変調フィルタリングによる自動音声区間検出とその多言語における比較*

◎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

インターネットの通信速度の高速化の背景により，様々な動画のコンテンツが世界中の人々に配信されるようになった。カンボジアでもインターネット上をはじめ，動画を楽しむ機会が増えてきている。他国の動画を視聴する場合，外国語を理解できない人にとっては字幕翻訳・吹き替えが不可欠である[1]。さらに，聴覚障害者や言語学習者のためにカンボジア語の字幕翻訳を付与する技術は必要である。しかし，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。この問題に対して，先行研究[2]では，音声データに対して，音声の特徴を分析することで音声の 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アルゴリズムを提案してきた。

先行研究[2]では，実験で使用されるデータの言語は日本語で，比較的雑音の少ない環境を前提として自動音声区間検出を行っている。それに対して本研究では，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新しい手法を提案した。さらに，提案法を用いて多言語に対する音声区間検出の比較を試みた。

2　変調フィルタリング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，入力音声データ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, 4]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，音声情報が多く存在する 2-8Hz[4]の変調周波数帯域を用いて音声区間検出実験を行った。

Fig.1 に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示す。まず，入力音声データに対して 125ms から 1000ms のフレーム長でフレーム化し，フレーム長の 1/3 ずつフレームシフトを行った。音声の情報の多くが 500-2000Hz 間の周波数帯域に存在するため[5]，その周波数帯域で制限を行った。

次に，500-2000Hz 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）をかけて，時間包絡を抽出した。この時間包絡は 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て 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 modulation index（RMS / 時間包絡の平均）とすることで変調スペクトルを求めた。

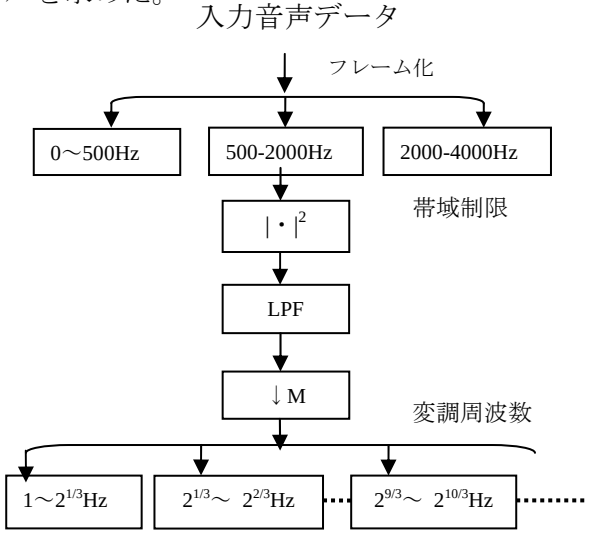

Fig.1 変調スペクトルに基づく特徴量の概要

* Voice activity detection by using modulation filtering and its multi-language comparison, by PEK, Kim Huoch, ARAI, Takayuki (Sophia University), KANEDERA, Noboru (Ishikawa National College of Technology) and YOSHII, Junko (Fujiyama Inc.)

3　予備実験

予備実験では，日本語の音声データベースCENSREC-1-C[6]を用いた。このデータベースのサンプリング周波数は8kHz, 量子化は16bit, 語彙は数字の 11 種類（1～9，ゼロ，まる），無音の 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 2 つの学生食堂と高速道路の雑音環境及び 2 つの SNR 環境（低 SNR，高 SNR）を用いている。ここで，低 SNR 環境とは-5dB から 5dB までの値で，高 SNR 環境とは 10dB から 20dB までの値となっている。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 20～-5dB(5dB 刻み)とクリーン環境である。

予備実験では，入力音声データのフレーム長(125ms，250ms，500ms，1s)を変化させながら，1 種類の雑音（データベースに収録された Subway の音）下での音声区間検出の正解率を調べた。まず，使用したデータベースを学習データと評価データに分けた。次に，音声・非音声のしきい値を求めるため，学習データの特徴量をデータベー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。評価データでは得られた閾値を基準にし，音声・非音声の判別実験を行った。

しかし，単純に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。そこで，先行研究[2]にならい，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。よって，本研究でもこの 300ms ルールを用いて実験を行った。

実験結果を Fig.2 と Fig.3 に示す。評価結果は以下の式を用いた。総誤り率，音声（非音声）誤り率，音声（非音声）再現率，適合率で算出した。

$$\text{総誤り率} = \frac{\text{誤ったフレーム数}}{\text{全フレーム数}}$$

$$\text{(音声・非音声) 誤り率} = \frac{\text{誤った (音声・非音声) フレーム数}}{\text{(音声・非音声) フレーム数}}$$

$$\text{(音声・非音声) 再現率} = \frac{\text{正解した (音声・非音声) フレーム数}}{\text{(音声・非音声) フレーム数}}$$

$$\text{(音声・非音声) 適合率} = \frac{\text{正解した (音声・非音声) フレーム数}}{\text{出力結果での (音声・非音声) フレーム数}}$$

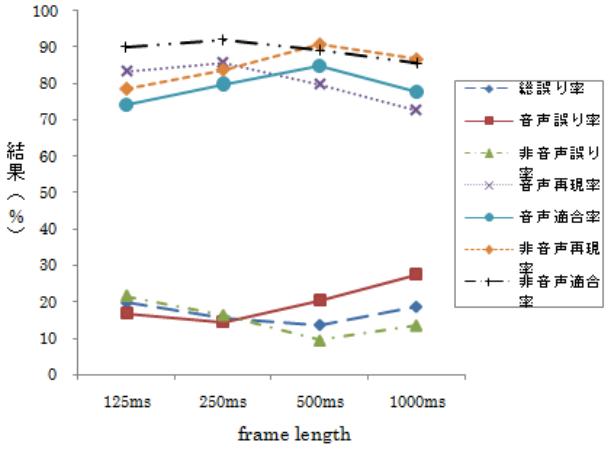

Fig.2 フレーム長の違いに対する結果 (SNR=5dB)

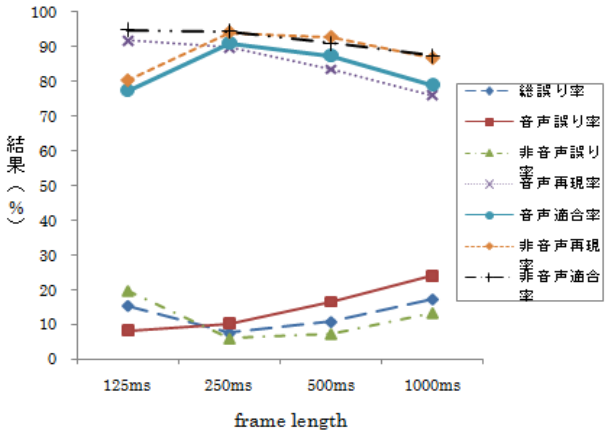

Fig.3 フレーム長の違いに対する結果(SNR=10dB)

Fig.2 と Fig.3 の上部は音声（非音声）正解率で, 下部は誤り率を表わしている。よって，上部では 100%に近い方が望ましい結果となり，下部は小さい（0%付近）ほど良い結果である。これらの結果より，フレーム長が 125ms-250ms の間で音声区間検出の正解率が高かった。その中で，250ms における音声区間検出の正解率の結果が最も高く，音声（非音声）誤り率が低かった。

4　音声区間検出の実験

本実験では，予備実験で音声区間検出の正解率の結果が高いフレーム長(125ms)を用いて，音声区間検出の実験を行った。予備実験では，CENSREC-1-C のデータベースを 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）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，動画には様々な言語があり，言語によって特徴が異なっている（例：閉音節，開音節)。

4.1 実験I

動画には音楽・バブルノイズ，車など様々な背景雑音が含まれることが多い。そこで実験 I は Noisex-92 [7]から 6 種類[White noise, babble，HF Radio Channel Noise (hfchannel), Passenger compartment (Volvo)，F-16 cockpit noise (f1)，Noise on floor of car factory (factory1)]の雑音下で変調スペクトルを特徴量として実験を行った。使用する音声としては JNAS[8]から男女各 12 名が，5 種類の日本語の文を読み上げた合計 120 文を用いた。データベースはサンプリング周波数 16kHz，16bit 量子化，平均データ長 10 秒である。また，SNR は 30～0dB (10dB 刻み) とクリーン環境条件で実験を行った。実験方法は予備実験と同じであった。実験結果は，従来法として国際規格である G.729[9]による結果と比較した。

Table 1 提案法による結果 （6 種類の雑音の平均, %）

	clean	30dB	20dB	10dB	0dB
総誤り率	9.61	6.23	6.26	6.68	12.36
音声誤り率	8.85	4.60	4.66	4.84	9.38
非音声誤り率	12.24	13.41	13.33	14.43	23.22
音声再現率	91.15	95.40	95.34	95.16	90.62
音声適合率	96.78	96.72	96.73	96.42	93.86
非音声再現率	87.76	86.59	86.67	85.57	76.78
非音声適合率	77.82	83.54	83.40	82.85	71.86

Table 2 G.729 による結果 （6 種類の雑音の平均, %）

	clean	30dB	20dB	10dB	0dB
総誤り率	12.49	13.50	18.43	25.76	46.38
音声誤り率	9.31	11.29	17.37	26.54	52.31
非音声誤り率	28.56	24.68	23.78	22.01	18.27
音声再現率	90.69	88.71	82.63	73.46	47.69
音声適合率	93.81	94.38	94.19	94.17	93.58
非音声再現率	71.44	75.33	76.22	77.99	81.73
非音声適合率	61.62	58.71	49.31	39.88	27.55

Table 3 提案法による結果 （バブルノイズのみ, %）

babble	clean	30dB	20dB	10dB	0dB
総誤り率	9.61	6.20	6.29	7.10	16.56
音声誤り率	8.85	4.59	4.66	4.87	11.99
非音声誤り率	12.24	13.32	13.48	16.27	32.54
音声再現率	91.15	95.41	95.34	95.13	88.01
音声適合率	96.78	96.75	96.70	95.97	91.24
非音声再現率	87.76	86.68	86.52	83.73	67.46
非音声適合率	77.82	83.58	83.43	82.63	65.41

Table 4 G.729 による結果 （バブルノイズのみ, %）

SNR	clean	30dB	20dB	10dB	0dB
総誤り率	12.49	14.19	19.54	27.40	46.48
音声誤り率	9.31	11.45	17.72	28.05	52.00
非音声誤り率	28.56	27.74	28.45	24.17	21.86
音声再現率	90.69	88.55	82.28	71.95	48.00
音声再現率	90.69	88.55	82.28	71.95	48.00
非音声再現率	71.44	72.26	71.55	75.83	78.14
非音声適合率	61.62	57.50	46.35	36.85	24.86

Table 1 と 2 に 6 種類の雑音に対する結果を平均したものを示す。これらの結果は従来法と比較して改善が見られた。また，全ての SNR について，音声再現率・音声適合率ともどの雑音に対しても 90%以上の結果が得られた。非音声再現率・非音声適合率とも従来法と比べて改善が見られたため，動画の音声区間検出に向いていると言える。

さらに，6 種類の雑音の中から 1 種類の雑音のみに注目し，バブルノイズ下での提案法と従来法の音声区間検出の結果を Table 3，4 に示す。従来法と比べてクリーン環境における提案法の音声誤り率は少ししか改善が見られなかったが, 雑音環境では大きく改善した。また，提案法の非音声誤り率は従来法と比較して全ての環境において，改善が見られた。よって，変調スペクトルを特徴量とした雑音下での音声区間検出について，低い誤り率の結果を得た。

4.2 実験II

言語によって特徴(例：閉音節，開音節)が異なるため，実験 II では多言語を対象とし，特徴量として提案法を用いてクリーン環境下で実験を行った。実験データは，多言語音声コーパス[10]から 5 言語(日本語・中国語・英語・フランス語・タイ語)を用いた。音声のデータはイソップ童話「北風と太陽」を読み上げたコーパスで，男女各 3 名がそれぞれ 1 種類の文を読み上げたものを用いた（サンプリング周波数 16 kHz, 16 ビット量子化，平均データ長約 1 分，防音室で収録)。さらに，筆者の母国語であるカンボジア語のデータ(「北風と太陽」をカンボジア語を母国語とする話者に読み上げてもらったもの）を加えて 6 言語間で音声区間検出の結果を比較した。実験の

方法としては 6 人のうち 5 人分を学習データとし，しきい値を求めて，残りの 1 人は評価データとして使用することで，全部で 6 通り行った。最後にそれらの結果を平均して，結果を算出した。

Table 5　各言語に対する結果（%）

言語	JPN	FRA	ENG	Thai	CHI	KHM
総誤り率	6.06	3.73	2.90	2.24	4.78	4.84
音声誤り率	5.70	3.07	2.71	1.71	4.08	5.16
非音声誤り率	6.81	5.83	3.77	3.68	6.88	5.35
音声再現率	94.30	96.93	97.29	98.29	95.92	94.84
音声適合率	96.78	97.86	98.89	98.51	97.53	97.98
非音声再現率	93.19	94.18	96.23	96.32	93.12	94.65
非音声適合率	89.10	92.91	91.85	96.03	90.41	91.27

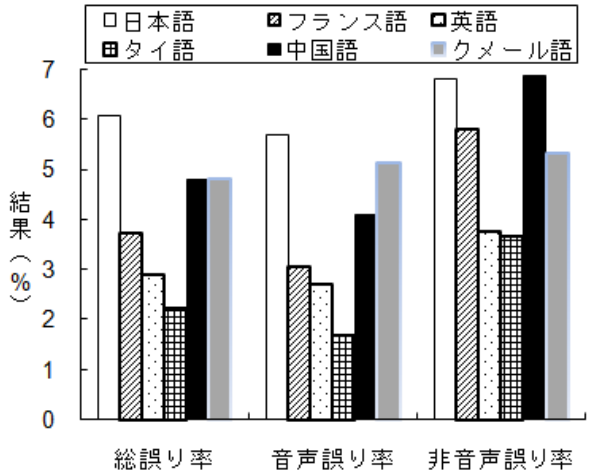

Fig.4 各言語に対する誤り率の結果

各言語に対する提案法の特徴量による音声区間検出結果を Table 5 と Fig.4 に示す。これらの結果より，各言語の総誤り率と音声誤り率が 6%以下で，非音声誤り率が 7%以下となっている。また，音声（非音声）再現率と音声（非音声）適合率は 6 言語とも 89%以上となっている。その中で，タイ語の結果が一番良く，続いて英語・フランス語・中国語・カンボジア語・日本語の順番になっている。英語･タイ語の誤り率が他の言語より低いのは，音声データに含まれる息つぎの回数が少ないことから，息つぎの部分を誤って音声区間として検出するケースが少なかったためだと考えられる。また，日本語・フランス語における音声区間検出の結果では，正解ラベルと比較して音声の out 点（音声の終了点）が長めに検出されてしまうことが多く見られた。それは，日本語・フランス語が母音で終わる開音節の言語で，音声の out 点が曖昧になると考えられる。今回音声・非音声の判別実験で用いた基準では，音声端点が 1 フレームでもずれた場合は誤りとされるため，誤り率が増加してしまうが，実用的には問題ないと考えられる。6 言語間での誤り率の結果は少し異なるが，いずれも誤り率が 7%以下となっているため提案法がどの言語に対しても適用できるものと結論付けられる。

5　おわりに

本研究では，雑音が付加された音声データの字幕翻訳作業の効率化を目的として，自動的に音声の開始部と終了部を決定する自動音声区間検出手法（変調スペクトルが特徴量）の提案と，提案法による多言語の音声区間検出の実験を行った。実験 I では，雑音が含まれる音声データの音声区間検出の正解率の結果は従来法と比べて，提案法が高かった。実験 II では，提案法による音声区間検出を多言語間で比較した結果，いずれの言語に対しても，音声区間検出の誤り率が低い結果が得られた。

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] 藤樫佑樹，音講論，pp. 33-34，2005.
[3] T. Arai et al., Proc. ICSLP, pp. 2490-2493, 1996.
[4] N. Kanedera et al. , Proc. Eurospeech, pp. 1079-1082, 1997.
[5] D. M. Jones, (New York , Wiley, 1983).
[6] 北岡教英 et al. , 音講論, pp. 103-104, 2006.
[7] A. Varga and H. J. M. Steeneken, Speech Communication, 12(3): 247 -251, 1993.
[8] http://www.mibel.cs.tsukuba.ac.jp/jnas/
[9] ITU-T, Annex B, V.70, ITU-T Recommendation 1996.
[10] 板橋秀一 et al. , 筑波大学, 2001.