Internal Use Only（非公開）

TR-SLT-0004

マルチストリーム特徴量による雑音にロバストな音声認識

Speech Recognition in Noisy Environment Using Multi-stream Features

梅田 将満　　伊田 政樹

Masamitsu Umeda　Masaki Ida

2002. 2.22

概要

近年、音声認識の性能は大きく改善されたが、さらに音声のＳＮＲが低い雑音環境での高い認識性能が求められている。実世界における雑音はスペクトル領域において偏りを持つものが多い。そこで本稿では、周波数領域で特徴量を分割して、１つの特徴量を複数のストリームで表すマルチストリーム特徴量を用い、有効性を検討した。最適な重みを手動で与えて認識をした場合、雑音の種類がairportでSNRが5dBのとき、27.71%向上した。また、GPDアルゴリズムによる重み自動推定を行った場合、雑音の種類がairportでSNRが5dBのとき、20.31%向上した。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288 京都府相楽郡精華町光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

1. はじめに

近年，音声認識の性能は大きく改善されたが，音声認識システムを実環境下で利用することを考えた場合，さらに入力音声の SNR が低い雑音環境での高い認識性能が求められている．

そこで，本研究では入力音声に混入する雑音そのものに注目する．ここでは，定常的な雑音について取り扱う．実世界に存在する雑音は図 1 に示すように，スペクトル領域において偏りを持つものが多い．

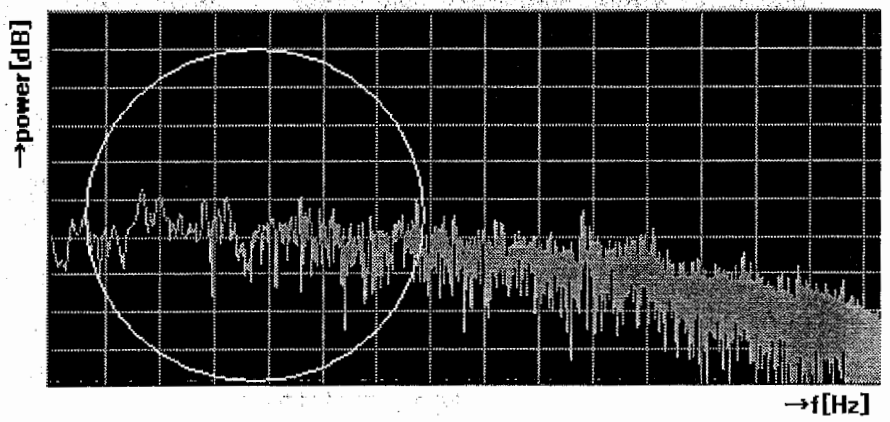

図 1 雑音の周波数特性（airport）

この場合，入力音声を $x(t)$，周囲の雑音を $n(t)$ とすると認識システムに入力される波形は，

$$y(t) = x(t) + n(t)$$

で表される．この関係は線形スペクトル領域では，

$$Y(\omega) = X(\omega) + N(\omega)$$

で表され，雑音が混入する影響を周波数帯域で分割して扱うことができる．そこで音声認識の際，短時間スペクトルを用いて特徴量の時系列を作成し，認識している．この特徴量を周波数領域で特徴量を分割して，音声の特徴量を複数のストリームで表し，マルチストリーム特徴量として用いることで，混入雑音の影響を帯域分割する．雑音の種類によって周波数的に偏りがある場合，雑音が多い周波数帯域の特徴量と雑音が多い周波数帯域の特徴量によるスコアに最適な重みを与えることで認識性能の向上を目指す[1]．マルチストリーム特徴量は，音声と画像情報を統合して扱う場合に用いられる[2].

さらに，雑音の種類や大きさによって最適な重みが異なるため，ストリーム重みの自動推定を行う．重み推定アルゴリズムとして GPD アルゴリズムを用い，推定されたストリーム重みでの認識により性能の確認を行った．

2. マルチストリーム特徴量を用いた音声認識

2.1. フィルタバンク出力を用いた実験

2.1.1. マルチストリーム特徴量の抽出

まず，特徴量として logFBE を用いる．logFEB は音声波形の短時間スペクトルにフィルタバンクをかけて算出したものである．

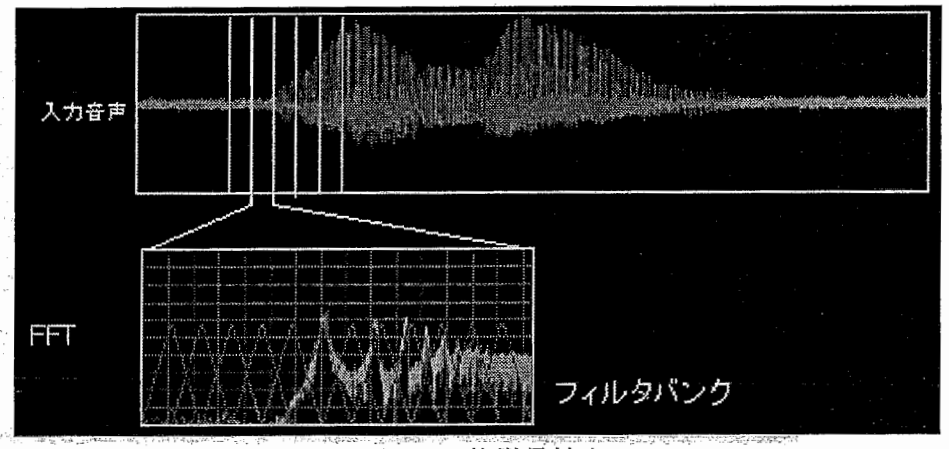

図 2.1 特徴量抽出

本実験では，図 2.2 に示す 12 次元の logFBE，パワーとそれぞれΔ，ΔΔを用いて合計 39 次元の特徴量を用いた．

ﾌｨﾙﾀﾊﾞﾝｸ 1-12	ﾊﾟﾜｰ	Δﾌｨﾙﾀﾊﾞﾝｸ 1-12	Δ ﾊﾟﾜｰ	ΔΔﾌｨﾙﾀﾊﾞﾝｸ 1-12	ΔΔ ﾊﾟﾜｰ

図 2.2 シングルストリーム特徴量

フィルタバンク係数を周波数の高い部分と低い部分に 6 個ずつ分割し，図 2.3 に示す 2 つのストリームに分割する．

パワーの項は，低周波数域の特徴量とあわせて用いる．

フィルタバンク 1-6	Δフィルタバンク 1-6	ΔΔフィルタバンク 1-6	パワー	Δ パワー	ΔΔ パワー	低周波域の特徴量

フィルタバンク 7-12	Δフィルタバンク 7-12	ΔΔフィルタバンク 7-12	高周波域の特徴量

図 2.3 マルチルストリーム特徴量

2.1.2. マルチストリーム特徴量を用いた HMM の学習および認識

前節で述べたマルチストリーム特徴抽出を用いて，HMM の学習を行う．このとき，各ストリーム重みは均一にして学習する．

そして，認識の際にはそれぞれのストリームごとにスコアが得られるが，各ストリームのスコアに対し，任意の重みを変えることで，ストリーム全体のスコアに対する各ストリームの影響を制御する．

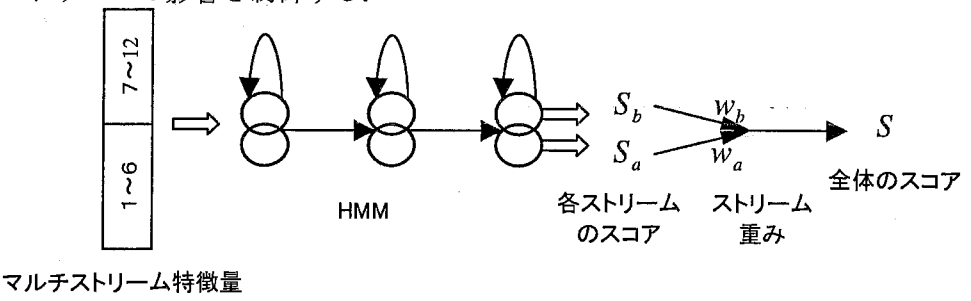

図 2.4 マルチストリーム HMM による認識

2.1.3. 実験条件

タスク：AURORA2(ＴＩ-digit)

特徴量：logFBE(12 次元)+power，Δ，ΔΔ

学習セット：雑音なし-8840 文

雑音：subway，babble，car noise，exhibition hall，restaurant，street，airport，train station

ＳＮＲ：clean 20dB 15dB 10dB 5dB 0dB -5dB

評価セット：各雑音，ＳＮＲにつき 1001 文

2.1.4. 認識結果

マルチストリームでストリーム重みを $w_{Low} : w_{High} = 0:10 \sim 10:0$ に変えて音声認識実験を行った．ノイズの種類ごとに SNR に対する認識性能とストリーム重みに対する認識性能を図 2.6～図 2.20 に示す．

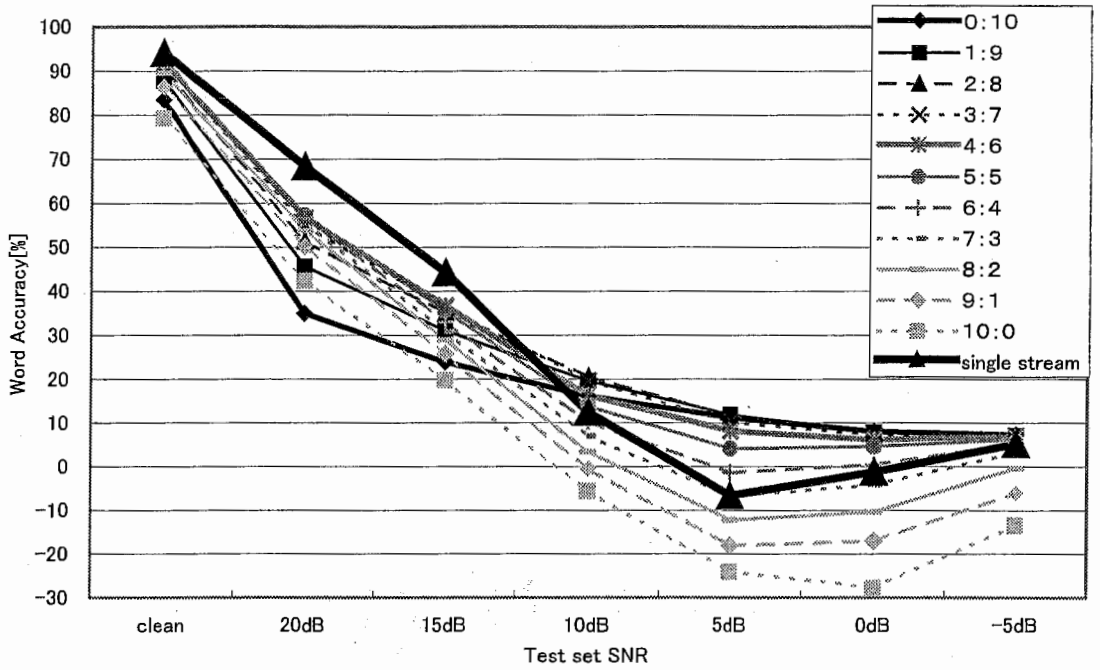

図 2.5 SNR に対する認識性能（subway）

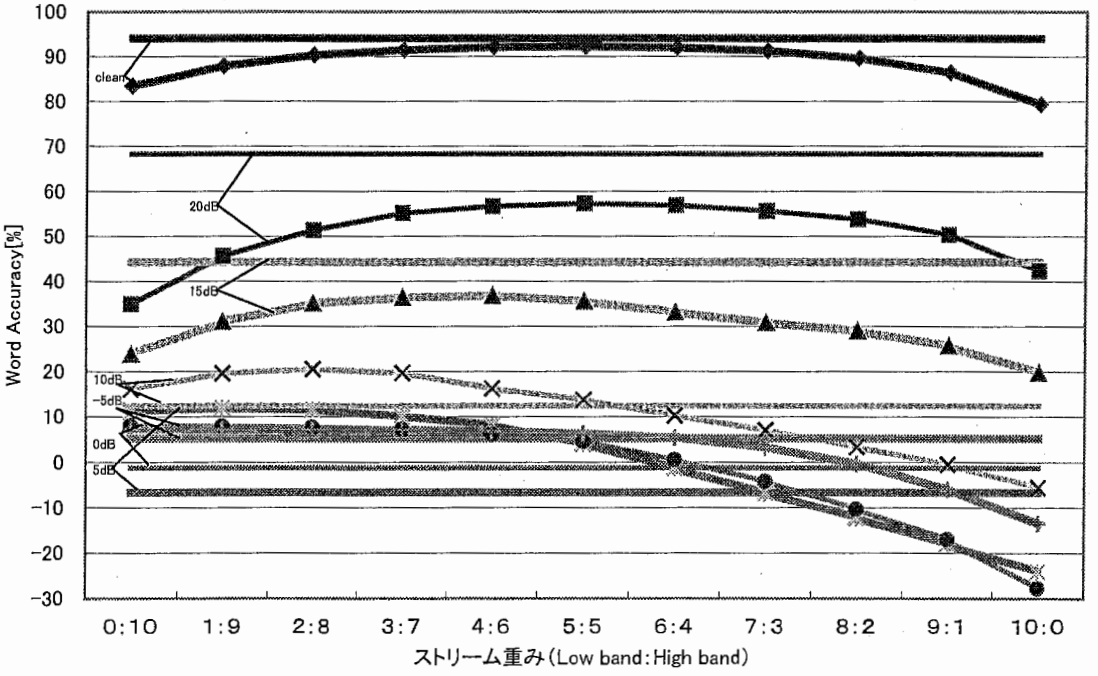

図 2.6 ストリーム重みに対する認識性能（subway）

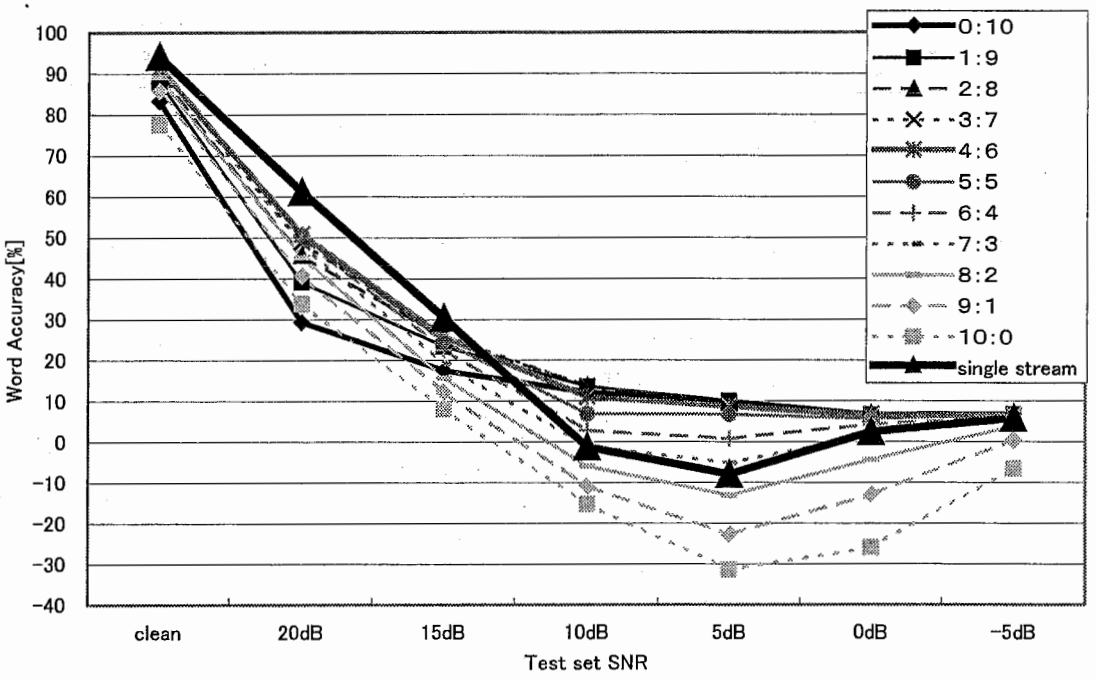

図 2.7 SNR に対する認識性能（babble）

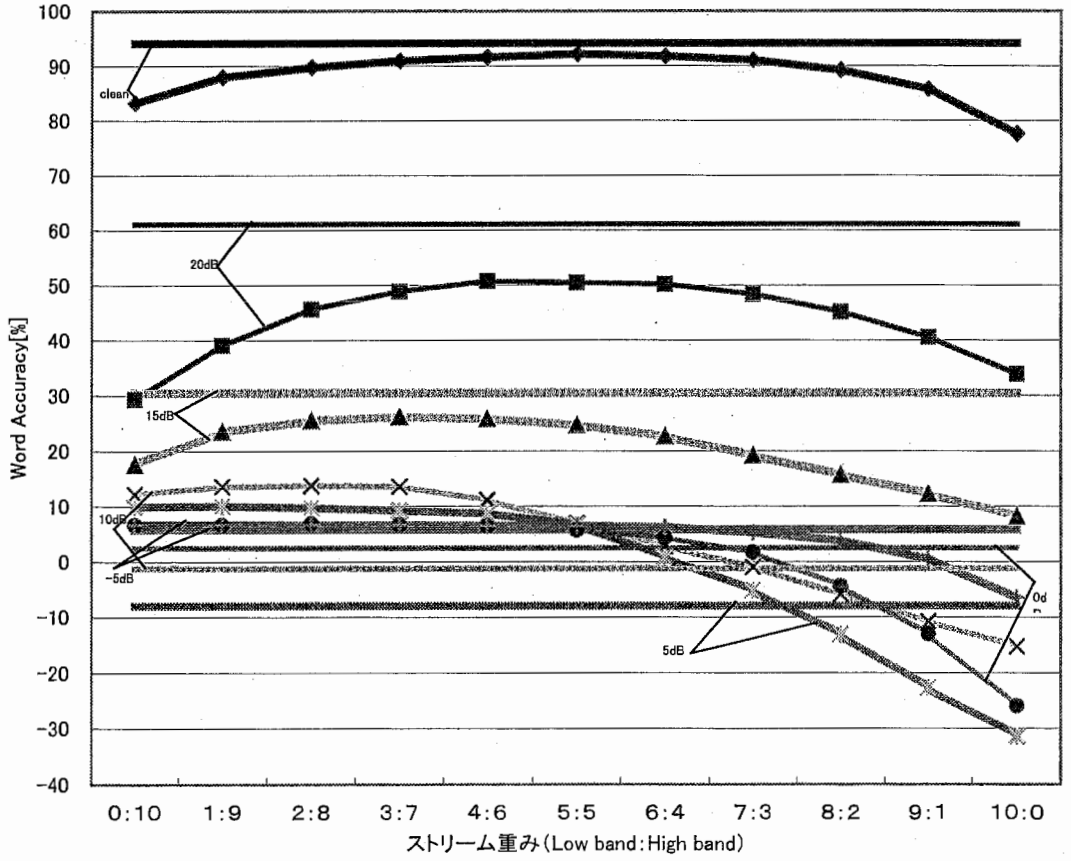

図 2.8 ストリーム重みに対する認識性能（babble）

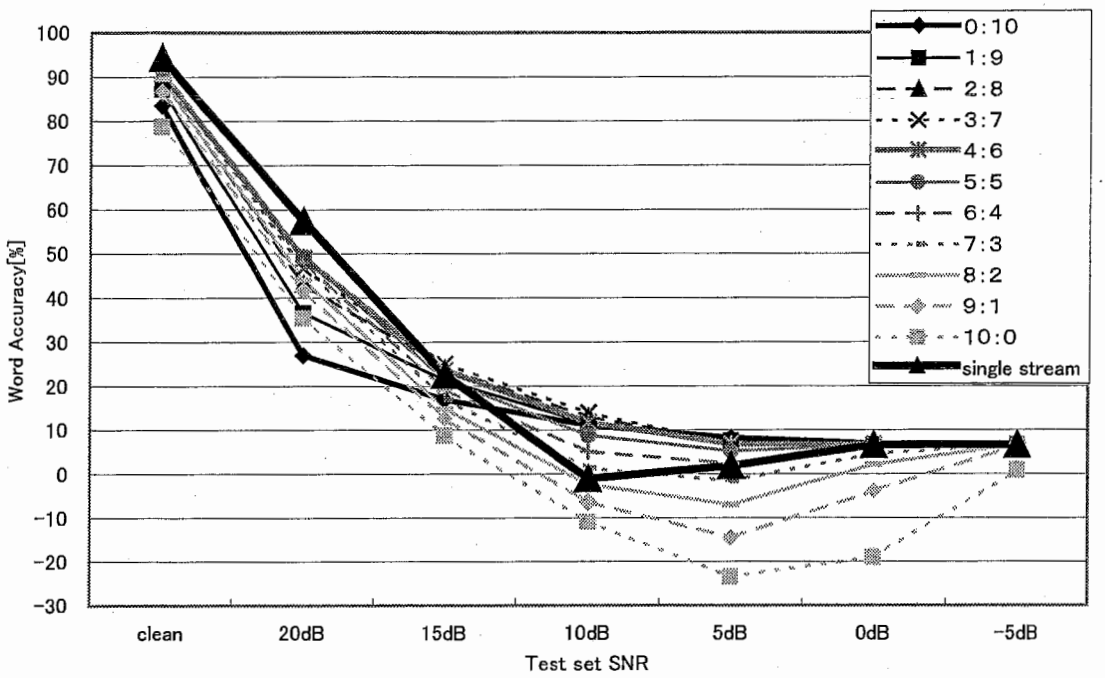

図 2.9 SNR に対する認識性能（car noise）

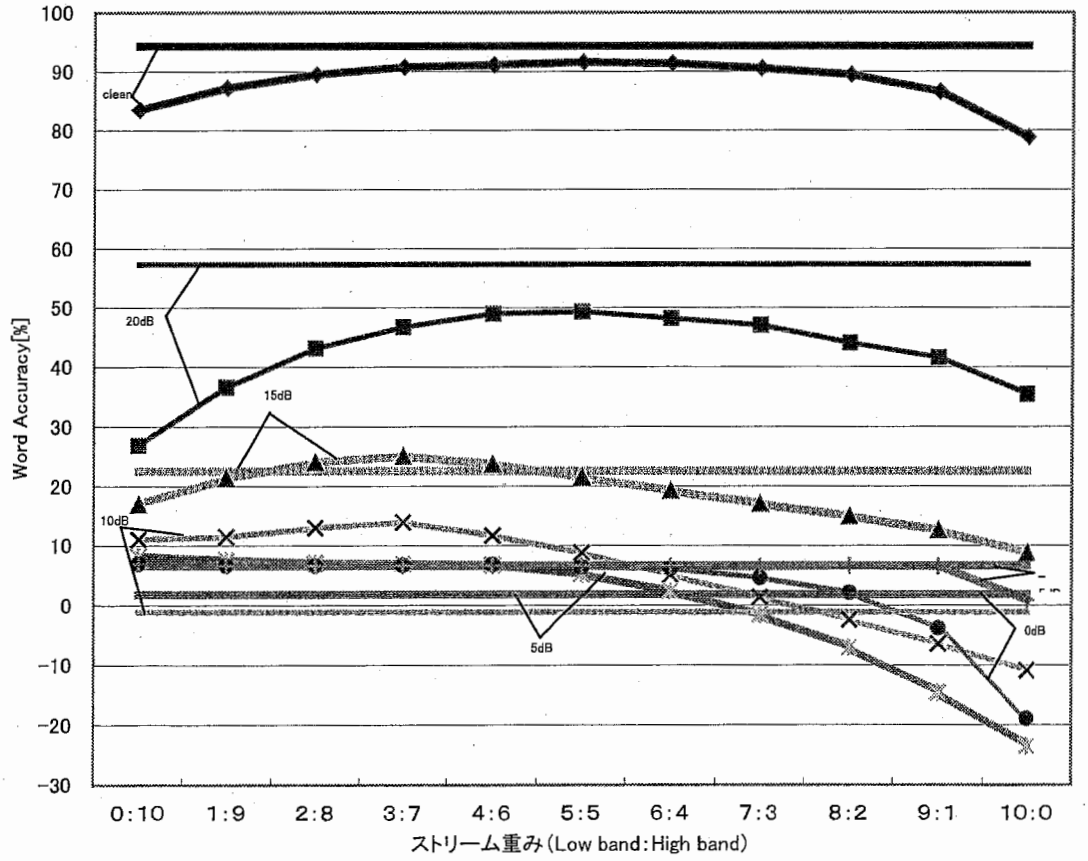

図 2.10 ストリーム重みに対する認識性能（car noise）

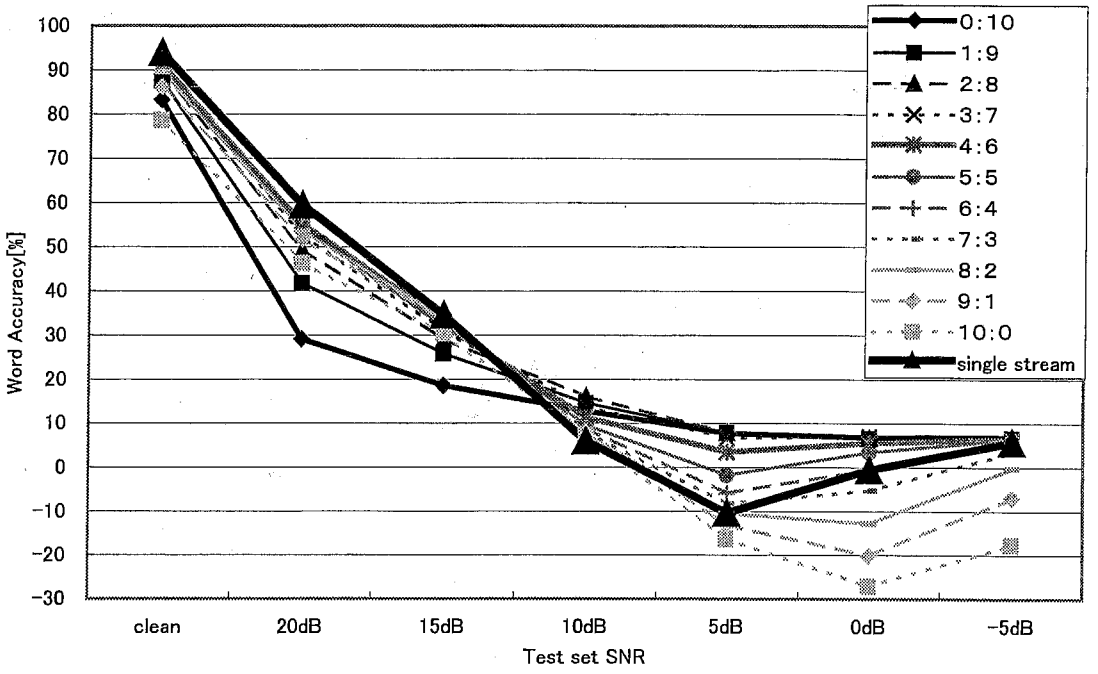

図 2.11 SNR に対する認識性能（exhibition hall）

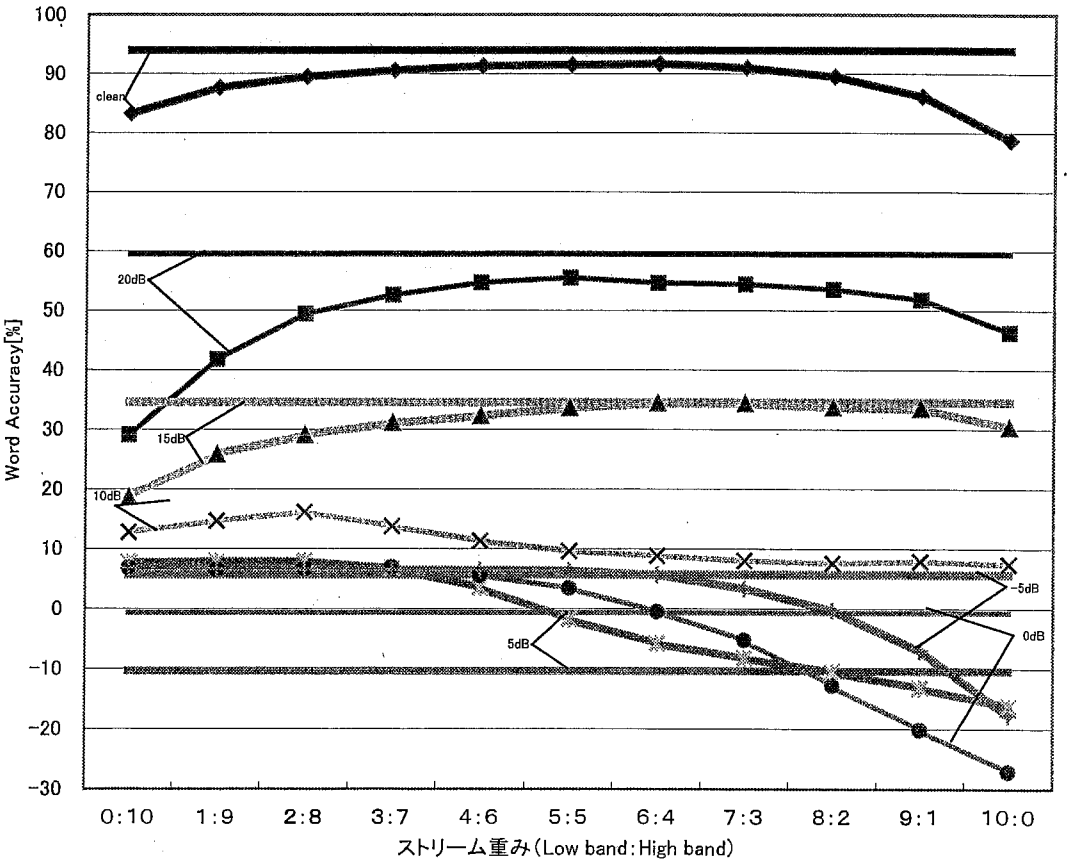

図 2.12 ストリーム重みに対する認識性能（exhibition hall）

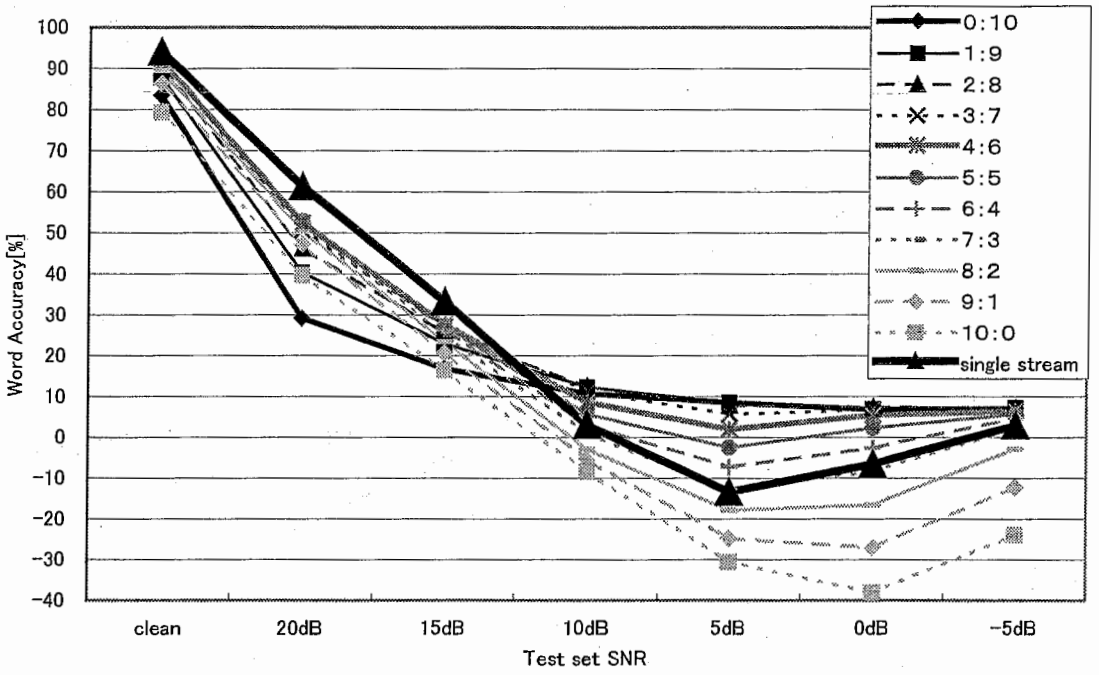

図 2.13 SNR に対する認識性能（restaurant）

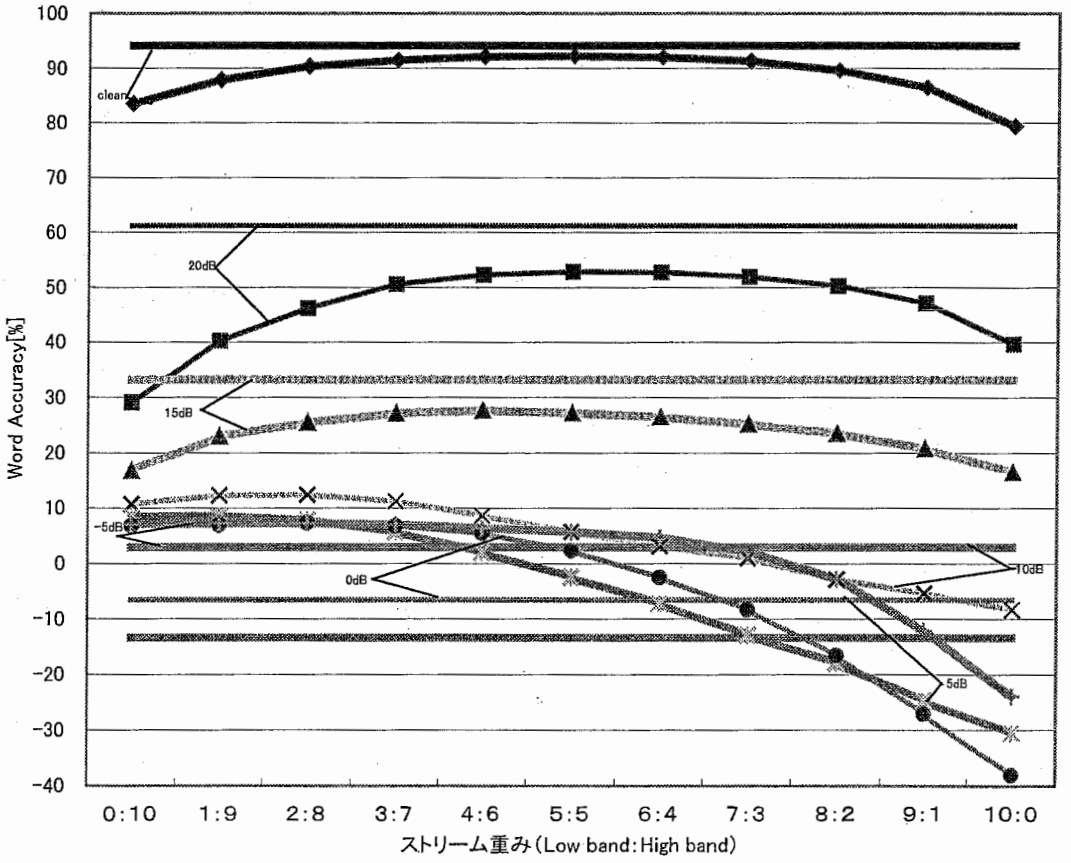

図 2.14 ストリーム重みに対する認識性能（restaurant）

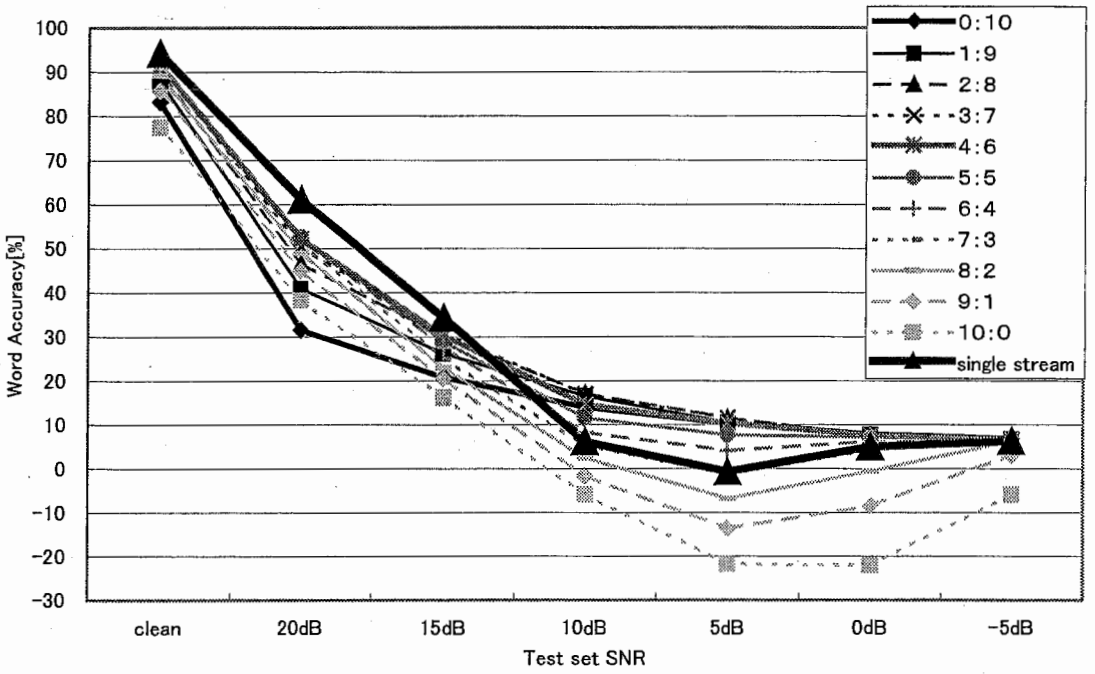

図 2.15 SNR に対する認識性能（street）

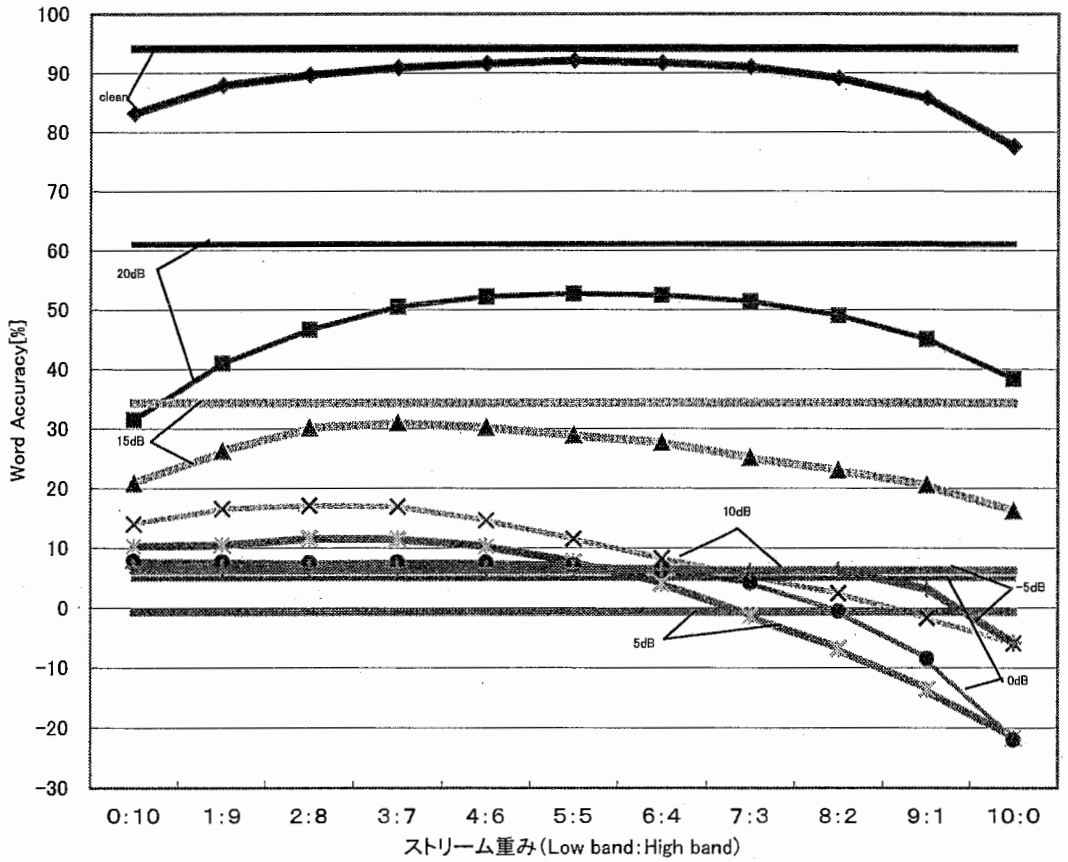

図 2.16 ストリーム重みに対する認識性能（street）

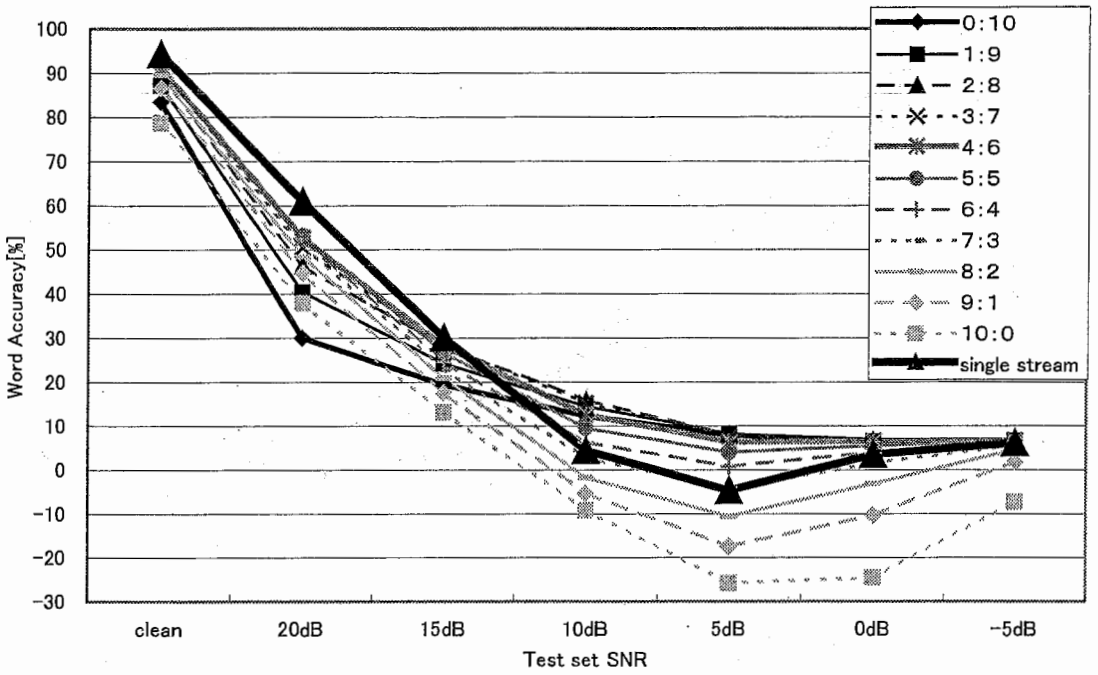

図 2.17 SNR に対する認識性能（airport）

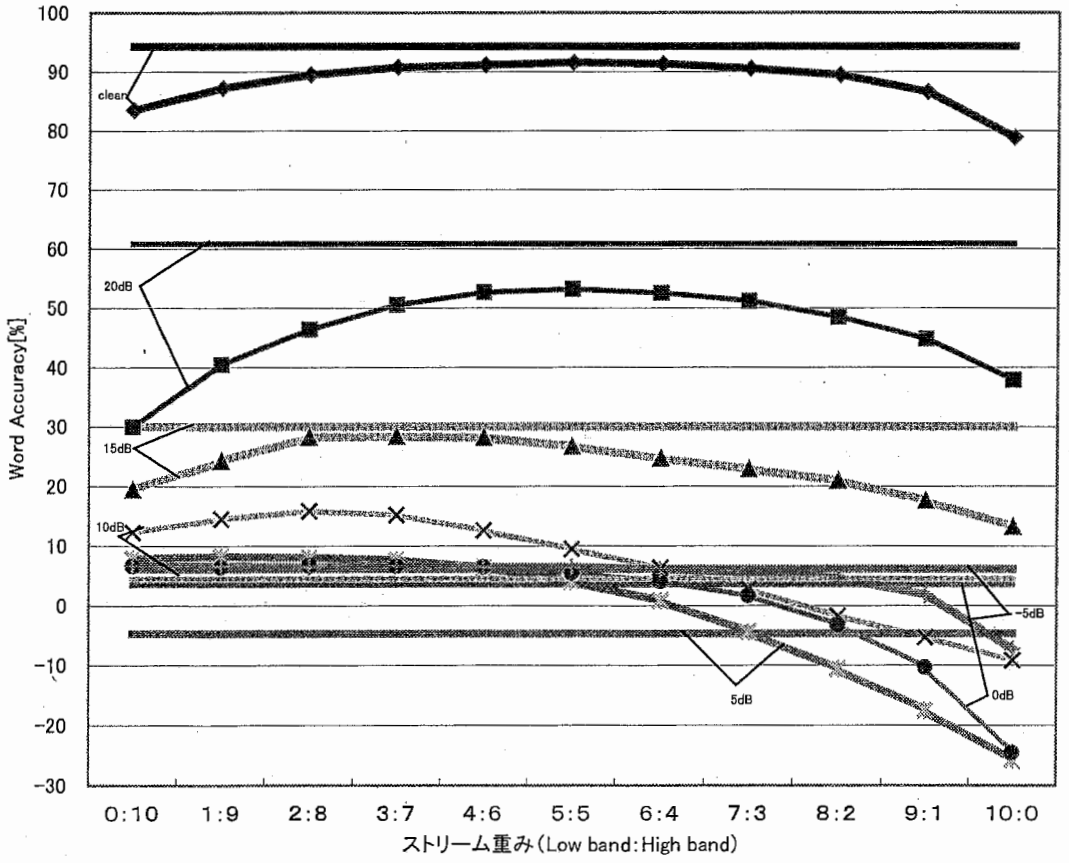

図 2.18 ストリーム重みに対する認識性能（airport）

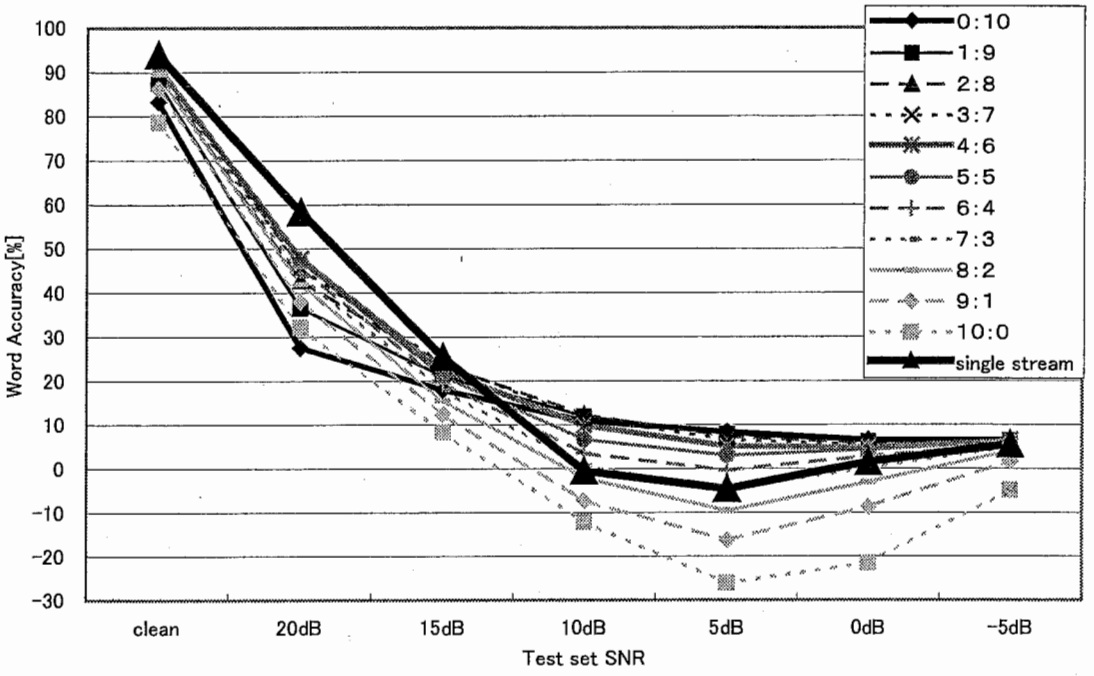

図 2.19 SNR に対する認識性能（train station）

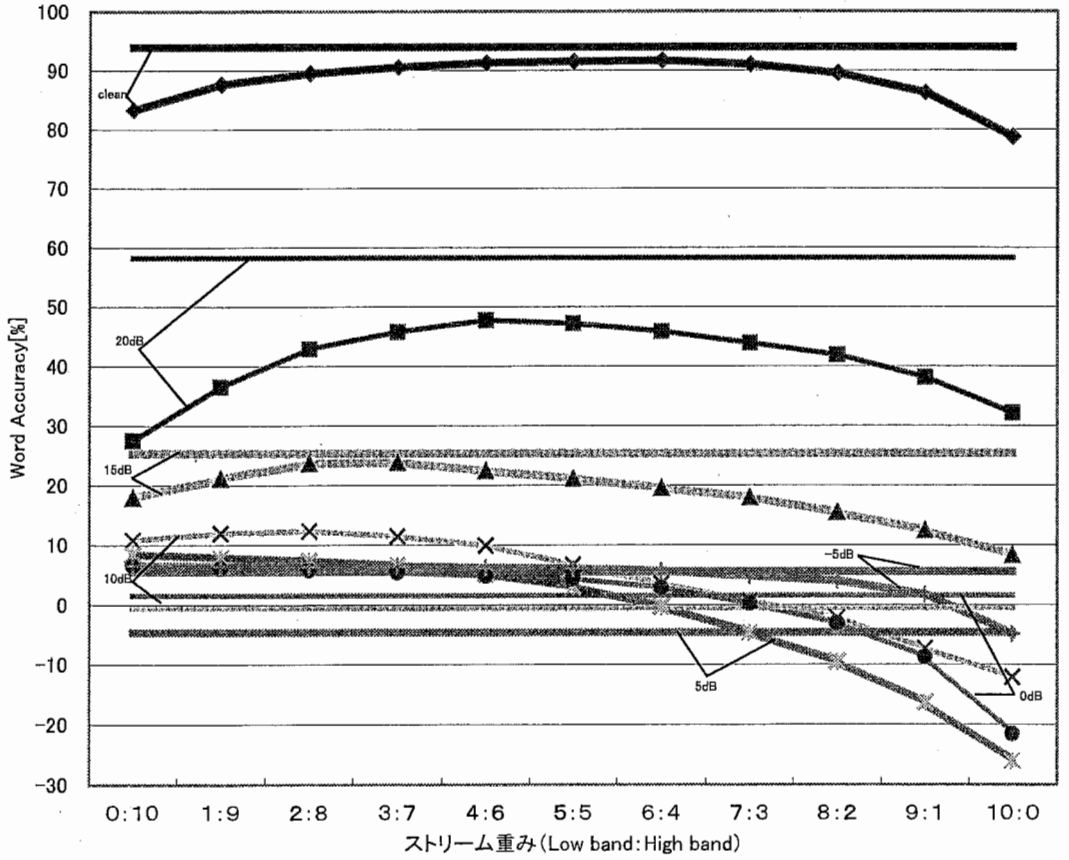

図 2.20 ストリーム重みに対する認識性能（train station）

2.1.5. 考察

まず，シングルストリームでの認識性能に着目すると，雑音が混入することによる認識率の低下が一般に音声認識で用いられているMFCCに比べて大きい．そこで，認識結果を分析すると，spの挿入誤りが多く，認識率が低下している．

そこでspが多く出現する原因を調べるため，spの特徴量の分析を比較する．clean speech HMMのspモデルと，20dBの雑音環境で学習したHMMのspモデルのパラメータについて，平均と分散を比較する．特徴量として，logFBEを用いたものを図2.21にMFCCを用いたものを図2.22に示す．

認識結果を見てみると，short pauseの出現率が多くて認識率の低下になっていた．

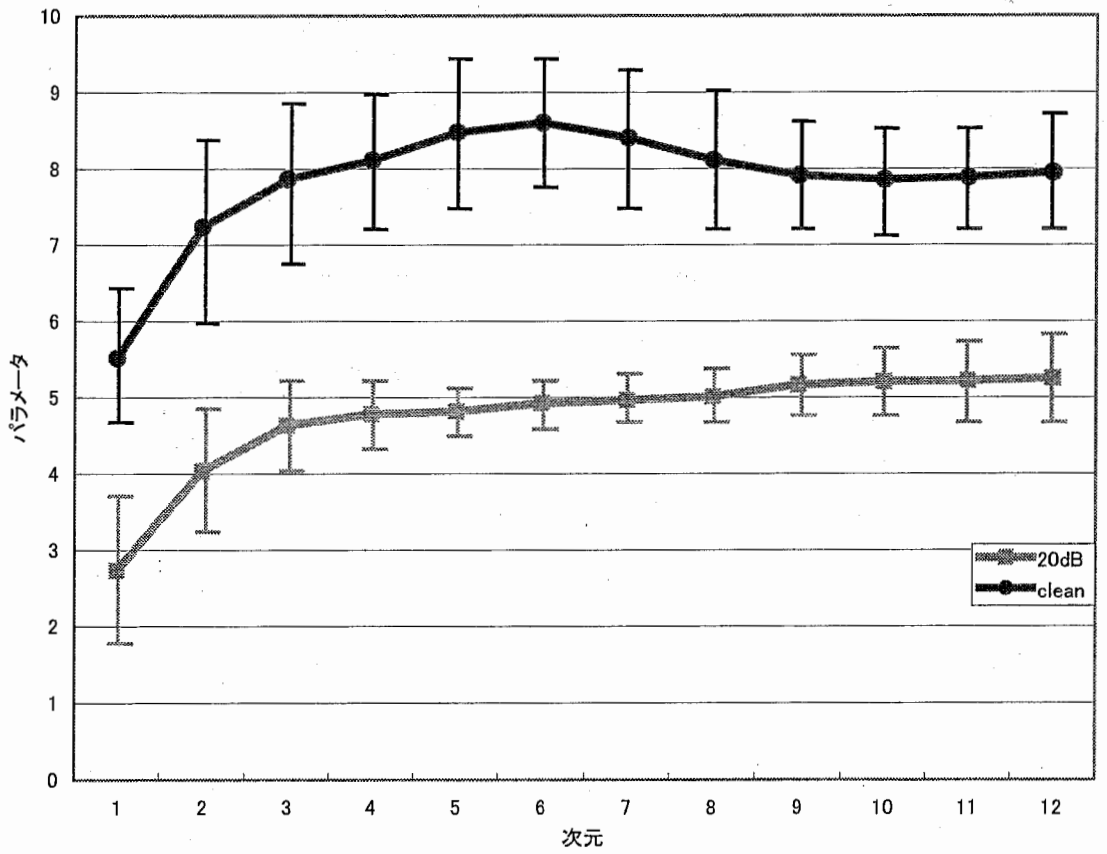

図2.21 フィルタバンクのHMMの一部

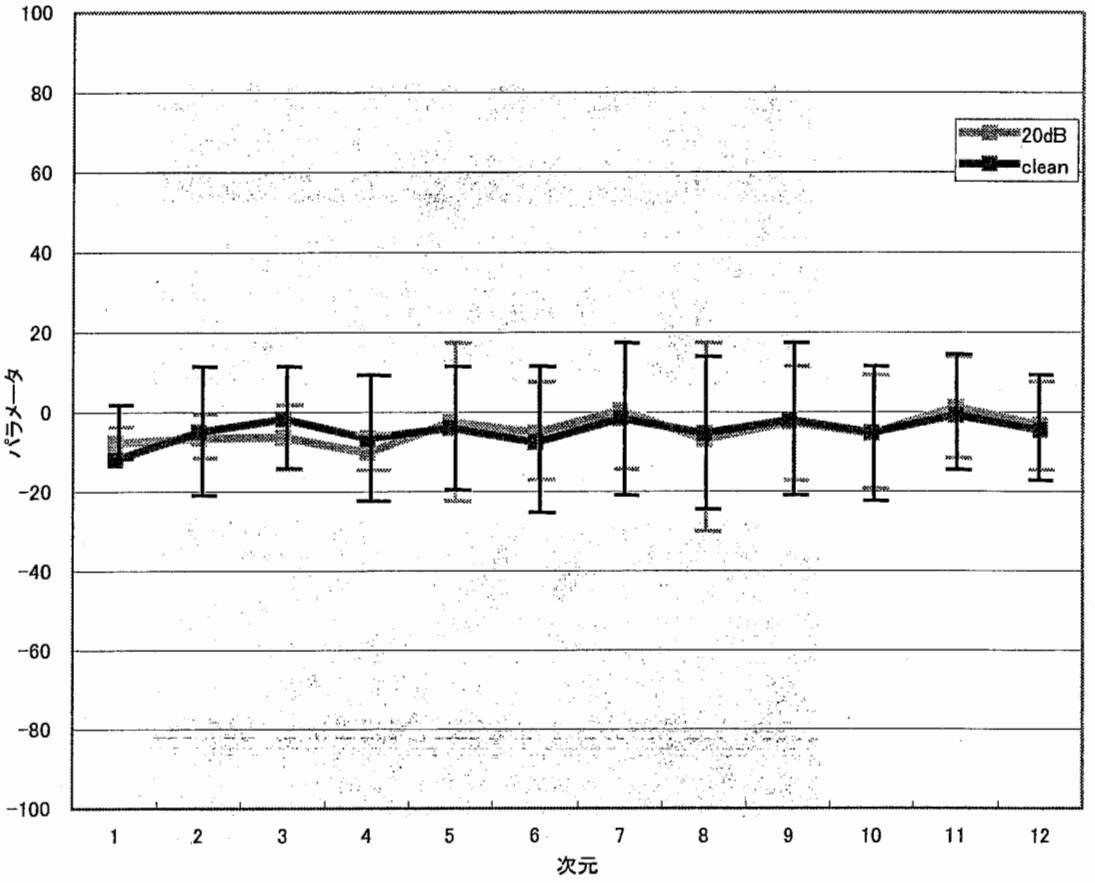

図 2.22 MFCC の HMM の一部

図から分かる通り，logFBE による sp モデルは雑音混入による差が大きく，MFCC では分布がほぼ一致していることがわかる．原因としては logFBE での特徴抽出ではパワーの正規化が行われていないため，特に sp でのミスマッチが大きいために，認識率低下につながっていると考えられる．

また，マルチストリーム特徴量を用いた場合，雑音の種類によって多少差はあるが，15～10dB 以下の雑音環境でストリーム重みが高周波域に重みを多く置いた場合の認識率がシングルストリームよりよくなった．最高で雑音の種類 restaurant で 5dB のとき，21.9%上昇した．このことは，本実験で用いた雑音が低周波数域に偏りがあることと一致する．

2.1.6. 高周波域にだけ雑音成分を含む評価データによる実験

AURORA2 の雑音データでの実験結果はどれも高周波域に重みを多く置いた方がよいという結果であるので，高周波域にだけ雑音を含む場合における有効性を検証する．高周波成分のみに帯域制限した白色雑音を生成し，クリーンな音声に重畳して，評価データを作成した．図 2.23，2.24 における破線より左側がフィルタバンクの 1

～6 次元にあたる部分でノイズがある右側が 7～12 次元にあたる部分である．認識結果を図 2.25 に示す．

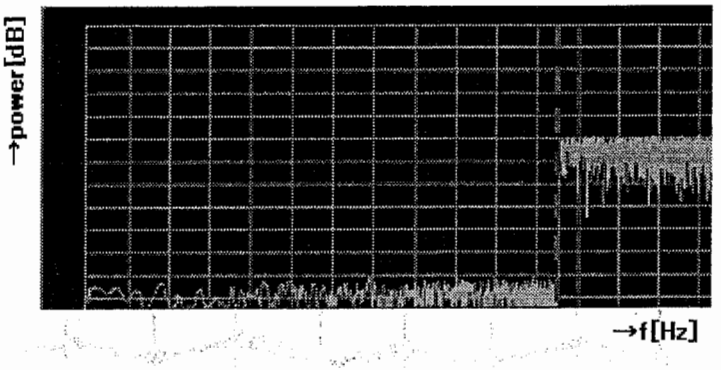

図 2.23 Band limited ノイズ

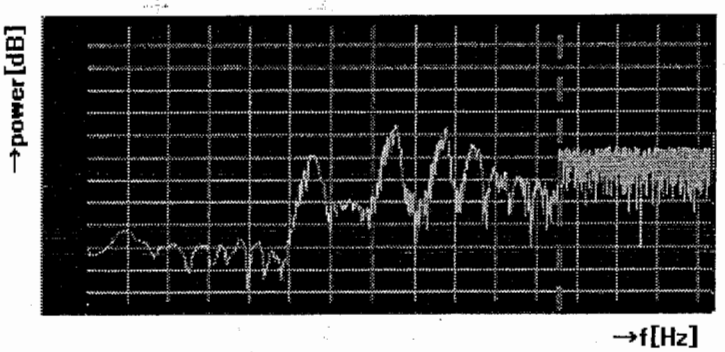

図 2.24 クリーンな音声に加えたもの

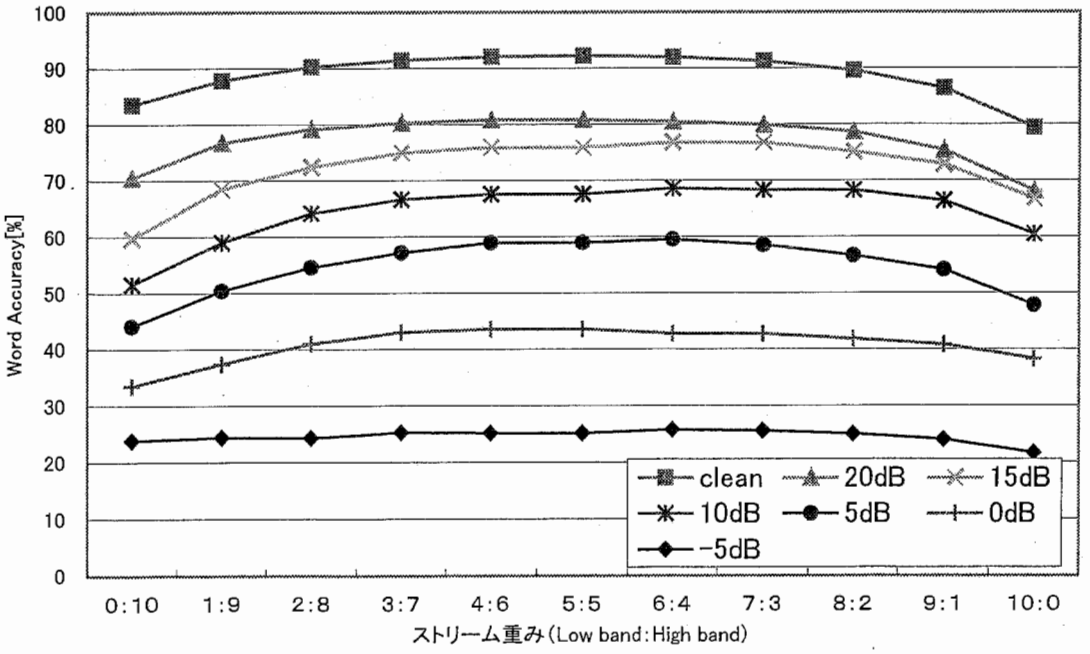

図 2.25 Band limited ノイズによる実験結果

この結果より，AURORA2 の雑音データと比べると最適なストリーム重みが高周波域に偏っている様子が分かる．

2.2. MFCC を用いた実験

フィルタバンクを用いた認識実験では，先にも述べたようにパワーの正規化が行われていないため，認識率が悪い．そこで，一般的に音声認識でよく用いられている MFCC(Mel Frequency Cepstral Coefficients)[3]を用いる．

2.2.1. MFCC への変換

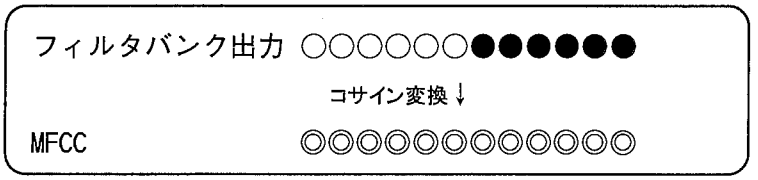

図 2.26 周波数分割せずに MFCC にした場合

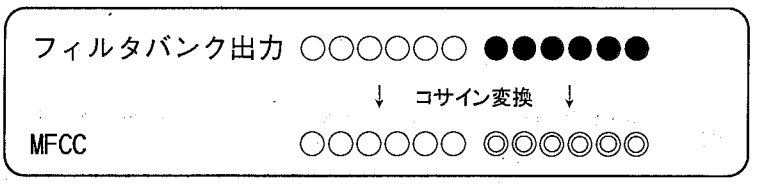

図 2.27 周波数分割して MFCC にした場合

MFCC に変換するにはフィルタバンクの出力をコサイン変換して求める．図 2.26 に示すように高周波域にだけ雑音があった場合，そのまま変換してしまうと MFCC の全部の成分に影響が及んでしまう．そこで，フィルタバンクの時点で分割し，それぞれで MFCC に変換する方法を用いる．これにより，一方の帯域に雑音が混入した場合においても，その影響を他の帯域に及ぼされないようにする．

2.2.2. 認識結果

前節の帯域分割した MFCC を用いて，マルチストリーム特徴量として認識した場合（各 SNR，マーカ付き実線）と帯域分割 MFCC によるシングルストリーム特徴量として認識した場合（single stream，太実線）との比較を行う．また，従来の方法である帯域分割を行わない MFCC による認識結果を Baseline（破線）で示す．

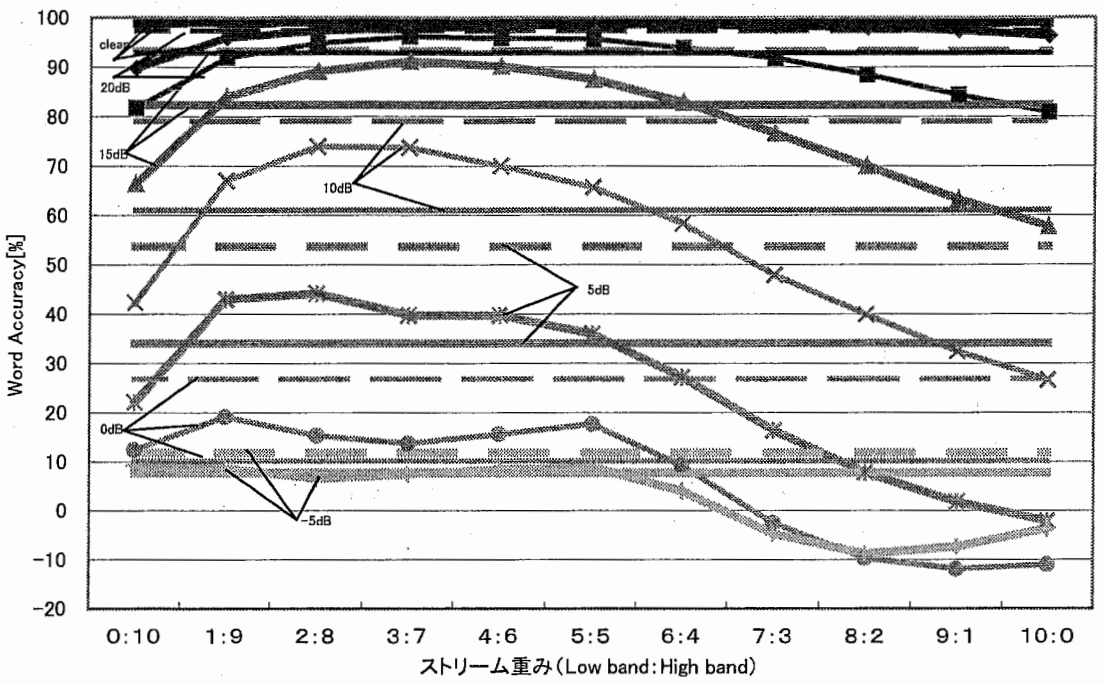

図 2.28 ストリーム重みに対する認識性能（subway）

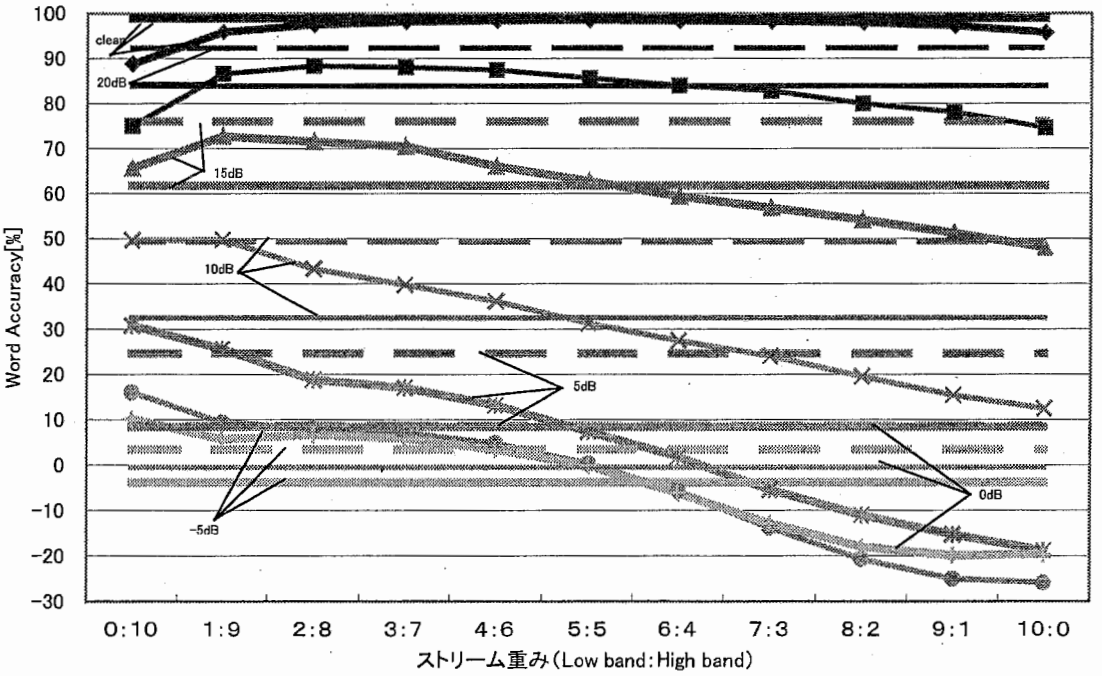

図 2.29 ストリーム重みに対する認識性能（babble）

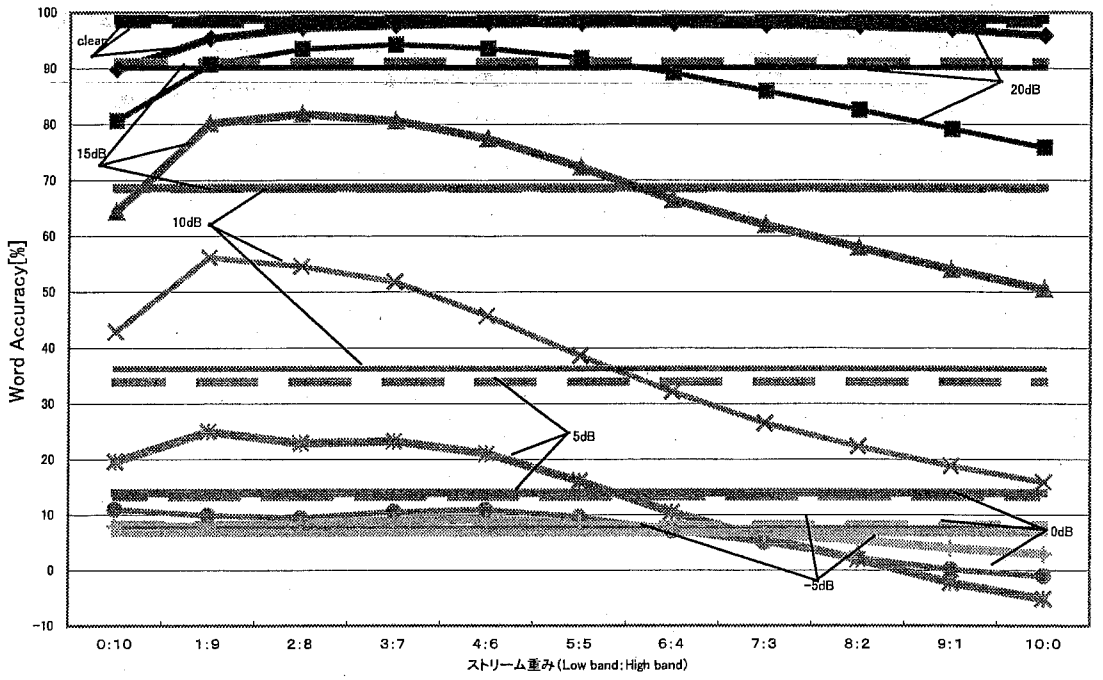

図 2.30 ストリーム重みに対する認識性能（car noise）

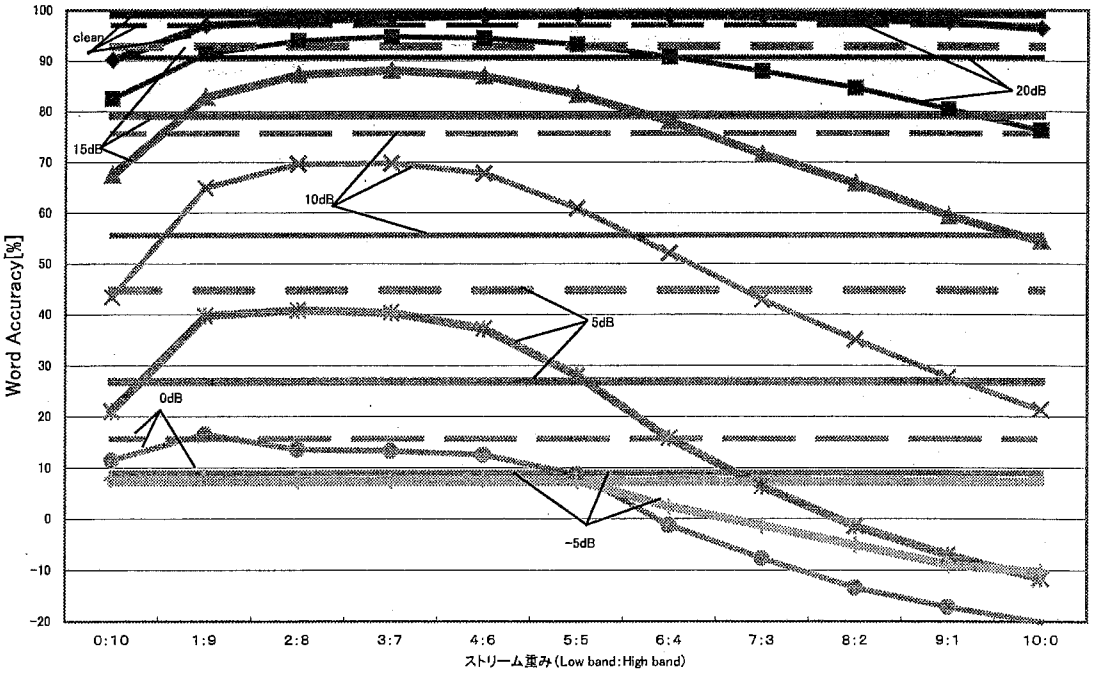

図 2.31 ストリーム重みに対する認識性能（exhibition hall）

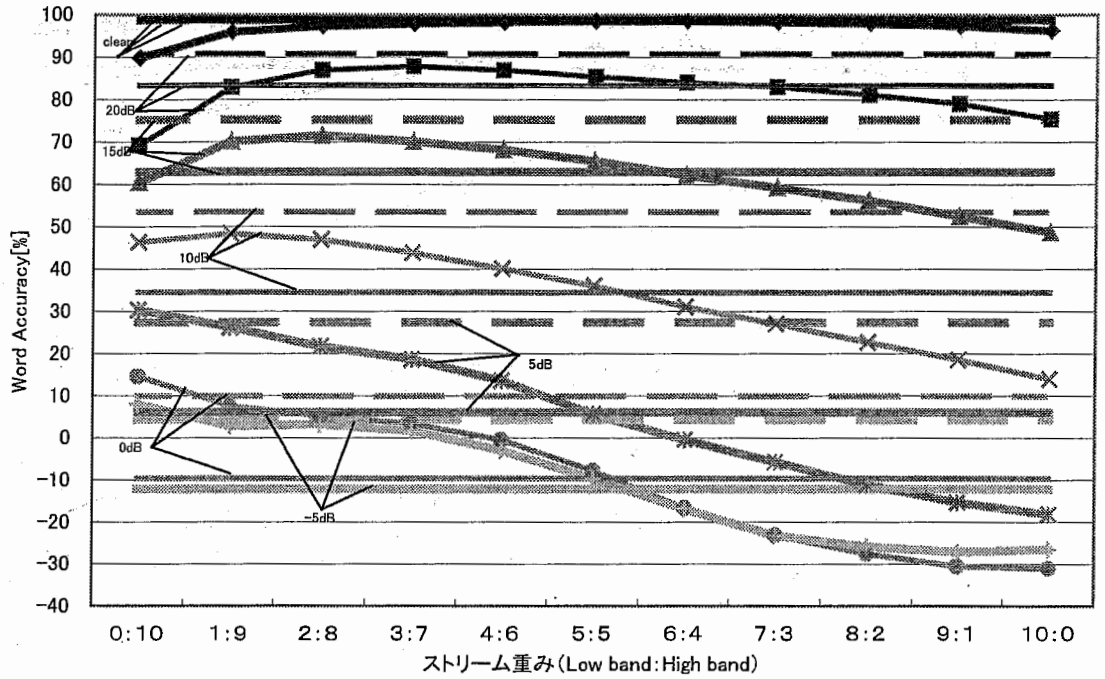

図 2.32 ストリーム重みに対する認識性能（restaurant）

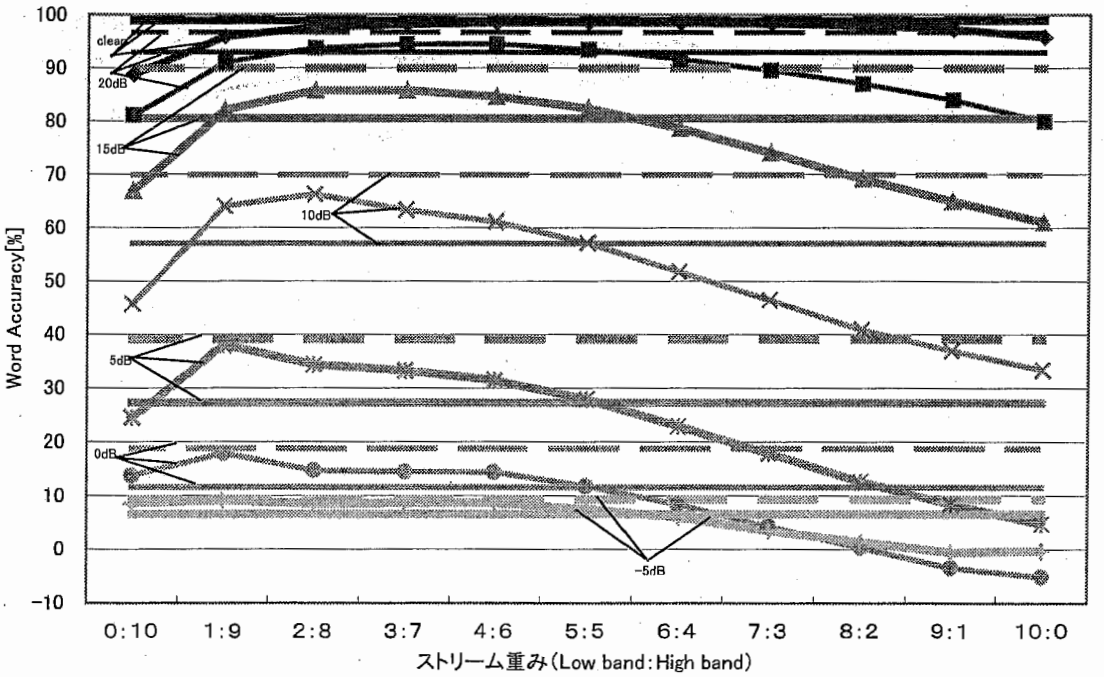

図 2.33 ストリーム重みに対する認識性能（street）

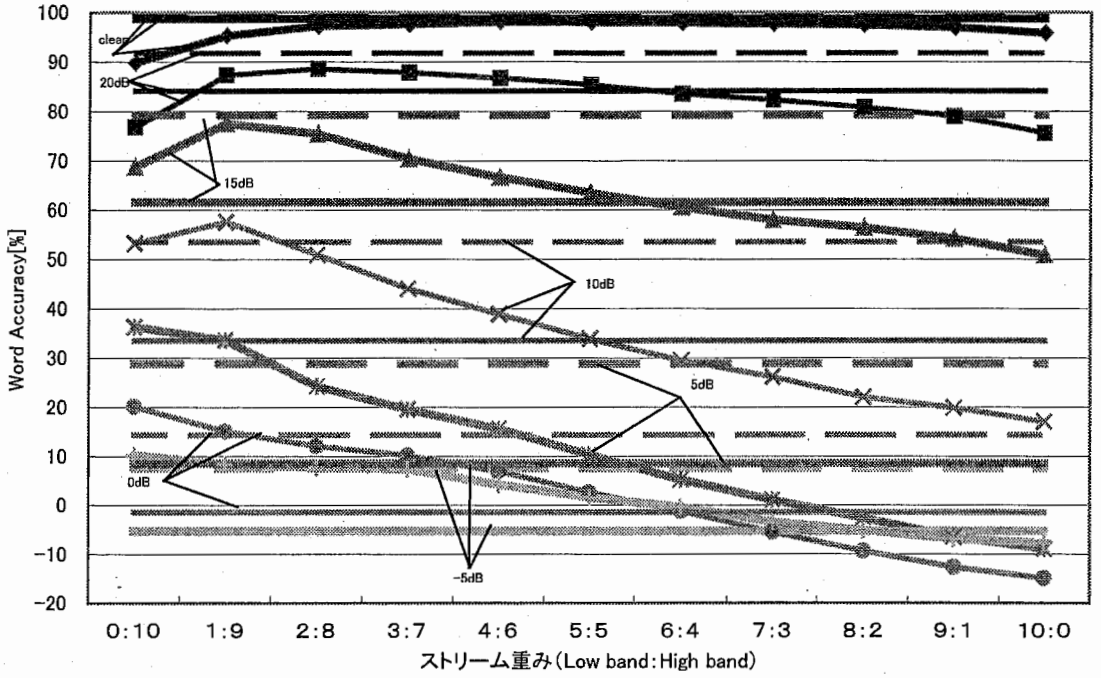

図 2.34 ストリーム重みに対する認識性能（airport）

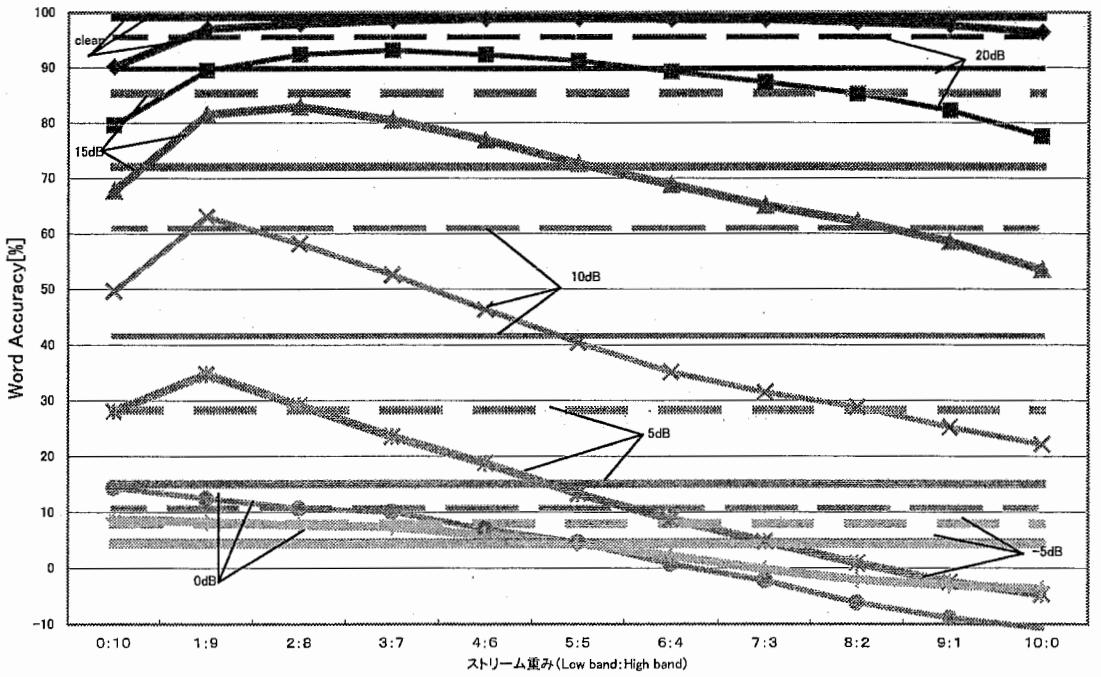

図 2.35 ストリーム重みに対する認識性能（train station）

2.3. 考察

マルチストリーム特徴量を用いることで，すべての雑音条件において適当な重みを与えることで認識率の向上が見られた．最高で雑音の種類 airport で 5dB のとき，27.71%上昇した．しかし，帯域分割をしない MFCC を用いるベースラインと比較した場合はわずかな向上にとどまった．また，シングルストリームとベースラインの比較から，帯域分割して変換したことで大きく認識率が落ちていることが分かる．これは帯域分割して変換する際の情報欠落によるものと考えられる．

3. GPD アルゴリズムによるストリーム重みの自動推定

混入する雑音の種類や SNR により，最適な重みが異なることは前節の結果からも分かる．そこで，実際に使用して認識率を向上させようとしたとき，最適な重みで認識することが重要である．そこで，最適な重みを自動推定する必要がある．今回，過去に画像と音声のマルチストリーム特徴量を用いたバイモーダル音声認識システム[2]で最適な重みの自動推定に用いられていた GPD アルゴリズムを用いた．

3.1. GPD アルゴリズム

GPD アルゴリズムとは雑音入りの音声データとその認識結果をセットとして用意する．そしてある重みからスタートして認識をし，認識率の傾きから新たな重みを計算し，その重みで認識をするといったことを繰り返し，最適な認識率に近づいていくアルゴリズムである．

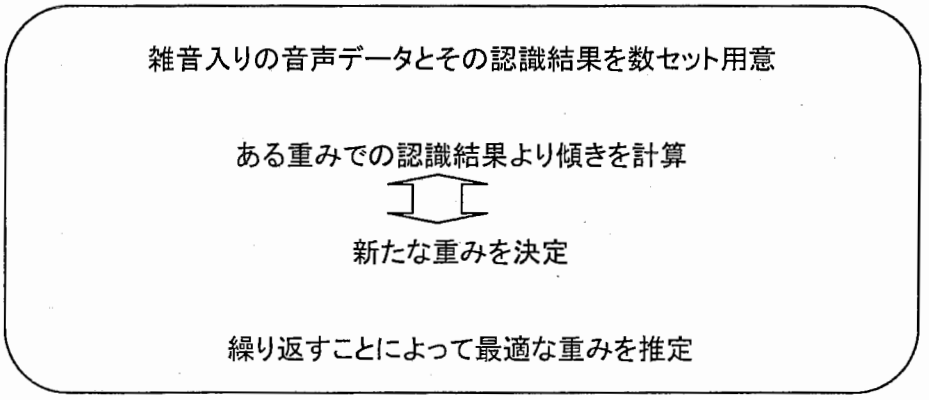

図 3.1 GPD アルゴリズム

3.2. 適応実験

MFCC を用いた評価実験において，比較的効果の大きかった airport と train station について GPD アルゴリズムによる重み推定を行い，推定された重みを用いた認識実験を行った．重み推定に用いた適応データ数を 10 発話と 100 発話の 2 種類で行った．この適応データは評価セットの一部である．

認識結果を図 3.2 に示す．手動重みによるマルチストリーム，シングルストリーム，ベースラインの結果は図 2.34，2.35 と同一である．

3.3. 認識結果

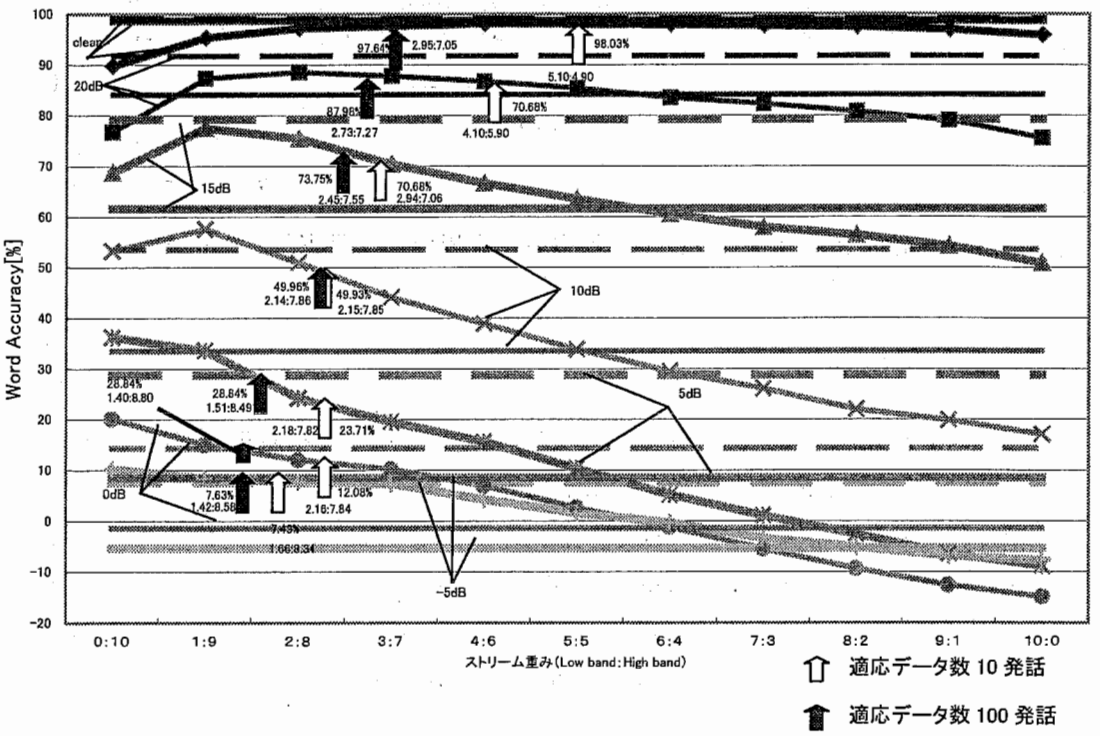

図 3.2 GPD アルゴリズムによる推定重みを用いた認識実験（airport）

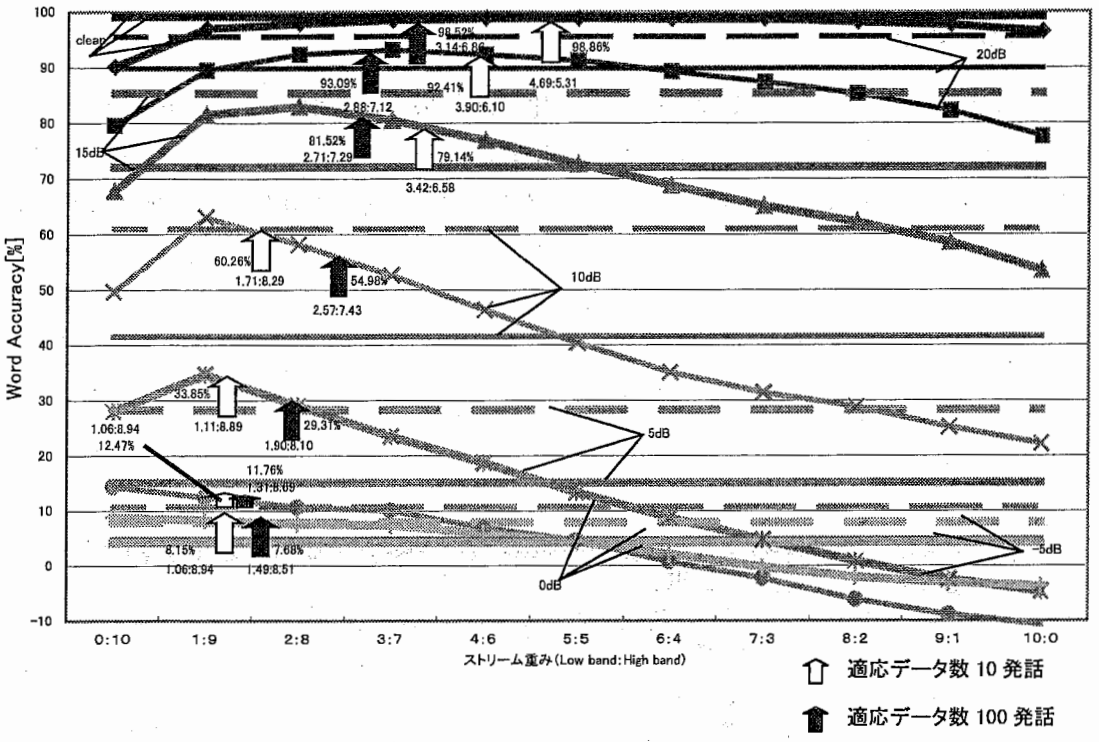

図 3.3 GPD アルゴリズムによる推定重みを用いた認識実験（train station）

3.4. 考察

推定された重みは手動による重み推定で求めた認識率の曲線の頂点に完全に一致していないが，比較的良好な推定結果を得た．推定重みによる認識性能は，すべての条件でシングルストリームの性能を上回った．雑音が airport で SNR が 5dB のとき 20.31%向上し，雑音が train station で SNR が 5dB のとき 18.76%向上した．また，適応データ数が多い方が重み推定精度，認識性能ともに向上する傾向が見られた．ベースラインとの比較では，train station の 5dB と 0dB の条件を除き，性能向上は見られなかった．

4. まとめ

マルチストリーム特徴量を用いた音声認識システムを構築し，フィルタバンク係数と MFCC の特徴量を用いた音声認識実験を行った．実験結果により，フィルタバンク係数を特徴量として用いた場合は特徴量としての性能が悪い面もあり，SNR が低い部分でのみマルチストリームの認識性能がシングルストリームに比べて上回った．MFCC ではすべての条件でシングルストリームと比べて認識性能が向上した．雑音の種類が airport で 5dB のとき，27.71%の向上を得た．

また，GPD アルゴリズムによるストリーム重みの自動推定を行い，推定された重みにより，認識性能が向上することが確認できた．雑音が airport で SNR が 5dB のとき 20.31%の向上を得た．

今後の課題として，MFCC の帯域分割する際の問題点があげられる．帯域制限雑音を用いた実験の際にあまり顕著な結果が得られなかったことから，音声の特徴の重要な部分がノイズと重なっていることが考えられる．本研究では，メルスケール上で等分した 6 次元+6 次元の 2 ストリームを用いたが，分割比や分割数について検討する必要がある．また，情報が欠落していると考えられる帯域分割した MFCC を用いることにより認識率が大きく低下してしまう問題を解決する必要がある．

GPD アルゴリズムによるストリーム重みの自動推定についても推定精度や適応データ数と推定速度の関係について検討する必要がある．

参考文献

[1]Jingdong Chen, Kuldip Paliwal, Satoshi Nakamura, Sub-Band Based Additive Noise Removal for Robust Speech Recognition, Eurospeech2001, Vol.1, pp.571-574, 2001 年

[2] 熊谷建一, 中村哲, 鹿野清宏 バイモーダル音声認識のためのモデル合成に基づく統合法と適応化 情処研報 200-SLP-34-12, 2000 年 12 月

[3]S.B.Davis, P.Mermelstein, Comparison of parametric representations for monosyllabic word recognition in contiuously spoken sentences, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-28(4)pp.357-366, 1980 年

[4]H.Ghirsch, D.Pearce, The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions, ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium", 2000 年 9 月