論文 / 著書情報
Article / Book Information

論題(和文)	VADの信頼度を利用した雑音に頑健な音声認識デコーダの検討
Title(English)	
著者(和文)	大西 翼, ディクソン ポール, 岩野 公司, 古井 貞熙
Authors(English)	Oonishi Tasuku, Dixon Paul, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2009年秋季講演論文集, , No. 1-1-16, pp. 49-50
Citation(English)	, , No. 1-1-16, pp. 49-50
発行日 / Pub. date	2009, 9

VADの信頼度を利用した雑音に頑健な音声認識デコーダの検討*

大西翼, ○ディクソン・ポール (東工大), 岩野公司 (東京都市大), 古井貞熙 (東工大)

1 はじめに

実環境で音声認識システムを利用する場合，システムにはユーザの発話の他にユーザが発話を行っていない区間の音 (非音声) が入力される．認識システムに長い非音声区間が入力されると無駄な計算コストが発生し，また本来認識したいユーザ発話の認識率を低下させる原因となる．そのため，ユーザの発話が行われている区間 (音声区間) と発話が行われていない区間 (非音声区間) を正確に判定する Voice Activity Detection(VAD) の技術が非常に重要となる．過去にはフロントエンドで VAD を行う手法として，零交差回数を利用した手法 [1] や GMM の尤度を利用した手法 [2] が提案されている．また一度音声認識を行うことで入力区間のアライメントを行い，その区間に対して GMM を用いた VAD を行うことで，頑健性を高める手法 [3] が提案されている．これらの手法ではシステムに入力された区間に対して，音声/非音声を判定するためのスコア（音声/非音声判定スコア）を計算し，そのスコアがある閾値を越えていたら，該当区間を音声と判定し，後段の認識処理を行う．またスコアが閾値を越えていなければ，非音声と判定し該当区間を棄却する．雑音などの影響により VAD の精度が低下する場合，入力された音声を厳密に音声，非音声の二値に判定することが難しくなる．そのためフロントエンドで誤って音声区間を棄却する可能性が増加する．

そこで本研究では音声/非音声判定スコアをフロントエンドの入力区間の棄却に利用するのではなく，スコアを探索時の仮説評価に利用することで，フロントエンドにおける音声区間の誤棄却の影響を低減させる音声認識手法について検討する．文献 [4] ではこのアプローチに基づく音声認識手法について提案を行っておりその有効性を確認している．本研究では，文献 [4] とは異なる仮説評価式と VAD の特徴量を採用した認識手法を提案し，その対雑音性をカーナビゲーションの音声操作タスクによって評価する．提案手法を従来のフロントエンドで VAD を行う認識手法と比較しその有効性を示す．

2 フロントエンド方式 VAD

本論文ではフロントエンドに零交差とパワーを用いた VAD [1] と GMM の尤度比を用いた VAD [2] を利用した認識手法と，提案手法とを比較する．以下ではそれぞれの VAD についての概要を述べる．

2.1 零交差とパワーに基づく VAD

零交差とパワーを用いた VAD 手法ではある区間に対して零の値を交差した回数 (零交差回数) を計算し，その回数がある閾値を越えた区間に対して，さらに平均エネルギーの値がある閾値を越えた場合に音声区間と判定し，それ外の場合を非音声区間と判定する．この手法は実装が容易で計算コストが小さく，SNR が大きいときには高い判定精度を得られるという利点がある．

2.2 GMM の尤度比に基づく VAD

音声と非音声の音響的な特徴を利用して VAD を行う手法として，GMM の尤度比を利用した手法が提案されている [2]．この手法では音声と非音声の音響的な特徴をあらかじめ GMM によりモデル化し，それらを用いて i 番目のフレームにおける音声/非音声のスコア L_{GMM} を以下の式により算出する．

$$L_{GMM} = \log \frac{p(X^i|H_1)}{p(X^i|H_0)} \tag{1}$$

X^i は，i フレーム目における観測ベクトルを表し，H_1, H_0 は，それぞれ音声，非音声であることを表す．$p(X^i|H_1), p(X^i|H_0)$ は，音声及び非音声を用いて学習された GMM により算出される．(1) 式の対数尤度比が閾値以上であれば音声と判定し，それ以下であれば，非音声と判定する．この手法では，SNR が小さい状況でも音声と非音声を精度よく判定できるという利点があるが，実際にシステムが利用される環境にモデルが適応していない場合には，精度が低下するという欠点がある．

3 提案手法

本章では，音声/非音声の信頼度を仮説の尤度の評価に利用する認識手法の詳細について説明する．

3.1 提案手法における音響尤度の評価式

提案手法では，あるフレームにおける仮説の音響尤度を以下の式により算出する．この時，探索している仮説が単語（有音）であれば (2) 式，無音であれば，(4) 式を用いて算出する．

$$\log \hat{p}_{am}(X^i|\theta) = \log p_{am}(X^i|\theta) + \alpha \log \bar{C}^i_{H_1} \tag{2}$$

$$\bar{C}^i_{H_1} = \frac{\sum_{i-n}^{i+n} C^i_{H_1}}{2n+1} \tag{3}$$

$$\log \hat{p}_{am}(X^i|\theta) = \log p_{am}(X^i|\theta) + \alpha \log \bar{C}^i_{H_0} \tag{4}$$

$$\bar{C}^i_{H_0} = \frac{\sum_{i-n}^{i+n} C^i_{H_0}}{2n+1} \tag{5}$$

上式，X^i が，i フレーム目の観測ベクトル，θ が仮説，$p_{am}(X^i|\theta)$ が音響モデルから算出される尤度，α がスケーリング係数，n が前後フレームの平滑化数である．なお，$C^i_{H_1}, C^i_{H_0}$ は，それぞれ i フレームにおける 0 ～ 1 の範囲で正規化された音声，非音声の信頼度であり，下記の式により計算される．

$$C^i_{H_1} = \frac{p(X^i|H_1)}{p(X^i|H_1) + p(X^i|H_0)} \tag{6}$$

$$C^i_{H_0} = \frac{p(X^i|H_0)}{p(X^i|H_1) + p(X^i|H_0)} \tag{7}$$

上式の $p(X^i|H_1), p(X^i|H_0)$ は音声及び非音声を用いて学習された GMM により算出される．(2),(4) 式に

*Robust Speech Recognition Using VAD-measure-embedded Decoder. by Tasuku Oonishi, Paul R. Dixon(Tokyo Institute of Technology), Koji Iwano(Tokyo City university), Sadaoki Furui(Tokyo Institute of Technology)

おいて，α を 0 とすれば VAD を行っていない場合と同様の音響尤度を仮説に与えることになる．また，$C^i_{H_1}$(または，$C^i_{H_0}$) が 1 となる区間では，有音 (無音) の仮説には 0，無音 (有音) の仮説には $-\infty$ のスコアが追加されることになるため，有音 (無音) の単語のみ認識されることになる．音声区間では単語，非音声区間では無音が正解となるため，$C^i_{H_1}$,$C^i_{H_0}$ のスコアが高精度に算出されている場合には，正解仮説に対して VAD を行っていない場合とほぼ同じ音響尤度を与える．

3.2 フロントエンド手法との組み合わせ

提案方式では，入力する全ての区間を音声認識するため，計算コストが増加するという欠点がある．そこで本手法の前段処理として，GMM の尤度比を利用したフロントエンドによる VAD 手法を利用する．この時，音声/非音声の判定スコアが十分に小さいとき，つまり，高い可能性で非音声区間を判定できるときのみ，フロントエンドで事前に区間を棄却する．これにより認識精度の劣化なく，認識処理に関する計算量を削減できることが期待できる．

4 実験

評価用データに，Drivers Japanese Speech Corpus in a Car Environment (DJSC) の高速道路走行におけるハンズフリーコマンド発話を用いた．これは音声認識によるカーナビゲーションの利用を想定し作成されたコーパスで，自動車走行中にカーナビゲーションを音声で操作するために発声されたコマンド発話を収録している．評価データに用いた音声の話者数は 40 人 (男性・女性各 20 人) で各話者は 41 個のコマンドを連続して発話している．各発話の前後には，1〜2 秒程度の無音区間が存在する．学習用データには，音響モデルに JNAS の男性 130 人 (25 時間)，女性 130 人 (27 時間) を用いた．音声 GMM の学習には，CSJ の 967 学会講演，非音声 GMM の学習には電子協騒音データベースの走行車雑音を用いた．音響特徴量には，フレームシフト 10ms, 分析窓幅 25ms の MFCC 12 次元+ΔMFCC 12 次元+ΔΔMFCC 12 次元+Δ 対数パワー+ΔΔ 対数パワーの計 38 次元を用いた．音響モデルには，2000 状態 16 混合のトライフォン HMM を用いた．数個程度の単語から構成されるコマンドを連続して受け付けるネットワーク文法を用いた．コマンドに用いられている語彙数は 83 であった．音声，非音声 GMM の混合数は 4 とした．デコーダは，東京工業大学で開発を行っている T^3 Decoder を使用した．

Fig.1 に各手法を用いて VAD を行った場合の認識精度を示す．図の baseline が VAD を行っていない場合，ZCR が零交差法及びパワーを用いて VAD を行った場合，GMM が GMM の尤度比を利用して VAD を行った場合，proposed が提案手法により VAD を行った場合，manual が，人手により VAD を行った場合を表す．なお，各手法のパラメータは，テストセット全体に対して最適な値を人手により設定している．各手法におけるパラメータ値は零交差法では，パワーの閾値が 0, 零交差回数が 10, ウインドウ幅が 25ms,GMM 法では，音声，非音声モデルの対数尤度比の閾値が-5, 提案手法では α が 10,n が 15 となった．図から VAD を行わない場合の単語正解精度は 43.1%, 零交差は 46.5%,GMM は 45.8%であり，VAD を行わない場合と比べて，それぞれ 3.4%,2.7%の認識精度の改善が見られることがわかる．一方，提案手法の単語正解精度は 53.1%と VAD を行わない場合と比べて，10.0%の精度の改善が見られ，従来手法である零交差法，GMM 法と比べて，大きな改善が見られる．また，人手で VAD を行った場合には 60.4%の認識精度が得られることから，本手法にはさらなる改善の余地があることがわかる．

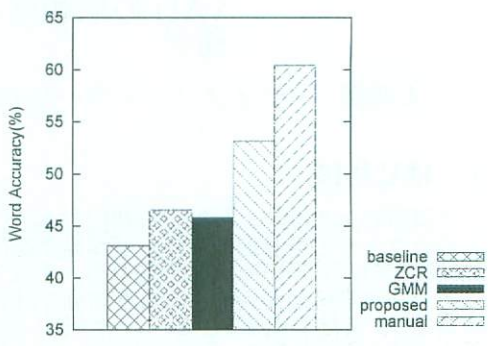

Fig. 1 *VAD* の効果

さらに，提案手法の前段処理として GMM を用いたフロントエンドによる VAD 手法と組み合わせた場合，認識精度の劣化なく最大で 60%程度の入力区間を棄却することができた．本実験で用いたテストデータでは，無音区間が全体の 70%程度を占めていることから，ほとんどの無音区間が除去されていることがわかる．

5 まとめと今後の検討

本論文では，音声/非音声の信頼度により仮説の尤度を調整する認識手法を提案した．本手法を用いることにより，VAD をフロントエンドで用いる手法と比べて，大きな認識精度の改善が得られることが確認された．さらに，フロントエンドに VAD を組み合わせることで，更なる認識精度の改善と計算コストの削減が両立できることが確認された．今後の課題として，関連手法との比較，様々な環境における有効性の評価及びパラメータの安定性の評価，音声/非音声信頼度の頑健な推定方法の検討があげられる．

謝辞 本研究は経産省「情報家電センサー・ヒューマンインターフェースデバイス活用技術開発・音声認識基盤技術」プロジェクトの支援により行った．

参考文献

[1] A. Benyassine et al. "ITU-T recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," IEEE Communications Magazine 35 (9): pp.64-73, 1997.

[2] R. Singh et al. "Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination," Proc. ICASSP, vol 1, pp.273-276, 2001.

[3] 酒井他, "実環境ハンズフリー音声認識のための音響モデルと言語モデルに基づく音声区間検出と認識アルゴリズム, " 信学技報, pp.13-18, 2007.

[4] 草水他, "音声認識における VAD の信頼度のデコーダへの組み込み, " 音学春季講論, pp.127-130, 2009.

[5] J. Ramírez et al. "Voice activity detection: Fundamentals and speech recognition system robustness," Robust Speech Recognition and Understanding, I-Tech Education and Publishing, pp.1-22, 2007.