論文 / 著書情報
Article / Book Information

論題(和文)	VADの信頼度を利用した音声認識デコーダの高精度化
Title(English)	
著者(和文)	大西 翼, 岩野 公司, 古井 貞熙
Authors(English)	Oonishi Tasuku, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2010年秋季講演論文集, , No. 2-9-3, pp. 39-42
Citation(English)	, , No. 2-9-3, pp. 39-42
発行日 / Pub. date	2010, 9

VADの信頼度を利用した音声認識デコーダの高精度化*

大西翼 (東工大), 岩野公司 (東京都市大), ○古井貞熙 (東工大)

1 はじめに

音声・非音声を判定する Voice Activity Detection(VAD) は，実環境下で頑健な音声認識を行う上で基盤となる技術である．この VAD の実装方式として，フロントエンドで非音声区間の棄却を行う方式が一般的である．この方式の問題点として，「確定的に音声・非音声を決定する点」と「音声・非音声の判定情報を認識に利用しない点」がある．高雑音環境下では，音声・非音声を判定するためのスコアを正確に算出することが難しくなるため，音声区間を棄却するといった判定誤りが増加する．一度棄却された音声区間は，音声認識デコーダで復元することができない．そのため音声区間の誤棄却が増えると大きな認識精度低下の原因となる．また，音声と判定されたフレームであっても，音響モデルと入力環境との乖離が大きい場合，後段の認識処理で誤って無音として認識する場合がある．そのため VAD の精度向上と認識精度の向上が，必ずしも結び付かない．

この様な問題を解決するため，過去に VAD の信頼度を仮説のスコア調整に利用する手法の検討を行った [1]．この手法は，非音声を表す仮説（文頭・文末の無音などを表す仮説）のスコアには非音声の信頼度を加え，逆に音声を表す仮説のスコアには音声の信頼度を加える処理を行う．この手法では，入力された区間を全て認識するため，音声フレームの誤棄却の問題を軽減することができる．また，音声 (非音声) の信頼度が高いフレームでは音声 (非音声) の仮説スコアが高くなる．このため音声と判定された区間をデコーダが誤って無音と認識する誤りを軽減することができる．

Drivers' Japanese Speech Corpus in a Car Environment(DJSC) [2] を用いた認識実験では，フロントエンドで VAD を行う認識手法と比べて，大幅な認識精度の改善が得られた．しかし，人手により付けられた情報を基に VAD を行った場合と比べて，認識精度の差があることが分かり，本手法のさらなる改善が必要であることが分かった．

音声・非音声が完璧に判定できた場合と比べて認識精度が低下した原因として，信頼度の計算に用いる GMM が入力環境の音響的特徴を十分に表現していないことが挙げられる．従って，GMM の音響適応を行うことで認識精度の改善が期待できる．しかし，リアルタイム性が求められるシステムでは，認識前にデータを収集し，それを適応するといったアプローチを取ることができない．そのため認識時に動的に GMM を学習するオンライン適応が必要となる．

そこで本論文では，音声・非音声の信頼度を探索に利用する手法に，GMM のオンライン教師なし適応を併用する効果について述べる．GMM のパラメータの推定手法として，Zhang らが行った MAP 推定に基づくオンライン適応手法 [3] を利用する．さらに本論文では，GMM の推定パラメータを学習データの信頼度により重み付けすることで頑健な教師なし適応を行う手法，GMM の適応回数を削減することで適応に関する処理を高速化する手法を提案する．

2 音声・非音声の信頼度を利用した音声認識

本節では，音声・非音声の信頼度を利用した音声認識手法 [4] について述べる．この手法では，あるフレームにおける仮説の音響尤度 $\hat{p}_{am}$ を以下の式により定義する．この時，探索仮説が非音声区間を表している場合には式 (1) を用い，音声区間を表している場合には式 (2) を用いて音響尤度を定義する．

$$\log \hat{p}_{am}(X_i|\theta_{uv}) = \log p_{am}(X_i|\theta_{uv}) + \alpha \log \bar{C}^i_{H_0} \quad (1)$$

$$\bar{C}^i_{H_0} = \frac{\sum_{k=i-l}^{i+l} p(X_k|H_0)}{\sum_{k=i-l}^{i+l}\{p(X_k|H_0)+p(X_k|H_1)\}}$$

$$\log \hat{p}_{am}(X_i|\theta_{v}) = \log p_{am}(X_i|\theta_{v}) + \alpha \log \bar{C}^i_{H_1} \quad (2)$$

$$\bar{C}^i_{H_1} = \frac{\sum_{k=i-l}^{i+l} p(X_k|H_1)}{\sum_{k=i-l}^{i+l}\{p(X_k|H_0)+p(X_k|H_1)\}}$$

上式 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．θ_{uv}, θ_v はそれぞれ非音声・音声を表す認識仮説，$p_{am}(X_i|\theta_{uv}), p_{am}(X_i|\theta_v)$ は音響モデルから算出される尤度，α はスケーリングファクタ，l は前後フレームの平滑化数である．$\bar{C}^i_{H_1}, \bar{C}^i_{H_0}$ は，それぞれ 0 ～ 1 の範囲で正規化された音声，非音声の信頼度であり，$i-l$ から $i+l$ フレームで平滑化された尤度により計算される．$p(X|H_1), p(X|H_0)$ は音声及び非音声を表す GMM により算出される．

式 (1)，(2) において，スケーリングファクタ α を 0 とすれば，信頼度による音響尤度の調整を行っていない場合と同じ尤度を仮説に与えることになる．また $\bar{C}^i_{H_1}$（または $\bar{C}^i_{H_0}$）が 1 となる区間では，音声（または非音声）を表す仮説には音響尤度の調整を行っていない場合と同じ尤度を与え，それ以外の仮説には $-\infty$ のスコアを与える．これは正しい信頼度が与えられた場合，非音声区間では非音声を表す仮説が，音声区間では音声を表す仮説のみが出力されることを意味する．これから，音声 (非音声) と判定された区間で，非音声 (音声) を表す仮説を認識する誤りが軽減できることが分かる．

3 GMM のオンライン教師なし適応

3.1 GMM のオンライン適応手法

ある D 次元のフレーム（特徴ベクトル）X に対する GMM の尤度は以下の式により算出される．

$$p(X|\lambda) = \sum_{m=1}^{M} w_m p_m(X) \quad (3)$$

ここで，λ は GMM のパラメータ，M は混合数，$p_m(X)$ は，m 番目のガウシアンの確率密度関数であ

*Improving Robustness of VAD-measure-embedded Decoder. by Tasuku Oonishi(Tokyo Institute of Technology), Koji Iwano(Tokyo City University), Sadaoki Furui(Tokyo Institute of Technology)

る．w_m はガウシアンの混合重みであり，$\sum_m w_m = 1$ となる．確率密度関数は以下の D 次元の正規分布により定義される．

$$p_m(X) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_m|^{\frac{1}{2}}} \cdot \exp\{-\frac{(X-\mu_m)^\top \Sigma_m^{-1}(X-\mu_m)}{2}\} \quad (4)$$

ここで μ_m は D 次元の平均ベクトル，Σ_m は $D \times D$ の分散共分散行列である．これから GMM のパラメータは $\lambda : \{\mu_m, \Sigma_m, w_m\}(m = 1, 2, \cdots, M)$ となる．

Reynolds らは適応的な GMM の学習手法として MAP 推定に基づく推定手法 [5] を提案している．この手法では適応データ $X_1, X_2 \cdots X_N$ が与えられた時の GMM のパラメータの推定を以下の二つのステップにより行う．

Step1: 観測ベクトル系列 $X_1, X_2, \cdots, X_N$ から n_m, $E_m(X)$, $E_m(X^2)$ を計算

$$Pr(m|X_i) = \frac{w_m p_m(X_i)}{\sum_{k=1}^{M} w_k p_k(X_i)} \quad (5)$$

$$n_m = \sum_{i=1}^{N} Pr(m|X_i) \quad (6)$$

$$E_m(X) = \frac{1}{n_m}\sum_{i=1}^{N} Pr(m|X_i)X_i \quad (7)$$

$$E_m(X^2) = \frac{1}{n_m}\sum_{i=1}^{N} X_i^\top Pr(m|X_i)X_i \quad (8)$$

$Pr(m|X_i)$ はフレーム X_i の尤度が，m 番目のガウシアンにより出力された確率を表す．この時 $\sum_m Pr(m|X) = 1$ となる．

Step2: GMM のパラメータの更新

$$\hat{w}_m = [\beta_m n_m/N + (1-\beta_m)w_m]\rho \quad (9)$$

$$\hat{\mu}_m = \beta_m E_i(X) + (1-\beta_m)\mu_m \quad (10)$$

$$\hat{\Sigma}_m = \beta_m E_m(X^2) + (1-\beta_m)(\Sigma_m + \mu_m\mu_m^\top) - \hat{\mu_m}\hat{\mu}_m^\top \quad (11)$$

上式 $\hat{w}_m, \hat{\mu}_m, \hat{\Sigma}_m$ は，パラメータ更新後の m 番目のガウシアンにおける混合重み，平均ベクトル，分散共分散行列を表す．ρ は，$\sum_m \hat{w}_m = 1$ とするための正規化パラメータである．また，w_m, μ_m, Σ_m は，パラメータ更新前の m 番目のガウシアンにおける混合重み，平均ベクトル，分散共分散行列を表す．更新後のパラメータは，Step1 で推定された統計量と更新前のパラメータの重み付き和により決定される．この重み β_m は以下の式により決定される．

$$\beta_m = \frac{n_m}{n_m + \gamma} \quad (12)$$

γ は GMM のパラメータの更新度合いを制御するパラメータである．$\gamma = 0$ の場合は適応データのみから GMM のパラメータを推定する場合に相当し，$\gamma = \infty$ の場合はパラメータの更新を行わない場合に相当する．

オンライン適応では，$i+1$ フレーム目において，過去のフレーム $X_1, X_2 \cdots X_i$ を用いて GMM のパラメータを更新する．そして更新された GMM を用いて，$i+1$ フレーム目の音声・非音声の信頼度を計算する．さらに $i+2$ フレーム目では，$X_1, X_2 \cdots X_{i+1}$ の適応データから $i+1$ フレーム目で適応された GMM を用いて統計量を計算し，再度パラメータの更新を行う．この処理をフレームの終端まで繰り返す．

3.2 教師なし適応手法

本手法では音声，非音声を表す GMM のパラメータを更新する必要がある．教師なし適応でこれらのパラメータを精度よく更新するため，Zhang らの手法 [3] と同様に信頼度による適応データの選択を行う．この手法では，i 番目のフレームの音声の信頼度 $\bar{C}^i_{H_1}$ が，事前に設定された閾値 τ を越えた場合に，それを音声 GMM の適応データとして利用する．また同様に，非音声の信頼度 $\bar{C}^i_{H_0}$ が閾値 τ を越えた場合に，それを非音声 GMM の適応データとして利用する．音声，非音声の信頼度が，共に閾値 τ を越えていない場合には，該当フレームを適応データとして利用せず棄却する．逆に両方とも閾値を越えている場合には，両方の適応データとして利用する．

閾値パラメータ τ は，認識精度に影響を与えるパラメータであるため，最適な値を用いることが望ましい．しかし，最適値は入力環境の雑音，GMM のパラメータなど様々な要因により変動するため，事前に最適なパラメータの値を予測することは難しい．そこで，認識精度に対する τ パラメータの変動に対する頑健性を向上させることが求められる．

そこで我々は，GMM の更新パラメータを適応データの信頼度により重み付けする手法を提案する．閾値付近の信頼度を持つデータと，閾値を大きく上回る信頼度を持つデータでは，後者のデータを用いて推定を行った方が，より正しい推定となることが予想される．そのため，大きな信頼度を持つデータから推定されたパラメータにはより大きな重みを付け，そうでないパラメータには低い重みを付けるように推定を行うことで，τ の変動に対する頑健性を向上させることができると考えられる．このような推定を実現するために，m 番目のガウシアンにより尤度が出力された確率 $Pr(m|X_i)$(式 (5)) に，そのフレームの信頼度を掛け合わせる．具体的には，$Pr(m|X_i)$ の代わりに，以下の $\hat{Pr}(m|X_i)$ を用いて GMM のパラメータを推定する．

$$\hat{Pr}(m|X_i) = C^i_{H_j} \times Pr(m|X_i), \quad j \in 0 \text{ or } 1 \quad (13)$$

非音声を表す GMM のパラメータを推定する場合は $C^i_{H_0}$ の信頼度を利用し，音声を表す GMM のパラメータを推定する場合は $C^i_{H_1}$ の信頼度を利用して式 (13) を計算する．このように信頼度を掛け合わせることで，信頼度に比例してパラメータの更新度合いを調整することができる．例えば，全ての適応データの信頼度が 0 となった場合，式 (6) の n_m は 0 となる．この時，モデルの更新パラメータ β_m は 0 となるため，GMM の更新後のパラメータは更新前と同じになる．

3.3 学習の高速化

3.1 節で述べた GMM のオンライン適応手法は，フレーム毎に全ての学習データの統計量を再計算するため，計算量が非常に多くなり，実用的ではない．そこで，GMM の適応を N フレーム毎に実行することで適応回数を削減し，高速化を行う．N の値が大きくなるほど，適応に関する計算時間が削減できるが，入力環境に適応することが遅くなるため，適応の効果が低下する．$N = \infty$ とすると，GMM の適応を行わなかった場合と一致する．

4 評価実験

評価用データに，Drivers' Japanese Speech Corpus in a Car Environment (DJSC) [2] 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．S/N 比は–8〜0dB の高雑音環境 [2] である．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．各コマンド発話の前後には，1〜2 秒程度の非音声区間が存在する．

学習用データには，音響モデルに JNAS [6] の男性 130 人 (25 時間)，女性 130 人 (27 時間) 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．言語モデルには，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 の学習には CSJ [7] の 967 学会講演を用いた．GMM の混合数は 4 とした．GMM のオンライン教師なし適応は話者毎に独立して実行した．各話者の発話は非音声区間を含めて 150〜200 秒程度であった．

デコーダは，東京工業大学で開発を行っている T^3 Decoder [8] を使用した．実験は，Interl Core 2 Quad 3.0GHz 4GB メモリの計算機を使用した．なお各種のパラメータの値は，$\alpha = 3$, $l = 15$, GMM のオンライン教師なし適応のパラメータは，$\gamma = 50$, $\tau = 0.7$, $N = 10$ とした．

4.1 提案手法における認識精度の評価

Fig. 1 に GMM の適応を行う前と後の単語正解精度を示す．図のグラフと実験条件の関係は以下の通りである．

baseline: VAD を行っていない場合

no adapt: VAD の信頼度を仮説スコアの調整に利用する手法を用いた場合 (GMM の適応なし)

adapt: VAD の信頼度を仮説スコアの調整に利用する手法を用いた場合 (GMM の適応あり)

manual: 人手で付けられた情報を基に音声・非音声の判別を行った場合（上限値）

GMM の適応を行わない場合の単語正解精度は，42.9% であり，VAD を行わない場合の単語正解精度 (43.1%) より低い単語正解精度となった．これは音声・非音声 GMM と入力環境との乖離が大きいため，信頼度が正しく算出されなかったためである．一方，提案手法と GMM のオンライン教師なし適応を併用した場合は 59.9% と大幅な単語正解精度の改善が得られた．これから，音声・非音声 GMM が入力環境の音響的特徴を適切に表現していない場合でも，GMM のオンライン教師なし適応を併用することで，上限値 (60.4%) に近い単語正解精度が得られることが分かった．以上の結果から，本論文で提案した音声・非音声の信頼度を利用する適応的な音声認識手法の有効性が確認できた．

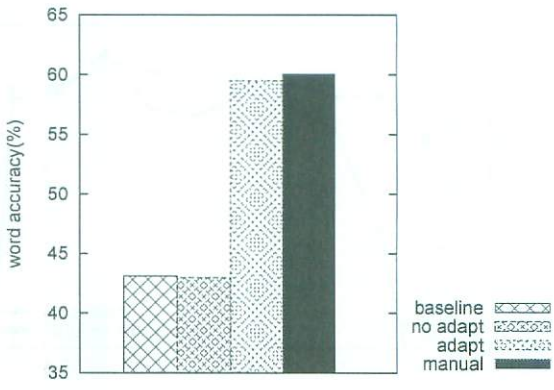

Fig. 1 GMM のオンライン教師なし適応の効果

4.2 閾値パラメータの頑健性に対する評価

GMM のオンライン適応において，適応データの信頼度により，GMM の更新パラメータを重み付けした場合の効果を Fig. 2 に示す．図の縦軸が単語正解精度，横軸が閾値パラメータ τ である．図の“threshold”が信頼度の閾値による適応データの選択を行った場合，“threshold + weight”が，さらに適応データの信頼度により，更新パラメータを重み付けした場合を表す．図の“threshold”から，τ が変化するに従い，単語正解精度が大きく変動していることが分かる．特に τ として小さい値を用いた場合，大きく認識精度が低下することが分かる．一方，学習データの信頼度を用いて更新パラメータを重み付けすることで，τ の変動に対する単語正解精度の変動が小さくなっていることが分かる．これから，3.2 節で述べた重み付けを行うことで，τ パラメータの変動に対する頑健性を向上できることが確認できた．

4.3 適応の高速化の効果

Fig. 3 に適応の高速化を行った場合の効果を示す．図の左縦軸が単語正解精度，右縦軸が実時間比 (RTF)，横軸が適応を行うフレーム間隔 N である．図から，N が 1 から 50 までは，大きな認識精度の低下は発生せず，N が 50 以上では，大きく認識精度が低下することが分かる．これは N が大きくなるほど，パラメータの更新回数が低下するため，入力環境への適応が遅れるためである．また，図の実時間比の結果から，N が 10 までで急速に認識時間が削減されていることが分かる．さらに N が 50 以上で，認識時間が収束していることから，GMM の音響適応に関する計算コストのほとんどが削減されていることが分かる．以

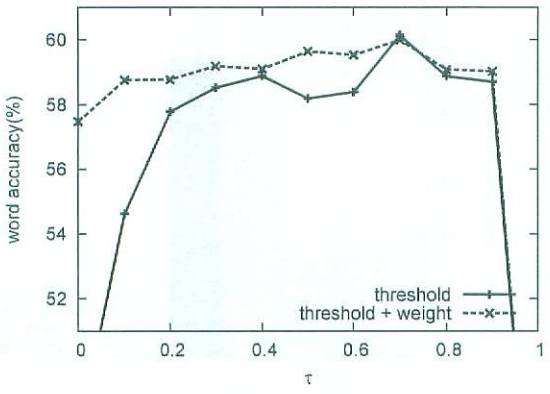

Fig. 2　閾値パラメータ τ に対する頑健性

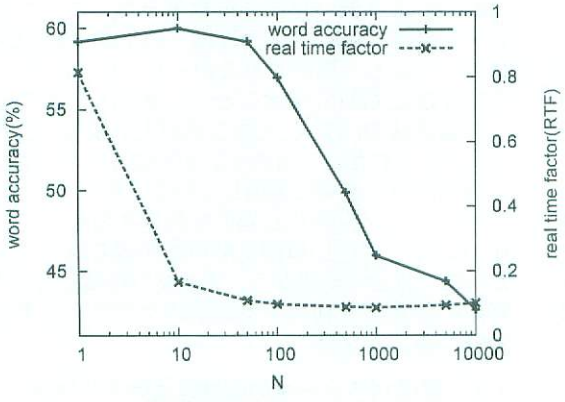

Fig. 3　適応高速化の効果

上より，今回の実験条件では N が 10 から 50 の範囲で，大きな認識精度の低下なく，計算時間を大幅に削減できることが分かった．

5　まとめと今後の検討

本論文では，信頼度計算の精度を向上させるため，GMM のオンライン教師なし適応法を音声・非音声の信頼度を探索に利用する手法と併用した場合の効果について検討した．この中で，適応データの信頼度を用いて，GMM の更新パラメータを重み付けし，閾値パラメータの変動に頑健な適応を行う手法と，GMM の適応回数を削減することで，適応に関する計算時間を削減する手法を提案した．DJSC を用いた認識実験では，GMM のオンライン教師なし適応を行うことで，大幅な認識精度の改善が得られることを確認した．これにより，人手で付けられた情報を基に音声・非音声を判別した場合と，ほぼ同程度の認識精度が得られることが分かった．また，適応データの信頼度に応じて GMM の更新パラメータを重み付けすることで，教師なし適応を行うためのパラメータの変動に対する頑健性が向上できることを確認した．さらに，適応回数を削減することで，大きな認識精度の低下なく，適応に関する計算量のほとんどが削減できることを確認した．以上の結果から，音声・非音声の信頼度を利用する適応的な音声認識手法を用いることで，雑音環境下における認識精度の大幅な改善を，少ない計算量で実現できることが確認できた．

今後の課題として，様々な雑音環境・SNR における本手法の効果及びパラメータの頑健性を評価することがあげられる．これにより多様な雑音環境への適応性を評価していきたい．

謝辞　Drivers' Japanese Speech Corpus in a Car Environment(DJSC) を使用させて頂いた旭化成 (株) に感謝致します．

参考文献

[1] 大西翼, ディクソンポール, 岩野公司, 古井貞熙. 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. 日本音響学会講演論文集, No. 1-1-16, pp. 49–50, 2009.

[2] Kousuke Hiraki, Takahiro Shinozaki, Koichi Shinoda, Agnieszka Betkowska, Koji Iwano, and Sadaoki Furui. Initial evaluation of the drivers' japanese speech corpus in a car environment. In *IEICE technical report. Speech*, Vol. 107, pp. 93–98, 2008.

[3] Yongxin Zhang and Michael S. Scordilis. Effective online unsupervised adaptation of gaussian mixture models and its application to speech classification. *Pattern Recognition Letters*, Vol. 29, No. 6, pp. 735–744, 2007.

[4] Tasuku Oonishi, Paul Dixon, Koji Iwano, and Sadaoki Furui. Robust speech recognition using VAD-measure-embedded decoder. In *Proc. Interspeech*, pp. 2239–2242, 2009.

[5] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, Vol. 10, No. 1-3, pp. 19–41, 2000.

[6] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyoshiro Shikano, and Shuichi Itahashi. JNAS Japanese speech corpus for large vocabulary continuous speech recognition research. *The Journal of the Acoustical Society of Japan*, Vol. 20, No. 3, pp. 199–206, 1999.

[7] Kikuo Maekawa. Corpus of spontaneous Japanese: Its design and evaluation. In *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.

[8] Paul Dixon, Diamantino Caseiro, Tasuku Oonishi, and Sadaoki Furui. The TITECH large vocabulary WFST speech recognition system. In *Proc. IEEE Workshop on ASRU*, pp. 443–448, 2007.