社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59（35）
2005／12／22

帯域フィルタ出力の時間変化特徴量を利用したニュース音声認識

尾上 和穂　佐藤 庄衛　小林 彰夫　本間 真一　今井 亨

ＮＨＫ放送技術研究所 〒187-8510 東京都世田谷区砧 1-10-11

E-mail: {onoe.k-ec, satou.s-gu, koabayashi.a-fs, homma.s-fc, imai.t-mq}@nhk.or.jp

あらまし　ニュース番組の自動字幕化のための音声認識では、雑音の混入した音声や対談調の音声の対策が課題となっている。筆者らは、人間の音声知覚に関する知見を考慮し、認識に重要とされる様々な音響特徴量を組み合わせることで、より頑健な音声認識を目指している。これまで、帯域フィルタ出力の時間変化量に着目した特徴量（BAnd-pass filtered outputs' Temporal property feature: BAT 特徴量）を提案している。BAT 特徴量は、知覚に重要な時間方向の変動成分だけを帯域ごとに独立して抽出することで、別の帯域に混入した雑音の影響などの軽減が期待される。本稿では、BAT 特徴量の最適な分析パラメータと、主成分分析や判別分析による次元数の削減効果を報告する。ニュース番組中の中継と対談調の音声の認識実験の結果、帯域の分割数 15、時間方向の抽出窓幅 150msec が最適であり、主成分分析による次元数削減によって従来特徴量と同等の認識精度が得られた。さらに、次元数を削減した BAT 特徴量と MFCC や PLP を組み合わせることで、従来特徴量に対して最大で 10.6%の誤り削減率が得られた。

キーワード　時間変化特性、バンドパスフィルタ、音声認識、音響特徴量、主成分分析、判別分析

Temporal Properties of Band-Pass Outputs for News Speech Recognition

Kazuo ONOE　Shoei SATO　Akio KOBAYASHI　Shinichi HOMMA　and　Toru IMAI

Science and Technical Research Laboratories Japan Broadcasting Corporation　1-10-11 kinuta, setagaya-ku, Tokyo, 187-8510 Japan

E-mail: {onoe.k-ec, satou.s-gu, koabayashi.a-fs, homma.s-fc, imai.t-mq}@nhk.or.jp

Abstract　Speech recognition with noisy background or spontaneous speaking style in news programs is an important issue for simultaneous closed-captioning. Considering knowledge about human's auditory perception, we are investigating an acoustic feature of temporal properties of band-pass outputs, which we call a BAT feature (Band-pass filtered outputs' Temporal property feature). Since it extracts perceptually important temporal components in each band-pass independently, less influence of noise in other bands is expected. This paper describes its optimum analysis parameters, effect of dimension reduction by the principal component analysis or the linear discriminative analysis, and its combination with a conventional acoustic feature. In recognition experiments of field reports and conversational speech in news programs, the proposed feature showed the best recognition accuracy with 15 bands and 150ms of a temporal window. The proposed feature with the principal component analysis gave the same recognition accuracy as a conventional feature and their combination yielded the maximum word error reduction rate of 10.6%

Keyword　Temporal Properties,　Band-Pass Filter,　Speech Recognition,　Acoustic Feature,　PCA,　LDA,　HDA

1. はじめに

近年、ニュース番組の自動字幕化[1]など、音声認識技術を用いたアプリケーションが多数実用化されている。しかし、雑音下の音声や発話スタイルが対談や講演調の音声に対して、音声認識精度が十分でなく、音声認識の実用化範囲は限定されている。一般に、雑音や発話スタイルによる認識精度の劣化に対して、信号レベルではスペクトルサブトラクションやカルマンフィルター、音響モデルでは適応化やモデル合成など様々な手法が提案されている。また、音響特徴量に関しても RASTA[2]や調音スペクトル[3]など多くの手法が提案されている。しかし、学習データとのマッチングの問題や、認識タスクによって効果が異なるなど、どの手法にも一長一短があり、十分とは言えない。

一方で、人間の場合を考えると、発話環境（雑音）やタスク（発話スタイル等）に依存することなく、ほとんどの場合問題なくコミュニケーション（認識）できる。これは、雑音に埋もれた音声や複数話者の同時発話など、音声の様々な部分が妨害音声によって劣化した場合にも、人間は劣化の影響の少ない様々な特徴量をうまく組み合わせて認識しているのが理由だと考えられる。そこで、人間と同様に、音声認識で

も複数の異なる特徴量を組み合わせることで、より頑健な音声認識が期待される。

音声知覚に関する種々の研究報告を参考にすると、周波数帯域ごとの時間変化特性にも知覚にとって重要な特徴量が存在する可能性がある。これまで、我々は帯域フィルタ出力の時間変化特性を特徴量（BAnd-pass filtered outputs' Temporal property feature : BAT 特徴量）として提案し評価している[18]。本稿では、第 2 章で音声知覚に関する知見について説明し、第 3 章で提案特徴量の分析手法と主成分分析や判別分析を用いた提案特徴量の次元数削減、MFCC・PLP を用いた従来特徴量との組み合わせについて説明する。第 4 章では、テストセットの詳細な説明と従来特徴量によるベースラインの評価、そして、第 5 章で提案特徴量の実験結果について報告する。

2. 人間の音声知覚

一般に聴覚器官が周波数分析器であることから、音声の知覚については、スペクトルの包絡形状（フォルマント構造）に関して議論する場合が多い。音声認識の分野では、MFCC や PLP を代表とする特徴量が用いられ、これらは基本的にフォルマント構造を数値化して表現している。しかし、フォルマント構造を全帯域にわたって表現する特徴量は、ある帯域だけに雑音が混入しても、1 次差分、2 次差分を含めて、すべての特徴量に雑音の影響がおよんでしまう問題がある。

フォルマント以外の知覚にとって重要な特徴量について、過去に様々な分野で調査・研究の報告がある。例えば、フォルマント等の周波数方向の特徴ではなく、時間方向の変化量に着目した研究が上げられる。脳内機能の研究[4]では、人間の音声知覚には、異なる 2 種類の時間分解能があり、それぞれ独立して知覚が行われていると報告している。2 種類の時間分解能の一つは、フォルマント等のスペクトルの早い変化を識別する数 10msec 程度の時間分解能で、もう一つは母音や子音を識別する 150~250msec 程度の時間分解能であるとの見解である。Rosen[5]らは、人工蝸牛の移植された人を対象に、周波数分析していない音声の時間波形を直接刺激として人工蝸牛に与える知覚実験を行った。実験の結果、この場合でも知覚が十分に可能で、音声知覚のための時間分解能が複数あると報告している。また、調音機構の研究[6]では、舌の動きは早くても 12Hz 以下であるとの発表がある。同様に聴覚器官の研究では、蝸牛の発火応答速度は、数 10Hz 以下であるとの研究報告がある。音声認識の分野では、このような時間変化特性を特徴量とした Hermansky[7][17]の研究や、重要な時間変化特性をスペクトルの段階で強調する聴音スペクトル手法[3]などが提案されている。

フォルマント構造が壊れた音声の知覚実験について着目すると、その研究の一つに、狭帯域フィルタを適用した音声の知覚実験があり、このような音声に対しても人間は頑健に認識できることがわかっている。[8]によれば、狭帯域フィルタを通した実験では、中心周波数が 0.8ー4kHz ではほぼ劣化なく音声を知覚でき、また、認識精度の良く無かった低い周波数と高い周波数を組み合わせることで、元音声と同等の認識精度を得られることが実験により明らかにされている。また、電話回線で音声を劣化なく伝送するために、音声知覚の実験に取り組んでいた Fletcher[9]は、“音声知覚は、周波数帯域ごとに認識し最終的に統合している”と結論づけている。音声認識の分野でも、これをモチベーションにマルチストリーム[10]の音声認識などが提案されている。

3. BAT 特徴量

3.1. BAT 特徴量の分析法

我々の提案する BAT 特徴量は、知覚にとって重要な信号の変動成分を帯域ごとに抽出し、数値化するものである。

図 1 に、提案する特徴量のフローチャートを示し、以下に、分析手順を説明する。

1） 入力音声を短時間周波数分析し、n 個の帯域にする。今回は対数振幅スペクトルによる n チャンネルのメルフィルタバンク出力を用いた。

2） それぞれの帯域で、処理時刻 t を中心として、時間方向に窓幅 W の信号を抽出する。同様にして、対数パワーからも信号を抽出する。

3） 2)で抽出した各帯域出力の平均値を除去し、中心時刻 t に重みを与えるためにハミング窓を適用する。

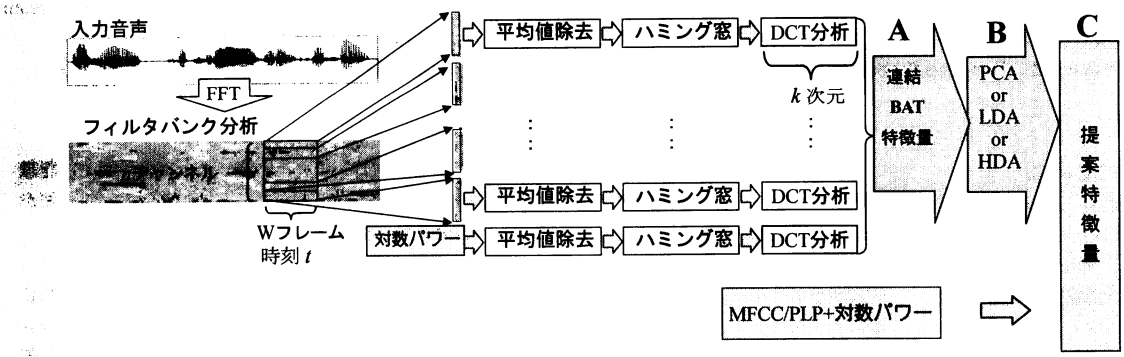

図 1.BAT 特徴量分析フローチャート

4） 3)で得られた各帯域の信号に対して DCT 分析を行い、1 次から k 次までを変動成分を表現する特徴量として出力する。なお、k は時間変化の表現できる周波数の上限値によって決定することができ、時間方向の窓幅 W によって変化する。今回のすべての実験では、先行研究を参考に、24Hz を表現する次数を上限値として用いて、時間方向の窓幅 W ごとに k を決定した。また、時間変化量のみを特徴量として抽出することを目的にしているため、直流成分の 0 次は、今回の実験では用いない。

5） 各帯域（n 分割）で得られた DCT 係数（k 個）を連結し、提案する特徴量（BAT 特徴量：(n+1)×k 次元）とする。（図１中のA）

6） 提案する特徴量に対して主成分分析(PCA)や判別分析(LDA/HDA)を行い、次元圧縮と対角化によって m 次元の特徴量（BAT-PCA/LDA/HDA 特徴量）を得る。（図１中のB）

7） 6)で得られた特徴量（BAT-PCA/LDA/HDA 特徴量）と従来特徴量（RASTA-MFCC、RASTA-PLP）を連結し、一つの特徴量とする。（図１中のC）

BAT 特徴量は、帯域ごとに特徴量を求めているため、別の帯域に混入した雑音の影響を受けないと想定される。ただし、帯域分割数 n と時間方向の抽出窓幅 W に比例して総次元数が増加するという問題がある。今回は BAT 特徴量の次元数削減による改善効果を検討し、最後に従来特徴量との組み合わせについて調査する。

3.2. 次元数削減と無相関化の検討

BAT 特徴量は、従来特徴量に比べ次元数が多く、学習データ量が問題となる。また、隣り合う帯域で高い相関がありメルフィルタのバンク間のオーバーラップによって、さらに相関が高くなっている。次元間の相関が高い特徴量は、次元間の無相関を仮定して対角成分のみの分散を扱う音響モデルとの整合性が低いという問題がある。

この問題の解決策として、提案特徴量の次元数削減と無相関化の検討を行う。削減手法として、主成分分析(PCA: Principal Component Analysis)、線型判別分析(LDA: Linear Discriminant Analysis)、不等分散性判別分析(HDA: Heteroscedastic Discriminant Analysis)について検討する。

3.2.1. 判別分析(LDA/HDA)[13]による次元数削減

LDA と HDA は、識別するクラス間とクラス内の分布を事前知識として用いることで、次元数削減による改善が期待できる。

クラス C_j (j=1.2,.....J)ごとの平均 μ_j と分散 Σ_j をそれぞれ

$$\mu_j = \frac{1}{N_j}\sum_{x\in C_j} x$$

$$\Sigma_j = \frac{1}{N_j}\sum_{x\in C_j}(x-\mu_j)(x-\mu_j)^T$$

とすると、クラス内分散 W は、

$$W = \frac{1}{N}\sum_{j=1}^{J} N_j \Sigma_j$$

クラス間の分散 B は、

$$B = \frac{1}{N}\sum_{j=1}^{J} N_j (\mu_j - \mu)(\mu_j - \mu)^T$$

となる。ただし、μ は全体の平均、J はクラス数、N と N_j はサンプル総数である。

LDA は、クラス間の分散 B を大きくして、次のようにクラス内の分散 W を小さくする軸に変換する。

$$\hat{\theta} = \arg\max_{\theta} \frac{\left|\theta^T B \theta\right|}{\left|\theta^T W \theta\right|}$$

HDA も同様に、次式でクラス内の分散 Σ_j を小さくし、クラス間の分散 B を最大にするような変換行列 $\hat{\theta}$ を求める手法である。

$$\hat{\theta} = \arg\max_{\theta} \frac{\left|\theta^T B \theta\right|^N}{\prod_{j=1}^{J}\left|\theta^T \Sigma_j \theta\right|^{N_j}}$$

これは、LDA での変換後の各クラスの分散が同じになってしまう制約を無くしたものである。

また、LDA と HDA では、PCA の様に変換後の軸の直交化の保証が無いので、MLLT[14][15]による直交化を行う。

今回は、事前知識としてトライフォン HMM の各状態をクラスとし、学習時の occupation counts をサンプル数とする。

3.3. 従来特徴量との組み合わせによる改善の検討

周波数方向の形状を表す MFCC や PLP と、時間方向の変化量を表す BAT 特徴量は、特徴量の間に相関が低いと想定できるため、これらを連結することで頑健な特徴量を構成できる可能性がある。また、MFCC や PLP の従来特徴量の１次差分と２次差分をより詳細に表現したのが BAT 特徴量だと考えることもでき、認識率の向上が期待できる。

RASTA-MFCC、RASTA-PLP の従来特徴量と BAT-PCA、BAT-LDA、BAT-HDA の提案特徴量を組み合わせた６種類の特徴量について、評価を行う。なお、従来特徴量の１次差分と２次差分は時間方向の特徴であるため、組み合わせ時には用いていない。また、RASTA-MFCC では１次差分と２次差分のために従来行っていたリフター処理も、BAT 特徴量と組み合わせる場合には行わない。

4. ベースラインの評価実験

ニュース番組のうち、中継と対談の認識精度の低い音声を選択しテストセットとする。表１にテストセットに使用した

表1.テストセット

対象ニュース番組	NHK おはよう日本6・7時 NHK 昼ニュース NHK ニュース7		
番組放送日	2001.06.01-06.14		
発話内容・環境	対談・中継		
文章数(総単語数)	459 文 (10758 語)	中継	396 文
		対談	170 文
		中継かつ対談	107 文

表2.テストセットの SNR（dB）

	全体	中継	対談	中継かつ対談
SNR	26.4	26.2	26.3	25.7

表3.テストセットの SNR ごと文章数

SNR	0～10dB	10～20dB	20～30dB	30～40dB
文章数	5	52	297	105

表4.テストセットの発話内容

今 の 状況 を 池 に なって お伝え し ます と ね みのも はま 時折 弱い 風 で 揺ら さ れる ぐらい で いたって 平穏な 朝 じゃ
けさ の 気温 十 一 度 な んです が やっぱり ちょっと 上 に 一 枚 羽織ら ない と 涼しい かな って いう 感じ です よ ねえ
学校 で は お金持ち に なる 方法 なんて 教えて くれ ません けれ ども やっぱり そういう こと が 必要な んでしょう か
眼鏡 とって アジサイ の 色 を 確かめ る

表5.音響モデル学習用データ

放送番組	NHK おはよう日本6・7時 NHK 昼ニュース NHK ニュース7
放送期間	2001.01.01-2001.05.31
データ量	約 200 時間
使用音声	番組内の雑音を含む音声全て

表6.音声認識装置[1]と言語モデル

2 パス デコーダ	1 パス：bi-gram による Viterbi サーチ 2 パス：tri-gram リスコア
言語モデル	放送日に適応、語彙 60k bi-gram と tri-gram

表7.音響分析条件

音声データ	16KHz, 16bit
分析窓幅，シフト長	25msec, 10msec
窓掛け	ハミング
プリエンファシス	0.97

表8.MFCC・PLP パラメータ

従来の MFCC 特徴量	39 次元 (12 次 MFCC+対数パワー+Δ+ΔΔ)
従来の PLP 特徴量[16]	39 次元 (12 次 PLP+対数パワー+Δ+ΔΔ)
音響モデル	64 混合・3 状態 (left to right)

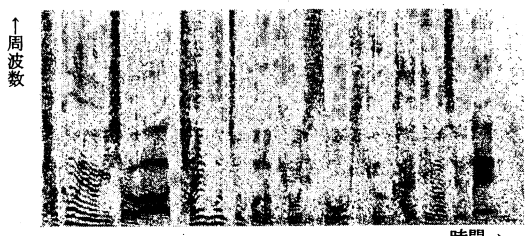

図2.評価音声のスペクトル

番組の詳細と文章数を示す。対象とした音声は、海外からの中継など、街頭騒音等の雑音がミックスされた音声や、スタジオ内での対談調の音声である。話題も経済問題から天気予報などの時事、海外からの突発事故など様々である。発話者は、中継部分は記者がほとんどで、対談部分では、アナウンサーと記者が約半分ずつの割合である。

次に、評価音声の音響的な特徴について述べる。図2.に、テストセット中の SNR が約 18dB の音声スペクトルを示す。この例は、ニュース番組中のヘリコプター中継の音声であり、雑音の混入で無音区間が埋まっており、認識はかなり困難である。NIST が提供する SPQA[11]の SNR 測定手法によってテストセットの SNR を調査したところ、テストセット全体で 26.4dB であった。また、各分類ごとの SNR を表2に示す。SNR の観点から見ると今回のテストセットでは、“中継かつ対談”が SNR25.7dB と一番認識が難しい分類となる。文章ごとに SNR を調査すると、数 dB 台から 40dB 近くのスタジオ環境のものまで、様々な条件の音声が含まれている。表3に SNR ごとの文章数割合を示す。

また、表 4 に対談調の発話の書き起こしを数例示し、言語的な側面を考察する。この発話内容の例は、番組“おはよう日本”内のお天気中継箇所などで、一般的なスタジオでのニュース原稿の読み上げと比べ、話題が多様で、発話スタイルも対談調に近いものが多く見受けられる。一般に原稿読み上げのパープレキシティーは約 10 程度であるが、今回のテストセットでは 62.2 と、言語的にも認識の難易度が高い。

4.1. 学習データと実験条件

今回の実験で用いた学習データは、テストセットの音声と同様に、ニュース番組中の男性話者（アナウンサー／記者）による原稿読み上げ音声と対談調の音声である（表 5）。また、音響環境においては、中継現場からの雑音の混入した音声なども含まれている。また、言語モデルについては、評価音声の番組ごとに、ニュース原稿によって適応した言語モデル[12]を用いて評価を行った。

その他の実験条件を、表6～表8に示す。

また、今回の認識実験では、実験ごとに特徴量や次元数が変わり、音響モデルの尤度出力値やレンジが異なるため、それぞれ認識処理時間が実時間の約 10 数倍程度になるようビーム幅を設定し、特徴量ごとに最適な言語重みを予備実験によって求めた。

表 9. 従来特徴量での評価（ベースライン単語正解精度:%）

従来特徴量	MFCC		PLP	
次元数	39		39	
RASTA	無し	あり	無し	あり
中継	**90.2**	91.0	90.0	**91.1**
対談	**76.3**	77.9	**76.3**	**78.3**
中継且つ対談	**77.5**	78.8	77.1	**79.6**
全体	**88.6**	89.5	88.5	**89.6**

4.2. ベースラインの実験結果

従来特徴量として MFCC と PLP、また、それぞれに RASTA を適用した 4 種類の特徴量による実験結果を表 9 に示す。通常スタジオの読み上げ部分では、認識精度で 95%を上回るが、今回の結果は最大で 89.6%(RASTA-PLP)とかなり難しいテストセットだとわかる。また、それぞれの特徴量を比較すると、テストセット全体では RASTA の効果による差が大きい。MFCC と PLP との比較では、MFCC の方が全体的によく、RASTA との組み合わせでは、PLP の方が改善効果が高い。今回は、これらの実験結果をベースラインとする。

5. 提案手法の評価

5.1. 分析パラメータの実験結果

BAT 特徴量を求めるには、帯域の分割数 n と時間方向の抽出窓幅 W フレームと DCT 分析後に出力する次数 k の 3 つのパラメータを決める必要がある。今回は、音声知覚に重要とされる変動成分の上限値を 24Hz に設定することで、抽出窓幅 W によって一意に k が決定でき、帯域の分割数 n と抽出窓幅 W の値を変化させ、認識実験によって最適値を求めた。

表 10 に、n と W を変化させた時の k の値と BAT 特徴量の総次元数を示す。表 10 からわかるように BAT 特徴量は、その特性から、n と W に比例して次元数が増加し、従来特徴量よりも次元数が多くなる。

認識実験の結果、W=15 で n=15 と n=20 の時に 86.8%と最大の認識精度になった(表 11)。最良の BAT 特徴量での認識結果を、MFCC(88.6%)や PLP(88.5%)と比較すると、このままでは 2 %ほど認識精度が低い。

帯域分割数については、n=10 よりも n=15 や n=20 の帯域を細かく分割した方が良い結果になっている。また、時間方向の抽出幅は、全体的には W=15 の場合が良い結果となった。

誤認識単語の傾向を調べると、W の違いにより誤認識単語の傾向が異なり、特徴量の識別しやすい音素に違いがある事がわかった。また、W が増加すると、挿入誤りが増加する傾向にあった。

これ以降の実験では、分析パラメータに W=15 フレーム、n =15 チャンネル（総次元数 128 次元）を用いた BAT 特徴量の評価を行う。

5.2. 次元数削減の実験結果

最初に、PCA による次元数削減を適用した特徴量(BAT-PCA)を、65 次元から 39 次元まで次元数を変化させて性能を調べた。

実験結果は表 12 に示すように、PCA による次元数削減によって、次元数が 52 次元の場合に 86.7%から、最大 89.3%まで改善した。ベースラインとの比較では、RASTA を適用した場合にはおよばなかったが、RASTA を適用しない場合の 88.6%（MFCC）、88.5%（PLP）よりも認識精度の高い特徴量となった。今回の評価では、次元数が 52 次元の場合で最良の結果となったため、これ以降の実験では、52 次元の BAT-PCA 特徴量を用いることにする。

表 13 に、LDA、HDA の両手法により次元数を削減した BAT 特徴量の認識結果を示す。BAT-LDA、BAT-HDA の削減後の次元数は、BAT-PCA 特徴量の 52 次元に合わせた。

実験の結果、それぞれ 89.1% (BAT-LDA)、88.9% (BAT-HDA) ととなり、削減前の 86.7%と比べて改善を確認した。ベースライン(88.6%:MFCC, 88.5%:PLP)との比較では、両手法とも認識精度が高かった。

表 10. 窓幅 W と帯域分割数 n を変化させた場合の k と総次元数

	W=9	W=15	W=20
k	5	8	11
n=10	55	88	121
n=15	80	128	176
n=20	105	168	231

表 11. 窓幅 W と帯域分割数 n を変化させた時の認識結果(単語正解精度:%)

	W=9/k=5			W=15/k=8			W=20/k=11		
認識音声	全体	中継	対談	全体	中継	対談	全体	中継	対談
n=10	85.6	87.5	71.6	84.3	86.3	69.9	84.4	86.4	69.6
n=15	86.6	**88.6**	72.9	**86.8**	88.5	**74.2**	85.2	87.0	71.0
n=20	86.5	88.3	73.2	**86.8**	88.5	73.5	85.5	87.6	70.9

表 12. PCA による次元圧縮の BAT-PCA 効果(単語正解精度:%)

次元数	39	52	65	128 圧縮なし
中継	**90.6**	**90.6**	**90.6**	88.4
対談	77.6	**78.6**	78.1	74.2
中継且つ対談	78.6	79.3	**79.5**	75.6
全体	89.1	**89.3**	89.1	86.7

表 13. 次元数を削減した特徴量の認識結果(単語正解精度:%)

圧縮手法	無し	PCA	LDA	HDA
次元数	128	52	52	52
中継	88.4	**90.6**	90.5	90.3
対談	74.2	**78.6**	**78.6**	77.9
中継且つ対談	75.6	79.3	**79.8**	79.0
全体	86.7	**89.3**	89.1	88.9

表 14.組み合わせた特徴量の実験結果(単語正解精度:%)

提案特徴量	BAT-PCA		BAT-LDA		BAT-HDA	
圧縮次元数	52	52	52	52	52	52
従来特徴量	RASTA					
	MFCC	PLP	MFCC	PLP	MFCC	PLP
次元数	13	13	13	13	13	13
総次元数	65	65	65	65	65	65
中継	91.9	**92.1**	91.6	91.1	91.4	91.3
対談	79.9	**80.5**	79.7	78.6	79.8	79.3
中継且つ対談	81.2	**81.4**	80.8	79.6	81.1	81.1
全体	90.4	**90.7**	90.3	89.7	90.0	89.9

しかし、PCA と LDA・HDA の性能比較では、事前知識を利用できる HDA と LDA にさらなる改善効果を期待していたが、一番単純な PCA による次元数削減がもっとも良い結果となった。今回、LDA や HDA には、判別するための事前知識としてトライフォン HMM の各状態をクラスとして与えたが、次元数削減で PCA より性能を上げるためには、より適切なクラスの検討が必要だと考えられる。

5.3. 従来特徴量との組み合わせの実験結果

従来特徴量と BAT 特徴量を組み合わせた実験を行った。この実験に用いた特徴量の総次元数は、それぞれ MFCC と対数パワーの 13 次元、あるいは PLP と対数パワーの 13 次元と BAT-PCA、BAT-LDA、BAT-HDA の 52 次元を足したものとなり、実験結果とあわせて表 14 に示す。

表 9 のベースラインとの比較をすると、今回の実験では、MFCC や PLP と BAT 特徴量を組み合わせることで、どの特徴量を単独で使用した場合よりも、改善が確認できた。詳しく見ると、ベースラインで一番認識率の高かった RASTA-PLP の 89.6%は、BAT-PCA 特徴量と組み合わせることによって単語認識精度が 1.1 ポイント向上して 90.7%となり、誤り削減率 10.1%が得られた。RASTA-MFCC(89.5%)との組み合わせでは、BAT-PCA との組み合わせによって 0.9 ポイント向上して 90.4%となり、誤り削減率では 8.7%であった。

PCA と比べて LDA と HDA で改善効果が低かったのは、分散に大きな差のある提案特徴量と従来特徴量を単純に 1 つのベクトルにしたことで、これらの音響モデルの分布に偏りが生じたためだと推測される。

6. まとめ

本稿では、音声信号の帯域フィルタ出力の時間変化特性である BAT 特徴量について、最適な分析パラメータや、次元数削減、従来特徴量との組み合わせ効果について検討した。BAT 特徴量は、帯域ごとに処理することで、他の帯域に混入した雑音の影響などの軽減が期待できる。今回の実験では、音声知覚に重要とされる 1～24Hz を表現するように DCT 係数 k を固定した場合、分析パラメータは帯域分割数 n=15 抽出窓幅 W=15(150msec)が最適であった。次元数削減の性能評価では、PCA が一番効果的で、これを従来特徴量の RASTA-MFCC や RASTA-PLP と組み合わせることにより、ニュース番組の中継や対談調の音声の評価実験では、改善効果を確認できた。今後は、より適切な従来特徴量との組み合わせ方法について検討をしていく。

謝　辞

LDA・HDA に関して協力して頂いた、BBN の F. Guo 氏と L. Nguyen 氏に感謝する。

文　献

[1] T. Imai, et al., "Speech Recognition for Subtitling Live Broadcast", Proc. ICA2004, vol. 1 pp. 165-168, Kyoto, 2004.

[2] H. Hermansky, et al., "Rasta processing of speech," IEEE Trans. on Speech and Audio Proc. vol. 2 no. 4, pp. 578-589, 1994.

[3] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram", Speech Communication, vol. 25, pp.117-132,1998.

[4] D. Poeppel, "The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'", Speech Communication, vol. 41, 2003.

[5] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects", Phil. Trans. R. Soc. Lond. B336(4), pp.367-373, 1992.

[6] C. L. Smith, et al., "Extracting dynamic parameter from speech movement data", J. Acoust. Soc. Am., 93(3), 1993.

[7] H. Hermansky, et al., "Temporal patterns (TRAPs) in ASR of noisy speech", Proc. ICASSP'98.

[8] R. M. Warren, et al., "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits", Perception and Psychophysics, 57(2), 1995.

[9] H. Fletcher, "Speech and Hearing in Communication", New York: Krieger, 1953.

[10] H. Bourlard, et al., "A new ASR approach based on independent processing and re-combination of partial frequency bands," Proc.ICSLP'96 vol. 4, pp. 426-429, 1996.

[11] NIST:SPQA (speech quality assurance package version 2.3), http://www.nist.gov/speech/index.html

[12] A. Kobayashi, et al., "Time Dependent Language Model For Broadcast News Transcription And Its Post-Correction", Proc. ICSLP'98, Vol. 6, pp.2435-2438, Sydney, 1998.

[13] N. Kumer and A. G. Andreou, "Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition", Speech Communication. Vol26, pp.283-297, 1998.

[14] M. F. J. Gales, "Semi-tied Covariance Matrices for Hidden Markov Models", IEEE Tran. On Speech and Audio Proc. Vol7, pp.272-281, 1999.

[15] R.A. Gopinath, "Maximum Likelihood Modeling with Gaussian Distoribution for Classification", Proc. of ICASSP'98, Seattle, 1998.

[16] H. Hermansky, "Perceptual Liner Predictive (PLP) Analysis of Speech", J. Acoust. Soc. Am., Vol.87(4), pp.1738-1752, 1990.

[17] H. Hermansky, "Should recognizers have ears?", Speech Communication, vol. 25, 1998.

[18] 尾上他, "帯域フィルタ出力の時間特性を特徴量としたニュース音声認識" ,音響学会春季講演論文集, vol. 1, pp.3-4, 2004.