音声言語情報処理 39－5
(2001. 12. 20)

環境音モデルとHMM合成による音声区間検出法の文章発話への適用

渡部 生聖[1] 山田 武志[2] 北脇 信彦[2] 浅野 太[3]

[1] 筑波大学大学院理工学研究科
[2] 筑波大学電子・情報工学系
〒305-8573 茨城県つくば市天王台 1-1-1

[3] 産業技術総合研究所
〒305-8568 茨城県つくば市梅園 1-1-4

[1]E-Mail:juni@jks.is.tsukuba.ac.jp

あらまし 著者らは，音声と環境音が重畳している場合にも頑健かつ効率的な音声区間検出を行うために，環境音モデルとHMM合成を用いる方法を提案している．提案法では，まず音声と環境音のモデルを用いてビタビアライメントを求め，音声に重畳している環境音を予測する．そして，音声と予測した環境音の重畳モデルをHMM合成により作成し，この重畳モデルを加えて再度ビタビアライメントを求める．その結果，音声と環境音が重畳している区間，重畳している環境音とそのSN比を推定する．本稿では，提案法を複数の音声区間が存在する文章発話の場合に拡張する．9通りの環境音を文章発話に重畳し，音声区間検出実験を行った結果，提案法の検出正解率は従来法と比べて最大で4.7％改善していることが明らかとなった．

キーワード 音声区間検出，ビタビアライメント，環境音モデル，HMM合成

Voice Activity Detection for Sentence Utterances Using Environment Sound Models and HMM Composition

Narimasa WATANABE[1] Takeshi YAMADA[2] Nobuhiko KITAWAKI[2] Futoshi ASANO[3]

[1]Master's Program in Science and Engineering, University of Tsukuba
[2]Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 JAPAN

[3]National Institute of Advanced Industrial Science and Technology
1-1-4 Umezono, Tsukuba, Ibaraki, 305-8568 JAPAN

[1]E-Mail:juni@jks.is.tsukuba.ac.jp

Abstract To realize a robust voice activity detection (VAD) method in real acoustic environments, we have proposed a VAD method using environment sound models and HMM composition. The proposed method predicts the environment sound that overlaps with the speech, then composes the speech model and the model of the predicted environment sound, and detects the mixture sound period by using the composed models. In this paper, the proposed method is applied to the case of sentence utterances. To evaluate the performance of the proposed method, experiments were conducted. These results showed that the VAD accuracy of the proposed method is improved by a maximum of 4.7 % compared to that of the conventional methods.

key words Voice activity detection, Viterbi alignment, environment sound models, HMM composition

1 はじめに

音声認識や音声符号化などの音声処理系では，音声が存在する区間を正確に検出することが極めて重要である．静かな環境ではっきりと発話されている場合，信号レベルに適当な閾値を設けることにより比較的容易に音声区間を検出できる．しかし，特にハンズフリーの状況で発話されている場合，周囲雑音や他の人の話し声などが混入してしまうために，音声区間を正確に検出することが非常に困難となる．音声の区間を誤って検出すると認識率の低下や品質の劣化などの深刻な問題が生じるので，頑健な音声区間検出法の開発が強く望まれている．

従来の音声区間検出法の中でもよく用いられているのは，エネルギーと零交差回数を用いる方法 (例えば [1]) である．この方法では，短時間エネルギーの継続時間に応じて確実に音声だとみなせる区間を検出し，さらに零交差回数に応じて語頭の摩擦音などを検出する．しかし，この方法では，雑音の信号レベルが大きい場合に音声区間だけを検出することが原理的に困難である．

一方，信号レベルにあまり依存しない音声区間検出法として，音声と非音声（以下では環境音と呼ぶ）の HMM を用いてビタビアライメントを求める方法がある（例えば [2]）．しかし，この方法では，音声と環境音が重畳した区間の検出が困難であるという問題点がある．この問題点に対処する 1 つの方法は，音声と環境音が重畳した信号で学習したモデルを複数用意することである．しかし，多種多様な環境音の各々に対して重畳モデルを用意することは，学習のコストや探索の効率，精度からも非現実的である．

著者らは，音声と環境音が重畳している場合にも頑健かつ効率的な音声区間検出を行うために，環境音モデルと HMM 合成を用いる方法を提案している．提案法では，まず音声と環境音のモデルを用いてビタビアライメントを求め，音声に重畳している環境音を予測する．そして，音声と予測した環境音の重畳モデルを HMM 合成 [3] により作成し，この重畳モデルを加えて再度ビタビアライメントを求める．その結果，音声と環境音が重畳している区間を検出すると同時に，重畳している環境音の種類や，その SN 比といった重畳区間情報を推定できると考えられる．

これまでに，単語発話を対象として，提案法の有効性を示している [4][5][6]．本稿では，提案法を，音声区間が複数存在する文章発話の場合に拡張し，その性能をシミュレーション実験により評価する [7]．

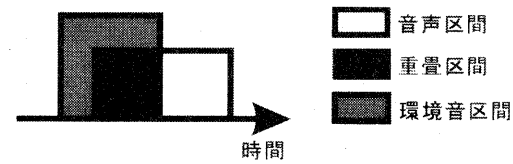

図 1: 音声と環境音の重畳例

以下，2 章で提案法について詳細に説明し，3 章で提案法の性能を評価する．

2 環境音モデルと HMM 合成を用いた音声区間検出法

提案法のアプローチを説明する．音声と環境音が重畳している例を図 1 に示す．図 1 は，音声の開始点の前後に環境音が存在し，音声と環境音が重畳する様子を示している．このような音声と環境音が重畳している区間は，音声と環境音の HMM を用いて単純にビタビアライメントを求めても，誤って検出されてしまうことがある．

この問題を解決するための一つの方法は，重畳区間に対応する HMM を用いることである．しかし，多種多様な環境音の各々に対して，音声との重畳を考慮した HMM を用意することは，探索の効率や精度という面で望ましくない．そこで，その都度重畳している環境音を予測することを考える．図 1 において，音声と環境音がそれぞれ単独で存在する区間を検出することは比較的容易である．それぞれの区間の間には，音声と環境音の重畳区間が存在する可能性があるので，音声区間の前後で検出された環境音が音声に重畳していると予測することができる．このようにして予測された環境音と音声の重畳モデルを作成し，再度ビタビアライメントを求めることにより，重畳区間を検出することができると考えられる．

提案法の詳細なアルゴリズムを説明する．

Step 1.

音声モデルと環境音モデルを図 2 のように連結し，入力信号のビタビアライメントを求める．ここで，図中の N_1 と N_2 は環境音モデル，S は音声モデルを表しており，説明の簡単化のために環境音モデルの数は 2，HMM の状態数は 1 としている．

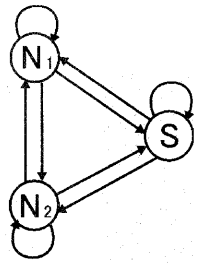

図 2: 音声モデルと環境音モデルの連結

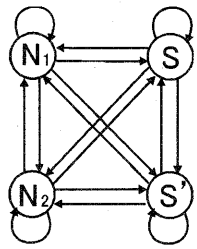

図 3: 音声モデル，環境音モデル，重畳モデルの連結

Step 2.

音声と Step 1. で音声区間の直前に検出された環境音の重畳モデルを HMM 合成により作成する．その際，あらかじめ設定した SN 比に応じて数通りの重畳モデルを作成する．

Step 3.

音声モデル，環境音モデル，Step 2. で作成した重畳モデルを図 3 のように連結し，再度入力信号のビタビアライメントを求める．ここで，図中の S' は重畳モデルを表しており，説明の簡単化のために SN 比は一通りとしている．

提案法では，各環境音に対する重畳モデルをあらかじめ複数用意するのではなく，重畳する環境音をその都度予測するので，組合せ爆発による探索の効率や精度の低下を防ぐことができると考えられる．

これまでに，単語発話の検出を対象とし，Left-to-Right 型で連結した HMM を用いて性能評価を行い，有効性を示している [4][5][6]．本稿では，複数の音声区間を含む文章発話の検出を対象とするために，ergodic 型で連結した HMM を用いるように拡張している．

表 1: 実験条件

音声 HMM	状態数 1，混合分布数 64
無音 HMM	状態数 1，混合分布数 64
環境音 HMM	状態数 1，混合分布数 16
学習データ	電総研単語音声データベース
(音声・無音)	話者 S0041,S0042 の各 1050 単語
	ASJ 研究用連続音声データベース
	vol1〜vol3 の全文章
(環境音)	RWCP 実環境・音声音響データベース
	candybowl, clock1, cymbals, pan, pipong, spray, toy, trash-box, whistle1 の偶数番号データ
評価データ	ATR 音声データベース
	SetC・4 話者の各 115 文章,
	candybowl, clock1, cymbals, pan, pipong, spray, toy, trash-box, whistle1 の奇数番号データ 1 個

3 音声区間検出実験

3.1 実験条件

評価実験に使用する HMM と音声，環境音のデータの条件を表 1 に示す．音声，無音モデルは 1 状態，64 混合分布であり，電総研音声データベース [8] の話者 S0041 と S0042 の各 1050 単語と，ASJ 研究用連続音声データベース [9] の全文章から学習している．各環境音モデルは 1 状態，16 混合分布であり，RWCP 実環境音声・音響データベース [10] の比較的継続時間の長い環境音の中から candybowl（金属箱を金属棒で叩く音），clock1（時計のベルの音），cymbals（シンバルの音），pan（鍋を金属棒で叩く音），pipong（電子音），spray（スプレーの噴射音），toy（ぜんまいの音），trashbox（ゴミ箱を金属棒で叩く音），whistle1（ホイッスルの音）を選択し，偶数番号のデータを用いて学習している．

評価用データは，ATR 音声データベース [11]SetC の 4 話者の各 115 文章と，上記の 9 種類の環境音の奇数番号データを各 1 個使用し，それぞれを計算機上で加算することにより作成している．その際，SN 比が 20，10，0 dB となるように環境音の信号レベルを調整している．また，1 文章には 1 種類の環境

表 2: 分析条件

標本化周波数	16 kHz
フレーム長	25 msec
フレーム周期	10 msec
高域強調	$1-0.97z^{-1}$
特徴量	メルケプストラム係数
	12 次元

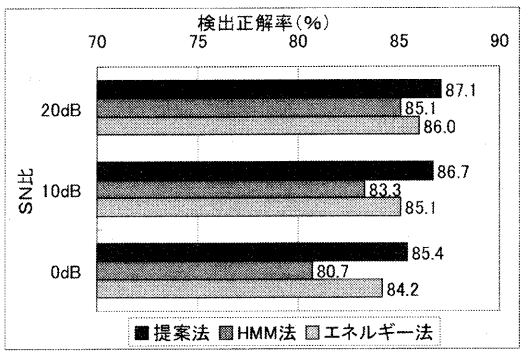

図 4: 検出正解率

音を 1.5 秒毎に加算しており，音声と環境音の様々な重畳パターンが含まれている．

分析条件を表 2 に示す．特徴量としては，標本化周波数 16 kHz，フレーム長 25 msec（ハミング窓），フレーム周期 10 msec で切り出されたフレームに高域強調（$1-0.97z^{-1}$）を行った後，12 次元のメルケプストラム係数を求めている．

3.2 実験結果と考察

3.2.1 検出正解率

本節では，音声区間と音声以外の区間における検出性能を総合的に評価する．図 4 に検出正解率を示す．図中の縦軸には評価用データの SN 比，横軸には検出正解率が示されている．また，図中のエネルギー法はエネルギーと零交差回数を用いる方法，HMM 法は音声と環境音の HMM を用いて単純にビタビアライメントを求める方法である．検出正解率の定義式は以下の通りであり，ATR 音声データベース中の音声区間ラベルを正解として用いている．

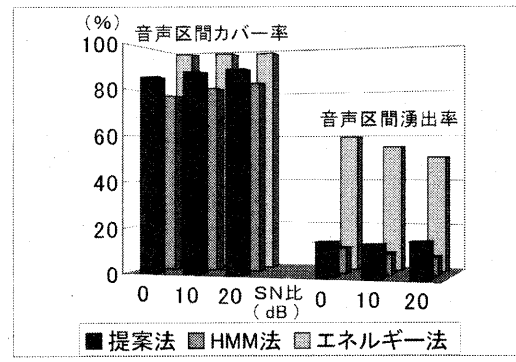

図 5: 音声区間の音声区間カバー率と音声以外の区間の音声区間湧出率

$$\text{検出正解率} = \frac{\text{音声区間で音声を検出したフレーム数} + \text{音声以外の区間で音声以外の音を検出したフレーム数}}{\text{総フレーム数}} \times 100\ (\%)$$

なお，1 データ当たりの平均フレーム数は 542.2 フレームである．

図 4 から，提案法はエネルギー法と比べて 1.1%～1.6%(6～9 フレームに相当)，HMM 法と比べて 2.0%～4.7%(11～25 フレームに相当) 改善していることが分かる．

3.2.2 音声区間カバー率・音声区間湧出率

本節では，音声区間と音声以外の区間のそれぞれにおける各手法の検出性能を詳しく調べる．図 5 に音声区間全体における音声区間カバー率と音声以外の区間全体における音声区間湧出率を示す．図中の縦軸には音声区間カバー率と音声区間湧出率，横軸には評価用データの SN 比が示されている．音声区間カバー率と音声区間湧出率の定義式は以下の通りである．

$$\text{音声区間カバー率} = \frac{\text{音声区間で音声を検出したフレーム数}}{\text{音声区間の総フレーム数}} \times 100\ (\%)$$

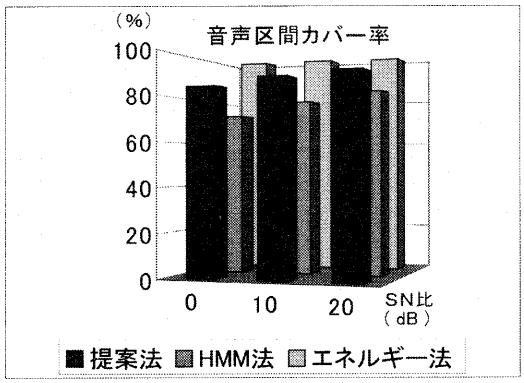

図 6: 重畳区間の音声区間カバー率

$$\text{音声区間湧出率} = \frac{\text{音声以外の区間で音声を検出したフレーム数}}{\text{音声以外の区間の総フレーム数}} \times 100\ (\%)$$

なお，1データ当たり，音声区間は平均411.0フレーム，音声以外の区間は平均131.2フレームである．

図5から，以下のことが分かる．

- 提案法の音声区間カバー率は，HMM法と比べて5.0%～7.3%(21～30フレームに相当)改善し，エネルギー法と比べて9.9%～12.9%(41～53フレームに相当)低下している．
- 提案法の音声区間湧出率は，HMM法と比べて3.8%～7.6%(5～10フレームに相当)低下し，エネルギー法と比べて34.9%～44.6%(46～59フレームに相当)改善している．

よって，提案法は，HMM法に比べてある程度湧出を伴うが，音声区間をより頑健に検出することができ，エネルギー法に比べて湧出を大幅に抑えることができることが分かる．

3.2.3 重畳区間の音声区間カバー率

提案法の最大の特徴は，HMM法で検出が困難であった，音声と環境音が重畳している区間の検出を可能にしていることである．そこで，重畳区間における検出性能を詳しく調べる．図6に，重畳区間における音声区間カバー率を示す．図中の縦軸には音声区間カバー率，横軸には評価用データのSN比が

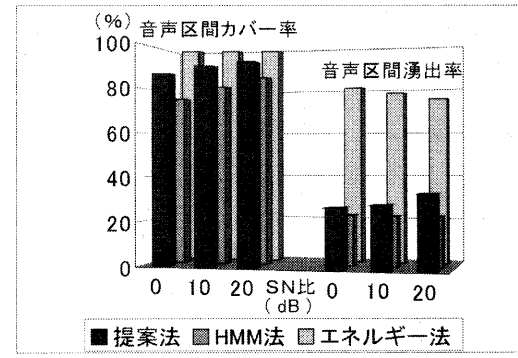

図 7: 音声開始点付近の音声区間カバー率と音声区間湧出率

示されている．なお，1データ当たり，重畳区間は平均181.8フレームである．

図6から，提案法の音声区間カバー率はHMM法と比べて7.2%～12.1%(13～22フレームに相当)改善していることが分かる．これは，音声区間全体で改善した21～30フレームのうちの，68.6%程度に相当するので，HMM法に対する提案法の改善量は重畳区間に集中していることが分かる．

3.2.4 音声開始点付近の音声区間カバー率，音声区間湧出率

本節では，音声区間検出において特に重要と考えられる音声端点の検出について詳しく調べる．図7に，音声開始点後10フレームの区間における音声区間カバー率と音声開始点前10フレームの区間における音声区間湧出率を示す．図中の縦軸には音声区間カバー率と音声区間湧出率，横軸には評価用データのSN比が示されている．なお，1データ当たり，両区間とも平均29.0フレームである．

図7から，以下のことが分かる．

- 提案法の音声区間カバー率は，HMM法と比べて5.6%～10.7%(2～3フレームに相当)改善し，エネルギー法と比べて8.2%～13.4%(2～4フレームに相当)低下している．
- 提案法の音声区間湧出率は，HMM法と比べて4.1%～10.6%(1～3フレームに相当)低下し，エネルギー法と比べて42.1%～53.9%(12～16フレームに相当)改善している．

よって，3.2.2節の結果と同様の傾向があることが分かった．

4 おわりに

本稿では，環境音モデルとHMM合成を用いた音声区間検出法を，複数の音声区間が存在する文章発話の場合に拡張し，その性能を評価した．9通りの環境音を文章に重畳し，音声区間検出のシミュレーション実験を行った結果，提案法の検出正解率は従来法と比べて1.1%～4.7%改善していることが分かった．また，提案法は，HMM法と比べて音声区間湧出率は低下しているものの，音声区間カバー率は改善されており，エネルギー法と比べて音声区間カバー率は低下しているものの，音声区間湧出率は大幅に改善していることが分かった．一方で，音声開始点(端点)付近の検出性能を改善する必要があることが明らかとなった．

今後，提案法で推定される重畳区間情報，すなわち重畳区間範囲，重畳環境音の種類，重畳区間のSN比などの情報を音声認識において利用し，音声と環境音が重畳する場合にも頑健な音声認識を行う方法について検討する予定である．

謝辞

本研究の一部は，科学研究費補助金（課題番号12780259），（財）サウンド技術振興財団の援助による．

参考文献

[1] 新美 康永, "音声認識," 共立出版, 1979.

[2] 古井貞熙 監訳, "音声認識の基礎," NTTアドバンステクノロジ, 1995.

[3] F. Martin, K. Shikano, Y. Minami, "Recognition of noisy speech by composition of speech and noise," Proc. European Conference on Speech Communication and Technology, pp. 1031–1034, 1993.

[4] 渡部生聖, 山田武志, 北脇信彦, 浅野太, "環境音モデルとHMM合成を用いた音声区間検出の検討," 信学技報, SP75–94, pp. 55–60, 2000.

[5] 渡部生聖, 山田武志, 北脇信彦, 浅野太, "環境音モデルとHMM合成による音声区間検出法," 日本音響学会春季研究発表会, 3-3-8, pp. 109–110, 2001.

[6] T. Yamada, N. Watanabe, N. Kitawaki, F. Asano, "Voice Activity Detection using Non-speech Models and HMM composition," Proc. HSC2001, pp. 131–134, 2001.

[7] 渡部生聖, 山田武志, 北脇信彦, 浅野太, "環境音モデルとHMM合成を用いた文章発話に対する音声区間検出の検討," 日本音響学会秋季研究発表会, 1-1-20, pp. 39–40, 2001.

[8] 田中和世, 速水悟, "電総研の研究用音声データベース," 日本音響学会誌, Vol. 48, No. 12, pp. 883–887, 1992.

[9] 小林哲則, 板橋秀一, 速水悟, 竹沢寿幸, "日本音響学会研究用連続音声データベース," 日本音響学会誌, Vol. 48, No. 12, pp. 888–893, 1992.

[10] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," Proc. International Conference on Language Resources and Evaluation, pp. 965–968, 2000.

[11] 武田一哉, 匂坂芳典, 片桐滋, 桑原尚夫, "研究用日本語音声データベースの構築," 日本音響学会誌, Vol. 44, No. 10, pp. 747–754, 1988.