Vol.2009-SLP-78 No.1
2009/10/23

長時間スペクトル変動と調波構造に基づく発話区間検出法の音声認識による評価

福田 隆[†1] 市川 治[†1] 西村雅史[†1]

発話区間検出（VAD）は音声認識を高精度化するための重要な要素の一つである．これまでに我々は車内環境を対象とした雑音に頑健な VAD 法を提案し，平均音素長以上の区間から抽出される長時間変動情報と調波構造情報に由来する特徴量が，VAD の性能改善に大きく寄与することを示した．しかし，過去の研究報告では発話単位の検出精度のみに注目していたため評価が限定的であった．本報告では，フレーム単位での音声／非音声識別能力，及び音声認識システムにおける提案法の効果を検証し，多方面からの考察を加える．CENSREC-2 を用いた音声認識実験において，提案法は ETSI-AFE で採用されている VAD と比較して認識誤りを 29.1%削減した．

Evaluation of VAD using ASR based on long-term spectro-temporal and static harmonic features

Takashi Fukuda,[†1] Osamu Ichikawa[†1]
and Masafumi Nishimura [†1]

Accurate voice activity detection (VAD) is important for robust automatic speech recognition (ASR) systems. We have proposed a statistical-model-based VAD using the long-term temporal and harmonic structure-related information in speech, which shows good robustness against noise in an automobile environment. But in our previous works, we focused on only the utterance-based speech segment detection performance. This paper further investigates frame-based speech/non-speech discrimination performance of VAD and ASR performance. In an experiment using CENSREC-2, the word error rate was reduced by 29.1% in a test that included an ASR system.

†1 日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research - Tokyo, IBM Japan, Ltd.

1. はじめに

音声の有声区間と休止区間を正確に検出する技術（VAD: Voice Activity Detection）は，音声認識システムの性能を左右する重要な要素である．近年では，雑音環境下での VAD 性能向上を目指して，統計モデルに基づく VAD 法が多数検討されている[1),2)]．一般に統計的手法では，利用環境で収録された音声データからモデル（主として GMM）を設計することによって高精度な VAD を実現している．しかし，雑音強度が大きくなった場合には，動作が不安定になるという問題があった．これに対して，我々は音声の長時間スペクトル変動に着目した VAD 法を検証し，低 S/N 環境下での性能改善を図った[3)]．提案法では長時間変動情報を，窓長を大きく取った動的特徴量として抽出している．その後，0 dB や −5 dB といった極めて雑音が大きい環境での性能改善を目指し，音声の調波構造情報に基づく特徴量を VAD に導入した[4)]．調波構造に由来する従来手法の多くは有声／無声判定と基本周波数の推定を必要としており，統計モデルの枠組みで調波構造を利用することは難しかった．提案法は，明示的な判定／推定処理を必要とせず，効果的に調波構造情報を取り込んでいる．

一般に，音声認識において VAD を利用する利点は次の二つが考えられる．

(1) 高性能な VAD は，スペクトルサブトラクションや Wiener フィルタなどの雑音除去技術における雑音スペクトルの推定精度を向上させる．
(2) VAD 情報に基づいて認識に不要な部分をカット（Frame dropping）する方式は，音声認識時の挿入誤りを削減する効果がある．しかし，逆に音声の検出漏れが発生すると，認識時には削除誤りが増えてしまう．

過去の研究報告では，音声認識システムにおける応答漏れを防ぐことを目標に，発話単位の検出精度に焦点をあててきた．しかし，認識精度の向上という目的においては，フレーム単位での VAD 性能も重要である．本報告では，フレーム単位での VAD 性能を比較評価した後，音声認識への影響について考察する．

2. 提案手法の概要

2.1 GMM に基づく発話区間検出

一般に，統計モデルに基づく VAD では，音声と非音声の GMM から計算される音響尤度比を基準として，フレーム単位での識別を行う．特徴ベクトル $\boldsymbol{x}_t$ の音声 GMM に対する尤度を $p(\boldsymbol{x}_t|M_1)$，非音声 GMM に対する尤度を $p(\boldsymbol{x}_t|M_0)$ とすると，時刻 t における音響

Vol.2009-SLP-78 No.1
2009/10/23

尤度比は次のように計算される．

$$L(\boldsymbol{x}_t) = \log p(\boldsymbol{x}_t|M_1) - \log p(\boldsymbol{x}_t|M_0) \quad (1)$$

ここで，M_i は音声／非音声 GMM（$i = 0$: 非音声，$i = 1$: 音声）である．式 (1) において $L(\boldsymbol{x}_t)$ が閾値より大きい場合は音声，小さい場合は非音声フレームと見なす．通常，この識別基準に加え，スペクトル強度の小さい発話末尾の検出漏れを防ぐため，発話末尾の音声区間を延伸する hangover 処理が適用される[1)]．

続いて，発話単位での音声区間検出には上述のフレーム単位の VAD 情報を利用する．現在のフレームを中心とした合計 N フレームから成るバッファを用意し，バッファ内の音声フレーム数が閾値を超えた時刻を音声開始時点と見なす．その後は，バッファを 1 フレームずつずらしながら非音声フレーム数を調べ，発話末尾検出用の閾値を超えた時刻を発話終了時刻とする．

2.2 長時間スペクトル変動情報の利用

本研究は GMM に基づく VAD において，音響特徴量を改良した方式に相当する．提案法では長時間スペクトル変動情報と，調波構造に基づく特徴量の二種類を利用している．このうち長時間変動情報は，窓長を大きく取った動的特徴量として抽出し，スペクトルの時間軌跡に対する線形回帰演算により計算する．

$$d_t(l) = \sum_{k=1}^{K} \left[k \cdot \left\{ c_{t+k}(l) - c_{t-k}(l) \right\} \right] \Big/ 2 \sum_{k=1}^{K} k^2, \quad (2)$$

ここで $c_t(l)$ は時刻 t における l 次元目のケプストラム係数，K は分析窓長（前後 K フレームを利用）を表す．音声認識における知見に基づいて，VAD では $K = 2 \sim 3$ 程度の短い窓長が使われることが一般的であった．しかし，VAD にとって有益な情報はさらに長い時間区間に内在しており，提案法では，発話の平均音素長を超える区間から計算される長時間スペクトル変動情報（以後，LongΔCep と略す）を VAD に利用する．なお，この長時間変動情報の抽出は変調周波数の観点から，2Hz 程度の比較的緩やかな時間変動成分を強調する処理と見なすことができる[3)]．

2.3 調波構造特徴量

次に，調波構造に基づく特徴量の抽出手順を示す（図 1）[4)]．

(i) 処理フレーム毎に，観測音声の対数パワースペクトル $y_t(j)$ を得る．j は bin 番号である．

(ii) 離散コサイン変換（DCT）により，そのケプストラム表現を得る．

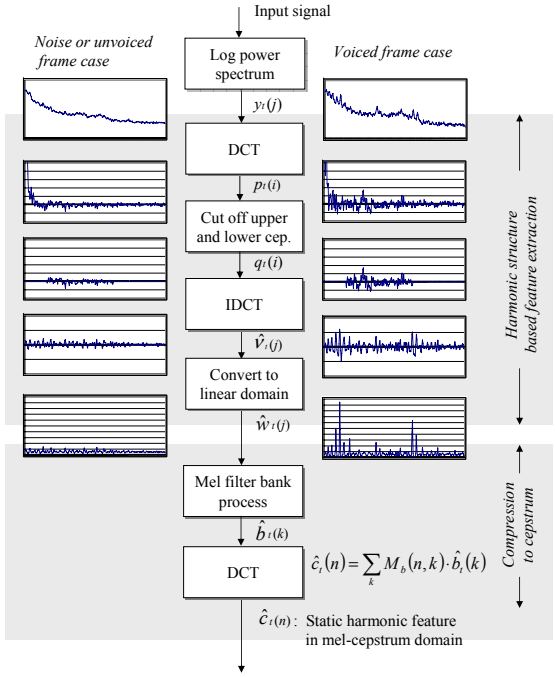

図 1 調波構造特徴の抽出過程

$$p_t(i) = \sum_j M_a(i,j) \cdot y_t(j), \quad (3)$$

ここで，$M_a(i,j)$ は離散コサイン変換行列である．

(iii) 音声の調波構造に対応した領域のみを残すべく，ケプストラムの上位項と下位項をカットする．

$$\begin{aligned} q_t(i) &= \lambda \cdot p_t(i) \qquad if\ (i < D_L)\ or\ (i > D_H) \\ q_t(i) &= p_t(i) \qquad otherwise, \end{aligned} \quad (4)$$

ここで，λ は 0 または非常に小さい定数である．D_L と D_H は調波構造としてとり得

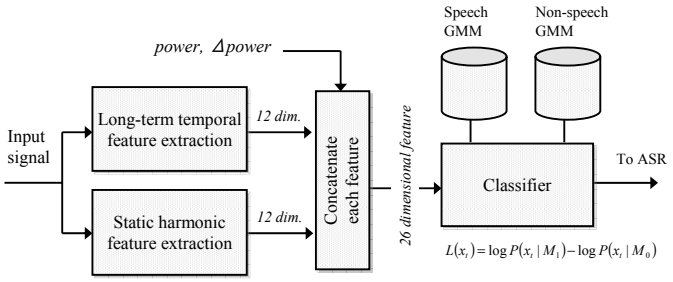

図 2 提案手法の概要

る範囲に対応する．例えば，音声の基本周波数が 100〜400 Hz の範囲に存在すると仮定すると，8 kHz のサンプリング周波数において，$D_L = 20$，$D_H = 80$ となる．

(iv) $q_t(i)$ について逆離散コサイン変換（IDCT）の後，指数変換を行い，そのスペクトル表現を得る．

$$\hat{\nu}_t(j) = \sum_i M_a^{-1}(j,i) \cdot q_t(i) \tag{5}$$

$$\hat{w}_t(j) = \exp\left\{\hat{\nu}_t(j)\right\}. \tag{6}$$

先行研究では，音声認識において $\hat{w}_t(j)$ を音声強調フィルタと見なし，パワースペクトルを強調する処理に用いていた[5]．

(v) $\hat{w}_t(j)$ を入力とするメルフィルタバンク処理を行い，その出力 $\hat{b}_t(k)$ を DCT 変換行列 $M_b(n,k)$ によって，ケプストラム $\hat{c}_t(n)$ に変換する．

$$\hat{c}_t(n) = \sum_k M_b(n,k) \cdot \hat{b}_t(k), \tag{7}$$

得られたケプストラムを調波構造特徴（Static Harmonic Feature）と呼ぶ．

3. VAD 単独の評価実験

3.1 実験概要

図 2 に提案手法の概要を示す．まず，フレームサイズおよびシフト幅をそれぞれ 25ms と

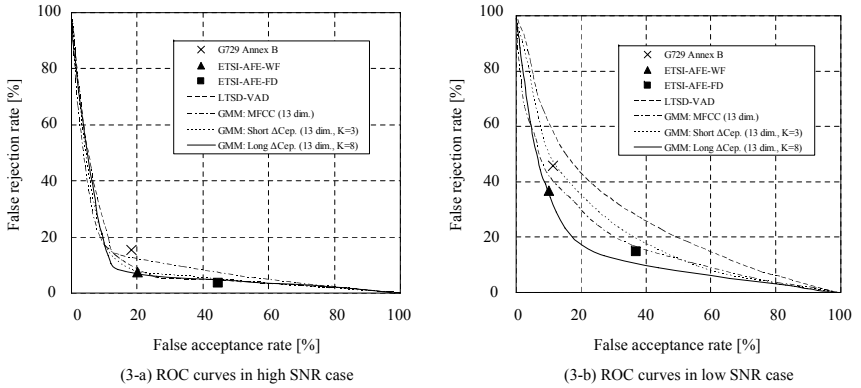

図 3 各 VAD における ROC 曲線

10ms とし，24 チャンネルのメルフィルタバンク処理を経由して，12 次元の調波構造特徴を抽出する．これと並行して，フレーム毎に 12 次元の（通常の）ケプストラムを求め，前後 10 フレームから計算される動的特徴量を，長時間スペクトル変動成分（LongΔCep.）として利用する．最後に，上記二種類の特徴量とパワー項（power，LongΔpower）を連結し，26 次元の特徴量として識別器に入力する．ただし，CMN 処理は行っていない．評価実験には情報処理学会 SIG-SLP 雑音下音声認識評価ワーキンググループから配布されている VAD の評価セット（CENSREC-1-C）の内，走行雑音の評価データ（シミュレーション環境データ：A Set，Car）を使用した．走行雑音はクリーン音声に対して 20 dB 〜 −5 dB の間で 5 dB 刻みに重畳されている[6]．本実験で利用する評価データは男女各 52 名による 6,986 発声であり，発話内容は連続数字，サンプリング周波数は 8kHz である． GMM の学習には，同ワーキンググループから配布されている AURORA2J ／ CENSREC1 の内，走行雑音のデータセットを利用した（評価データと同じ雑音環境）[7]．学習データ数は，男女各 55 名による 1,668 発話である．混合数は音声／非音声 GMM 共に 32 とした．

3.2 フレーム単位での音声識別実験

本節では ROC 曲線を用いて，長時間変動特徴量単独の識別性能を調査する．図 3 に実験結果を示す．図では，長時間情報に着目した従来手法である LTSD-VAD[8]，および標準化手法である G.729 Annex B[9]，ETSI AFE で利用されている 2 種類の VAD（ETSI-AFE-WF for Wiener Filter，及び ETSI-AFE-FD for Frame Dropping）[10] との比較を示している．

Vol.2009-SLP-78 No.1
2009/10/23

表 1 フレーム単位識別実験における各特徴量の等誤り率

Name of VAD algorithm	Equal error rate [%]		
	High SNR	Low SNR	Average
LTSD-VAD	13.4	31.6	22.5
GMM: MFCC	14.0	24.6	19.3
GMM: Short ΔCep. (K=3)	12.2	27.4	19.8
GMM: Long ΔCep. (K=8)	11.7	18.6	15.2

GMM に基づく VAD（以後，GMM-VAD）の特徴パラメータにはパワー項も含んでいる．ここで，ShortΔCep は 7 フレーム（K=3）から，LongΔCep は 17 フレーム（K=8）から計算している．また，図において High SNR は Clean，20 dB，15 dB，10 dB の平均を，Low SNR は 5 dB，0 dB，−5 dB の平均を表している．

G.729-VAD と ETSI-AFE-WF は High SNR 環境において比較的高い識別性能を示したが，Low SNR 環境では誤棄却率（False rejection rate）が増加した．一方，ETSI-AFE-FD は Low SNR においても良好な誤棄却率を示すが，G.729-VAD や ETSI-AFE-WF と比べて High SNR 環境でさえも誤受理率（False Acceptance Rate）が増加してしまう．これは，ETSI-AFE-FD が音声区間の欠落による音声認識の削除誤りを避けるべく，やや余裕を持たせた設計になっているためである．他方，LTSD-VAD は High SNR で高い識別性能を示すが，Low SNR では低い性能にとどまった．

次に，GMM-VAD の特徴量間で比較すると，LongΔCep は顕著な性能改善を達成し，Low SNR 環境において等誤り率 18.6%を示した（表 1 参照）．これは，同環境の MFCC と比べて 24.4%の等誤り率の削減である．また，LongΔCep は High SNR 環境での改善は少なかったものの，種々の方式の中で一番低い等誤り率を示した．これらの実験結果は，長時間変動情報が雑音環境下において頑健な性質を持っていることを示している．なお，発話単位の識別性能については文献[3] を参照されたい．

3.3 調波構造特徴の効果

ここでは，調波構造特徴の効果を検証する．図 4 は Low SNR における ROC 曲線を示している．等誤り率の観点では，調波構造特徴は ”MFCC+LongΔCep ” とほぼ同等の性能であるものの，誤受理率と誤棄却率のどちらか一方に注目した場合，他の特徴量よりも低い値を示すため，一番優れた特徴量であると言える．一方，発話単位での性能評価（図 5 参照）において，調波構造特徴と LongΔCep の組み合わせは，特に Low SNR 環境での正解精度

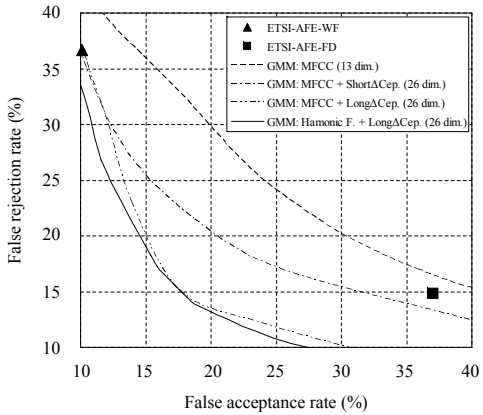

図 4 フレーム単位識別実験における調波構造特徴とその他特徴量の比較

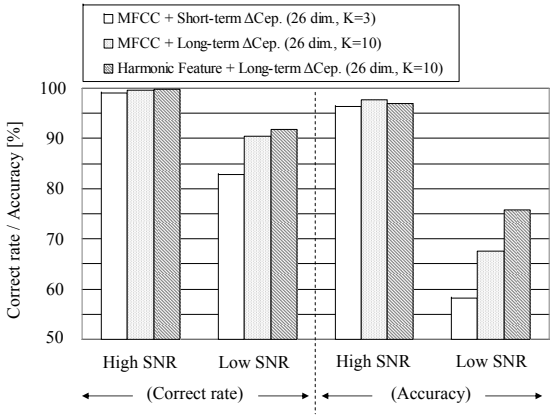

図 5 発話単位識別実験における長時間変動と調波構造特徴の効果

（Accuracy）の改善に関して貢献が大きく，”MFCC+LongΔCep ” からは 23.6%，”MFCC + ShortΔCep ” と比べると 41.7% の誤り削減となった．調波構造特徴は LongΔCep を補助する形で効果的に性能を高めていることがわかる．なお，調波構造特徴は単独では性能の改善がなく，平均正解率 64.1%，平均正解精度 41.7%であった．

4. 音声認識による評価実験

4.1 実験の概要

本節では VAD 性能の音声認識への影響を検証する．ここでは，雑音抑圧手法と Frame Dropping 処理に対する効果を中心に，図 6 に示す 3 種類の音声認識システムを用いて評価を行う．図 6-(a) は Baseline システム，同図-(b) は Wiener フィルタに基づく雑音抑圧処理を含んだシステムである．実験では，ETSI-AFE で採用されている 2 段階 Wiener フィルタリングによる雑音抑圧手法を利用している．ETSI-AFE において，VAD は雑音抑圧処理の 1 段目のみに影響する．最後に図 6-(c) は，(b) のシステムにさらに Frame dropping 処理を追加したシステムである．

音声認識実験には CENSREC-2 の連続数字発声を用いた[11]．評価カテゴリは 0 にあたる．本データベースは，アイドリング，市街地走行，高速走行時の車内において収録された音声であり，雑音条件とマイクロフォンの組み合わせにより 4 種類のデータセットが用意されている．実験では，学習時と評価時で収録条件が一致しているもの（Condition1, 雑音環境／マイクロフォン位置が同じ）を利用する．なお，Condition1 ではマイクロフォンはドライバーの頭上付近に取り付けられており，車内環境としては，エアコン，窓の開閉，オーディオが含まれている．ただし，一般的な車載音声認識システムでは，音声認識開始と同時にオーディオがミュートされる仕様となっているため，オーディオ音声が混じっているデータは評価の対象外とした．音響モデルの学習データは男性 33 名，女性 40 名による 7,195 発話であり，評価データは男性 19 名，女性 12 名による 2,074 発話である．サンプリング周波数は 16kHz で，フレームサイズおよびシフト幅はそれぞれ 25ms と 10ms である．評価音声は実際の車載音声認識システムを模擬すべく，発話終了以降の非音声部分を拡大して，各音声ファイルの長さが 5 秒になるように調整した．これは，発話末尾が検出されなかった場合のタイムアウト値に相当する．

音声認識には，12 次元の MFCC と ShortΔCep（$K = 3$），および Δ パワーを結合した 25 次元の特徴量を用いる．MFCC には ETSI-AFE の Blind equalization 処理を施している．音響モデルは 16 状態 20 混合分布の数字単位 HMM であり，5 状態の silence モデルと 3 状態の pause モデルを持つ．Wiener フィルタによる雑音除去は HMM の学習と評価の両方に用い，Wiener フィルタ用の VAD も学習と評価で統一する．Frame dropping 処理は評価時にのみ適用する．実験では，Wiener フィルタと Frame dropping の二種類について VAD を用いるが，認識性能が最も高くなるように VAD の閾値をそれぞれ調整した．

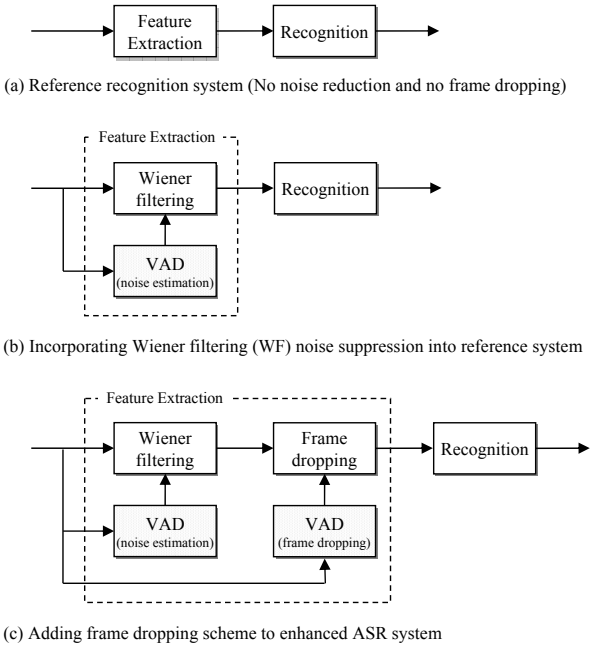

図 6 VAD 評価のための音声認識システム

4.2 実験結果

表 2 に実験結果を示す．はじめに雑音抑圧処理への影響を考察する．Baseline システムは発話末尾以降の長い非音声区間によって挿入誤りが増えるため低い性能となった．一方，ETSI-AFE-VAD や LTSD-VAD を用いた Wiener フィルタによる雑音抑圧手法は，認識精度の改善を与えた．両者を比較すると，LTSD-VAD の方がやや高い性能を示している．GMM-VAD の比較においては，”MFCC + LongΔCep ” が全環境で最も高い性能を達成し，”調波構造特徴 + LongΔCep ” の組み合わせもほぼ同等の性能を示した．

図 2 の右側は Frame dropping 処理を含めた場合の結果である．Frame dropping 処理を加えると，認識率が全体的に向上し，特に GMM-VAD で大きな効果があった．”MFCC + LongΔCep ” は平均で 87.8%の性能を達成し，ETSI-AFE-VAD と比較して，22.8%の認識誤り削減となった．また，”調波構造特徴 + LongΔCep ”の組み合わせは最も高い性能

表 2 車載音声コーパスに対する音声認識精度 (Audio-off case)

VAD System	Word accuracy [%]							
	Base + Wiener Filter				Base + Wiener Filter + Frame Dropping			
	Idling	Low Speed	High Speed	Average	Idling	Low Speed	High Speed	Average
None (WF-less)	75.6	65.2	59.1	66.6	-	-	-	-
ETSI-AFE-VAD	87.7	80.5	78.7	82.3	91.2	82.1	79.3	84.2
LTSD-VAD	89.4	82.6	79.4	83.8	90.9	82.9	78.1	83.9
GMM: MFCC (13 dim.)	87.8	83.3	78.9	83.4	89.2	82.5	77.7	83.1
GMM: MFCC + ShortΔCep. (26 dim.)	88.4	82.6	80.8	83.7	91.1	83.9	79.7	84.9
GMM: MFCC + LongΔCep. (26 dim.)	91.4	84.9	80.9	85.4	94.2	86.8	82.5	87.8
GMM: Harmonic Feature + LongΔCep. (26 dim.)	90.6	84.6	80.0	85.1	94.7	87.0	84.5	88.8

を示し，ETSI-AFE-VAD から 29.1%の認識誤りを削減した．

5. おわりに

本稿では，長時間スペクトル変動情報と調波構造成分に基づく VAD の評価を中心にまとめた．評価の結果，提案法は低 S/N 環境下において顕著な改善を示し，音声／非音声の識別能力の向上と共に，音声認識精度にも改善を与えることを報告した．今回は目的話者以外の声を含む雑音源は評価の対象外としたが，実際の車載音声認識では，オーディオ装置からの人間の声が議論の対象になることがある．今後は，その他雑音抑圧技術との組み合わせを含め，車内環境のあらゆる雑音に対する頑健な方式を検討したい．

謝辞 本研究では，(社) 情報処理学会 音声言語情報処理研究会雑音下音声認識評価ワーキンググループ提供のデータベース CENSREC-1-C，及び CENSREC-2 を利用した．

参考文献

1) J.Sohn et al, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, pp. 1-3, 1999.
2) Y.D.Cho and A.Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, Vol. 8, No. 10, pp. 276-278, 2001.
3) T.Fukuda, O.Ichikawa, and M.Nishimura, "Phone-duration-dependent long-term dynamic features for stochastic model-based voice activity detection," *Proc. Interspeech*, pp. 2262-2265, 2008.
4) 福田 隆，市川 治，西村雅史，"長時間スペクトル変動情報と調波構造特徴量を併用した発話区間検出法，" 情報処理学会研究報告，2008-SLP-73 (1)，pp.1-6，2008.
5) O.Ichikawa, T.Fukuda, and M.Nishimura, "Local Peak Enhancement Combined with Noise Reduction Algorithms for Robust Automatic Speech Recognition in Automobiles," *Proc. IEEE ICASSP*, pp. 4865-4868, 2008.
6) N.Kitaoka et al, "Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance," *Proc. IEEE Workshop on ASRU*, pp. 607-612, 2007.
7) S.Nakamura et al, "Data collection and evaluation of AURORA-2 Japanese Corpus," *Proc. IEEE Workshop on ASRU*, pp. 619-623, 2003.
8) J.Ramirez, J.C.Segura, C.Benitez, A.Torre, and A.Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, Vol. 42, pp. 271-287, 2004.
9) ITU-T recommendation G.729-Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," 1996.
10) ETSI ES 202 050 recommendation, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm," 2002.
11) S.Nakamura, M.Fujimoto, and K.Takeda, "Censrec2: Corpus and evaluation environments for in car continuous digit speech recognition," *Proc. Interspeech*, pp. 2330-2333, 2006.