動画コンテンツにおける音声の端点と字幕表示タイミングの検討*

◎古賀綾子，松浦杏子，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

様々な動画コンテンツがインターネットを介して世界中に配信されるようなった今日において，字幕翻訳の現状は，その過程のほとんどが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翻訳作業の効率化を目指し，自動音声端点検出を提案してきた[1-4]。自動音声端点検出でタイムコードを作成することにより，翻訳者は効率よく翻訳作業をすることができ，動画と字幕を別のサーバから配信して同期させることで，字幕の多言語対応も可能となる[5]。本研究では，字幕付与システムにおいて自動音声端点検出を実用化するため，動画コンテンツに字幕を付与する際の字幕表示タイミングに関する調査を行った。

2　先行研究[3-4]

2.1 スペクトル遷移による音声端点検出

先行研究では，音響的な変化を特定する子音性ランドマークをスペクトル遷移により検出し，対数エネルギーと組み合わせて 2 段階処理を行うことによって自動音声端点検出を試みた。

2.2 音声・非音声判別実験

実験では実際にネットワーク配信されている動画コンテンツの字幕付与区間決定を試みた。実験結果は従来法として国際規格である G.729 Annex B による結果との比較を行った。実験結果を Table 1 に示す。表に示すように先行研究での提案法は従来法と比較して，特に音声誤り率で改善が見られ，実環境に向けた手法として効果を示した。

3　字幕表示タイミング

先行研究で行った音声端点検出の実験では，検出結果が少しでもずれた場合は誤りとされた。しかし，本研究の究極的な目的は字幕付与システムにおける自動音声端点検出であり，実際に字幕を表示する際には，視聴者にとって実際の音声区間との数フレーム程度のずれは問題にならないと考えられる。そこで，視聴者にとって字幕表示タイミングはどの程度許容されるか，その範囲を調査するための主観評価実験を行った。実験では，外国語の動画の字幕翻訳を表示させる例も想定した。

Table 1 Results of end-point detection of speech [%] (from [3])

	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

4　実験

4.1 刺激

実験データは屋外において収録されたある地域を紹介する動画データ（データ長：約 120 秒，話者 1 名，SNR：約 10 dB）で，音声は日本語・英語のものを使用した。動画データからいくつかの区間を切り出し，手動で日本語の字幕を付与した。刺激は字幕の表示タイミングを in 点（音声区間の始まり）では正解ラベルよりも-1000 ms から 1000 ms まで，out 点（音声区間の終わり）では-600 ms から 1400 ms まで 200 ms ごとにずらしたものを使用した。従って，in 点・out 点，日本語・英語音声それぞれ 66 刺激[（区間 6 種類）×（ずれ 10 種類＋ずれなし）]を用意した。

4.2 実験参加者

参加者は日本語を母語とする健聴者 20 名（男性 11 名，女性 9 名，年齢 19-24 歳）であった。

4.3 手順

実験はコンピュータを用いて，参加者 1 名または 2 名ごとに防音室内で行われた。刺激

*Study on end-point detection of speech and timing of caption for video contents, by KOGA, Ayako, MATSUURA, Kyoko, ARAI, Takayuki (Sophia Univ.), KANEDERA, Noboru (Ishikawa Natl. College of Technol.) and YOSHII, Junko (Fujiyama Inc.)

は動画がコンピュータの画面に表示され，それに対応した音声がヘッドフォン（STAX SR-303，SENNHEISER HDA 200）から提示された。音量は参加者に適した大きさに調整した。実験の前半では日本語・英語音声それぞれの in 点の字幕表示タイミング，後半では out 点の字幕表示タイミングに注目し，回答してもらった。各試行において字幕付きの動画・それに対する音声が 1 度提示され，参加者に字幕表示タイミングが「許容できる」「許容できない」のどちらかを選択させた。選択後，次の刺激が自動的に提示され，各参加者に対する刺激の提示順はランダムであった。

4.4 実験結果

日本語と英語，in 点と out 点それぞれにおいて，95%以上の参加者が許容できる字幕の表示タイミングの範囲を Table 2 に示す。結果は in 点・out 点で大きく異なっていた。

5　考察

5.1 in 点・out 点の比較

表に示すように，参加者は動画における話者が話し始めると同時・またはそれよりも前に字幕を表示し，話し終わってからもしばらく字幕が残っていることを好んでいる。つまり，字幕表示は必ずしも実際の音声区間で表示する必要はなく，in 点と out 点では異なる許容範囲内で表示しても視聴者に違和感なく字幕を提供できると考えられる。

5.2 言語の比較

実験では日本語・英語音声の 2 種類の動画を使用した。日本語を母語とする参加者に対し，許容される字幕表示タイミングの範囲は日本語・英語音声ともにほぼ同じ傾向が見られた。しかし細かく考察してみると，in 点においては話者が話し始めた後に字幕が表示される場合，日本語音声よりも英語音声の方がその許容範囲が広いことがわかる。例えば，話者が話し始めた後 800 ms 後に字幕が表示された場合，日本語音声の場合は 16.7 %の参加者しか許容しないのに対し，英語音声の場合は 50.8 %の参加者が許容する結果になった。これは，参加者が日本語母語話者であったため，日本語音声の話し始めには敏感であったことが理由として考えられる。また，音声が字幕表示と同じであるため，より話し始めに注目しやすかったのではないかと考えられる。

out 点に関しては音声の言語による顕著な差は表れなかったが，in 点と同様に話者が話し終わった後も字幕表示が残っている場合，日本語音声よりも英語音声の方がその許容範囲が広くなる傾向が見られた。参加者が日本語母語話者であったために，日本語音声の話し終わりにも敏感であることがわかる。

この結果から，動画コンテンツの音声の言語が違っても，許容される字幕表示タイミングの傾向は同じであるが，母語話者が動画を視聴する可能性が高い場合や，話者が話す内容と同じものを字幕で表示する場合はより正確に音声端点を検出し，視聴者が許容できるタイミングで字幕を表示する必要があるということがわかった。

Table 2 Results of timing for caption

音声		95%以上の参加者が許容できる字幕表示タイミング
日本語	in 点	-470 ms から 70 ms まで
	out 点	120 ms から 660 ms まで
英語	in 点	-330 ms から 140 ms まで
	out 点	180 ms から 690 ms まで

6　おわりに

本研究では，動画コンテンツに付与する字幕表示タイミングに関する主観評価実験を行い，音声端点と視聴者に許容される表示タイミングとの関係，対象とする音声の言語による違いを調査した。今後は，視聴者に許容される字幕表示タイミングを考慮した字幕付与システムの構築を目指したい。

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] 藤樫他，音講論（秋），33-34，2005.
[2] 向他，音講論（秋），263-264，2006.
[3] 古賀他，音講論（秋），261-262，2006.
[4] Koga *et al.*, J. Acoust. Soc. Am., 120 (5), 3215, 2006.
[5] 株式会社フジヤマの Web Page http://www.fujiyama1.com