字幕付きテレビニュース放送を題材とした CALL システムにおける副音声の利用

田中 敬志† 小林 聡†† 中川 聖一†††

†豊橋技術科学大学大学院 情報工学専攻

††豊橋技術科学大学 情報処理センター

†††豊橋技術科学大学 情報工学系

E-mail: {ttakashi, koba, nakagawa}@slp.ics.tut.ac.jp

1. まえがき

語学学習には繰返し学習が必要であるが[1]，既存の教材では話題が限定されている．そのため，学習者が興味を持ち続けられるか否かという問題がある．それに対し，より広い話題や時期に即した話題に触れられるならば，学習者が外国語学習に興味を持ち続けることが可能になると考えられる．しかし，適当な話題のニュース放送などを用いて，学習者個人や語学教師が語学学習教材を作成するのには限界があり，また，幅広い話題の内容を教材として利用するためには手間がかかるので，教材を半自動的に手軽に作成できるツールが必要となる．

これまでの研究において，字幕付きのテレビニュース放送をもとに，自然な発話による豊富な話題の提供，時期に即した話題の提供，教材の繰返し視聴，単語の語義表示，聞き取りテストの作成などの機能を持つ，市販のソフトウェア[2]の機能を強化した語学学習教材を，半自動的に作成するシステムを構築した[3]．本システムは日本語学習用教材，英語学習用教材ともに作成することができる．本システムで作成される語学学習教材では，ニュースの副音声(翻訳音声)は学習者にそのまま提示されるのみであり，このままの機能でも有用ではあるが[4]，学習時の副音声の利用方法は限られてしまっている．そこで本研究においては，音声言語処理技術を用いて副音声と字幕との対応付けを行い，副音声を用いた学習をより効果的なものにすることを検討している．本稿では，この字幕と副音声の対応付けの方法，その評価結果について述べる．

2. 本システムの概要[3]

現在，NHK の「ニュース 7」では音声認識技術を用いてアナウンサーの発話内容の字幕がリアルタイム(実際は数秒遅れる)で字幕放送として送られてくる[5]．また，他の番組においても字幕放送や Web を通して字幕を取得できる番組が増えてきている．さらに字幕を用いた教材として，映画の字幕を用いた研究もされている[6]．本研究で構築しているシステムは，この TV 番組の映像と音声，字幕から，語学のリスニング教材を作成するものである．

2.1. 教材化手順

先に述べたように，本システムの教材は NHK による字幕付きニュース放送を基に，自動的な教材化を目指している．日本語版教材作成システムの構成と処理の流れを以下に示す．英語版システムも本質的には同じである[1]．

(1) 2 枚の TV チューナーボードを使用し，MPEG1 形式のビデオと字幕のテキストを PC に取り込む．英語版の場合は Web[7]から取得する[2]．

(2) (1)で得た字幕放送のログを，無用なスペースを除くなど整形すると共に，ログの文字化けの修正と確認を行なう．

(3) 字幕の形態素解析を茶筅[8]を用いて行なう．また，形態素解析の結果の確認・修正を行なう．英語版では Brill’s Tagger[9]を使用して品詞解析を，読み付与は CMUDictionary[10]を使用して行なう．

(4) 手順(3)の形態素解析の結果として得られる読みを利用し，音節(音素)系列を作成する．

(5) 録画から得られる音声を分析し特徴パラメータ系列に変換する．

(6) 音声認識技術[11]を利用して，手順(4)で得た字幕の音節(音素)系列と音声の特徴パラメータ時系列との照合により，字幕と音声との同期情報を作成する．

(7) 手順(6)の結果から単語単位での音声と字幕の同期情報を得る．

(8) 手順(3)で得た品詞，発音情報，および手順(6)で得た同期情報を，XML のデータ形式を用いて字幕テキスト中に埋め込む．

(9) 最終的に，MPEG1 形式の映像+音声ファイルを QuickTime 形式に変換し，それと手順(8)で得られる様々な情報を含む字幕により，Java で開発した専用のプレイヤを用いて再生する語学学習教材となる．この語学学習教材の画面は図 1 のようになる．

上記手順(1)～(9)中，手順(2)を除き，個々の手順においては自動化がなされている．手順(2)では，字幕取得プログラムの制約により字幕が正確に収録されていない，字幕中に発話されていない部分が存在する，などの問題

[1] 日本語版と英語版の両者の主な違いは，字幕の取得方法，同期用の形態素解析(英語版では品詞 Tagger と発音辞書を使用)，音響モデル(日本語版では音節モデルを，英語版では音素モデルを使用)，辞書(和英/英和)などである．

[2] PBS ニュースの場合，字幕が画像として音声と同期して表示されている．そのため，パソコンにテキストとして読み込むためには OCR 等の処理が必要である．現在は，Web に公開されているニュースクリプト（字幕と同等）を利用している．

点があるため，確認および修正を人手で行なっている．その後，教材作成ツールに音声データと字幕を入力すると，教材用のデータが作成されるようになっている．また，その他に，音声データを MPEG1 形式のビデオから分離，ビデオのファイル形式を MPEG1 形式から QuickTime ビデオ形式に変換，などの作業については現時点では手作業で行なわなければならない．本システムを用いて教材を作成する作業にかかる時間は，コンピュータの習熟度合いにもよるが，数分のニュース放送のデータであれば，ファイル形式変換や計算時間を含めて全体で 15～30 分程度，このうち教材作成者の作業時間は 10～20 分程度である．

2.2. 字幕と主音声の同期

現在，NHK による字幕放送は，音声から数秒から十数秒の遅れを伴なっている．PBS の場合，発話と字幕はほぼ同時に提示されているが，字幕は画像情報であり，Web 上のニューススクリプトには発話との同期情報までは含まれていない．また，いずれの場合も，単語単位での同期情報は放送からは得られない．本システムでは，単に放送されたタイミングでの字幕の提示だけではなく，単語単位での字幕表示を目指しているので，音声と字幕の同期を新たに取る必要がある．

字幕のかな文字列(あるいは音素記号列)からそれに対応する音節(音素)単位 HMM 系列を構成し，HMM 系列と音声の特徴ベクトルを Viterbi アルゴリズムを用いて最適な時系列同士のマッチングを行ない，音声と字幕の同期情報を得ている．ただし，この際，無音区間をパワーの閾値により除く処理を行なっている．特徴ベクトルの分析条件はサンプリングが 12kHz，フレーム周期が 8 ミリ秒，LPC14 次による分析であり，日本語版は LPC メルケプストラム 10 次元を 4 フレームまとめて KL で 20 次元に圧縮したものとΔケプストラムとΔΔケプストラム，Δパワー，ΔΔパワーの計 42 次元，英語版は LPC メルケプストラム 10 次元とそのΔの 20 次元を使用している．

この方法により得られる同期の精度を正確に測定するため，各名詞の終端点での同期のずれを集計したところ，日本語の音声データで平均 14 ミリ秒(サンプル数 39)，英語の音声データで平均 92 ミリ秒(サンプル数 25)となった．英語の音声に対しては，背景音などの比較的大きなノイズが含まれるため(英語音声の S/N は 16dB～20dB，日本語音声の S/N は 30dB～33dB[1])，また音響モデルの精度が不十分(音響モデル作成のためのトレーニングデータ量が少ない)であることが，日本語音声の精度よりも悪くなる原因となっている．

2.3. 学習教材の各機能

本システムで作成される学習教材の動作例を図 1 に示す．

図 1 では，英語ニュース番組の映像と字幕，辞書検索が示されている．

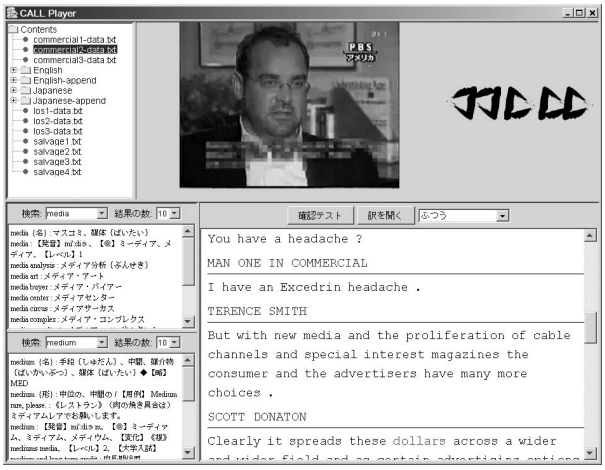

図 1 学習教材の動作例 ―英語版―
（日本語版でも同様に作成可）

2．3．1． ビデオの再生

字幕中の任意の単語をクリックすることで，その単語の時間までビデオの再生位置を戻すことができ，それにより聞きたい箇所を繰り返し視聴することが容易となっている．

2．3．2． 字幕

字幕の提示は，漢字かな混じりでの提示と，ルビ(日本語)や発音記号(英語)を付加した提示，特定の品詞のみの提示，および提示なしなどの提示方法が選択可能である．また，前節で求めた字幕と主音声の同期情報を用いて，ビデオの再生とともに，字幕中の単語がカラオケ様にハイライトしていく．

2．3．3． 辞書引き

通常の電子辞書と同じように任意の語句をキーボードからタイプする以外にも，字幕中の単語にマウスカーソルをあわせることで，その単語(と単語の原型)が自動的に辞書引きされ表示される．

2．3．4． 翻訳音声(副音声)の再生

副音声をそのまま再生することで，日本語版では英語の，英語版では日本語の翻訳音声をコンテンツ毎に聞ける．

2．3．5． ディクテーション問題

字幕中の単語を入力フィールドとすることにより，自動的にディクテーション問題を作成し，採点することができる．

2．3．6． 選択問題（内容把握）

各コンテンツの理解度を確認するためのテストとして，学習中任意に内容把握を行うための選択問題を実行できる．これは他の機能とは異なり，自動的に TV 番組と字幕から作成することが現時点で困難なため，現時点では語学教師等，指導者や教材開発者が問題を作成することになる．

[1] 音声自動認識では，S/N が 20dB より悪くなると急激に認識率が低下する．研究段階でよく用いられているクリーンな音声の S/N は 40～50dB 程度である．

3. 副音声の利用

本システムにより作成される学習教材は，TV ニュース放送より得られる情報を最大限に引き出して，学習者に提示するものである．TV ニュース放送より得られる情報としては，映像，主音声，副音声，字幕を挙げることができる．日本語ニュースの場合，主音声が日本語で，副音声がその対訳の英語音声，英語ニュースの場合，主音声が英語で，副音声がその対訳の日本語音声が吹き込まれており，現在の本システムではこの主音声と字幕の情報を主に利用して学習教材を構築している．副音声については，現状の学習教材では字幕とは独立に学習者に提示しているにすぎず，まだ改善の余地があった．

そこで本節では，副音声と字幕との対応付けについて取り上げ，学習教材上でより高度に副音声を利用するための手段を検討する．

3.1. TV ニュース放送の副音声

現在の TV ニュース放送では，2 カ国語放送を行っているものがある．例えば，本教材で利用している NHK の「ニュース 7」，「ＢＳ２３」では，主音声が日本語で，その対訳の英語音声が副音声で放送され，また，BS 等で放送されている PBS，CNN，ABC 等では，主音声が英語で，その対訳の日本語音声が副音声で放送されている．

これら TV ニュース放送を数本ずつ録画し，日本語ニュース，英語ニュースそれぞれ 10 分程度，主音声と副音声の関係を調べた結果，以下のようなことがわかった．

- 主音声よりも早く，それに対応する副音声が発声されることがある
- 主音声と副音声の文毎のずれは，平均で 3 秒程度である
- 主音声と副音声は文単位で対応しており，英語 1 文に対して対訳日本語 2 文，また，英語 2 文に対して対訳日本語 1 文など，最小の対訳が複文対複文であることも多い(1 対 1 の対応は全体の 6 割程度)
- 主音声のうち，まれに副音声の対訳が付かないものがある

3.2. 字幕と副音声の対応付け

字幕と副音声の文単位での対応付けは図２のように行った．以下に，その手順を示す．以下では，英語ニュース(主音声：英語，副音声：日本語，字幕：英語)を考えている．

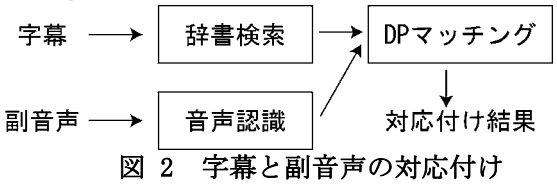

図 2 字幕と副音声の対応付け

(1) 字幕中の各単語を英和辞書で検索し，その単語を日本語の語彙(辞書の内容をそのまま利用)で置き換える． その際，前置詞は除外し，単語の原型でも検索を行っている．

(2) 副音声を音声認識し，単語列に変換する．

PBS の副音声中には主音声が重畳されているため，副音声から主音声の成分をスペクトルサブトラクションを用いて除去した音声で音声認識を行った．

t を時間とし，主音声のデータ系列を Xmain(t)，副音声のデータ系列を Xsub(t)，副音声中の日本語のクリーンな音声を Xclean(t)としたとき，PBS の副音声のデータ系列は以下のように表すことができる．

$$Xsub(t) = Xclean(t) + m(t) \times Xmain(t)$$

m(t)は主音声の重みで，日本語の発話がない時には m(t)は 1 程度であり，日本語の発話中は m(t)が 0.1 程度に変化する．

そのため，各フレーム毎に m(t)を計算(Xclean(t)の各パワースペクトル成分が 0 以上となる m(t)の値を利用)して，スペクトルサブトラクションを行い，Xclean(t)を求めている．

(3) (1)と(2)の DP マッチングを文単位で行い，対応付け結果を求める．DP パスは図３のものを用いる．

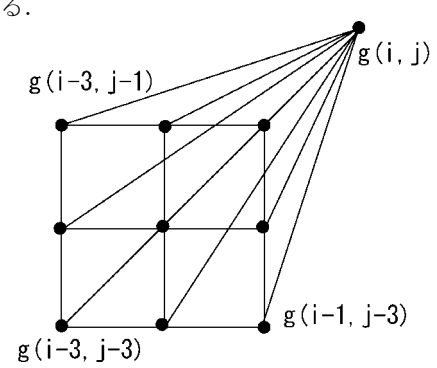

図 3 DP パス（i 軸：副音声，j 軸：字幕）

g(i, j)は以下の式を用いた．

$$g(i,j) = \min_{0 \le x,y \le N} \left\{ \begin{array}{l} g(i-x, j-y) \\ +(x+y)^{\alpha} \times d(i-x,\cdots,i, j-y,\cdots,j) \end{array} \right\}$$

N：DP パスのサイズ．図では N＝３

α ：文の長さに対する重み． $\alpha \geqq 1$

また，距離 d(i-x,…,i,j-y,…,j)は副音声の i-x から i 番目までの文と字幕の j-y から j 番目までの文の距離で，次のように定めた．

① 副音声を音声認識した結果の i-x 番目の文から，i 番目の文までを取り出し，単語毎にまとめ，i-x 番目から i 番目の文までの単語のリストWを作る．

② 字幕の j-y 番目の文から j 番目の文までの中で，①で作成した単語リストWが含まれているかどうかを調べ，含まれていた単語のリストWinc を作る．

③ 距離 d を以下の式で求める．

$$d(i-x,\cdots,i, j-y,\cdots,j) = \exp\left(-\frac{|Winc|}{|W|}\right)$$

|W|：Wのサイズ，|Winc|：Winc のサイズ，

3.3. 字幕と副音声の対応付け結果

実験条件として表1のものを用いた.

表 1 実験条件

形態素解析	ChaSen ver. 2.1
辞書	英辞郎 ver. 0.50
音声認識	
デコーダ	Julius
音響モデル	話者依存, 対角音節モデル, ASJ-JNAS, ASJ-DBを使用
言語モデル	bi-gram(1pass), tri-gram(2pass) 毎日新聞データベースを使用 2万語彙
特徴ベクトル	MFCC+ΔMFCC+ΔPOW(計25次元)

テストセットとして, PBS の「News Hour」の 4 分弱の4つのニュースを収録し, 主音声と副音声の対訳をそれぞれ 25 対ずつ抜き出し, それをテストセットとした. それぞれのニュースのトピックと, 英文と和文の文数を表2に示す.

上記テストセットの副音声(日本語)の認識結果を表 3 に示す.

表 2 テストセットの各トピック毎の文数 (括弧内は対応する和文がない英文の数)

トピック	英文	和文
アメリカの軍事	25	36
航空問題	27(2)	28
薬事問題	27(1)	33
西ナイルウイルス	30	28

表 3 副音声の単語認識率[%] (SS:スペクトルサブトラクション)

SS	正解率	正解精度
なし	63.5	45.9
あり	72.8	67.6

SS なしでは, 前節(2)で示したように副音声中に主音声の成分が含まれているために, 挿入エラーが増え, 正解精度が顕著に低くなる結果となっている.

表 4 に字幕と副音声の対応付け結果を示す. 音声認識については表 3 と同じもので, 手動書き起しは, 副音声を人手で書き起したものである. また, ずれの平均は, 対応付けの結果が表 2 のテストセットの対応付けからどの程度ずれているかを, 文毎にずれた文数を計算し, 平均をとったものである. DP 計算時のパスのサイズ N, 文の長さに対する重みαについては, それぞれ, 2〜5, 1.0〜2.0 の中から, 最も対応付けの結果が良かったものを選択した.

表 4 対応付け結果

単語列への変換方法	単語正解率	ずれの平均
音声認識(SSなし)	63.5%	0.41文
音声認識(SSあり)	72.8%	0.23文
手動書き起し	100%	0.14文

手動で書き起したものを用いて対応付けを行った結果, ずれの平均は 0.14 文となっており, 対応付けがうまく行われていることがわかる. 音声認識を用いた結果では, SS なしで 0.41 文, SS ありで 0.23 文となっており, 音声認識率が大きく影響していることがわかる.

具体的には, 平均のずれが 0.1 文程度であれば, かなり正確に対応付けがとれており, 学習での使用も問題ないと考えられる. 今回最も良かった結果は 0.2 文程度のずれであり, このままでも学習に利用することは可能であるが, 認識率の向上, 未知語への対応(固有名詞が多いため)など, さらなる改良が必要であろう.

また, 図 4 には N とαを変化させたときの音声認識(SS あり)の結果を示している. 図に示すように, αが 1.4 前後では N に因らず安定した結果となっている. これは, 音声認識(SS なし), 手動書き起しのそれぞれの方法についても同様な結果となった.

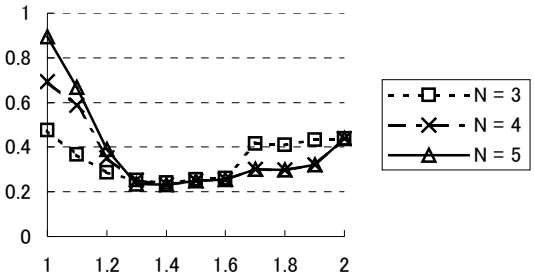

図 4 対応付け結果と DP 時のパラメータの関係 音声認識(SS あり) 横軸:α, 縦軸:ずれの平均

4. まとめ

本稿では, TV 放送とその字幕から語学学習教材を作成するシステムについて, また, 副音声と字幕との対応付けを行う方法, その評価結果について述べた. それより, 改良の余地は残るものの, 副音声と字幕との対応付けを行う本手法が有効であることを示すことができた.

参考文献

[1] Darwing, Tracey M.: "Information type and its relation to non-native speaker comprehension", Language Learning, 39, pp.157-172(1989).

[2] 豊橋技術科学大学中川研究室: "英会話 CAI ソフトアンケート結果",http://www.slp.ics.tut.ac.jp/CALLsoft/

[3] 小林聡, 田中敬志, 森一将, 中川聖一: "字幕付きテレビニュース放送を素材とした語学学習教材作成システム", 人工知能学会論文誌, vol.17, no.4, pp.500-509

[4] 田中敬志, 小林聡, 中川聖一:"テレビニュース放送を利用した語学学習システムの評価", 電子情報通信学会 技術研究報告 音声研究会 (2003 Jan.)

[5] 安藤彰男 , 今井亨, 小林彰夫, 本間真一, 後藤淳, 清山信正, 三島剛, 小早川健, 佐藤庄衛, 尾上和穂, 世木寛之, 今井篤, 松井淳, 中村章, 田中英輝, 都木徹, 宮坂栄一, 磯野春雄: "音声認識を利用した放送用ニュース字幕製作システム", 電子情報通信学会論文誌, Vol.J84- D-II, No.6,pp.877-887 (2001).

[6] 五十里慎吾, 佐野輝希, 緒方淳, 有木康雄:"ユーザー発話のセグメンテーションと発話評価機能をもつ英語学習支援システム", 情報処理学会 研究報告 SLP 2001 No.040, pp.7-12

[7] The Online News Hour ,http://www.pbs.org/newshour/

[8] 形態素解析システム: "茶筌", http://cl.aist-nara.ac.jp/lab/nlt/chasen/

[9] 品詞解析プログラム: Brill's Tagger,: http://www.cs.jhu.edu/~brill/

[10] The CMU Dictionary ,http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[11] 甲斐充彦, 中川聖一: "冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価" , 信学論, Vol.J80-D-II, No.10, pp.2615--2625 (1997)