字幕付与システムを目的とした線形予測に基づく音声端点検出*

◎藤樫佑樹，古賀綾子，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

近年におけるブロードバンドネットワークの急速な発展は、ユーザの求める情報を従来のような文字情報だけでなく、音声を伴う動画データとして配信することを可能にした。それと同時に動画配信のユーザ層も様々になりつつあり、音声の多言語化 [1] や聴覚障害者への支援 [2] など、映像への字幕による文字情報付加というニーズも高まっている。

本研究ではこの字幕付与を行う区間を決定するため、線形予測分析に基づく音声特徴量を提案し、実際にネットワーク配信されている動画コンテンツに対する処理の評価を行った。

2　音声・非音声判別法

2.1　音声特徴量 α

特徴量による音声信号からの端点検出は様々提案されている [3]。本研究では音声生成の音源フィルタ理論を基に、フォルマント周波数におけるスペクトルのピーク性に着目した特徴量 α を提案し、音声検出の指標とした（Fig. 1）。

これにより求められた α 値は、音声知覚において重要とされる 3 kHz までのフォルマント周波数成分を用いていることや、処理フレーム内部のエネルギーも反映していることにより、無音部との間に大きな値の差を持つ。この違いは特に有声音において顕著であるが、3 kHz 以上の成分を多く含んだり、あるいは比較的平坦なスペクトル構造を持つ無声音についてはやや低い値を示した。音声と非音声判別の閾値は、それぞれの度数分布に対するマハラノビス距離を用いた学習により決定している。

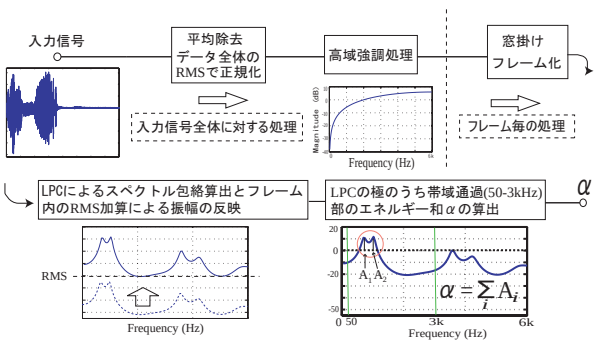

Fig. 1　α 処理のフローチャート

2.2　メディアンフィルタと階層的な判定法

無声音や息継ぎなどに起因する、短時間の α 値の変動を含んだ状態で判別を行うと、判別結果もまた短い間隔での誤判別を起こしてしまう。これに対し音声・非音声区間はある程度の時間間隔をもって交替されるべきものである。そのため本研究では目的に応じた判別処理の階層化を行うことにした。この判別の階層化処理は 1 次処理と 2 次処理から構成され、1 次処理はあきらかに音声と考えられる区間を決定すること、2 次処理は音声・非音声区間の境界を微調整することを目的とする。以下に各処理について述べる。

1 次処理

前節で定義した α 値の時系列を平滑化するため、メディアンフィルタを用いた。これにより得られた $\hat{\alpha}$ 値を用いて音声・非音声の判別を行う。

2 次処理

1 次処理による $\hat{\alpha}$ 値では、音声・非音声区間の境界における α 値の鋭敏な変化もまた抑えられてしまう。しかし字幕付与区間の決定においては区間の端点こそ細かく定められるべきものであるため、このような影響は好ましくない。そこで 1 次処理で非音声とされた区間に対して、2 次処理では平滑化前の α 値による再判別を行う。

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 16 kHz、16 ビット量子化、平均データ長 10.5 秒、SNR 約 30dB）を用いた。また提案手法には、12 次 LPC および 45 次メディアンフィルタに基づく階層的な判別法を用いた。実験の評価は、男女各 3 名による計 30 文を 1 セットとして data1 から data4 の 4 セットに分類し、学習データ 3 セットと評価データ 1 セットでのジャックナイフ法による平均で行う。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。

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

*Linear–prediction based end–point detection of speech for captioning system by FUJIKASHI, Yûki, KOGA, Ayako, ARAI, Takayuki (Sophia University), KANEDERA, Noboru (Ishikawa National College of Technology), YOSHII, Junko (Fujiyama Inc.)

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

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

1 次処理で非音声、かつ 2 次処理で音声と評価されたフレームの最終的な出力は、その前後フレームの判別結果に依存して決定される。これは、前後フレームのどちらかが 1 次処理において音声と判別されていれば音声とし、それ以外の場合は非音声とするものである。

4 実験結果

結果を Table 1 に示す。提案したメディアンフィルタ処理や階層化システムが音声端点検出性能向上にどの程度貢献したかを調べるため、α のみ、および $\hat{\alpha}$ のみによる判別結果と共に階層化処理による結果を比較した。

表に示すように α に対してメディアンフィルタをかけることで、特に誤り率、適合率に大きな改善が見られた。これはフィルタによって判別結果の細かな変動を抑制したことの効果を表している。また、階層化システムに関しても検出精度の向上が見られ、音声・非音声区間の境界に対する端点補正の効果があったことが確認された。

5 映像コンテンツへの応用に向けて

現在の映像コンテンツに対する字幕制作は、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。そのため翻訳者には長時間の作業とそれに伴う負担が生じている。この負担軽減と処理の高速化のため、自動処理によるサポートの導入は有用な方策と考えられる。

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。ここで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 300 ms 以上の非音声部を非音声区間とし、それ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。

評価用データは手作業による正解ラベルつき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 30dB であった。結果は 1/30 秒フレーム単位で判別を行い、評価尺度には誤り率、適合率を用いた。

Table 1　それぞれの手法による処理結果 [%]

	特徴量 α	$\hat{\alpha}$(45 次)	階層化システム
総誤り率	13.9	6.3	5.6
音声誤り率	14.6	5.2	4.1
非音声誤り率	12.9	11.8	12.5
音声再現率	85.4	94.9	96.0
音声適合率	95.3	97.0	96.8
非音声再現率	87.1	88.3	87.6
非音声適合率	65.8	82.6	85.6

Table 2　実データの 300ms ルール適用と評価 [%]

	総誤り率	音声 誤り率	非音声 誤り率	音声 適合率	非音声 適合率
適用前	4.4	5.9	2.0	98.0	91.6
適用後	2.6	3.0	2.0	98.0	95.5

実験結果を Table 2 に示した。300 ms ルール適用前の結果はメディアンフィルタを用いたとはいえ、正解ラベルの音声・非音声区間と比較すると短い間隔で各区間が区切られていた。これに対しルール適用後の結果では、短い時間間隔の非音声部が音声区間の一部として取り入れられ、より字幕に適切な検出結果を得ることができた。これは誤り率や適合率の向上にも示されている。また、端点決定を誤った部分についても、端点におけるずれの大半は動画の 1/30 秒フレームで 1, 2 フレームであり、字幕表示の自然性を考えれば許容範囲であると考えられた。この結果から 300 ms ルールによる区間端点の決定が実環境に向けた手法として効果を示したことがわかる。

6 むすび

字幕付与を行う区間の決定のため、線形予測分析に基づく音声特徴量 α を提案し、ネットワーク配信に用いられる実際のデータでの実験を行った。今後は背景雑音や複数話者による音声、あるいは効果音などの含まれたデータについても研究を進めたい。

謝辞　字幕制作の現状における問題や課題に関する情報を提供いただきました株式会社フジヤマに感謝致します。また共に研究を行った 2004 年度卒業生の三好徹君、浅井健司君、栗山奏君、深見政君に感謝致します。

参考文献

[1] 株式会社フジヤマの Web Page
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] ITU-T, “Annex B A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70,”, ITU-T recommendation, 1996.