視覚障害者への動画の音声解説提示インタフェースの試作

An Experimental Interface to Present Audio Description of Video for the Blind

佐藤 大介　　高木 啓伸　　浅川 智恵子*

Summary. The number of video files distributed via the Internet is increasing in recent years, because of the wide adoption of broadband. At the same time, the number of video files that are not accessible to the blind is also increasing. The W3C WCAG 2.0 working draft seeks to make video data understandable by providing audio descriptions and extended audio descriptions. However, it is too expensive to create alternatives for video data and also calls for expert knowledge and skills. In addition, it has not been clarified whether or the extended audio descriptions help blind users. Therefore, we developed an interface which provides interactive audio descriptions generated from text descriptions by using a Text To Speech (TTS) engine. The users are able to get audio description easily by interacting with the interface. We also conducted a pilot study and a listening test with two blind users, and discuss the impact of audio descriptions.

1　はじめに

ブロードバンド環境の普及により動画など大容量のデータがインターネットで配信され，より視覚的でインタラクティブなマルチメディアコンテンツが増加する一方，そのアクセシビリティが大きな問題となっている．視覚障害者は視覚的な情報を十分に得ることができないため，代替の手段によって視覚的な情報が提供される必要がある．このため W3C Web Accessibility Initiative (WAI) [6] が策定中の Web Content Accessibility Guidelines (WCAG) 2.0[7] では，動画の代替のアクセス手段として，テキストによる書き下しやオーディオディスクリプション[1](本論文では音声解説と呼ぶ) を提供するように求めている．

音声解説は動画中で視覚的にのみ表現される情報を音声で説明するものである．動画中の台詞など意味を持つ音声以外の部分 (無声区間) に挿入され，人や物の見た目，動作，画面に表示される文字，場所や状況，シーンの切り替わりなどを説明する [4].

しかし，音声解説を作成するためには専門的な知識・技術や経験が必要であり多くのコストがかかるため，音声解説を提供している動画はほとんど無いのが現状である．また，音声解説によってコンテンツの内容がきちんと理解されるのかどうかや，動画中の無声区間に収まりきらない音声解説[2] (本論文では拡張音声解説と呼ぶ) を視覚障害者にどのように提示するべきかといったことはきちんと議論されていない．

本論文では，音声解説および拡張音声解説を提示するためのインタフェースについて述べ，どのように視覚障害者へ提示を行うべきか分析する．試作したシステムではジョグダイアルによる話速制御等の基本的な音声コントロールを行うことができ，テキストで記述した動画の音声解説を Text To Speech (TTS) エンジンにより音声に変換し音声解説としてユーザに提示することができる．また，視覚障害者を対象に予備実験とヒアリングを実施した結果を踏まえ，拡張音声解説や試作システムのインタフェースについて議論をする．

2　音声解説

音声解説を伴う放送は解説放送と呼ばれ，NHK の連続テレビ小説や大河ドラマが良く知られている．総務省の報告書 [8] によれば，平成 17 年度の解説放送の年間放送時間は，NHK 総合が 306 時間 06 分で全体の 3.5%，NHK 教育が 701 時間 40 分で全体の 8.1%，民放キー 5 局[3]が 76 時間 20 分で全体の 0.2%，在阪準キー 4 局[4] が 184 時間 32 分で全体の 0.5%である．平成 12 年度の調査と比べると放送時間は増加しているものの，テレビ放送においても音声解説の普及が進んでおらず，インターネットで配信される動画においてはさらに深刻な状況である．

また，映画の音声解説も国内ではほとんど提供されていない状況にあり，ボランティアや NPO 法人

* Daisuke Sato, Hironobu Takagi and Chieko Asakawa, 日本アイ・ビー・エム株式会社　東京基礎研究所

[1] WCAG2.0 の原文では “audio description”．副音声や音声ガイドと訳されることもある．

[2] WCAG2.0 の原文では “extended audio description”.

[3] 日本テレビ放送網 (株)，(株) 東京放送，(株) フジテレビジョン，(株) テレビ朝日，(株) テレビ東京

[4] 讀賣テレビ放送 (株)，(株) 毎日放送，関西テレビ放送 (株)，朝日放送 (株)

による映画の音声解説の作成，および上映会が各地で行われているが，著作権上の問題から小規模での活動にとどまっている．欧米諸国においては映画製作の段階から音声解説を作成する作品も多く，映画館で音声解説を提供するためのヘッドフォンの設置や，映画の DVD に音声解説を収録するなどの活動が積極的に行われている．このような活動が広がり，インターネット上のコンテンツに対しても音声解説が普及することが強く望まれる．

2.1 拡張音声解説

音声解説を提供しないとコンテンツの意味が分からなくなってしまう部分に，音声解説を挿入するのに十分な長さの無声区間がない場合に，動画の再生を一時的に止めるなどして音声解説を提供する仕組みが拡張音声解説である．

教育用の動画コンテンツなどのように視覚的に多くの情報を伝えている場合，無声区間に音声解説を挿入する手法だけでは時間の制限から十分に視覚情報を補うことができない場合が多いことが知られている．そのため現状では，別のテキストファイル等の形で視覚情報を補うことが一般的である．

これに対して拡張音声解説ではこのようなコンテンツに対しても音声解説を提供できる可能性がある手法として近年注目されている．従来のテレビや DVD においては拡張音声解説を提供することは技術的に困難であった．しかし，動画のインターネット配信や PC でのテレビの視聴が一般化するにつれ，高度な動画制御とインタラクションを実現できる端末が動画の再生装置として利用されるようになってきたため，実現可能になってきた．WEB ページのアクセシビリティ標準である WCAG2.0 において拡張音声解説の手法が具体的に検討され，記述されていることはそのような動きを象徴的に示しているといえよう．

しかし，WCAG2.0 では拡張音声解説を “説明を追加する時間を取れるよう動画を一時停止して視覚的表現へ付加する音声解説” と定義[5] し，提示方法については “一時停止して” としか言及していないため，具体的な提示方法についてユーザインタフェースの観点からの議論が必要である．

2.2 メタデータの提供モデル

視覚障害者のための音声解説等，動画に対する代替のアクセス手段を提供する情報は動画に対するのメタデータとして捉えることができる．ここでは音声解説作成のモデルをメタデータの提供モデルとして検討する．

[5] WCAG2.0 の原文では “audio descriptions that are added to an audiovisual presentation by pausing the video so that there is time to add additional description”

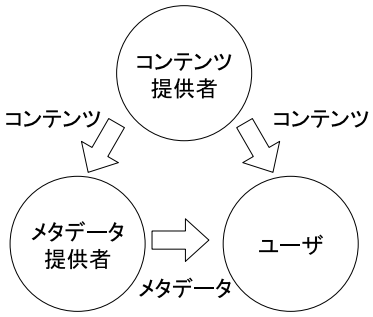

図 1. メタデータの提供モデル

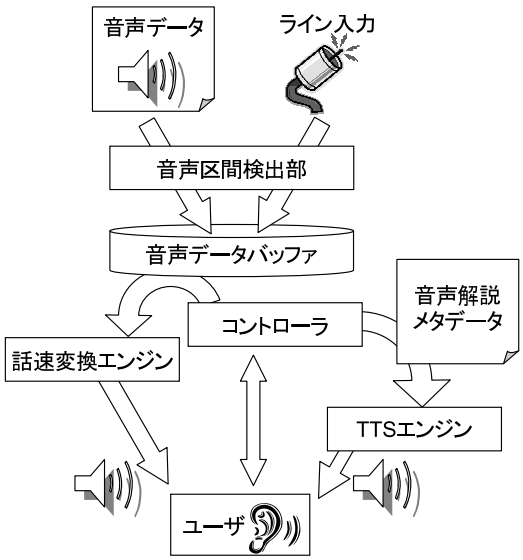

図 2. システム構成図

音声解説を作成するには，それぞれ専門的な知識・技術や経験が必要である．そのため多くの場合，動画コンテンツの作成者はそれらのメタデータを作成することができないため，メタデータを専門的に提供する機関が必要である (図 1)．メタデータ提供者は動画のコンテンツ提供者と同じであっても構わないが，一般のユーザはコンテンツ提供者からコンテンツの提供を受け，視覚障害を持つユーザはコンテンツとは別に音声解説をメタデータ提供者から得ることができるようにすることが必要である．

3 視覚障害者向け動画ブラウズシステム

視覚障害者は日常的に音声データからの情報取得を行っているため，音声を聞き取る能力が健常者より優れていることが知られている [10]．そのためピッチ固定で再生速度を速くしたり，無声区間を削除して情報密度を高くすることで音声データブラウズのユーザビリティを大きく向上可能である [1]．さらに動画の音声を再生する場合には音声解説によって視覚的な情報を補う必要があるため，動画の再生に同

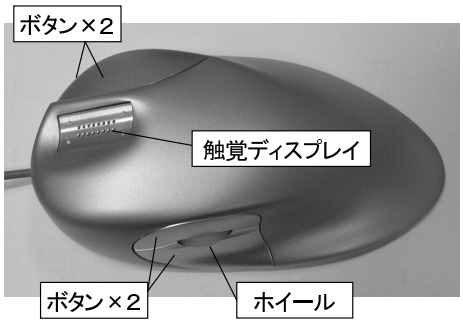

図 3. TAJODA

期して音声解説を提示しなければならない.

本システムは視覚障害者を対象としているため, 動画中の音声のみをコントロールする (システムの説明において, “動画を再生する” という表現は動画中の音声のみを再生することを意味する). 図 2 に実装したシステムの構成を示す. システムは動画の音声データを音声区間検出部で音声パワーによって音声区間と無声区間に別けてバッファリングする. 音声データはライン入力によって提供されても良い. この場合はすでにバッファリングした位置までを自由にコントロールすることができる. ユーザは音声の再生速度を調整したり, 音声区間ごとに再生位置をジャンプしたりすることができる. 話速変換アルゴリズムは PICOLA[9] をベースにして, 再生中の再生速度の変更にスムーズに追従するよう改良を加えた. 音声解説メタデータは同期情報とテキストを持ち, 同期情報に基づいて所定の時間にテキストを TTS エンジンで音声に変換して出力する.

3.1 再生のコントロール

再生のコントロールにはジョグダイアルと触覚ディスプレイを組み合わせた TAJODA[2, 3] (図 3) を用いた. TAJODA には 2 x 8 の微小な振動子マトリクスを持つ触覚ディスプレイ, ジョグダイアルとして動作するホイール, および 4 つのボタンがある. 触覚ディスプレイでは音声解説がある時間に合わせて, 指先側の部分を 1 回振動させて音声解説の存在を提示した. またホイールの回転によって, 音声解説がある時間ごとに再生位置をジャンプさせることができるようにした. 各ボタンには再生・一時停止・再生速度変更などのを機能を割り当てた.

3.2 音声解説提示機能

ここで拡張音声解説の提示方法について考える. WCAG2.0 では拡張音声解説を一時停止して付加する音声解説であると定義しているが, どのように一時停止するのかについては言及していない. そこで本システムによって実現可能な拡張音声解説の提示方法を以下の 3 つに分類した (図 4).

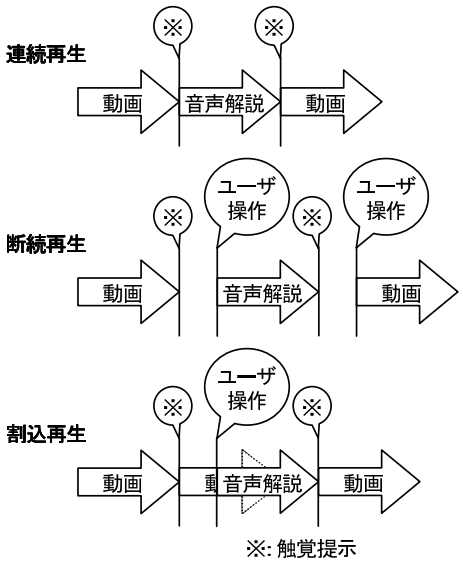

図 4. 音声解説の提示方法

連続再生 音声解説の位置で動画を一時停止し, 音声解説を再生したのち, コンテンツの再生を再開する. 触覚提示によって動画と音声解説の切り替わりを提示する.

断続再生 音声解説の位置で触覚提示し動画を一時停止する. ユーザの指示により音声解説を再生したのち, 解説の終わりを触覚提示する. 次のユーザの指示で動画の再生を再開する.

割込再生 音声解説が挿入されている時間位置でユーザに触覚提示をしつつ, 動画の再生を継続する. ユーザの操作で動画の再生を一時停止して音声解説を再生し, その後音声解説の挿入位置まで戻して動画の再生を再開する.

4 予備実験

実験では適切な拡張音声解説の提示方法について基礎的な情報を収集した.

4.1 実験手順

被験者にはインタフェースを簡単に説明したあと, 2 種類の動画を以下の順で提示し, 各提示方法の後にユーザの理解度を確認した. 1, 2 の方法における再生は通常の速度で行い, 3 の方法ではユーザが自由に再生速度を選べるものとした.

1. コンテンツをメタデータ無しで提示. ユーザは何も操作を行わない.
2. 連続再生によって動画と音声解説を提示. ユーザは何も操作を行わない.
3. 断続再生もしくは割込再生のどちらかで提示. ユーザが操作しながら自由に再生を行う.

表 1. 実験に用いた動画

	ヨガ[6]	料理[7]
動画の再生時間	3 分 26 秒	4 分 18 秒
無声区間の長さ	1 分 54 秒	1 分 52 秒
メタデータ文字数	386 文字	700 文字
音声解説の数	10ヶ所	24ヶ所
連続再生での再生時間	4 分 26 秒	6 分 05 秒

4.2 実験題材

提示した 2 種類の動画はヨガのポーズを説明するものと，料理の作り方を説明するものである．ヨガのポーズでは動画を提示したのち，実際にポーズをとってもらいその理解度を確認した．料理の作り方では，Web 上で動画と一緒に提供されていた材料の一覧を点字に印刷して動画を提示する前に一読してもらい，動画を提示した後に口頭で理解度を確認した．使用した動画はインターネット上で公開されているものであり，ヨガが 3 分 26 秒，料理が 4 分 18 秒の長さであった．(表 1)

4.3 音声解説メタデータ

音声解説のメタデータは，著者が動画を見て音声だけでは分かりにくいと判断した部分に対して記述した．音声解説では，すでに行われた動作の説明や，動画中の音声による説明の補足を行い，動画中の無声区間にメタデータの時間位置を指定するように心がけた．作成した音声解説のデータは全て拡張音声解説として扱った．メタデータの量は，ヨガの動画で 10ヶ所 386 文字，料理の動画で 24ヶ所 700 文字であり，TTS の標準速度で音声解説を読上げた場合の連続再生における再生時間は 4 分 26 秒，6 分 05 秒であった．(表 1).

メタデータはテキストエディタを用いて手作業で作成した．音声解説が必要だと思われる部分を何度も繰り返し再生し，プレーヤの再生時間を参照して時間位置を決定した．最初から通しで動画を見ながらメタデータを音声出力し，確認・修正しながら完成させた．実験で使用したメタデータは付録として表 2 に示す.

4.4 被験者

コンピュータ使用経験のある視覚障害者に依頼をした．被験者はスクリーンリーダを日常的に使用している視覚障害者 2 名で，共にコンピュータの使用経験は 10 年以上であり，音声合成をもちいた高速音声からの情報取得に十分に習熟したユーザである．また，2 名とも全盲であり 1 名は先天性 (被験者 B) である．2 名ともヨガの経験は無いが，普段から料理をする機会がある．

4.5 実験結果

4.5.1 ヨガのポーズを説明する動画

1 の提示方法で提示した場合，被験者 A は理解できなかったと答えた．被験者 B は理解できたと答えたが，被験者 B に実際にポーズをとってもらったところ間違った理解をしていた．2 の提示方法で提示した場合，両被験者とも 1 の提示方法よりも理解できたと答えたが，やはり間違った理解をしていた．3 の提示方法で何度かコンテンツを聞きなおすことによって最終的に両被験者とも正しいポーズを理解できた．3 の提示方法ではそれぞれ約 10 分間再生を行っていた．再生速度は被験者 B が途中で 0.8 倍速を用いた以外は全て通常の速度であった．

4.5.2 料理の作り方を説明する動画

1 の提示方法で提示した場合，両被験者とも雰囲気は分かるが作るのは難しいという内容の答えだった．次に 2 の提示方法で提示すると，両被験者ともどのように調理し盛り付けを行うかということを理解したと答えた．この題材では 2 の提示方法によって十分に理解できたため 3 の提示方法は実施しなかった．

4.5.3 ヒアリング

実験終了後，実験の感想やインタフェースに対するリクエスト，また普段動画を見ていて感じることなどを自由に答えてもらった中で有用と思われる意見を以下に示す.

- いくつかの提示モードは自由に切り替えられるようにして欲しい．たとえばヨガのポーズでは前半と後半は左右が入れ替わっただけなので，後半は簡単に聞き流して聞きたかった.
- コンテンツの長さが分からないうえに，今どのあたりを再生しているかが分からないので分かるようにして欲しい．またコンテンツの内容の目次や概要の説明があるとより理解しやすいと思う.
- コンテンツのボリュームとメタデータの読上げのボリュームのバランスを調整できるようにして欲しい.
- 動画のロード中は何も音がしないので不安になる.
- メタデータ無しでは気が付かなかった事実がいくつもあって驚いた.
- 動画だけでは理解できないことが多いが，理解したとしても多くの場合は動画の音声で得

[6] Yahoo! (R) ビューティー，http://diet.beauty.yahoo.co.jp/special/200607/exercise/lesson20.html

[7] Live-kitchen，http://www.live-kitchen.jp/recipe/00000105/index.html

られる情報だけで理解した気になっているのではないか.

5 考察

まず音声解説が動画の代替のアクセス手段に成り得るかどうか考える．実験におけるヨガと料理の 2 種類のコンテンツについては，音声解説が無い場合と有る場合では理解に大きな違いが見られた．ドラマや映画などこれまで音声解説が作成されてきたコンテンツにおいても，音声解説は重要な代替のアクセス手段になっていると考えられる．

どの提示方法が適切かどうかという議論は今回の実験だけでは十分にすることはできないが，今回の実験において，料理の場合に 2 の提示方法によって，ヨガの場合に 3 の提示方法によって理解することができたことを考えると，音声解説による理解に必要な情報量の増加と，情報を理解するために必要な時間によって適切な提示方法は変わるのではないかと推測できる．また，これは動画の内容に関するユーザの経験や，内容の複雑さによっても変わると考えられる．ドラマや映画などのコンテンツの場合には，健常者と同じ時間を共有することが重要であるため，時間的な制約の中で提供できる音声解説を用意するか，時間制約を考慮した新しいインタフェースを考える必要がある．

次にメタデータの編集について考える．今回作成したメタデータは全て手作業によるものであったが，メタデータを挿入する位置を探したり，通常の音声解説が無声区間にきちんと収まるかどうかを手作業で行うのは手間がかかる．音声区間検出や，シーン検出 [5] などを使ってメタデータの作成を補助するオーサリングツールの開発が今後必要になるであろう．

また，社会環境の整備も必要である．現在，第三者であるメタデータ提供者がコンテンツの音声解説などを作成するためには，著作権上，コンテンツ提供者の許諾が必要になる．しかし，障害を持つユーザのアクセシビリティ向上のためには，第三者がメタデータを作成する提供モデルは必要不可欠である．用途を障害を持つユーザを対象に限定し，作成者も登録制にするなどの制限によって，このようなモデルが実現できるような著作権法の改正が望まれる．

6 まとめ

本論文では，音声解説および拡張音声解説を提示するためのインタフェースについて述べた．試作システムではジョグダイアルによる話速制御等の基本的な音声コントロールと，動画の一時停止を伴う 3 つの提示方法によって拡張音声解説を提供することができる．試作システムを使い，視覚障害者を対象に予備実験とヒアリングを実施した結果，拡張音声解説によって動画の理解度を向上できる可能性が示された．しかし適切な拡張音声解説の提示方法はコンテンツの内容やユーザの経験によっても変わると考えられる．

今後，音声解説とその提示インタフェースについてより詳細な実験を行い，コンテンツの種類・被験者の経験等条件に応じた提示インタフェースについて検討を行う予定である．最終的にはシステムを実装して提供することで動画のアクセシビリティ向上へ貢献していきたい．

謝辞

本研究の一部は NiCT ((財) 情報通信研究機構) の助成を得て進められている「視覚障害者向けマルチメディアブラウジング技術の研究開発」で実施した内容を含んでいます.

参考文献

[1] B. Arons. SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Trans. Comput.-Hum. Interact.*, 4(1):3–38, 1997.

[2] A. Chieko, T. Hironobu, I. Shuichi, and I. Toru. A Proposal for a Dialing Interface for Voice Output Based on Blind users' Congnitive Listening Abilities. In *Universal Access in HCI (UAHCI) 2003*, 2003.

[3] A. Chieko, T. Hironobu, I. Shuichi, and I. Toru. TAJODA: Proposed Tactile and Jog Dial Interface for the Blind. *Transaction of IEICE*, E87-D(6), 2004.

[4] P. J. Piety. Audio Description, a Visual Assistive Discourse: An Investigation Into Language Used to Provide the Visually Disabled Access to Information in Electronic Texts. Master's thesis, Faculty of the Graduate School of Arts and Sciences of Georgetown University, 2003.

[5] M. A. Smith and T. Kanade. Video Skimming for Quick Browsing Based on Audio and Image Characterization, July 1995.

[6] W3C. Web Accessibility Initiative. http://www.w3.org/WAI.

[7] W3C. Web Content Accessibility Guidelines 2.0 (Working Draft), Apr. 2006. http://www.w3.org/TR/WCAG20/.

[8] 総務省 情報通信政策局 情報通信利用促進課. 平成１７年度の字幕放送等の実績, 2006.

[9] 森田直孝. 音声の時間軸での圧縮・伸長に関する研究. Master's thesis, 名古屋大学, 1987.

[10] 浅川智恵子, 高木啓伸, 井野秀一, 伊福部達. 視覚障害者への音声提示における最適・最高速度. ヒューマンインタフェース学会論文誌, 7(1):105–111, 2005.

表 2. 付録　動画のメタデータ (漢字など TTS の読上げが適切でない部分はひらがな表記にしてあります)

ヨガ

00m02s00,	まんじのポーズです，最初は仰向けできをつけの姿勢です
00m12s00,	左膝を右側に倒して体を右向きにします，左足は歩く時のモモを上げた状態です
00m23s00,	左右のふとモモはそのままの位置で，右ひざを曲げて右足首を左手でつかみます．その後右足を大きく後ろにひいて，左右の膝を 90 度ぐらいにたもち，足を前後に広げます
00m33s00,	そのままの姿勢を維持し，音楽に合わせて腹式呼吸をします
01m26s00,	そのまま 15 秒くらい静かに呼吸を続けます
01m43s00,	仰向けできをつけの姿勢に戻ります
01m57s00,	さきほどと反対に，右膝を左側に倒して体を左向きにします，右足は歩く時のモモを上げた状態です
02m07s00,	左右のふとモモはそのままの位置で，左ひざを曲げて左足首を右手でつかみます．その後左足を大きく後ろにひいて，左右の膝を 90 度ぐらいにたもち，足を前後に広げます
02m16s00,	そのままの姿勢を保ち音楽に合わせてゆっくり腹式呼吸をします
03m23s00,	stylife beauty のロゴマークが表示され，再生終了です

料理

00m00s00,	メニューはぎゅうロース肉のステーキ，そら豆と，干し貝柱ソース．まずソースを作ります，オリーブオイルは 60cc です
00m23s00,	エリンギ 1 本，赤ピーマン 3 分の 1 個は 5 ミリかくに切ってあります
00m35s00,	干し貝柱 1 から 2 個は包丁の側面で叩いてほぐします
00m45s00,	そら豆 16 粒はあら刻みにしてあります．次のシーンではあら刻みにした，にんにくひとかけ分を炒めています
00m51s00,	エリンギと赤ピーマンを入れました
01m05s00,	そら豆も入れました
01m35s00,	バター 30 グラム，薄口醤油，小さじ 1，唐辛子の粉少々を入れました
01m43s00,	軽く塩コショウをし，混ぜながら炒めています
01m46s00,	画面が変わり，ぎゅうロース肉が用意されています，1 人前 60 から 70 グラムで 4 人分の大きなステーキ肉です
01m58s00,	両面に塩コショウをふります
02m10s00,	焼き目が縞模様に付く波型の鉄板で肉を焼きます
02m16s00,	肉と一緒にくし型に切った竹の子を焼いています
02m22s00,	竹の子に軽くオリーブオイルをかけました
02m27s00,	肉と竹の子をひっくりかえしました
02m40s00,	竹の子に軽く塩をふりました
02m55s00,	肉と鉄板の間にスプーンやフォーク 4 本を挟んで，鉄板から肉を離しました
02m58s00,	画面が変わり，付けあわせを炒めています
03m15s00,	アスパラと菜の花に軽く塩をふりました
03m18s00,	盛り付けをはじめました
03m21s00,	まず竹の子を置き，それを箸置きのようにしてアスパラと菜の花を置きました，位置はお皿の真ん中より少し奥側です
03m35s00,	焼いた牛肉をまな板に移し，1.5 から 2cm 幅に切りました
03m45s00,	肉は各皿 3 切れずつで，お皿の真ん中に，竹の子，アスパラの上から流れるように，切り口を上に向けておうぎ形に広げて盛り付けました
03m52s00,	最初に作ったソースを温め直し，その中に刻んだおおば 3 枚分を入れ，かき混ぜました
04m05s00,	最後に肉の上からソースをかけて完成です．肉の赤みとアスパラやそら豆の緑がきれいです