2008－SLP－73 (1)
2008／10／24

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

福田 隆　　市川 治　　西村 雅史

日本アイ・ビー・エム株式会社 東京基礎研究所
Email: {fukuda1, ichikaw, nisimura}@jp.ibm.com

あらまし 高精度な発話区間検出（VAD）の実現は，音声認識性能に直結する重要な課題であるが，高騒音下ではいまだ性能が乏しい．本報告では，スペクトルの長時間変動に着目し，低 S/N 環境下における VAD 性能の改善を目指す．提案手法は，平均音素長以上の区間から長時間変動を抽出することにより，検出性能を大幅に改善することを示す．続いて，低 S/N 環境における過剰な湧き出し誤りの削減のため，音声の調波構造に基づく特徴量を VAD システムに導入する．この特徴量は，基本周波数（F0）の明示的な推定を必要とせず，スペクトルの長時間変動情報と併用することで高い性能を実現する．提案システムは，CENSREC-1-C を用いた評価実験において，雑音環境下での性能を顕著に改善し，標準化手法である ETSI AFE-VAD に対して 77.7% の誤り削減を達成した．

Voice activity detection using long-term spectro-temporal information and harmonic structure-based features

Takashi Fukuda, Osamu Ichikawa, and Masafumi Nishimura
Tokyo Research Laboratory, IBM Japan, Ltd.
Email: {fukuda1, ichikaw, nisimura}@jp.ibm.com

Abstract Accurate voice activity detection (VAD) is important for robust automatic speech recognition (ASR) systems. However the VAD system can often fail to detect speech present segments in low S/N environments. This paper first proposes a noise-robust VAD system using long-term temporal information in speech. Long-term temporal information has been an ASR focus recently, but has not been investigated sufficiently for VAD. This paper describes an attempt to incorporate long-term temporal information into a feature parameter set by using a longer window length than average phoneme duration. Next, harmonic structure-based feature extraction is applied to the VAD system in order to reduce false alerts in low S/N environments. The proposed feature extraction doesn't need an explicit fundamental frequency estimation. The VAD system combining long-term features with harmonic structure-based features led to considerable improvements in noisy environments and had 77.7% error reduction as compared to the standardized ETSI AFE-VAD.

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（発話区間検出（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．高精度な VAD の実現は，後続の雑音除去技術等における雑音スペクトルの推定精度を高めるほか，ユーザの発話ではない音声区間をあらかじめ棄却することにより，認識システム全体としての計算負荷を抑えることにも役立つ．音声認識システムによっては明示的な VAD 機構を備えない場合も多いが，そのようなシステムでは無音区間（非発話区間）のスペクトル特性を学習した GMM（Gaussian Mixture Model）や Garbage モデルにより非発話区間の入力音声を吸収している[1]．一方，VAD と音声認識処理が分離しているシステムでは，入力音声全てを音声認識の対象とするのではなく，VAD システムによって得られた発話区間のみを認識対象とするのが一般的である．従って，VAD を前処理として持つ認識システムでは，雑音等の阻害要因によって発話区間の検出に失敗すると正しく音声認識処理が行えないほか，雑音を人間の声と誤って判定した場合，システムは誤動作を引き起こしてしまう．それゆえ，悪条件下でも高精度に動作する方式が望まれる．

従来，VAD ではパワー基準やゼロ交差情報などが広く用いられてきたが，雑音環境においては性能が著しく劣化する問題があった[2]．これを受け，近年では，

統計モデルに基づく VAD 法が提案され，雑音環境下での性能が全体的に向上しつつある[3, 4]．統計モデルとしては GMM が広く用いられ，実際の利用環境で収録した音声からモデルを学習することにより高精度な VAD を実現している．しかし，雑音強度が強くなった場合には，依然として動作が不安定であった．

この問題に対して，雑音環境下における頑健性向上を目標とした音声認識の研究事例[5, 6]を受け，VAD において長時間変動の利用が検討されるようになった[7]．一方，音声知覚の分野では，人間は音声言語を聴取するために，スペクトルの短時間変動（20ms – 40ms）を重要視しているとされてきたが，これに加えて，近年，音声言語聴取にはスペクトルの長時間変動（150ms – 250ms）をも利用している可能性が示唆された[8]．これらの先行研究は，長時間スペクトル変動情報が頑健な VAD を実現する可能性を示している．

他方，VAD の高精度化のために，調波構造に由来する情報の利用も検討されている[9, 10]．調波構造は人間の声に含まれる特有の性質の一つであり，VAD の性能向上に役立つことが期待されている．しかし，調波構造に基づく従来手法の多くは，基本周波数（F0）の検出を前提としており，VAD 性能が F0 の検出性能に依存するという問題があった．

本報告では，まず，統計モデルに基づく VAD において，長時間変動情報を特徴パラメータに組み込むことを検討する．長時間変動は，平均音素長を超える長い窓幅から抽出することにより，高い性能を達成できることを示す．続いて，さらなる頑健性の向上のため，F0 の検出を必要としない調波構造特徴量の抽出法を提案し，長時間変動情報と併用するシステムを検証する．評価実験では，自動車内雑音環境を対象に，特徴パラメータの比較，及び標準化手法との比較を行う．

本報告は，以下のように構成される．2. で長時間変動情報について説明した後，3. で評価実験を実施する．続いて，4. で調波構造特徴量の抽出方法を示すと同時に，評価結果を述べ，5. で結論をまとめる．

2. 長時間スペクトル変動

2.1 長時間スペクトル変動成分の利用

音声のスペクトル変動情報は，動的特徴量として音声認識の分野で長らく用いられてきた[11]．VADにおいても，動的特徴量は性能の改善に寄与することが示されており，通常，MFCCと組み合わせて利用される．一般的に，動的特徴量 d_t は，次式に示すスペクトルの時間軌跡に対する線形回帰演算によって計算される．

$$d_t = \sum_{k=1}^{K} \left\{ k \cdot \left(c_{t+k} - c_{t-k} \right) \right\} / 2 \sum_{k=1}^{K} k^2 \qquad (1)$$

ここで c_t は時刻 t におけるケプストラム係数，Kは分析窓長（前後Kフレームを利用）を表す．音声認識では，個々の音韻をモデル化するという観点からK = 2～3，すなわち合計 5～7 フレームが用いられることが多く，この知見を基づいてVADでもK = 2～3 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，本報告では，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，Long Δ Cepと略す）をVADに利用する．便宜的に，K = 3 で求められる従来の動的特徴量を Short Δ Cepと呼ぶことにする．

音声及び非音声の識別には GMM を用いる．識別部では，次式に示すように，両モデルが出力する対数尤度比 $L(x)$ について閾値判定をすることによって音声／非音声を決定する．

$$L(x) = \log P\left(x \mid \Lambda_{Sp}\right) - \log P\left(x \mid \Lambda_{Sil}\right) \qquad (2)$$

ここで，Λ_{Sp}，Λ_{Sil} は音声，非音声のGMMである．

2.2 対数尤度比の分布

本節では，長時間スペクトル変動の効果を対数尤度比の観点から考察する．図 1 に，音声区間のみ（Speech）と非音声区間（Silence）のみを入力とした場合の尤度比 $L(x)$ の分布図を示す．図では，特徴量として MFCC 単独と Short Δ Cep 単独（K = 3），および Long Δ Cep 単

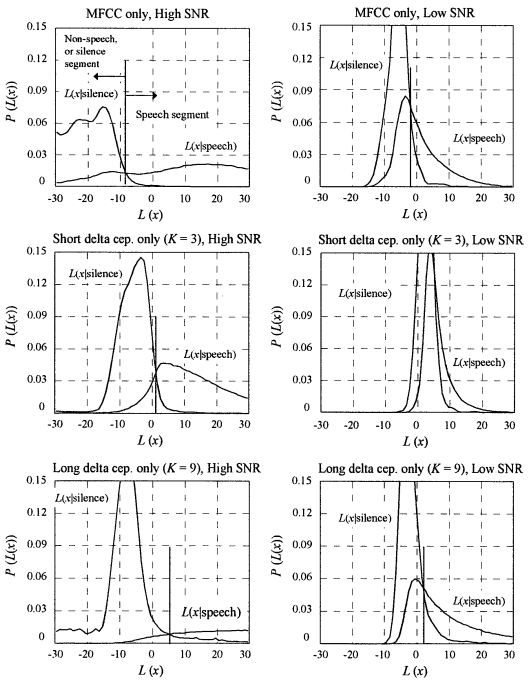

図 1 各特徴量による尤度比分布

独（K = 9）を比較している．図の左側は High SNR（Clean，20dB，15dB，10dB の平均）の分布を，右側は Low SNR（5dB，0dB，-5dB の平均）の分布を示している．VAD の観点から，音声／非音声区間の尤度比分布は完全に分離していることが理想であり，両分布の重なり部分が小さいほど識別誤りが少ないといえる．図から，長時間スペクトル変動情報を保有する Long Δ Cep は，MFCC や Short Δ Cep と比較して分布の重なり部分が小さいことがわかる．MFCC や short Δ Cep は，特徴量として子音と雑音の区別がつきにくい一方，長時間スペクトル変動情報には母音区間も含まれるため，高騒音下においても高い識別能力を持つ可能性が示唆される．

2.3 フィルタ処理としての動的特徴抽出

スペクトルの時間変動成分を周波数次元で表したものは変調スペクトルと呼ばれる．式 (1) に示す線形回帰演算は，スペクトルの時間軌跡に対するフィルタ処理と等価であり，変調スペクトル成分を強調する処理と見なすことができる．図 2 は短時間窓（K=3，計 7 フレーム）及び長時間窓（K=8，計 17 フレーム）の線形回帰演算による周波数応答を示している．図から，短時間窓の線形回帰演算は 10Hz 付近の変調スペクトルを強調する一方，長時間窓の線形回帰演算は 2Hz 付近の変調スペクトルを強調していることがわかる．文献[8]において，Poeppel らは人間が音声言語を理解する際に，スペクトルの短時間変動，及び長時間変動の双方を利用していることを心理実験により導いた．この知見に基づき，我々は短／長時間変動特徴量の組み合わせが音声認識の性能改善に寄与することを示している[12]．VAD においては，長時間変動特徴量がゆるやかに変化するスペクトル変動を捉えると共に，雑音環境でよく観測される変調周波数の高い変動成分を効果的に抑制することにより，頑健な VAD を実現することを意味している．

3. 評 価 実 験

3.1 実 験 概 要

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 20dB～－5dB の間で 5dB 刻みに重畳されている．本実験で利用する評価データは男女各 52 名による 6986 発声であり，発話内容は連続数字である．サンプリング周波数は 8kHz である．フレームサイズおよびシフト幅はそれぞれ 25ms と 10ms とし，入力音声はフレーム毎に

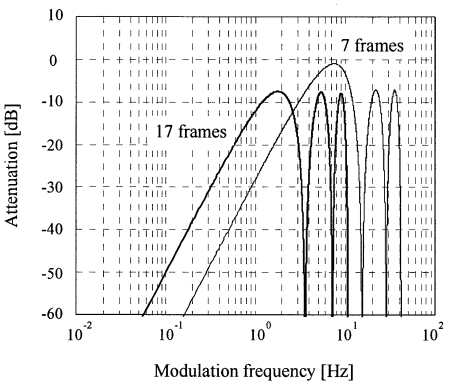

図 2 線形回帰演算の変調周波数応答

$1-0.97z^{-1}$の高域強調を行った．そして，ハミング窓掛け処理と 24 チャネルのメルフィルタバンク分析を行った後，12 次元のMFCCを抽出し，式 (1) を用いてΔケプストラムを求めた．ただし，CMNは行っていない．特徴量にはパワー項も含めている．

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）．学習データ数は，男女各 55 名による 1668 発話である．混合数は音声／非音声 GMM 共に 32 とした．VAD の評価は発話単位で正解／不正解を判定する方法を用い，正解率と正解精度により各特徴量の性能を比較する．

3.2 実 験 結 果

まず，長時間スペクトル変動情報の効果を検証するため，窓長の違いによる性能の推移を調査した．図 3 に実験結果を示す．図には，全 S/N 比における VAD 結果の平均値をプロットしている．実験結果から，K = 3 以下では性能が低く，MFCC 単独よりも劣っていることがわかる．一方，窓長を K = 4 以上にすると MFCC よりも高い性能が得られ，正解精度を基準に見ると，K = 10 のときに最も高い性能 82.7% を達成した．

次に特徴量の違いによる性能と，従来手法との比較結果を示す．従来手法としては，ITU-T の標準化手法である G.729 Annex B[13]，及び ETSI（European Telecommunication Standards Institute）の標準化手法である ETSI AFE-VAD[14], そして長時間変動に着目した方法である LTSD（long-term spectrum divergence）[7] を用いた．表 1 に実験結果を示す．表において，High／Low SNR の定義は 2.2 節と同様であり，括弧内の数値は特徴量の次元数を表している．ここで Short Δ Cep は窓長 K = 3 から，Long Δ Cep は窓長 K = 10 から求めている．まず，従来手法の間で比較を行うと，G.729

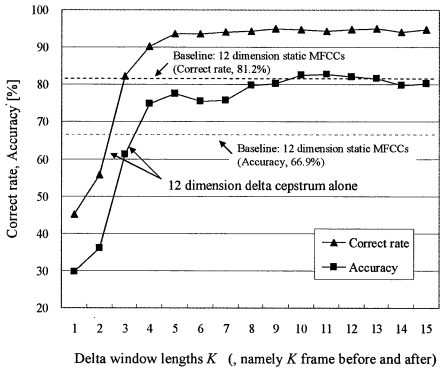

図 3 分析フレーム長による性能の推移

表 1 各種手法との比較

Conventional methods (top), Feature parameters (middle, bottom)	Correct rate [%]		
	High SNR	Low SNR	Average
G.729 Annex B	92.9	49.5	74.3
ETSI AFE-VAD (ES 202 050)	87.3	79.4	83.4
Long-term spectral divergence (LTSD)	96.5	56.4	79.3
MFCC (13 dim.)	94.6	63.2	81.2
Short-term ΔCep. (13 dim. K=3)	93.9	66.8	82.3
Long-term ΔCep. (13 dim. K=10)	99.7	88.1	94.7
MFCC + ShortΔCep. (26 dim. K=3)	99.1	82.9	92.2
MFCC + LongΔCep. (26 dim. K=10)	99.7	89.1	95.2
Conventional methods (top), Feature parameters (middle, bottom)	Accuracy [%]		
	High SNR	Low SNR	Average
G.729 Annex B	88.6	20.1	59.2
ETSI AFE-VAD (ES 202 050)	69.8	32.0	50.9
Long-term spectral divergence (LTSD)	88.7	22.6	60.3
MFCC (13 dim.)	90.5	35.5	66.9
Short-term ΔCep. (13 dim. K=3)	86.5	27.5	61.2
Long-term ΔCep. (13 dim. K=10)	97.8	62.4	82.7
MFCC + ShortΔCep. (26 dim. K=3)	96.3	58.3	80.0
MFCC + LongΔCep. (26 dim. K=10)	97.2	68.2	84.8

Annex B と LTSD-VAD は Low SNR 環境において低い性能にとどまる一方, ETSI AFE-VAD は Low SNR 環境でも比較的高い性能を示した．しかし, ETSI AFE-VAD は High SNR では性能が劣る結果となった．

続いて，GMM-VAD において 13 次元特徴量を比較すると，Long Δ Cep は High／Low SNR 双方の環境で，MFCC や Short Δ Cep と比較して顕著に性能を改善していることが見てとれる．通常，音声認識や VAD で Δ ケプストラムが単独で利用されることはまれであるが，実験結果からもわかるように Long Δ Cep は単独でも性能改善に大きく貢献する．26 次元特徴量の比較においては，音声認識でもよく使われる“MFCC + Short Δ Cep”の組み合わせが，MFCC 単独よりも高い性能となった．分析窓長が短いとはいえ，“MFCC + Short Δ Cep”は時間変動成分を含んでいるため，改善が得られたと考えられる．しかし，Short Δ Cep の代わりに，今回提案の Long Δ Cep を併用することでさらに高い

表 2 雑音環境の違いによる比較

Feature Parameter	Correct rate [%]			
	Subway	Babble	Car	Exhibition
MFCC + ShortΔCep. (26 dim.)	92.2	85.4	92.2	93.8
MFCC + LongΔCep. (26 dim.)	95.8	88.0	95.2	96.1
Feature Parameter	Accuracy [%]			
	Subway	Babble	Car	Exhibition
MFCC + ShortΔCep. (26 dim.)	77.1	65.9	80.0	78.5
MFCC + LongΔCep. (26 dim.)	83.7	67.4	84.8	82.0

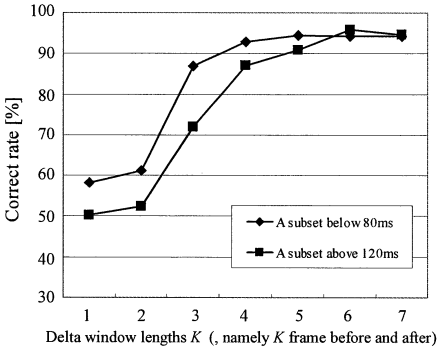

図 4 話速と窓長の関係

性能を実現することができた．これは，長時間スペクトル変動成分が湧き出し誤りを効果的に抑えつつ，正確に発話区間を検出したためと言える．

表 2 にその他環境下（subway, babble, subway）での VAD 性能を示す．Long Δ Cep を用いた提案システムは全ての雑音環境で Short Δ Cep を併用した従来手法と比較して高い性能を示した．しかしながら，人ごみ雑音（Babble）のような目的話者以外の発声が入力音声に混入するような環境では改善率が低くなってしまう．このような環境では，マイクロフォンアレイなど，その他技術との併用が望ましい．

3.3 考　察

ここでは窓長と話速の関係を考察する．まず，全評価データの平均音素長を調べたところ 98.6 ms であった．平均音素長はコーパス中の正解発話区間情報を利用し，平均音素長 ＝ 発話時間／音素数とした．図 3 を見ると，K = 5（シフト幅 10ms の条件で，時間にして 100ms の分析窓長）以上で平均して高い性能が得られており，この時間区間は概ね平均音素長に対応している．これをうけ，平均音素長が 80 ms 以下の発話と，120ms 以上の発話のみでそれぞれ評価セットを再構成し，実験を行った．特徴量は Δ ケプストラム単独である．図 4 に結果を示す．図に示すように 80ms 以下の評価セットの場合は K = 4, 120ms 以上の評価セットの場合は K = 6 で性能の上限値に近づいており，最低限必要な分

析窓長と平均音素長の関係が一致している．この事実は，話速が遅い発話に関しては，窓長を十分に取る必要があり，平均音素長以上の区間からスペクトル変動を求めることの重要性を意味している．

4. 調波構造特徴量

4.1 調波構造の利用

前節までに，長時間スペクトル変動情報が VAD 性能改善に大きく貢献することを示した．しかしながら，正解精度を基準に見ると，最も高い性能を示した“MFCC + Long Δ Cep”でも 84.8 % の精度であり，依然として改善の余地がある．本節では，さならる性能改善を目指して，音声の調波構造の利用を試みる．

調波構造に注目した従来手法は，有声音／無声音の判定と，正確な F0 の検出を前提としていることが多く，F0 の推定が難しい環境では動作が不安定になることがあった[9, 10]．これまでに我々は，音声認識性能向上のために，F0 の検出を陽に行わず，観測音声そのものから音声強調フィルタを推定し，低 S/N 条件での認識性能を改善する方式（LPE : Local Peak Enhancement）を提案している[15]．この手法は，入力音声の対数パワースペクトルにおけるケプストラム表現の一部が，調波構造情報を保持していることに基づいている．設計されたフィルタは，有声音区間では調波構造部分に重みのある特性となり，調波構造がない無声音区間ではフラットに近い性質を示す．この重み係数は，有声音区間と無声音区間で異なる性質を示すため，音声／非音声識別のための有用な情報になり得る．本報告では，LPE フィルタをメルケプストラムに変換した後，VAD のための特徴量として利用する．この特徴量を LPW メルケプストラムと呼ぶことにする．

4.2 LPW メルケプストラム

図 5 は提案手法の概要である．処理手順を以下に示す．

I. 処理フレーム毎に，観測音声の対数パワースペクトル $X_T(j)$ を得る．T はフレーム番号，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．

$$P_T(i)=\sum_j \hat{M}(i,j)\cdot X_T(j) \qquad (1)$$

ここで，$\hat{M}(i, j)$ 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．

$$Q_T(i)=\varepsilon\cdot P_T(i) \quad \text{if } (i<D_L) \text{ or } (i>D_H)$$
$$Q_T(i)=P_T(i) \quad \text{otherwise} \qquad (2)$$

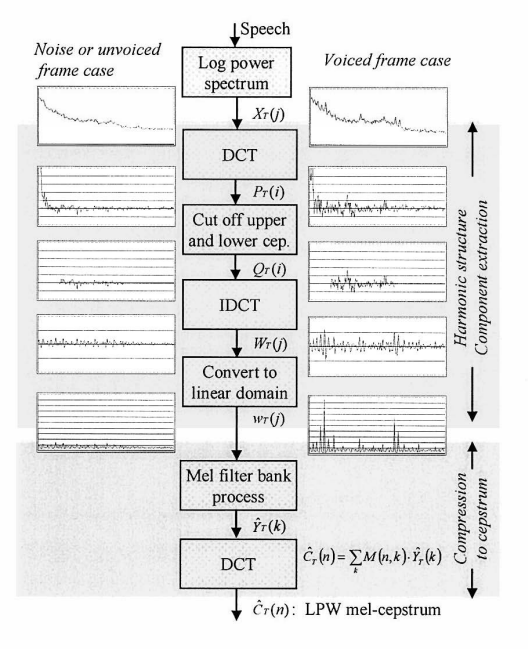

図 5 調波構造特徴量の抽出

ここで，ε 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る範囲に対応させる．

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．

$$W_T(j)=\sum_i \hat{M}^{-1}(j,i)\,Q_T(i) \qquad (3)$$

$$w_T(j)=\exp(W_T(j)) \qquad (4)$$

なお，LPE法では $w_T(j)$ を音声強調フィルタと見なし，パワースペクトル $x_T(j)$ を強調していた．

V. $w_T(j)$ を入力とするメルフィルタバンク処理を行い，その出力 $\hat{Y}_T(k)$ をケプストラム $\hat{C}_T(n)$ に変換する．

$$\hat{C}_T(n)=\sum_k M(n,k)\cdot\hat{Y}_T(k) \qquad (5)$$

得られた $\hat{C}_T(n)$ をLPWメルケプストラムと呼ぶ．

4.3 実験概要

図 6 に提案システムの概要を示す．フレームサイズおよびシフト幅をそれぞれ 25ms と 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 LPW メルケプストラムを抽出する．これと並行して，フレーム毎に 12 次元の（通常の）MFCC 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

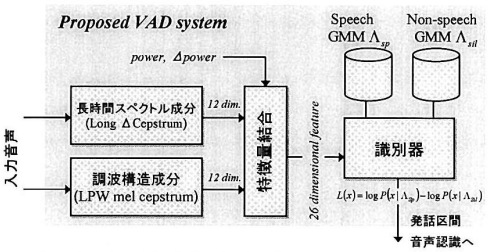

図 6 調波構造特徴量を組み込んだ VAD システム

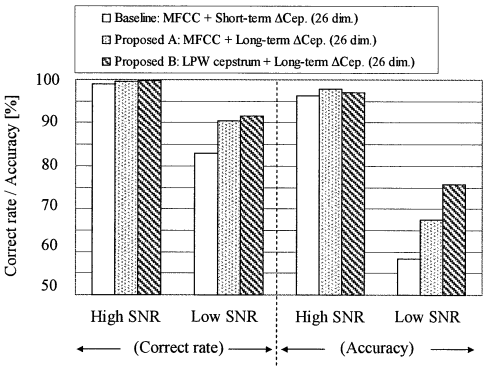

図 7 LPW メルケプストラムの効果

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．評価データおよび分析条件は 3 節と同様である．

4.4 実 験 結 果

図 7 に実験結果を示す．ここでは，MFCC と ShortΔCep を結合した 26 次元の特徴量を Baseline とし，“MFCC + LongΔCep”（Proposed A）および，“LPW メルケプストラム + Long Δ Cep”（Proposed B）の組み合わせを比較した．図中の High／Low SNR の定義は 2.2 節と同様である．Proposed B を見ると，特に Low SNR 環境での正解精度（Accuracy, Low SNR）改善に関して貢献が大きく，Proposed A からは 23.6 % の誤り削減であり，Baseline と比べると 41.7% の誤り削減となった．LPW メルケプストラムは Long Δ Cep 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．前節で比較対象とした標準化手法（G.729 annex B，ETSI AFE-VAD）と比べると，ETSI AFE-VAD から 77.7%，G.729 Annex B からは 85.6%もの誤り削減を達成した．なお，LPW メルケプストラム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．

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 を利用した．

参 考 文 献

[1] C. Torre and A. Acero, “Discriminative training of garbage model for non-vocabulary utterance rejection,” *Proc. ICSLP '92*, pp.9-12 (1992).

[2] S. V. Gerven and F. Xie, “A comparative study of speech detection methods,” *Proc. Eurospeech '97*, vol. III, pp.1095-1098 (1997).

[3] J. Sohn, N. S. Kim, W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, Vol. 6, pp. 1-3 (1999).

[4] 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).

[5] H. Hermansky and S. Sharma, “TRAPS – Classifiers of Temporal Patterns”, *Proc. ICASSP '99*, Vol. I, pp. 289-292 (1999).

[6] B. Chen, Q. Zhu, and N. Morgan, “Learning long-term temporal features in LVCSR using neural networks”, *Proc. ICSLP*, pp. 612-615 (2004).

[7] 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).

[8] D. Poeppel, “The analysis of speech in different temporal integration windows: cerebral lateralization as asymmetric sampling in time,” *Speech Communication*, Vol. 41, pp. 245-255 (2003).

[9] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, “Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio,” *Interspeech'07*, pp. 230-233 (2007).

[10] Y. Guo, Q. Fu, and Y. Yan, “Robust voice activity detection based on adaptive sub-band energy sequence analysis and harmonic detection,” *Interspeech'07*, pp.2949-2952 (2007).

[11] S. Furui, “Speaker-independent isolated word recognition using dynamic features of speech spectrum”, *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. ASSP-34, No. 1, pp. 52-59 (1986).

[12] T. Fukuda, O. Ichikawa, and M. Nishimura, “Short- and Long-term Dynamic Features for Robust Speech Recognition,” *Interspeech 2008*, pp.2262-2265 (2008).

[13] ITU-T recommendation G.729-Annex B, “A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70,” (1996).

[14] ETSI ES 202 050 recommendation, “Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm” (2002).

[15] O. Ichikawa et al., “Local peak enhancement for in-car speech recognition in noisy environment,” 信学論, Vol. E91-D, No.3, pp. 635-639 (2008)