E-055

VADが音声認識性能に与える影響

Influence of Speech Activity Detection for Speech Recognition Performance

草水智浩† 山本一公† 北岡教英‡ 中川聖一†
Tomohiro Kusamizu Kazumasa Yamamoto Norihide Kitaoka Seiichi Nakagawa

1 はじめに

実環境下で音声認識を行う場合、全入力区間を認識対象とすると、湧き出し誤りが発生するために認識率が低くなる場合がある。その対処として、認識したい音声区間と雑音である非音声区間とを識別する手法がある。これを音声区間検出（VAD：Voice Activity Detection）と言い、実環境下の音声認識では必須である。また認識率向上のために耐雑音手法も併用するのが一般的である。本稿では、VADの精度が認識性能に与える影響について調査した結果について述べる。CENSREC-1-C[1]を使用して、VADが正解音声区間の「始端・終端が脱落した場合」、「前後を余分に検出した場合」を想定し、検出区間長を変えたデータを擬似的に作成して、認識実験を行った。また、無音モデルの有無、耐雑音手法の一つであるSS（Spectral Subtraction）を併用する場合についても同様に調査した。

2 音声区間検出と認識性能

現在、雑音に頑健な様々なVADの手法が研究されている。VADの手法にはパワー、ゼロ交差数[2]、MFCCなどを特徴量として用いたり、これらそれぞれの特徴量に重みをつけて統合することにより高い検出結果を得ようとする方法[3]などが試みられている。

一般に音声認識をターゲットにしたVADは、正解音声区間より前後を長めに検出する傾向がある。これは音声の始端と終端が切れ、音声情報が失われることにより認識率が大幅に低下するのを防ぐためである。しかし、これは経験的に行われていることで、実際の認識精度に与える影響は詳細には分かっていない。そこで本稿では、擬似的に20ms刻みで200msまで正解音声区間の両端を脱落、延長したデータを作成し、認識実験を行った。また、余分な区間は無音モデルにより無音として扱われることが期待される。そこで雑音モデルの有無による認識率の低下傾向の違いも調査した。

また雑音下でのVADでは耐雑音手法を併用するのが一般的である。こうした手法を用いると、雑音が音声認識率へ与える悪影響を抑圧することができるので、認識率の傾向はクリーンな音声に近づくはずである。そこで耐雑音手法(SS)の有無による認識率の傾向の違いも調査した。

3 実験データ

本稿では実験データとしてCENSREC-1-C[1]を使用している。これは雑音加算によるシミュレーションデータと、実際の雑音環境下で収録された実環境データの2種類からなる。シミュレーションデータと実環境データの認識率の傾向が同じであれば、シミュレーションデータの正当性が確認できる。音声データは連続数字を間隔をあけて発声したものからなり、個々の発話内容は、CENSREC-1(AURORA-2J)[4]に準じている。

シミュレーションデータのSNRは−5〜20dB(5dB刻み)及びクリーン環境であり、雑音は8種類である。実環境データは2種類のSNR環境(High,Low)、2種類の雑音からなる。認識にはクリーン音声で学習したクリーンHMM, 雑音重畳音声で学習したマルチHMMの2種類を使用した。

4 実験

4.1 実験条件

本稿では、HTK（HMM Tool Kit）[5]を用いてCENSREC-1ベースライン[4]と基本的に同じ実験条件で、検出区間毎に認識を行う。ベースラインより変更した部分は、特徴パラメータの次元数を39次元から38次元(powerを使用しない)にしている点である。表1に実験条件を示す。

表1 実験条件

サンプリング周波数	8kHz
プリエンファシス	$1-0.97z^{-1}$
分析窓/窓長	ハミング窓/25ms
フレームシフト	10ms
メルフィルタバンク数	23
特徴パラメータ	12MFCC+12ΔMFCC +12ΔΔMFCC+Δ*log*power +ΔΔ*log*power　計38次元
HMM状態数	16状態(無音モデルは1状態)
混合数	20混合(無音モデルは36混合)

また今回使用したSSは文献[6]の手法であり、サブトラクション係数$\alpha=1.0$、窓長32ms、フレームシフト6.25ms、雑音推定フレームは音声ファイルの最初の30フレームである。

4.2 実験結果

図1から6に実験結果を示す。ここで"sp"は無音モデルのことであり、"baseline"は、クリーンHMM・sp有り・SS無しの条件での認識結果である。各図はこのベースラインの条件より、使用HMM・sp・SSの各条件だけを変えた場合の実験結果である。図の横軸は正解音声区間の前後をどれだけ脱落・延長させたかを表しており、例えば−200、200の点はそれぞれ200ms脱落、延長した場合を示している。図中の認識率は各SNRの全ての種類の雑音の平均値である。またシミュレーションデータはクリーン環境と10dB、0dBの結果のみプロットしている。

実験結果全体から、正解音声区間の前後を脱落・延長したどちらの場合も、基本的に認識率は低下するが、脱落した場合のほうが低下率が大きいことが分かり、経験則が正しいということが確認できる。クリーンデータに対しては、認識率のピークは+40ms前後であり、正解音声区間の前後を多少延長した方が認識率は良くなっている。また、SNRが悪くなるほどピークは0

† 豊橋技術科学大学 工学部情報工学系
‡ 名古屋大学大学院情報科学研究科メディア科学専攻

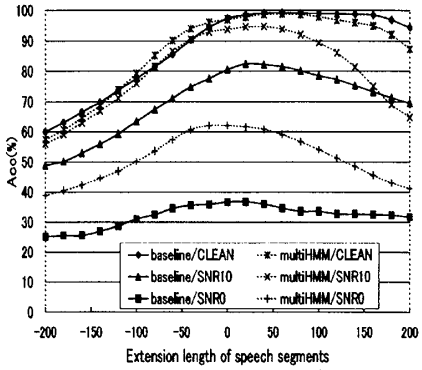

図 1 シミュレーションデータ・HMM の比較 (sp 有り・SS 無し)

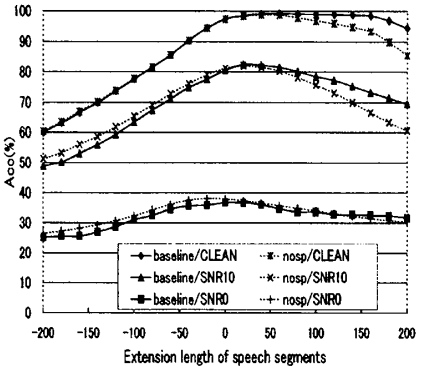

図 2 シミュレーションデータ・sp 有無の比較 (クリーン HMM・SS 無し)

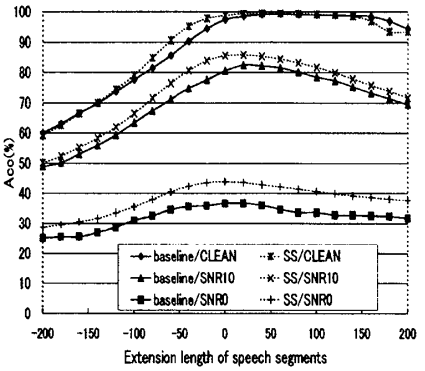

図 3 シミュレーションデータ・SS 有無の比較 (クリーン HMM・sp 有り)

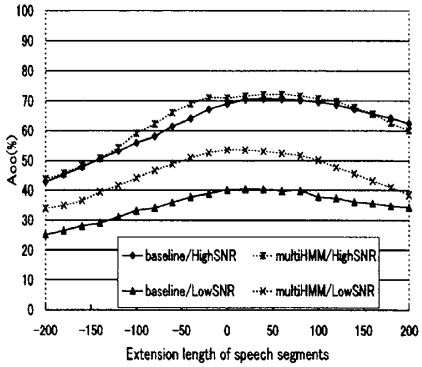

図 4 実環境データ・HMM の比較 (sp 有り・SS 無し)

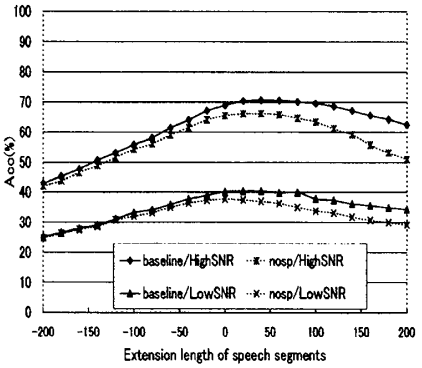

図 5 実環境データ・sp 有無の比較 (クリーン HMM・SS 無し)

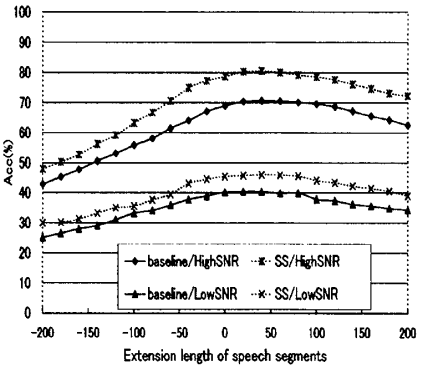

図 6 実環境データ・SS 有無の比較 (クリーン HMM・sp 有り)

の点 (正解音声区間通りに検出された場合) に近づき、正解音声区間との誤差が敏感に認識率の低下に繋がっている。このことから SNR が悪い程 VAD の精度が高くなければならないことが分かる。またシミュレーションデータと実環境データの認識率の傾向が同じであることから、シミュレーションデータの正当性が確認できる。よって、実環境データを大量に使用することが困難な場合には、代わりにシミュレーションデータを使用しても手法の比較評価ができることが分かる。

図 1 と図 4 からは、使用 HMM の違いによって認識率の低下の傾向が変わらないことが分かる。

図 2 と図 5 からは、sp を用いないと、前後を余分に検出した場合に用いた場合より認識率が下がりやすく、sp が雑音を吸収していることが確認できる。しかし、SNR が悪くなると認識率の傾向に差が無くなっており、sp が雑音を吸収しきれていないことが分かる。

図 3 と図 6 からは、SS を行うと認識率が上がってピークが +40 に近づき、クリーン音声の認識率の傾向に近づくことが分かる。

本稿の認識単位は 1 数字であり、1 数字の HMM 状態数はすべて 16 状態で統一しているので、認識結果を出力するには最低 16 フレーム (175ms) 必要である。ここで例えば認識単位が音節であったとすると、1 音節の HMM の状態数が 3 状態であれば、認識結果を出力するのに最低 3 フレーム (45ms) と本稿より短くなる。つまり、正解音声区間の前後が延長した場合に、延長した区間で本稿より挿入エラーが起きやすくなり、認識率が低下することが考えられる。認識単位が違う場合には注意が必要である。

5 まとめ

本稿では、VAD の精度が認識性能に与える影響について調査した。実験の結果、VAD で正解音声区間の前後を余分に検出するよりも、脱落して検出した場合の方が認識率が低下しやすいことが確認でき、SNR が悪くなるほど VAD に高い精度が求められることが分かった。また、sp は前後を余分に検出した場合に、ある程度雑音を吸収する役割をしていること、SS を使用するとクリーン音声に認識率低下の傾向が近づくことも確認できた。

今後は、SNR が悪くなると正解音声区間の前後を余分に検出した場合に sp が雑音を吸収しきれていないことから、sp を雑音環境で適応化することで認識率の低下を防ぐことについて検討する。

参考文献

[1] 北岡ら, "CENSREC-1-C： 雑音下音声区間検出手法評価基盤の構築," 情報研報, 2006-SLP-63-1, pp.1–6, 2006.

[2] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," ITU-T/Recommendation G.729-Annex B. 1996.

[3] 本田, 河原, "複数特徴の重み付き統合による雑音に頑健な発話区間検出," 信学論誌, Vol.J89-D, No.8, pp.1820–1828, 2005.

[4] S. Nakamura et al., "AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition," IEICE Trans. Inf. Syst., Vol.E88-D, No.3, pp.535–544, 2005.

[5] http://htk.eng.cam.ac.uk/, HTK Speech Recognition Toolkit.

[6] 北岡, 赤堀, 中川, "スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識," 信学論文誌, Vol.J83-D-II, pp.500–508, 2000.