字幕付与システムを目的としたスペクトル遷移による音声端点検出*

◎古賀綾子，藤樫佑樹，荒井隆行（上智大・理工），金寺登（石川高専），
△吉井順子（フジヤマ）

1　はじめに

近年，通信技術の急速な発展によって，世界へ向けた動画コンテンツの配信が盛んになってきており[1]，それらへの字幕翻訳の必要性が増してきている。また，同一言語内においても字幕放送による聴覚障害者へのニーズは高い[2]。しかし，音声・非音声の決定を含む字幕付与の過程は現在，ほとんど字幕翻訳者による手作業で行われている。そこで私たちは，字幕付与を効率化するために，自動音声区間検出を提案してきた[3]。本研究では，音響的な変化を特定する子音性ランドマークに基づき，音声端点検出を試みた。

2　音声・非音声判別

2.1 子音性ランドマーク

子音性ランドマークは, 音声信号において, 子音と母音のような音響的な変化が起こる時間を特定するものである[4]。本研究では，子音性ランドマークを音声端点検出に応用するため，スペクトル遷移によるランドマークの検出を試みた。

2.2 スペクトル遷移

スペクトル遷移を表すDの算出は基本的にはFuruiが行った方法[5]と同じであるが，本論文ではArai *et al.* [6-7]によって提案されている帯域分割された信号の時間包絡に対する回帰係数を複数帯域に渡って二乗平均したものを用いた（Fig. 1）。このようにして求められたDは音声部と非音声部，母音部と子音部などの境界において顕著になる。そこでDのピークを音声･非音声の境界の候補として検出し，検出されたピークで挟まれる区間の対数エネルギーの平均で閾値処理を行うことによって音声区間を判定した。音声・非音声判別の閾値は，評価データの一部を手作業で切り出したものを学習することにより決定した。

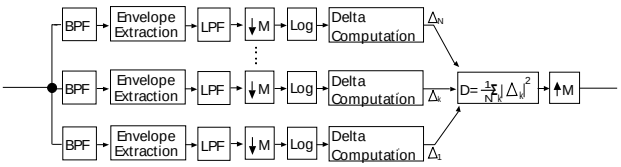

Fig. 1 Block diagram for calculating *D*.

2.3 2 段階処理

本研究では，音声・非音声区間の境界を微調整することを目的として，2 段階処理を行った。2 次処理では，無声子音などが原因の誤判定を考慮し，音声・非音声区間の境界において 1 次処理よりも対数エネルギーの閾値を下げることにより調整した。

3　実験方法

実験データには，日本音響学会編『新聞記事読み上げ音声コーパス（JNAS）』から男女各12名がそれぞれ5種類の文を読み上げた計120 文（サンプリング周波数 16kHz，16 ビット量子化，平均データ長 10.5 秒，SNR 約 30 dB）を用いた（音声・非音声に関する正解ラベルは，本コーパスに付属する音素ラベルを参考に作成した）。本実験では, 動画コンテンツのインターネット配信の中でも需要の多い企業の IR 広告に近い環境で実験を行うため, 同時に 1 話者のみが発話する音声データを用いるものとし，背景雑音があまりないものを選んだ。また，動画コンテンツへの応用を考慮し，音声を WMA 方式に圧縮した後，再び PCM 方式に伸長した。

本研究では，字幕表示タイミングの自然性の観点から，字幕付与区間を決定するルールとして 300ms 未満の長さの短い非音声部は音声区間に含めるものとした。この評価結果を以下に示す誤り率，再現率，適合率で算出した。

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

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

* Detecting end-points of speech for captioning system using spectral transition, by KOGA, Ayako, FUJIKASHI, Yûki, ARAI, Takayuki (Sophia Univ.), KANEDERA, Noboru (Ishikawa Natl. College of Technol.), YOSHII, Junko (Fujiyama Inc.)

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

4　実験結果

結果を Table 1 に示す。従来法として，対数エネルギーのみを特徴量としたものの結果を比較した。また，提案した 2 段階処理の効果を調べるため, 1 次処理のみを行ったものと 2 段階の処理を行ったものの結果を併記した。

表に示すように提案法は従来法に比べて，特に音声誤り率に大きな改善が見られた。これは音声・非音声区間の境界をスペクトル遷移 D を用いて判定したことで，より正確な音声端点検出が行われたことを表している。また，2 段階処理に関しても音声区間の検出精度に向上が見られ，補正の効果があったことが確認された。

5　動画コンテンツへの応用に向けて

現在の動画コンテンツに対する字幕付与作業は，字幕付与区間の決定から時間長に適合した文字数での翻訳まで，そのほとんどが翻訳者自身による手作業である。そのため作業の効率化のため，字幕付与区間決定の自動化は有用であると考えられる。

本研究では提案した音声・非音声判別手法を用い，実際にネットワーク配信されている動画コンテンツへの字幕付与区間の決定を試みた。

評価用データはインターネット配信用企業 IR 広告の動画データで，話者 1 名，SNR は約 30 dB であった（このデータに対する音声・非音声に関する正解ラベルは，波形とスペクトログラムの目視と聴取によって手動で行った)。評価結果には誤り率，適合率を用いた。

実験結果をTable 2 に示す。本実験では,従来法として国際規格であるG.729 Annex B [8]

Table 1 Results for a speech corpus [%].

	対数エネルギー	提案法	
		1 次処理のみ	2 段階処理
総誤り率	20.4	6.6	5.8
音声誤り率	24.1	5.0	4.0
非音声誤り率	8.3	13.9	14.2
音声再現率	75.9	95.0	96.0
音声適合率	97.2	96.4	96.4
非音声再現率	91.7	86.1	85.8
非音声適合率	51.2	83.0	86.1

Table 2 Results for an example of real data [%].

	G.729	提案法
総誤り率	5.5	3.9
音声誤り率	8.6	4.4
非音声誤り率	0.7	3.1
音声適合率	99.5	97.9
非音声適合率	88.3	93.5

による結果との比較を行った。

表に示すように提案法は従来法に比べて，特に音声誤り率に改善が見られた。これにより，提案法は実環境に向けた手法としても効果を示したといえる。

6　おわりに

字幕付与の区間決定のため，スペクトル遷移によって子音性ランドマークを決定し，対数エネルギーと組み合わせることによる音声端点検出手法の提案をした。実際にネットワーク配信されているデータでの実験を行い，実用化への可能性を確認した。今後は背景雑音や効果音などの含まれたデータについても研究を進めていきたい。

謝辞

字幕製作における情報を提供いただきました株式会社フジヤマに感謝いたします。

参考文献

[1] 株式会社フジヤマの Webpage http://www.fujiyama1.com

[2] 四日市章, “ 聴覚障害児の字幕の読みに関する実験的研究, ” 風間書房，2002.

[3] 藤樫他，日本音響学会秋季研究発表会講演論文集, 33-34, 2005.

[4] K. N. Stevens, *Acoustic Phonetics*, 1998.

[5] Furui, *J. Acoust. Soc. Am.*, 80(4), 1016-1025, 1986.

[6] 荒井他，日本音響学会秋季研究発表会講演論文集, 1, 449-450, 2001.

[7] Arai *et al.*, *Acoust. Sci. Tech.*, 23, 229-232, 2002.

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