変調スペクトルによる雑音下における自動音声区間検出：音声周波数帯域及び変調周波数帯域の検討*

◎Pek Kimhuoch，荒井隆行（上智大），金寺登（石川高専），△吉井順子（フジヤマ）

1　はじめに

音声認識・音声情報処理をはじめとする分野では音声区間検出の技術を前処理として用いることが多く，近年では字幕翻訳の分野への応用も注目されている。現在の字幕翻訳では，ほとんどが翻訳者の手作業で行われており，中でも特に時間がかかるのが字幕付与区間の決定である。よって，動画の字幕翻訳・吹き替えをする際、音声の in 点（開始点）と out 点（終了点）の音声区間の時間を記述したタイムコードを自動的に作成し（自動音声区間検出)，翻訳作業を支援できる技術が必要となっている[1]。

動画の多くは背景雑音や BGM が含まれている。先行研究[2]では，音声に雑音が多く含まれる環境で自動的に音声部・非音声部を検出するための変調フィルタリング手法を提案した。しかし，先行研究の手法では音声周波数帯域と変調周波数帯域を固定したが，それに対して本研究では，上記の二つの帯域を変化させて，音声部・非音声部を検出するために重要な音声周波数帯域と変調周波数帯域を検討した。

2　変調スペクトルによる音声区間検出

本研究では，変調スペクトルを用いたアルゴリズムを提案し，自動音声区間検出を試みた。変調スペクトルとは，入力音声特徴量の時間変化を周波数領域で表したものであり，その周波数領域は変調周波数と呼ばれている。先行研究[3-5]より，雑音環境下において変調周波数が 2 Hz 以下や 16 Hz 以上の変調スペクトル成分が音声認識性能を劣化させることが報告されている。本研究では，音声周波数帯域と変調周波数帯域を変化させて，音声区間検出に有効な周波数帯域を検討した。さらに，その周波数帯域情報を用いて音声区間検出実験を行った。

本研究の音声区間検出に使用する特徴量の計算の流れを Fig.1 に示す。まず，入力音声データ全体に対して音声周波数帯域で帯域制限を行った。提案法に対して，どの音声周波数帯域が音声区間検出に有効かを調べるため，いくつかの帯域分けを行った。

次に，制限した帯域の振幅を 2 乗（$|\bullet|^2$）した後，カットオフ周波数が 30Hz のローパスフィルタ（LPF）をかけ，時間包絡を抽出した。この時間包絡は主に 30Hz 以下の周波数成分しか持たないため，低いサンプリング周波数で時間包絡を表現することができる。そこで，時間包絡を 80Hz にダウンサンプリング（↓M）した。ダウンサンプリングした時間包絡に対して，低域遮断周波数 f_L と高域遮断周波数 f_u の範囲でバンドパスフィルタをかけた。その結果に対してフレーム化を行い，各フレームの RMS (root mean square) を計算した。そして，横軸を変調周波数（$(f_u + f_L)/2$）に変換して，縦軸は変調指数（RMS / 全時間包絡の平均）とすることで変調スペクトルを求めた。

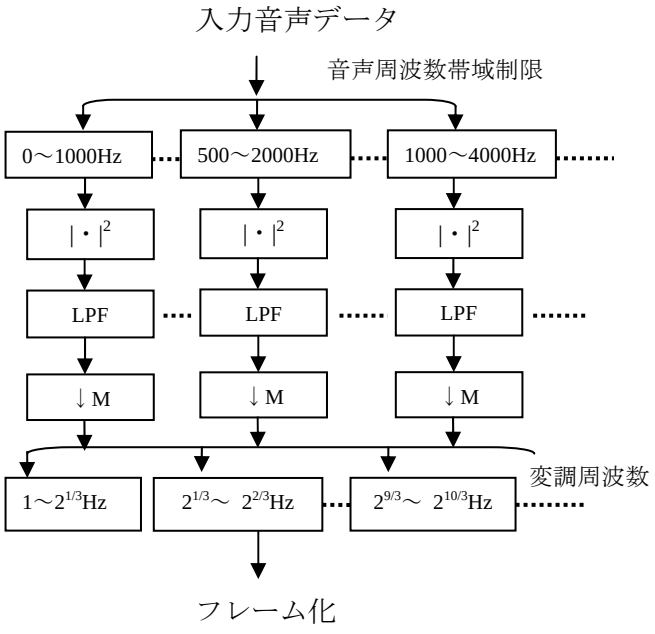

Fig.1 変調スペクトルに基づく特徴量の概要

* Voice activity detection by using modulation spectrum in noise : Investigation on speech frequency band and modulation frequency band, by PEK, Kimhuoch, ARAI, Takayuki (Sophia University), KANEDERA, Noboru (Ishikawa National College of Technology) and YOSHII, Junko (Fujiyama Inc.).

3　実験I

実験 I では，日本語の音声コーパス CENSREC-1-C[5]の数字音声を用いた。このコーパスのサンプリング周波数が 8kHz, 量子化が 16bit, 語彙は数字の 11 種類（1～9，ゼロ，まる），無音の計 12 種類である。収録データの雑音環境はシミュレーション環境と実環境である。実環境では学生食堂と高速道路という 2 つの雑音環境及び 2 つの SNR 環境（低 SNR，高 SNR）が存在する。ここで，低 SNR 環境とは-5dB から 5dB, 高 SNR 環境とは 10dB から 20dB までの範囲となっている。シミュレーション環境では Subway，Babble，Car，Exhibition，Restaurant，Street，Airport，Station を付加雑音として用い，SNR は-5～20dB(5dB 刻み)とクリーン環境である。

実験 I では，1 種類の雑音下（コーパスに収録された Subway の音）での音声区間検出の結果を調べた。まず，入力音声データの周波数帯域を 0～2000Hz に固定して変調周波数帯域を調べた。次に，最も結果が良かった変調周波数帯域に固定して，音声の周波数帯域(例えば，150～1000Hz，150～2000Hz，150～3000Hz)を変化させながら, 音声区間検出の正解率を調べた。フレーム長は 112.5ms，フレームシフトはフレーム長の 1/3 を用いた。

3.1 変調周波数帯域の実験

変調スペクトルの例を Fig. 2, 3 に示す。音声の変調スペクトルでは 2～8Hz 付近に局所的なピークがあるのに対して，白色雑音における変調スペクトルのピークは変調周波数の広い範囲に分布し，複数のピークが存在している。

バブルノイズは複数の話者が同時に発話した雑音であり，音声のエネルギ変化が似ているため，バブルノイズの変調スペクトルのピークが音声の変調スペクトルのピークと重ねる場合が多いと言える。しかし，変調周波数の 5～7Hz 付近のバブルノイズの変調指数は音声に比べて低いところに見られる。

これらの結果から，5～15Hz 間の変調周波数帯域における変調スペクトルを用いることにした。5Hz 以下の変調スペクトル成分は音声情報も存在するが，雑音情報や無音区間の情報も含まれる。よって, 本研究では 5～15Hz の変調周波数帯域の変調指数を特徴量として，

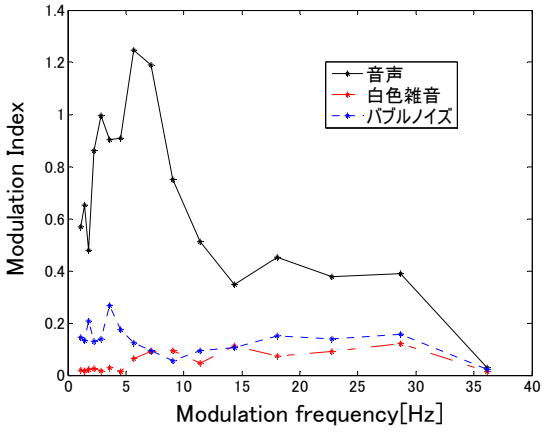

Fig. 2　音声，白色雑音，バブルノイズの変調スペクトル

音声の周波数帯域の実験を行った。

3.2 音声周波数帯域(BPF)の実験

音声・非音声を判別するためにはしきい値を決定する必要がある。しきい値を求めるために，学習データにおける特徴量を求めコーパスの正解ラベルをもとに各フレームで音声と非音声に分類し, ヒストグラムを作成した。そのヒストグラムからマハラノビス距離を用いて，しきい値を求めた。評価データに対して得られたしきい値を基準とし，音声・非音声の判別を行った。

学習データは Subway 雑音が付加された音声（SNR=10dB, 15dB, 20dB）とし，しきい値を求めた。得られたしきい値は 0.07～0.09 の間となった。この実験では，しきい値（0.075）を固定した。それは，企業の IR 広告動画の SNR が 10～30dB 間の環境で収録されることが多く[1]，上記のしきい値を用いることで，音声区間がより多く検出できると考えるためである。

また，単純なしきい値判定ではある種の無声子音において音声区間が切れ，非音声区間として判定されてしまうなどの不都合が生じる。そこで，先行研究[6]にならい，本研究では音声区間に挟まれた非音声区間が 300ms 以下の場合には音声区間としてつなげることにした。

先行研究[7]より，字幕表示タイミングは音声が終了後もしばらく残った方がわかりやすいという報告がある。本研究では，動画の字幕翻訳を主な目的としている。よって，コーパスで作成される音声の正解ラベルの out 点

より長め(最長 250ms)に検出されても正解とした。評価方法は以下の式，フレームベースの FRR (False Rejection Rate) と FAR (False Acceptance Rate)を用いた。

$$\text{音声誤り率}(FRR) = \frac{\text{誤った音声フレーム数}}{\text{音声フレーム数}}$$

$$\text{非音声誤り率}(FAR) = \frac{\text{誤った非音声フレーム数}}{\text{非音声フレーム数}}$$

Subway (SNR=5dB)のデータを用いたときの音声の変調帯域の実験結果を Fig. 3-5 に示す。BPF が 150～1000Hz 間に比べて 150～2000Hz 間のとき，非音声区間を音声区間として誤って検出されることが多く見られる。また，BPF が 150～3000Hz 間のときも同じ傾向が見られる。つまり，音声周波数帯域の幅を広げるほど，音声情報がより多く抽出される一方，音声・非音声の判別が難しくなっている。本研究の提案法においては，音声周波数帯域が 3000Hz 以上になると、音声・非音声の判別が困難であることが明らかになった。

Table 1, 2 に BPF を変化させた時の音声（非音声）誤り率の結果を示す。列は BPF の上限周波数であり，行は BPF の下限周波数を示す。Table 1,2 の結果より，BPF の上限周波数として 1000～2000Hz の間，下限周波数として 0～250Hz の間を組み合わせた場合に音声・非音声誤り率が低くなることが明らかになった。

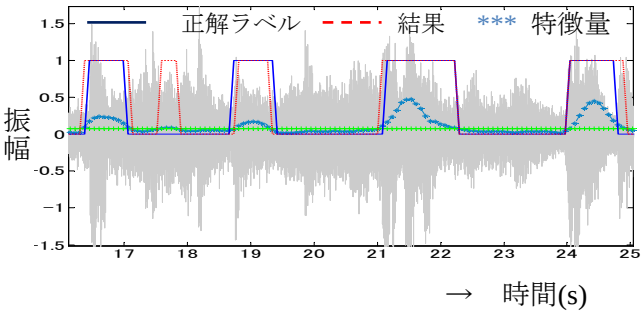

Fig. 3 BPF=150～1000Hz (Subway, SNR=5dB)

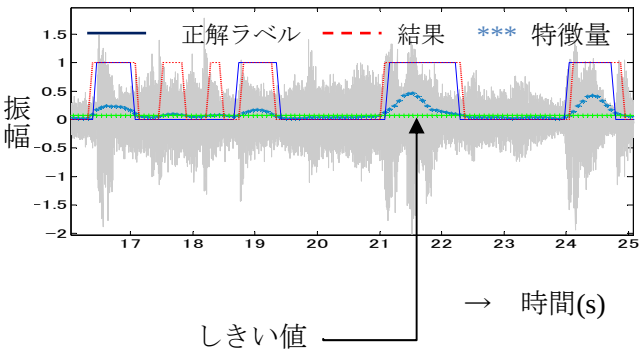

Fig. 4 BPF=150～2000Hz (Subway, SNR=5dB)

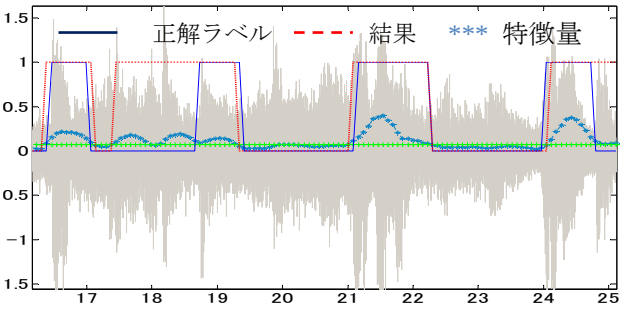

Fig. 5 BPF=150～3000Hz (Subway, SNR=5dB)

Table 1 音声誤り率 (Subway, SNR=5dB, %)

上 / 下 (Hz)	125	250	500	1000	2000	3000
0	5.63	6.18	6.34	6.18	6.50	6.57
125		6.26	6.33	6.19	6.50	6.57
250			7.18	6.76	7.09	7.01
500				14.16	14.64	12.27
1000					13.59	15.57
2000						25.54

Table 2 非音声誤り率 (Subway, SNR=5dB, %)

上 / 下 (Hz)	125	250	500	1000	2000	3000
0	68.83	29.22	22.29	13.77	13.35	20.71
125		29.50	22.28	13.75	13.35	20.70
250			27.85	14.90	14.38	22.60
500				16.84	17.22	34.22
1000					39.20	51.39
2000						62.45

4 実験II

本実験では，実験 I で得られた音声区間検出の正解率の結果が高い音声周波数帯域と変調周波数帯域を用いて音声区間検出の実験を行った。実験 I では，CENSREC-1-C のコーパスの中で 1 種類の雑音（Subway）だけ用いて実験を行った。しかし，雑音の種類によっては音声と性質（例：エネルギの時間変化）が似ているものもあるため，音声・非音声を判別するのが難しくなることがある。また，動画の多くはクリーン環境で収録されることが少ない。そこで，動画の音声データにより近い環境で実験を行うため，本実験では CENSREC-1-C コーパスの実環境のデータ（remote microphone）用いて実験を行った。

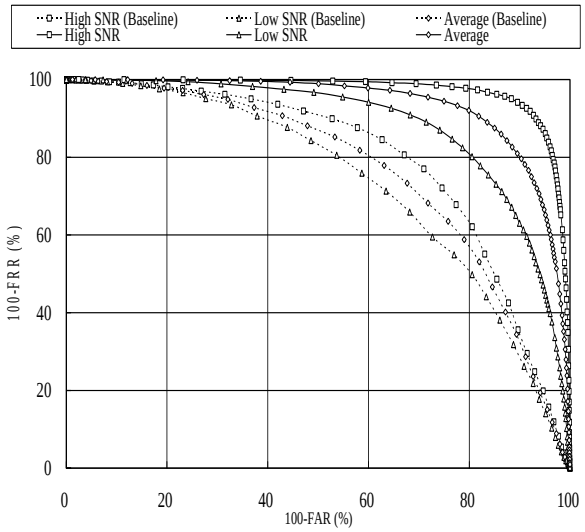

Fig. 6　実環境における SNR 別の ROC 曲線

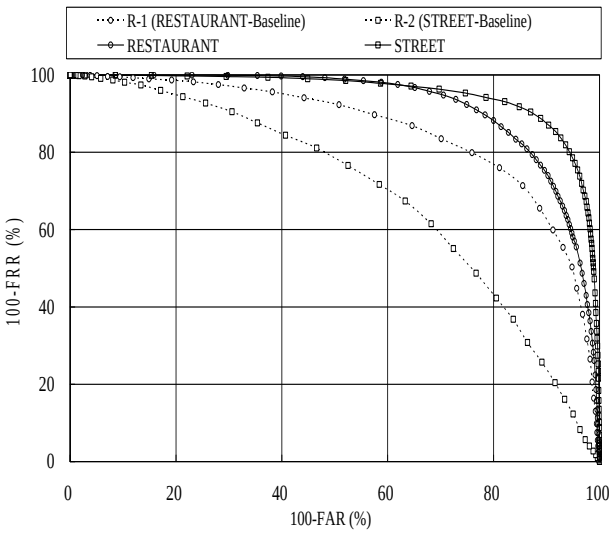

Fig. 7　実環境における雑音別の ROC 曲線

4.1 実験方法と結果

フレーム長は 112.5ms，フレームシフトはフレーム長の 1/3 を用いた。音声周波数帯域幅 BPF は 125～2000Hz に固定し，5～15Hz 間の変調周波数帯域における変調スペクトルを特徴量として実験を行った。評価方法はしきい値（0.005～1.605，0.005 間隔）を変化させながら FRR と FAR で算出し, ROC (Receiver Operating Characteristic)曲線をプロットした。実験結果は，コーパスに収録されている Baseline result による結果と比較した。実験 I と同様に, 正解ラベルの out 点より最長 250ms まで長めに検出されても正解とした。実環境における音声区間検出の結果を Fig. 6-7 に示す。Fig. 6 は SNR 別の結果であり，Fig. 7 は雑音の種類別で示す。提案法による FRR, FAR の結果は低 SNR と高 SNR において, Baseline の結果に比べて改善が見られた。また，雑音別の音声区間検出の結果においても，本研究の特徴量は従来法に比較して改善が見られる。特に，Street 雑音が付加されたときの結果が従来法と比べて 20%程改善が見られた。

5　おわりに

本研究では，変調スペクトルによる音声周波数帯域と変調周波数帯域の条件を変化させた場合の音声・非音声誤り率を調べた。実験 I より得られたパラメータを用いて，実環境における音声データに対して音声区間検出の実験を行った。変調スペクトル成分中で，5～15Hz の変調周波数バンドにおける変調スペクトルが高かった。BPF の上限周波数は 1000～2000Hz 間と下限周波数が 0～250Hz 間を組み合わせた場合に音声・非音声誤り率が低かった。また，実験 II では，実環境における音声データの音声区間検出の誤り率については，従来法と比べて提案法がより低い結果となった。

今回の研究では音声データのフレーム長を固定し，日本語の音声データのみ用いて実験を行った。今後の課題としては，多言語の音声データを用いて，フレーム長を変化させながら実験を行いたい。

謝辞

この研究の一部は，文部科学省私立大学学術研究高度化推進事業　上智大学オープン・リサーチ・センター「人間情報科学研究プロジェクト」の支援を受けて行われた。

参考文献

[1] http://www.fujiyama1.com.
[2] K. Pek et al.，音講論，pp. 33-34，2009.
[3] T. Arai et al., Proc. ICSLP, pp. 2490-2493, 1996.
[4] N. Kanedera et al. , Proc. Eurospeech, pp. 1079-1082, 1997.
[5] T.Arai et al., JASA,105(5), pp. 2783-2791, 1999.
[6] 北岡教英他，音講論, pp. 103-104, 2006.
[7] 藤樫佑樹他，音講論，pp. 33-34，2005.
[8] 古賀綾子他，音講論，pp.445-446，2007.