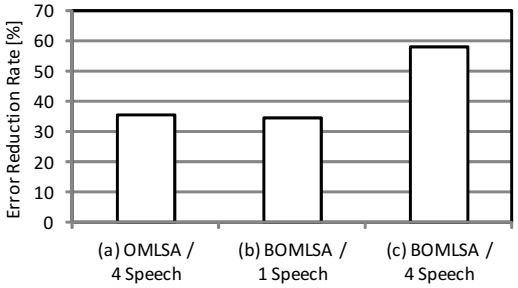

図 5 雑音抑圧および認識結果統合による認識誤り削減率

Fig. 5 Error reduction rate by noise suppression with integration of recognized result.

本研究では、1 個のクラウド型音声認識サービスでのみ効果を確認した。しかし、他のサービスを利用する場合であっても、各サービスに合わせて雑音抑圧強度や結果統合方法を調整することにより、音声認識の誤りを削減できると考えられる。これにより、ユーザに対して、高い品質の音声認識サービスを提供することができると考えられる。

4. まとめ

本研究では、クラウド型音声認識サービスを車載機で利用することを目的とし、走行雑音に対応するための発話区間モジュールおよび雑音抑圧モジュールの技術を検討した。開発したモジュールの性能評価をした結果、走行雑音がある環境において、高い性能を発揮できることを確認した。今後は、これらの技術を実際の製品に適用し、ユーザにとって利便性が高い音声 HMI を提供していく予定である。

謝辞 本研究を遂行するにあたり、ご議論、ご協力いただいたクラリオン (株) スマートアクセス開発部、システム企画部、および第一先行開発部の皆様に感謝いたします。

参考文献

[1] 小窪浩明, 天野明雄, 畑岡信夫: 車載用音声認識における騒音対策とその評価, 電子情報通信学会論文誌 D-II, Vol.J83-DII, No.11, pp.2190–2197 (2000)
[2] Obuchi, Y., Takeda, R. and Kanda, N.: *Voice Activity Detection Based on Augmented Statistical Noise Suppression*, Proc. APSIPA Annual Summit and Conference (2012)
[3] Cohen, I. and Berdugo, B.: *Speech Enhancement for Non-stationary Noise Environments*, Signal Processing, Vol.81, pp.2403–2418 (2001)
[4] Obuchi, Y., Takeda, R. and Togami, M.: *Noise Suppression Method for Preprocessor of Time-Lag Speech Recognition System Based on Bidirectional Optimally Modified Log Spectral Amplitude Estimation*, Acoustical Science and Technology, Vol.34, No.2, pp.133–141 (2013)