音声言語情報処理 34－10
(2000. 12. 21)

環境音モデルとHMM合成を用いた音声区間検出の検討

渡部 生聖[1] 山田 武志[2] 浅野 太[3] 北脇 信彦[2]

[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]電子技術総合研究所
〒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 本稿では，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する．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比を検出できる．9通りの環境音を音声に重畳し，音声区間検出の実験を行った結果，提案法の音声区間検出率は従来法と比べて数%から最大で40%程度高い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 Using Non-speech Models and HMM Composition

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

[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]Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki, 305-8568 JAPAN

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

Abstract This paper proposes a new voice activity detection (VAD) method using non-speech models and HMM composition. The proposed method predicts a non-speech sound that overlaps with speech by a Viterbi algorithm with a speech model and non-speech models. Then, a Viterbi algorithm with composed models of the speech model and the predicted non-speech sound models is applied again. As a result, the overlap segments are detected. Furthermore, the non-speech sound that overlaps with speech and its SNR are identified. To evaluate the performance of the proposed method, preliminary experiments are conducted. These results showed that the VAD accuracy of the proposed method is improved by a maximum of 40 % compared to that of the conventional method.

key words Voice activity detection, Viterbi alignment, non-speech 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である（例えば[1]）．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，エネルギーを用いる方法では，雑音の信号レベルが大きい場合に音声区間を検出することが原理的に困難である．

一方，音声区間検出を音声認識（パターンマッチング）の枠組みで行う方法が提案されている（例えば[2]）．その中でも代表的なものは，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（入力フレームとHMMの状態との対応）を求める方法である．一般に音声認識ではメルケプストラムなどのスペクトル包絡情報を利用することから，環境音の信号レベルに依存しない安定した音声区間検出が可能となる．しかし，様々な環境下で安定した性能を得るためには，

- 多種多様な環境音の適切なモデル化
- 音声と環境音が重畳した区間への対応

という問題に対処する必要がある．前者については，最近になってRWCPのプロジェクトにより環境音のデータベースが整備されつつあり[3]，モデルの単位や構造などに関する研究が活発に行われている[4, 5]．後者の問題に対処するための一つの方法は，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．

以上から，本稿では，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する．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[6]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比を検出できる．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．

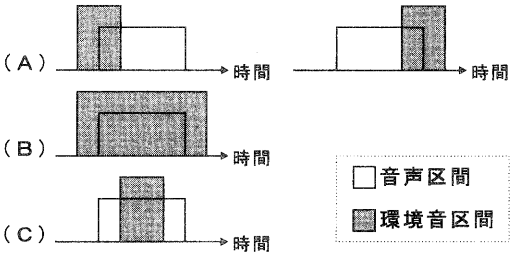


図 1: 音声と環境音の重畳パターン

2 環境音モデルとHMM合成を用いた音声区間検出法

音声と環境音のHMMを用いてビタビアライメントを求めることにより音声区間を検出する方法（以下では従来法と呼ぶ）には，音声と環境音が重畳した場合に検出精度が低下するという問題がある．この問題に対処するための一つの方法は，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．本稿では，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する．

音声と環境音の重畳パターンのうち典型的なものを図1に示す．図中の（A）では音声区間の前後で環境音が重畳しており，（B）では音声区間を覆うように環境音が重畳している．また，（C）では音声区間の中で環境音が重畳している．ここで，（A）と（B）には環境音が単独で存在する区間があることに着目する．このような区間は従来法でも検出できると考えられるので，音声の直前（直後）の環境音を予測することができる．そこで，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

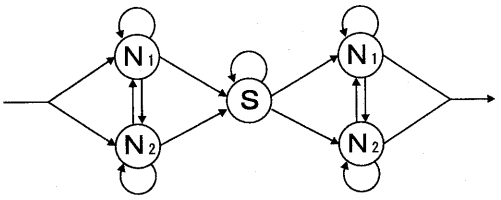


図 2: 音声モデルと環境音モデルの連結

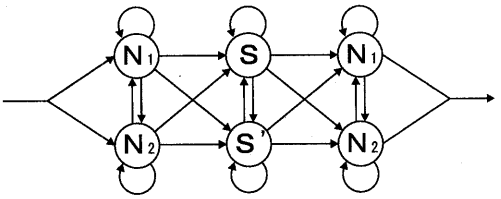


図 3: 音声モデル，環境音モデル，重畳モデルの連結

モデルを加えて再度ビタビアライメントを求めることにより，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比を検出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．なお，本稿では音声区間が1個所だけ必ず存在すると仮定している．

Step 2.

音声とStep 1.で音声区間の直前（直後）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．

表 1: 実験条件

音声データベース	電総研単語音声データベース
学習データ	話者S0001の1050単語
評価データ	話者S0001の492単語
環境音データベース	RWCP実環境音声・音響データベース
学習データ	candybowl, clock1, cymbals, pan, pipong, spray, toy, trashbox, whistle1 の偶数番号データ
評価データ	candybowl, clock1, cymbals, pan, pipong, spray, toy, trashbox, whistle1 の奇数番号データ
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
音声モデル	状態数1，混合分布数64
環境音モデル	状態数1，混合分布数16
無音モデル	状態数1，混合分布数16

3 音声区間検出実験

3.1 実験条件

実験条件を表1に示す．音声データベースは電総研単語音声データベース（ETL-WD-I&II）[7]であり，話者S0001の1050単語を音声モデルの学習用，残りの492単語を評価用とする．環境音データベースとしてRWCP実環境音声・音響データベース[3]を用いる．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環境音モデルの学習用，奇数番号のデータのうち1つを評価用とする．サンプリン

グ周波数は16 kHz，フレーム長は25 msec（ハミング窓），フレーム周期は10 msecである．この条件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る．音声のモデルは状態数1，混合分布数64，環境音と無音のモデルは状態数1，混合分布数16である．

評価用データは，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，音声と環境音を計算機上で加算することにより作成している．その際，環境音の開始点が音声の開始点から5フレーム以上前になるようにしている．その結果，環境音は音声区間の前で単独で存在し，かつ音声区間の一部あるいは全てと重畳している（図1の（A）と（B）を参照）．また，提案法のStep 2.では，SN比を20, 10, 0, -10, -20 dBの5通りに設定しており，音声区間の直前に検出された環境音についてのみ重畳モデルを作成している．

3.2 実験結果と考察

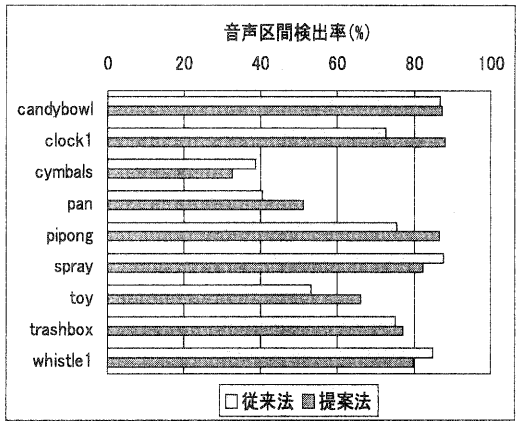
図4にSN比が20, 10, 0 dBのときの音声区間検出率を環境音別に示す．ここで，図中の従来法は提案法のStep. 1に相当する．また，音声区間検出率の定義は次の通りである．

音声区間検出率

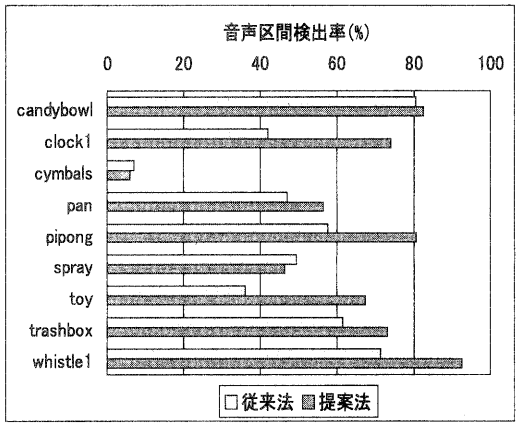
$$= \frac{\text{音声開始フレームの検出に成功した単語数}}{\text{全単語数}} \times 100\,(\%)$$

なお，音声データベースに添付されている音声区間ラベルを正解とし，そこから±5フレームの誤差を許容している．

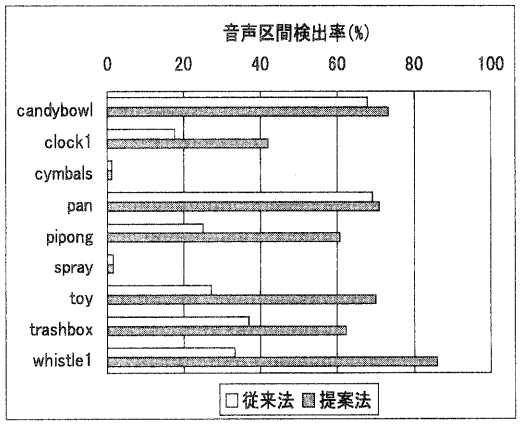
図4より，全体的に従来法よりも提案法の方が音声区間検出率が高く，従来法と比べて数%から最大で40%程度改善していることが分かる．特に，SN比が低いときほど，従来法に対する改善量が大きくなっている．SN比が高いときには，従来法でも音声と環境音が重畳している区間を音声としてある程度検出できる．しかし，SN比が低いときには，従来法では音声と環境音が重畳している区間を環境音として検出することが多くなる．よって，提案法による改善の効果が大きくなっていると考えられる．図5に従来法と提案法による音声区間検出例を示す．図中の波形は「るいぎご」という音声に環境音（whistle1）がSN比0 dBで重畳しているときのものである．波形



(1) SN比20 dB



(2) SN比10 dB



(3) SN比0 dB

図 4: 音声区間検出率

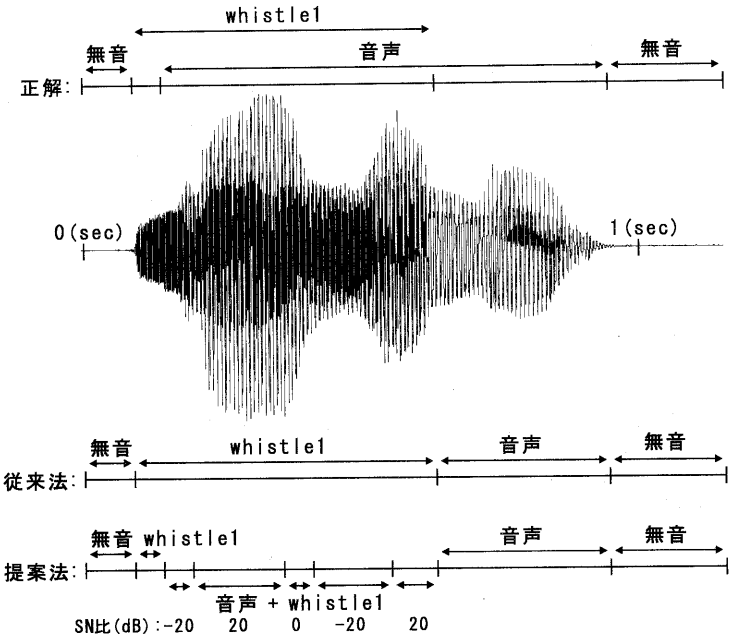


図 5: 従来法と提案法による音声区間検出例

の上部には正解区間を示しており，音声開始フレームから音声の中ほどまでwhistle1が重畳していることが分かる．波形の下部には従来法と提案法により検出された区間を示している．従来法では音声と環境音が重畳している区間をWhistle1と誤って検出しているが，提案法では音声とWhistle1の重畳区間として検出できていることが分かる．

一方，図4より，提案法の方が従来法よりも音声区間検出率が低い（あるいは同じ）場合があることが分かる．その理由としては，音声区間の直前に検出された環境音が実際に重畳している環境音と異なることが考えられる．図6にSN比が20, 10, 0dBのときの直前環境音検出率を環境音別に示す．直前環境音検出率の定義は次の通りである．

直前環境音検出率

$$= \frac{\text{直前の環境音を正しく検出した単語数}}{\text{全単語数}} \times 100\,(\%)$$

図6より次のことが分かる．

- 直前環境音検出率がある程度高い場合には，音声区間検出率も概ね改善されている．一方，音声区間検出率の改善につながっていない場合がいくつかある．その主な理由は，環境音が単独で存在している区間を音声と環境音が重畳している区間として誤って検出していることにある．この問題に対処するための方法としては，音声モデルと環境音モデルを合成するときのSN比を一律に設定するのではなく，環境音の特徴に応じて個々に設定することなどが考えられる．

- 直前環境音検出率が極端に低いときに音声区間検出率が改善されている場合がある．これは，音声区間の直前に検出された環境音と実際に重畳している環境音が音響的に似ていることによる．

- SN比が低いときほど直前環境音検出率は高くなる傾向がある．これは，音声と環境音が重畳している区間で環境音の特徴が支配的になるからである．

- 直前環境音検出率が極端に低い場合がある．音声と環境音が重畳している区間が長くなると，その区間には存在しない環境音が湧き出すことがある．この問題に対処するための方法と

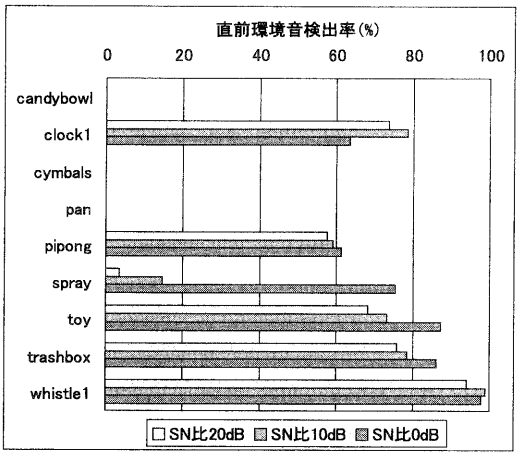


図 6: 直前環境音検出率

しては，Step 1.において音声区間の直前に検出された環境音が重畳していると仮定するのではなく，音声区間のある程度前に検出された環境音の中からフレーム平均尤度の高いものを選択することなどが考えられる．

4 おわりに

本稿では，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た．9通りの環境音を音声に重畳し，音声区間検出の実験を行った結果，提案法の音声区間検出率は従来法と比べて数%から最大で40%程度高いことが明らかとなった．その一方で，音声区間の直前の環境音の予測に改善の余地があることが分かった．

今後，音声区間の直前の環境音を正確に予測するために，音声区間のある程度前に検出された環境音の中からフレーム平均尤度の高いものを選択する方法について検討する予定である．また，環境音の高精度なモデル化や提案法の文章発話への適用について検討していく．さらに，提案法では重畳している環境音とそのSN比を知ることができるので，これらの情報を頑健な音声認識のために利用するという枠組で研究を進める予定である．

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] 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.

[4] 比屋根一雄, 澤部直太, 飯尾淳, “擬音語表現に基づく衝突音認識システム,” 日本音響学会秋季研究発表会, pp. 135–136, 1998.

[5] 三木一浩, 西浦敬信, 中村哲, 鹿野清宏, “HMMを用いた環境音識別の検討,” 信学技報, SP99-106, pp. 79–84, 1999.

[6] 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.

[7] 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.