K-020

リアルタイム字幕放送における字幕の提示タイミングずれに対する補正方法の主観評価
Cognitive Experiments on Compensation Method of Timing Differences for Real-Time Broadcast Captioning

小川 修太† 大盛 善啓†
Shuta Ogawa Yoshihiro Ohmori

1. はじめに

ワンセグ放送の普及に伴い電車やバスなど公共の乗り物でテレビ放送が視聴されるようになった．電車やバスではヘッドフォンや字幕を利用する必要があるが，ヘッドフォンは装着が煩わしいため字幕を利用する機会が増えている．リアルタイムに字幕を付与しながら放送されるニュースやスポーツなどのライブ中継番組では（リアルタイム字幕放送と呼ぶ）字幕が遅れて表示されるので放送内容の理解を妨げられる．そこで，音声と字幕を同期する手法が研究されている．

リアルタイム字幕放送の音声と字幕のずれは，送信側でも受信側でも補正できる．丸山らは音声認識と動的計画法を用いて送信側で自動的にずれを補正する手法を提案し，1 秒以内のずれを正解とする場合に 99%の精度を達成した[1]．送信側での補正には台本の利用やアナウンサーのマイク音のみを使って精度を上げられる利点があるが，放送局で大規模な設備投資が必要で早期の実現は難しい．これに対して受信側での補正には受信端末の改良だけで現状の放送設備のまますぐに利用できる利点がある．そこで本稿では受信側での補正を考える．

丸山らは音声と同期がとれた字幕を固定時間ずらして視聴する場合に±2 秒程度のずれが許容される事を主観評価実験で明らかにした[2]．金澤らも音声よりも字幕が 2 秒遅い表示から 4 秒早い表示が許容される事を主観評価実験で示した[3]．しかしずれが変動する実際のテレビ放送番組でもこの結果が有効かどうか明らかではない．さらに金澤らは受像装置でずれを手動補正する方法を示したが，この方法については客観評価も主観評価も行われていない[3].

このように，現状ではテレビ放送番組の字幕を受信側で手動や自動で補正する場合を主観評価実験した研究は見当たらない．そこで本稿では，字幕のずれを手動補正する方法と音声認識を用いて自動補正する方法を主観評価する．評価尺度として番組内容を理解したという感覚（以下，理解感）を用いる．12 本の評価用映像を 20 人の被験者が評価した結果を集計して，手動補正は精度が低いにも関わらず中程度の理解感を得られ，自動補正は精度も理解感も実用的な高い水準である事を示す.

2. 補正方法

手動補正と自動補正の処理の概要を図 1 に示す．4 本の右向き矢印は時系列を表していて，左側が過去に放送した内容である．i 番目の字幕に対するリアルタイム字幕放送の発話時刻を $T_s(i)$，元の字幕の表示時刻を $T_c(i)$，$T_s(i)$と $T_c(i)$の間のずれを $d(i)=T_c(i)-T_s(i)$とすると，$d(i)$はほとんどの場合正の値で字幕毎に異なる．

†（株）東芝 研究開発センター, Toshiba Corporation Corporate R&D Center

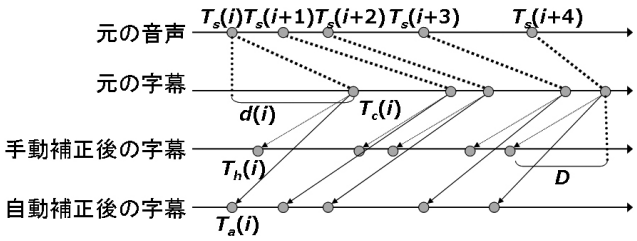

図 1 補正のイメージ

表 1 実験条件

視聴条件	：音声無，照明有
画面サイズ	：12～17 インチ，カラーモニタ
被験者	：健聴者 20 名
評価データ数	：12
視聴時間	：2 分×12 評価データ = 約 24 分
字幕の表示形式	：全入れ替え

本稿で用いる手動補正では，全ての字幕で $T_c(i)$に一律に固定値 D を加える．D を正にすれば字幕は補正前よりも一律に遅く表示され，負にすれば一律に早く表示される．$D=-d(i)$の時に $T_c(i)$と $T_s(i)$は一致する．図 1 は D を負にして，全字幕の表示時刻が一律に過去の方向へシフトした状態を表している．

本稿で用いる自動補正では，音声認識によって音声から取り出した単語列と形態素解析によって字幕から取り出した単語列を動的計画法によって対応付ける．字幕単位でずれを補正するため，図 1 に示すように字幕毎に補正値 $T_c(i)-T_a(i)$が異なる．1 秒以内のずれを正解とする場合に，生放送のニュース番組に対するこの自動補正手法の精度は 70%程度である．

3. 実験

3.1 実験方法

字幕のずれの補正方法と理解感の関係を調べる主観評価実験を行った．補正方法は「手動のみで補正する」，「手動と自動を併用して補正する」の 2 種類とした．

表 1 に実験条件を示す．ニュース，トーク，スポーツから抽出した各 2 分の評価用映像 12 本を 20 名の被験者に視聴してもらい，理解感を 5 段階の尺度で答えてもらった．12 本の評価用映像全てについて字幕の表示タイミングが無補正の映像と自動補正したものを作成し，そのうちの 6 本ずつを視聴してもらった．視聴中ずれに気づいたら手動補正をするよう実験前に被験者に指示した．評価尺度は 5（全て理解できた），4（ほとんど理解できた），3（半分理解できた），2（あまり理解できなかった），1（まったく理解できなかった）とした．

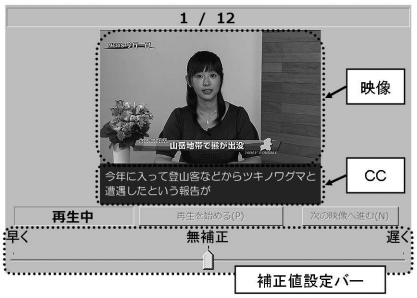

図 2 評価ツールの外観

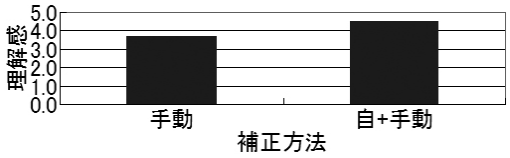

図 3 補正方法と理解感

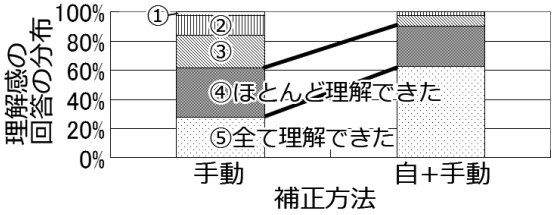

図 4 補正方法と理解感の回答の分布

評価のために開発したツールの外観を図 2 に示す．この評価ツールは字幕と映像の再生中に補正値を自由に調整できる．映像再生領域の下の矩形領域に最大 40 文字（1 行 20 文字で 2 行）の字幕を表示する．表示する文字は全て同時に切り替える．

自動補正済みかどうかを知る事で評価前のデータに被験者が先入観を持たないよう，全ての評価データを視聴する際に手動補正を可とした．そのため，本稿の自動補正の評価は自動補正した結果を手動補正したものであり，自動補正のみの場合とは異なる事に注意が必要である．無補正の字幕を手動で補正しながら視聴する視聴方法を「手動モード」，自動補正済みの字幕をさらに手動補正しながら視聴する視聴方法を「自+手動モード」と呼ぶ．映像 1 本をいずれかの方法で視聴する事を 1 試行と呼ぶ．

3.2 実験結果と考察

手動モードと自+手動モードについてアンケートで得られた理解感の平均を図 3 に示す．手動モードの理解感は平均 3.7，自+手動モードの理解感は平均 4.5 であり，有意水準 5%で手動モードよりも自+手動モードの方が高かった．理解感の回答の分布を図 4 に示す．手動モードでは評価尺度 4 と 5 の比率合計が約 60%であり，被験者の半数以上で受け入れられる事がわかったが実用的な水準ではない．自+手動モードでは理解感の評価尺度 4 と 5 の比率合計が約 90%だった．今回，自動補正のみで視聴した場合の理解感は調べていない．しかし自動補正のみで視聴した場合，自+手動モードよりも精度が高いことがわかっている．そのため，自動補正のみで視聴した場合も自+手動モードと同様に理解感が高いと考えられる．

精度と理解感の関係についても調べた．予め人手で $T_s(i)$を作成しておいて$|T_h(i)-T_s(i)|\leqq 2$ または$|T_a(i)-T_s(i)|\leqq 2$ の

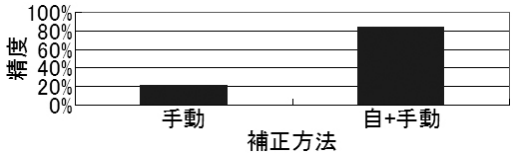

図 5 補正方法と精度

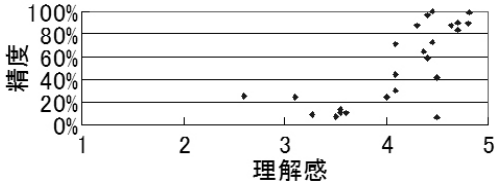

図 6 精度と理解感の分布

場合に正解とした．字幕の総数に対する正解の数の比率を精度とした．手動モードと自+手動モードについて精度の平均を図 5 に示す．手動補正の精度は約 20%と低いにも関わらず理解感の評価尺度 4 と 5 の比率合計は約 60%と高い事がわかった．精度が低くても理解感を得やすいと言える．全ての試行についての精度と理解感の分布を図 6 に示す．精度が 20%程度と低くても中程度の理解感が得られる事と，精度が 80%程度で高い理解感が得られる事がわかった．

提案手法で用いた自動補正手法を PC（CPU:2.66GHz, Memory:1GB）で実行すると，30 秒の映像を映像長の半分以下の 12 秒程度で処理できる．したがって，PC 上にこの自動補正手法を実装すれば地上デジタル放送番組をリアルタイムに処理できると考えられる．一方，モバイル機器の計算資源は PC よりも格段に限られるため，ワンセグ放送番組をリアルタイムに自動補正する事は困難と考えられる．そこでリアルタイムに補正するには手動補正を使い，録画した番組の補正には今回用いた自動補正手法を使う事が考えられる．

4. おわりに

音声と字幕のずれが変動するテレビ放送番組の字幕を手動や自動で補正する場合の理解感を主観評価した．20 人の被験者による 12 本の映像の主観的な理解感の評価を集計した結果，手動補正は精度が 20%と低いにも関わらず半分以上理解できた割合が 60%の中程度の理解感を得られる事がわかった．音声認識による自動補正は精度が 80%程度で理解感も半分以上理解できた割合が 90%以上で実用的な高い水準である事がわかった．本稿で評価した自動補正手法はモバイル機器でリアルタイムに動作しないため，モバイル機器でも動作する高速な自動補正手法の開発が今後の課題である．

参考文献

[1]丸山 一郎, 阿部 芳春, 江原 暉将, 白井 克彦, “ワードスポッティングと動的計画法を用いたテレビ番組に対する字幕提示タイミング検出法”, 信学論 D-II, Vol. J85-D-II, No.2, pp.184-192, Feb. 2002.

[2]丸山 一郎, 阿部 芳春, 沢村 英治, 三橋 哲雄, 江原 暉将, 白井 克彦, “ニュース字幕の提示タイミングずれに対する許容特性”, 信学技報 HCS, Vol.99, No.123, pp.21-28, Jun. 1999.

[3]金澤 章, 磯野 春雄,“ニュース字幕の提示タイミングずれの主観評価と補正方法”, 映像情報メディア学会年次大会講演予稿集, pp.89-90, Aug. 2001.