International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Report ITU-R BT.2447-1

(03/2021)

Artificial intelligence systems for programme production and exchange

BT Series

Broadcasting service (television)

International Telecommunication Union

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from http://www.itu.int/ITU-R/go/patents/en where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Reports

(Also available online at http://www.itu.int/publ/R-REP/en)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	Broadcasting service (sound)
BT	**Broadcasting service (television)**
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management

***Note**: This ITU-R Report was approved in English by the Study Group under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2021

REPORT ITU-R BT.2447-1*

Artificial intelligence systems for programme production and exchange

(2019-2021)

Summary

New broadcasting technologies driven by artificial intelligence (AI) are being introduced to the broadcasting workflow. These technologies are intended to increase productivity, efficiency and creative opportunities during programme production, and to convey information to viewers quickly, accurately and automatically.

This Report discusses current applications and efforts underway and evaluated that are relevant to broadcast programme and production pathway. Relevant applications and efforts are categorized into the following topical descriptions for areas of technological benefit: Workflow Optimization, Bandwidth Optimization, Automated Content Creation, Content Creation from Legacy Archives, Content Selection for Targeting Audience Demographics, Optimization of Asset Selection – Metadata Creation, Dynamic Product Placement and Advertising for Broadcast and Content Personalization.

1 Introduction

The use of Artificial Intelligence (AI) for broadcasting applications has moved past in-lab research demonstrations and theoretical constructs. In the past year, there have been numerous demonstrable practical applications of Machine Learning (ML) and Artificial Intelligence (AI) in working broadcast programme and production efforts. These efforts have successfully used ML/AI to target opportunities in the broadcast programme and production pathway that would provide improvements in production efficiency and correlated cost reduction, targeted higher value content distribution to audience demographics, effective solutions for transition of legacy content into modern feature delivery, and improved or optimized content quality at lower bandwidths. The advance and investment by many broadcast and industry organizations to successfully operationalize and evaluate these methods for internal and external content creation and distribution is a strong indication of the interest and relevance. Further, it is believed that the application and inclusion of ML/AI algorithms will be an integral part of the future broadcast programme and production pathway.

For the purpose of this Report, only applications and efforts underway and currently evaluated are discussed for their relevance to broadcast programme and production pathway. Nonetheless, it is important to recognize that there are numerous efforts underway and in development within multiple research and industry organizations that directly consider the role of future AI algorithms in content generation, optimization and quality evaluation. Only some are discussed in this Report, but many of these innovations will have direct relevance in the near future. In that way, the purpose of the current Report is to highlight areas where ML/AI algorithmic approaches are already affecting creation, process and distribution within the broadcast programme and production pathway. At this time, current efforts underway, evaluated, and captured within this Report include the following areas of technological benefits:

– Workflow optimization;
– Bandwidth optimization;
– Automated content creation;
– Content creation from legacy archives;
– Content selection for targeting audience demographics;

* The revision of this Report should be brought to the attention of IEEE-SA, ISO/IEC JTC 1/SC 42, ITU-T SG 11 and ITU-T SG 16.

– Optimization of asset selection – Metadata creation;
– Dynamic product placement and advertising for broadcast;
– Content personalization.

A description of notable efforts underway for each category is provided.

2 Workflow optimization

The former Head of Marketing and Communications, Media Solutions, Sony Professional Solutions Europe said: “For many, the primary driver of adoption of AI technology is the opportunity to automate routine workflows that are manually executed”.

“Netflix, for instance, estimates that its use of AI to automate workflows and reduce customer churn saves the company around $1 billion annually. This not only increases the quality of experience and quality of service for users, but also reduces the number of bits required to achieve the same quality stream”, he said.

Current efforts to improve workflow optimization span a broad range of serial points of intersection in the programme and production timeline. Companies are incorporating ML/AI into optimized content programming, methods of metadata generation, tagging, and mining, as well as efficient creation of production content including the automation of compliance closed captioning through automated transcription and integration, and extensions in videography and cinematography through automating camera capture techniques and use of virtual camera views. In some of these instances, ML/AI generation, capture or description of content has further been paired with ML/AI algorithms to create new content that has subsequently been distributed to live audiences.

Below are a few notable efforts to describe the integration of these applications.

2.1 Content programming

Companies such as Accenture are working with broadcast clients to incorporate AI into optimization of programming schedules. The BBC has multiple efforts underway to use AI/ML to automate and optimize content programming. These efforts have been used in live broadcast and are also discussed further in later sections. A description of the BBC programming effort follows.

The BBC has multiple projects underway that specifically target workflow improvements aimed towards cost and time savings in production and delivery. To implement and validate their research efforts underway and potential application in a modern broadcast, BBC Four used ML/AI algorithms to mine through thousands of hours of legacy archived content dating back to 1953 to generate programming across two full days, branded as “BBC 4.1”.

This programming was defined by ML/AI algorithms that used information from past scheduling, content metadata and other programme attributes to learn and mine the archived content for optimal targeted demographic programming. This programming then went live on BBC Four paired with other ML/AI content generated efforts.

2.2 Virtual video angle capture and automation

It is possible to take an image sequence and use AI processing to identify or re-frame areas in order to extract a new image with for example, a different aspect ratio or focus of interest. “Virtual-Video Angle Capture” (V-VAC) techniques driven by AI algorithms are being investigated by broadcasters.

BBC R&D

The BBC has implemented efforts through their research and development group that could provide substantial benefit and cost savings during production. For multiple evaluations they have set up several fixed, ultra-high-definition cameras for capture during live events. The capture from these cameras has been used as a feed to a highly reduced or even single-operator human-driven system.

Their efforts allow fewer cameras and camera operators to capture a richer scene and environment. See Single Operator Mixing Application, SOMA:

– https://www.bbc.co.uk/rd/blog/2017-07-compositing-mixing-video-browser

– Lightweight Live https://www.bbc.co.uk/rd/projects/ip-studio-lightweight-live.

These are examples of video projects that allow a single operator to generate a vast number of virtual camera views from the capture of high-resolution videos to create a high-quality edited output. For example, the human operator can interact with the editorial process by adjusting the frequency in the automation cuts between different crops. This ability gained by enabling generation of virtual camera views from a reduced set of cameras during capture in an automated manner provides the potential for notable cost savings during production.

TV Asahi

TV Asahi, a commercial broadcaster in Japan, has developed a V-VAC system based on a lightweight AI engine designed for general-purpose object detection.

To help prevent unnatural framing, the system employs motion processing on each scene to identify areas of interest and a range of object framing movements, which determine the most appropriate focal point coordinates and frame sizes for particular subjects (e.g. MCs, newscasters and anchors). The system can handle up to three individual virtual angles simultaneously.

Fuji Television

Fuji Television, another commercial broadcaster in Japan, has implemented a system that can automatically identify “natural” image framing options. This approach focuses on the balance between focal objects and the space around them. The model is generated through machine learning (ML) with a training dataset of thousands of 2K still images selected from a aired programme.

Figure 1 shows an example of automatic framing.

FIGURE 1

Automated framing operation

a) Detection of two faces (yellow boxes) and cropping a frame in appropriate angle (green)

b) Cropped image

2.3 AI edited content generation for optimization and operational efficiency

Wimbledon and IBM partnered for an effort to rapidly generate highlight clips for distribution to audiences. These highlight clips were identified and edited together entirely by ML/AI algorithms that assessed the exuded facial and emotional response of the players. This assessment was then used to automate a ML/AI driven process on the original content to generate increased emotionally impactful highlight reels that were distributed to public audiences.

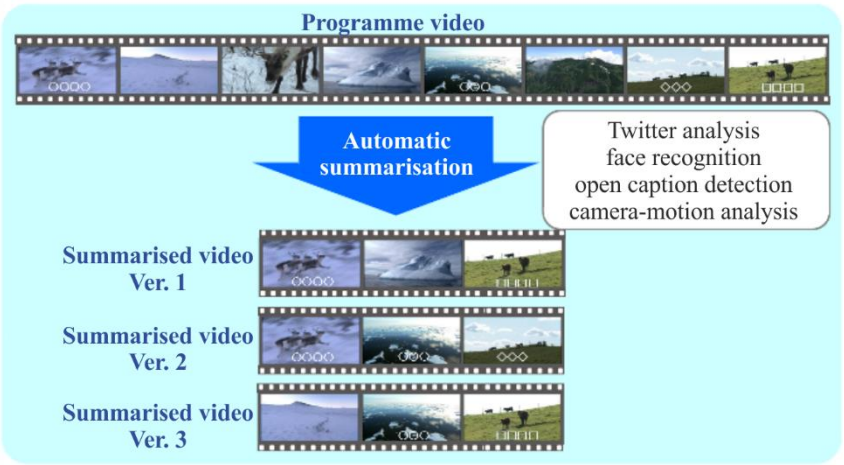
In other instances, ML/AI algorithms have been used extensively by sports production teams through partnerships with companies such as Aspera to rapidly analyse all video scenes and metadata generated as content from a live sports event to put together effective montages of play highlights in a matter of seconds. This has enabled production teams to have significant gains in efficiency and timely generation of content for distribution.

2.4 Automated video digest

Previews of programmes and digest videos are important types of content to give viewers brief introductions to programmes. NHK[1] has developed image analysis technologies that identify the characters and performers shown in a programme. These descriptive extractions can be used for the creation of previews and digests (see Fig. 2). Additionally, public viewer comments on social media regarding the programme, performers and other programme features, may be taken into account for content modification. Furthermore, by allowing programme producers to arbitrarily weight the importance of the analysed information, various types of summary video can be automatically generated.

FIGURE 2

Process of video summarisation



Report BT.2447-01

2.5 Live content optimization for programming development

Endemol Shine Group (ESG) is using a Microsoft Azure AI workflow to replace an entirely manual selection process in the Spanish version of the reality show “Big Brother”. Their efforts make use of ML/AI algorithms to learn patterns of interactions happening in the relationships and dynamics of the house members. The output of these learnings is used to infer and anticipate relationship dynamics of the group interactions and direct resource content development efforts. In some instances, this automation enables improved demographic targeting for content shown to different demographics.

---

[1] Japan’s national public broadcasting organization.

2.6 Compliance tracking and content creation

Multiple companies have targeted creating ML/AI driven workflow improvements to help facilitate and improve FCC mandated compliance in production and delivery. For example, TVU networks has created a transcriber service that is deployed and available for use by Call-letter stations. Their service assures that all video content is FCC compliant prior to on-air broadcast or delivery through any other digital platform. They have integrated AI algorithms into their workflow solution that detect the need for closed captioning in content and automatically transcribe absent dialog as closed captions. Additionally, they make use of ML/AI algorithms to mute audio during any profanity or excluded speech.

3 Bandwidth/Quality optimization

ML/AI is being used in industry applications to improve encoder efficiency and optimization. Efforts by companies like BitMovin are successfully using AI to enable algorithmic learning of information about the content complexity and other features from prior encodes to improve quality and efficiency during later stage encoding. After a few iterations, the resulting encode is considerably closer to a targeted optimum quality at maximum bandwidth efficiency.

4 Automated content creation

Automated content creation to improve workflow efficiency is discussed in §§ 2.3, 2.4 and 2.5. Additional efforts making use of AI generated content creation are becoming ubiquitous. For example, multiple news agencies have been able to break stories well-ahead of their peers by leveraging data mining of online public social comments. In these instances, some of the news agencies also make use of an AI algorithm to generate the captured news story for distribution.

In many ways, news programme production can benefit from big data analysis. This algorithmic approach allows news agencies to automatically sift through massive amounts of diverse information and identify the most relevant themes or trends to present to their producers and subsequent viewers. The information sources range from social media posts, reports and alerts issued by public organisations, to the broadcast station’s own archived programming.

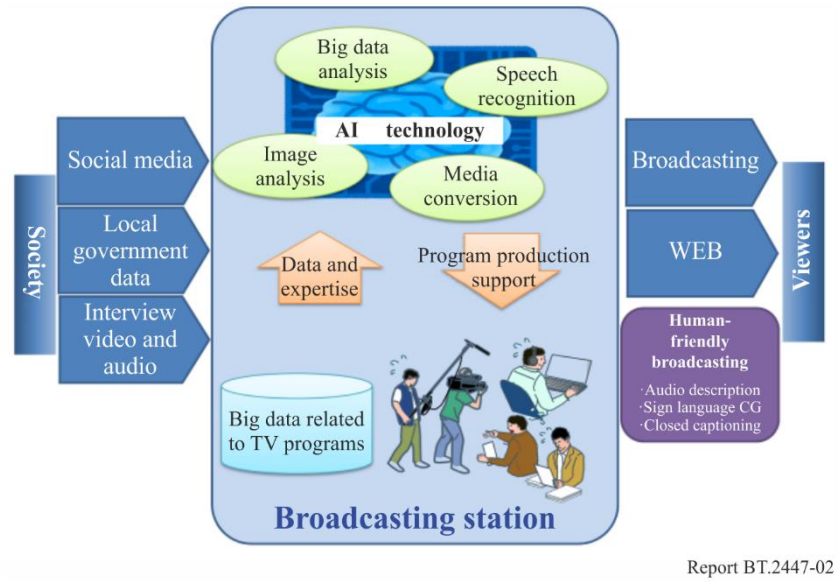
In other instances, speech recognition helps to efficiently produce closed captions for live TV programmes. These are especially useful for aging and hearing-impaired individuals. Speech recognition can also be used to transcribe speech in video content. This method and approach have become indispensable for producing programmes from huge amounts of video materials faster and more accurately. Further, speech synthesis is used to translate written information into aural information where it is more convenient. This has enabled broadcast programmes to identify and employ alternate styled voices where beneficial to the broadcasted content.

As other examples using ML/AI, broadcasters have been developing technologies for automatically translating broadcast data into forms that can be easily understood by all viewers including non-native language speakers or hearing-impaired individuals (i.e. foreign languages or computer graphics (CG)-based sign language).

Figure 3 shows an outline of an example AI-driven programme production system as it could be used for content generation within the programme and production pathway. The AI technology first analyses diverse information (video, audio and text data) gathered from publicly sharable data sets. It then selects information identified to be useful for the produced programme. The media conversion technology is then used for human-friendly (improved accessibility) broadcasting to assure that the produced programme content is successfully available to all viewers including visually or hearing-impaired.

FIGURE 3

Configuration of AI-driven programme production



4.1 Social media analysis system

It is becoming common for news programming to make use of social media data from people geographically located near to where a known incident or event may have occurred. This data can influence and become a critical part of the breaking news story. In this way, productivity and efficiency of programme production can be increased through use of systems that automatically search for trends and notable information for news production from a massive amount of Twitter data (tweets) and subsequently judge the authenticity of such posts [1]. Recursive neural networks (RNN) can be used to determine the presence of any target terms typically associated with news-worthy breaking events. These common terms are used by the RNN to categorize information into categories that represent several types of relevant event classes such as “fire” or “accident”. The RNN then uses these pre-defined classes of information that are representative of types of information often featured in news programmes, to “learn” and categorise the existing information.

Figure 4 shows the interface for the tweet presentation system developed by NHK. The producer checks the utility relevance of the information in each tweet identified by the system. By incorporating feedback from the producer, the system can find new learning data and use it to maintain and improve its ability to extract useful information from the tweet content. Additionally, research is under way to improve the accuracy of the system’s image recognition classification. This will allow the algorithm to improve image identification of relevant events and objects. For example, it would be better at differentiating between a fire and a fire engine.

FIGURE 4

Social media analysis system

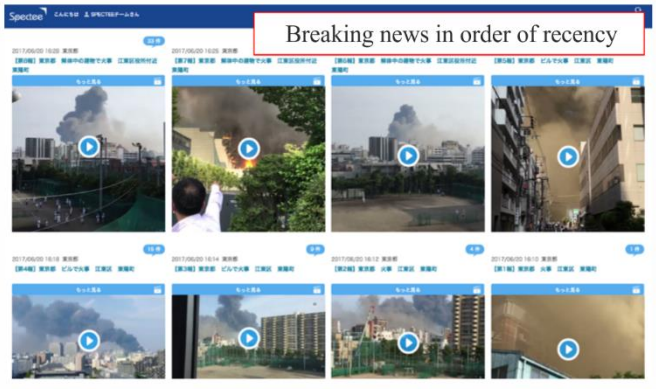
Report BT.2447-03

As an example, since 2016, Fuji Television has been making use of “Spectee”, a service that offers real-time data gathering using a social network service (SNS) to inform their news programming and production. Figure 5 shows the interface of “Spectee”. The AI installed in the system enables the service to automatically analyse, select, and stream newsworthy topics posted on SNS by machine learning algorithms. The service has been adopted by other commercial broadcasters in Japan.

The developed convolutional neural network (CNN) is used for image categorisation based on events and regions. It makes use of keywords to perform image analysis and is accessible through the interface to enable real-time assessment of news data. Deep learning algorithms are successfully applied within the application. Since social media may contain news that are not true or “fake”, an algorithm is used to evaluate the legitimacy and reliability of potential news content. This algorithmic assessment is followed by a human review to verify the information. Authenticity of content and fact checking of sources and details is critical to the success of any automated ML/AI content generation system.

FIGURE 5

Real-time SNS breaking news service



Report BT.2447-04

4.2 Automatic generation of manuscripts describing state of rivers

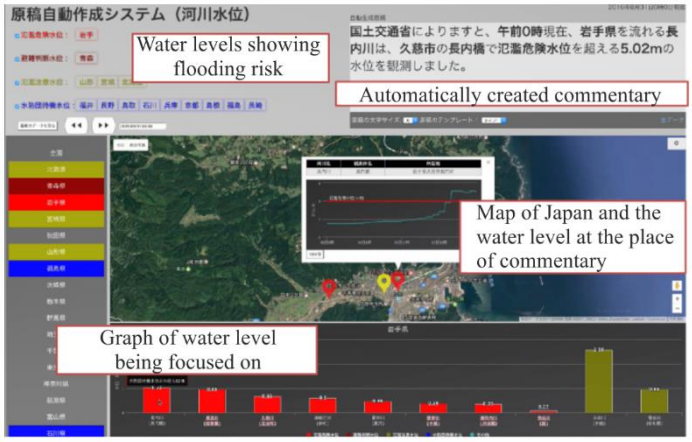
Broadcasting stations gather and analyse various sensor data released by public institutions and use this data to shape programmed broadcasting. Creating a broadcast manuscript in a timely manner, however, places an extra burden on programme producers, who also have to monitor massive amounts of data being released to the public. As a solution, a support system has been developed by NHK that automatically creates manuscripts on the state of rivers based on data from water level sensors deployed along rivers and from previously broadcast news manuscripts [2].

Numerical river water level data is made available every ten minutes by a public institution. This data includes the names of observation posts, current river water levels and water level thresholds categorised into four levels related to the danger of flooding.

Broadcasting stations maintain a database of past broadcast manuscripts. Templates for manuscript generation are created by extracting specific expressions in a manuscript matched to river names and water levels using a neural network. Figure 6 shows the system interface. A draft manuscript based on the water level data and the manuscripts accumulated so far is displayed. The system also allows a news reporter to change the manuscript or create a new manuscript incorporating the latest river level data.

FIGURE 6

Automatic manuscript generation system



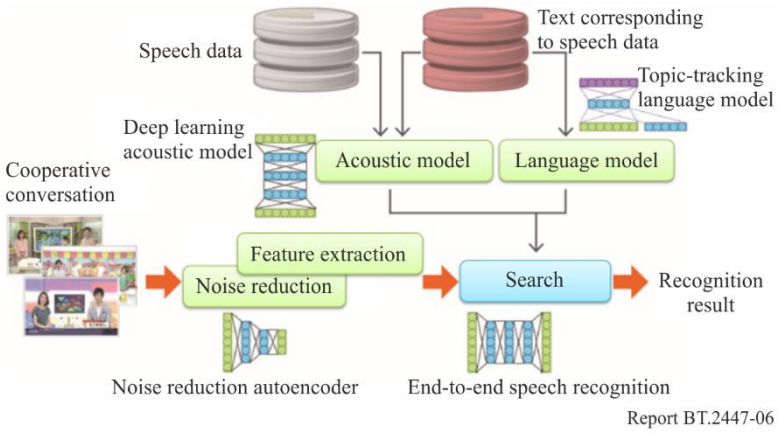
Report BT.2447-05

4.3 Captioning

Speech recognition technology is expected to help quickly and accurately produce captions in live broadcasts. The accuracy of speech recognition in a TV programme deteriorates if there is background noise and inarticulate speech during a conversation. Conventional captioning systems use humans who work in a quiet room to simultaneously provide a clean speech stream from the original noisy or inarticulate speech (re-speaking) that serves as input to the automated recognition algorithms. Using this clean speech as input enables the automated captioning system to reach sufficient accuracy for live captioning. A way of automatically recognizing programme audio without the need for the added human re-speaking is desired. For the direct recognition of such programmes, Deep Neural Network (DNN)s have been adopted to develop an acoustic model that estimates vowels and consonants of input speech to increase recognition accuracy. Furthermore, to recognise inarticulately or ambiguously pronounced words that cannot be incorporated manually into a pronunciation dictionary of phoneme sequences, an end-to-end speech recognition system that does not require a pronunciation dictionary has been developed [3]. This system trains a DNN that maps an input speech signal to a symbol sequence (see Fig. 7).

FIGURE 7

Elemental technologies for speech recognition that can be replaced by DNN



Speech recognition is used in different areas of broadcasting operations; however, the required recognition accuracy differs depending on the application and its purpose. Table 1 shows the subjective usability for Japanese speech recognition accuracy reported in [4].

TABLE 1

Subjective usability for automatic Japanese speech recognition accuracy

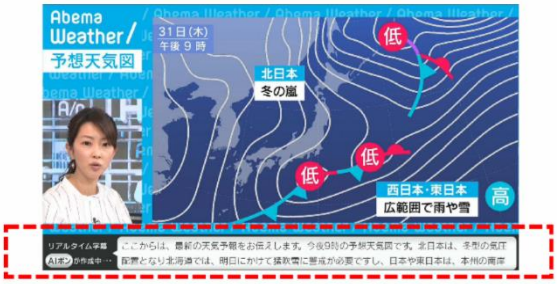
Speech recognition accuracy	Subjective usability
60%-80%	Capable of understanding topics/keywords
75%-90%	Capable of understanding the contents
85%-95%	Errors are scattered in the recognition text
95%-	Almost perfect

Where captions are delivered directly to viewers, high accuracy without errors is essential. Currently, speech recognition technology has yet to achieve 100% accuracy and therefore, errors in transcribed text still requires correction by human operators. In the case of Japanese “real-time” news captions, operators can correct approximately one word in six seconds. This means, for news captions at a rate of about 200 words per minute, a word error rate (WER) of 5% or less is required to produce error-free captions.

TV Asahi has developed a system that automatically attaches captions on TV images (see Fig. 8). This system has been used for live broadcast programmes since December 2018. The Announcer’s conversations or comments are first transcribed and then converted into captions through automatic AI-based proofreading with the following three functions: (1) to provide punctuation, (2) to delete unnecessary words such as “um”, and (3) to correct inadequate words and phrases for broadcasting. This AI application uses open-source programmes and application programming interfaces (APIs) for morphological analysis and speech recognition.

FIGURE 8

A system attaching captions on the image



Report BT.2447-07

NOTE – The inside of the dashed line indicates Japanese captions created by AI.

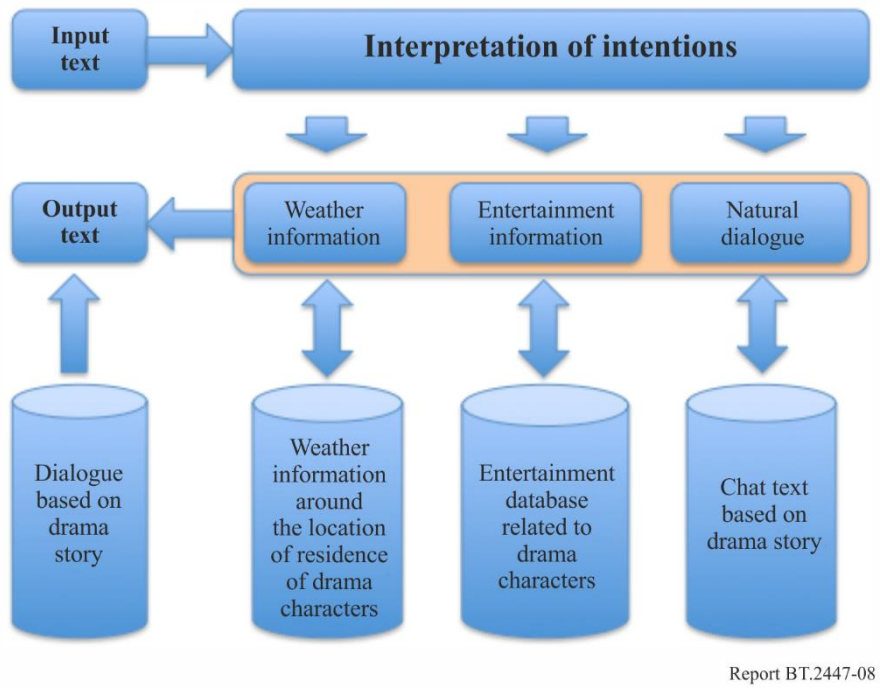
4.4 Dialogue systems

TV Asahi has developed a mascot character with a dialogue system driven by AI technology. This system is enabled by technology with three primary algorithmic functions. These include algorithms that understand human words and phrases, enable cooperation with an external system and perform generative speech synthesis. The algorithms that understand human words and phrases function by first analysing similar words with converted text from human voices to detect inconsistent spellings and differences in pronunciation. The algorithm then calculates the closeness of the meaning of the converted text to that of the original human dialogue by comparison with the script of the previous human dialogue. The second class of algorithms enables cooperation with an external system. For example, if the system encounters a question such as “what is the day today?”, for which the correct answer changes dependent on a variable, the algorithm must be able to query an external system for the correct content to answer. In that way, it is necessary to have algorithms that cooperate with a multitude of external systems. Finally, the third class of algorithms performs generative speech synthesis and enable candidate texts to be generated by the dialogue system and converted to a human voice by speech synthesis. After further developing these base technologies, TV Asahi is introducing appearance of this mascot character into TV programmes as a substitute for a caster or announcer.

In 2017, Nippon TV, a commercial broadcaster in Japan, provided an SNS application service using a chatbot (chat robot). This chatbot took the persona of a heroine character in a popular TV drama series and conducted text message conversations with users of the SNS application where the users were placed in the role of the heroine. The conversation database was fed with input content from the evolving scenarios of the weekly drama. This allowed the supporting long short-term memory (LSTM) network to be successfully trained and produce appropriate conversation responses. The chatbot was able to become more friendly to the SNS application users and automatically chat in a manner corresponding to the story. In addition, to generate a more natural conversation, the linguistic context of the user’s text conversation was reflected in the responses of the chatbot. This service led to 440 000 users joining the chat that engaged in 100 000 000 chats in three months. In 2019 the service was extended to enable users to communicate with at least four characters in other TV drama series through implementation of a Natural Language Processing (NLP) algorithm that enables natural conversation among all characters and users. With these updates each dialogue helped to change users’ impressions of the characters and the relationship between the characters and users.

FIGURE 9

Outline of AI chatbot



4.5 AI-driven announcer

An AI-driven announcer has been successfully used in one of the NHK news programmes since April 2018. In this news programme a CG-generated announcer with a high-quality synthesised voice generated through machine learning algorithms (see Fig. 10) reads out news manuscripts about topics being discussed on social networks. The success of the AI-driven announcer is expected to be refined through ongoing machine learning algorithms and feedback from the audience. The news read by the AI-driven announcer system is also provided on-demand for streaming on smart speaker devices.

FIGURE 10

AI-driven announcer in a news programme

Text data collected from a large number of past news manuscripts and audio data as read by professional announcers serve as input data to machine learning and DNN algorithms. The AI-driven announcer system learns the characteristics of the speaker's voice quality and methods of news reading, and creates a statistical model to reproduce them. The model, which continues to be improved

through machine learning, facilitates automated reading of news manuscripts in a natural cadence and prosody, reflecting the characteristics of professional newsreaders. For example, it effectively learns delivery traits common to announcer's delivery such as lowering the pitch at the end of a sentence.

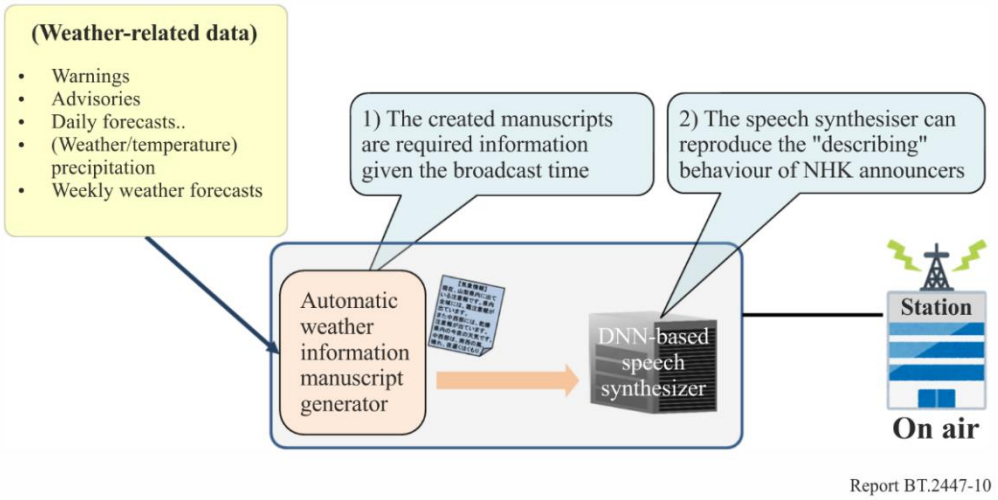
Currently, human follow-up modifications to the synthesised voice are necessary before broadcast to preserve the level of broadcast quality desired for news reporting. Nonetheless, with additional development efforts, it is anticipated that the step requiring human correction will soon be unnecessary to enable a fully automated AI-driven announcer system that reaches an accepted quality for delivery.

It is expected that the use of AI-driven announcers will not only enhance the effect of programme production but also reduce the burden on human announcers. Simple and formulaic tasks can be entrusted to AI-driven announcers while human announcers can devote themselves to more creative professional work.

NHK has also developed an automatic voice generation system for weather information programmes. Using this system, a trial of automatically producing radio programmes started in March 2019. Weather information including daily and weekly weather forecasts and temperature and precipitation probability is spoken by the system in a style similar to a professional announcer. In-house developed DNN-based speech synthesis technology has been used to learn a number of voice styles from weather reports read by a large and diverse population of professional announcers. This system can also generate a suitable length manuscript for the programme by choosing necessary items from various weather data and arranging them into a logical order considering the programme length and knowledge from the professional announcers. By employing the automatic voice generation system, the announcers of the broadcasting station can focus on other areas such as news coverage and programme production.

FIGURE 11

Automatic generation of vocal weather forecast programme



4.6 Automated commentary

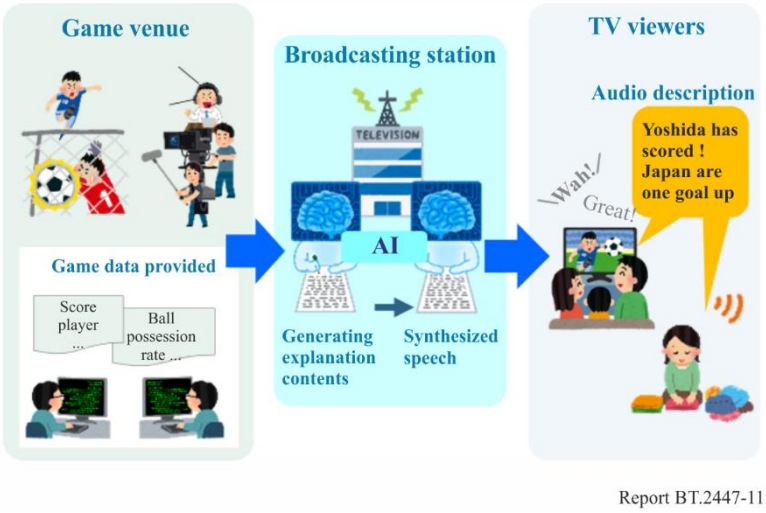
Broadcasting services provide a second sound channel in programmes for commentary. For example, by adding descriptive speech commentary to video, visually impaired individuals can gain an improved understanding of the program content. It is highly desirable that this second sound channel be enabled for both pre-recorded content such as dramas as well as live broadcasts such as sports events. Developments underway in laboratories suggest a "barrier-free service" with automated commentary produced for live programmes through speech synthesis can be anticipated in the near future [5].

In live broadcasts of sports, player names, scores, elapsed times, rankings and past scores are displayed along with the game images. However, not all data on the screen is necessarily spoken. It is common that the sportscaster does not refer to everything appearing on the broadcast screen and often makes comments about things that do not appear on the screen.

An automated audio description system (see Fig. 12) converts data that is not provided in a broadcast into speech that overlaps with the sportscaster's voice. Often during a sporting event, the organiser and manager provide real-time game data to the broadcasters, such as "who", "when" and "what". This includes, scores, goals and other descriptive information. This data is used to generate a script describing the ongoing game by using a template that has been prepared in advance. The script is fed to a voice synthesiser to generate an audio description. The audio description must be presented in a manner that does not interfere with the voice of the sportscaster.

FIGURE 12

Process of audio description for live sports programme



Using the same technology, it is possible to produce an automated commentary without a sportscaster [6]. A large-scale experiment was successfully conducted at the Rio Olympics and Paralympics, where a live broadcast of all of the games was generated by using the distribution feed from the organiser. In real-time this system successfully described the progress of each game in terms of shots on the goal, serving success, ball possession rate, and so forth. Both automatic audio description and automated commentary services will be used in a wider range of services in future.

4.7 Foreign language translation

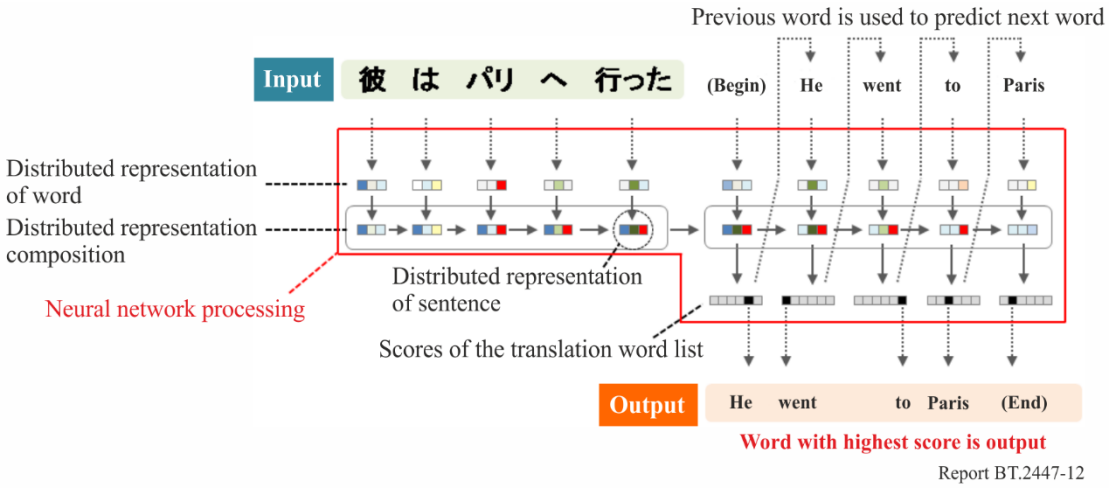
Broadcasters aim to strengthen information dissemination in foreign languages. For example, NHK broadcasts TV and radio programmes in 17 languages. Machine translation technologies are becoming highly valued for translation into foreign languages for generation of captions and scripts.

Machine translation systems are constructed by having AI algorithms learn to translate between parallel sentences. An example would be from Japanese to English. Neural network machine translation technology (NMT) has made remarkable progress over the past few years.

Figure 13 shows the process of NMT. Each word of an input sentence is first converted into a sequence of numerical values called a distributed representation. The numerical sequences are then composed to generate the distributed representation of the sentence. Once this word input process is completed, the translation scores of the word list are calculated, and the word with the highest score is the output. The calculation of the scores and the output of a translated word are repeated.

FIGURE 13

Overview of machine translation using neural network



NHK WORLD-JAPAN, the 24-hour international English TV channel of NHK, extended the automated multilingual captions translation service in 2020 (see Fig. 14). The service allows viewers to select captions in one of eight languages. The captions on the live stream are provided either by the website or the official app. According to native language speakers involved in the service, current automatic translation for News programmes is almost perfect for English to French or Spanish but translation from English to Asian languages contains occasional mistakes.

FIGURE 14

Multi-lingual caption image (Spanish) on international live streaming service

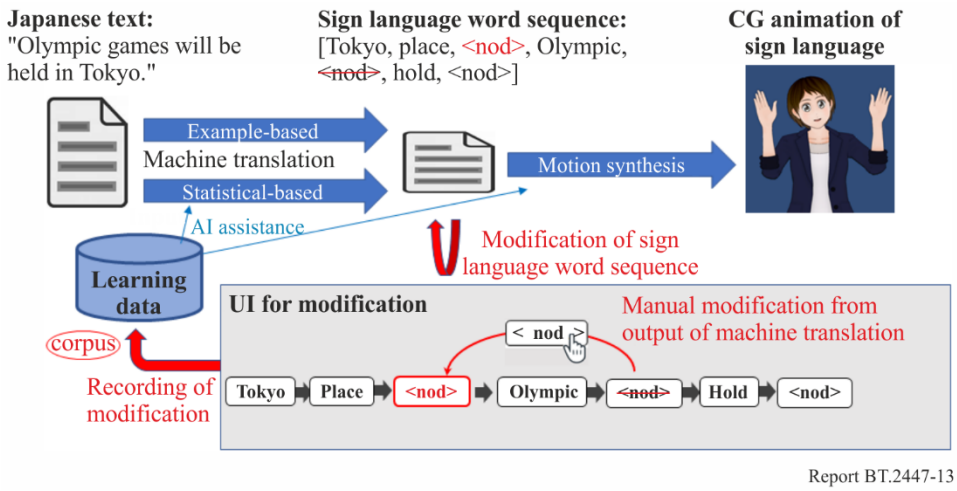
4.8 Sign language CG synthesis

Many hearing impaired individuals would benefit from having Sign Language accessible on programming. To transfer spoken/written information to visual Sign Language, machine translation technology is applied to synthesise a CG animation of Sign Language. A machine translation system translates text to Sign language word sequences using two technologies, an example-based translation and a statistical machine translation [7] [8]. The machine learning translation is enabled through acquisition of training data from a parallel corpus of Sign Language news content. However, it is often the case that the size of the corpus available is small, which can cause the output of the machine translation to be insufficiently accurate. To improve the success of the algorithmic translation, approaches currently include an added step by a human evaluator that manually modifies the word sequence.

Figure 15 shows a computer-assisted Sign Language CG production system. The user interface allows Sign Language words to be modified manually. In this system, the more Sign Language word sequences modified and recorded, the larger the corpus becomes, and the improved accuracy of machine translation can be expected through the use of the system.

FIGURE 15

Computer-assisted sign language CG production system



4.9 Automatic colourisation of monochrome images

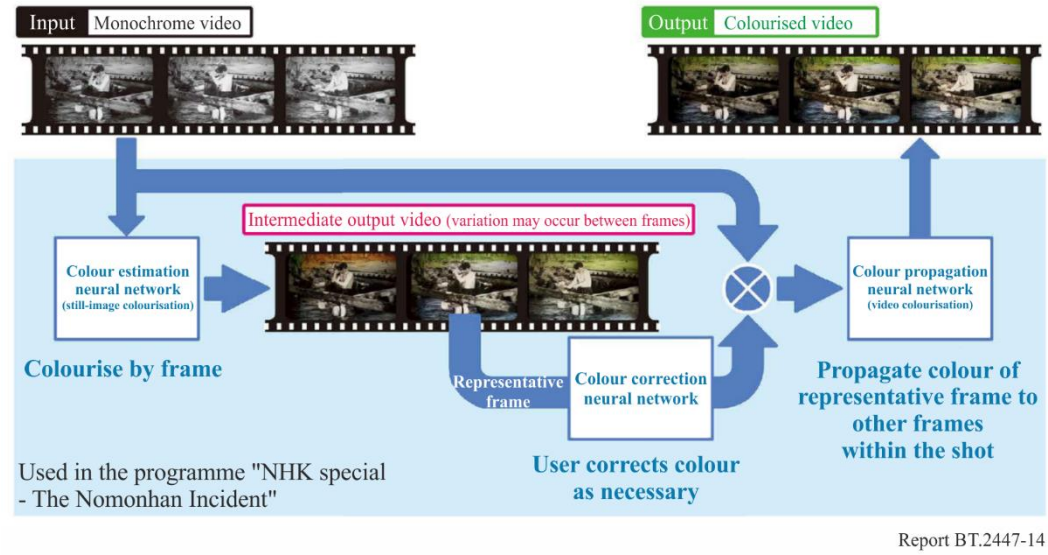
Colourisation can be beneficial to the use of archival monochrome images in historical or documentary programmes. It can enable more realistic, engaging, and immersive content production. In the past, the only way to colourise monochrome video was for specialists to colour each frame manually, requiring considerable time and effort. NHK has developed a colourisation technology for monochrome video images (see Fig. 16) that uses AI with DNNs.

For AI-based colourisation, it is necessary to learn the colours of various objects in advance. For example, the algorithm can learn the colours statistically prevalent in the sky, mountains, or buildings. A colour estimation neural network was trained using about 20 000 past programmes stored in archives. Using the trained neural network, the time required for colourisation was significantly reduced from several days by hand to ranges in the seconds and minutes.

The colour estimation neural network may not always colourise images correctly. Once a producer manually specifies colours at several points in the target area, a colour correction neural network can subsequently provide the correct colours. A colour propagation neural network adjusts the colours of entire scenes for consistency, resulting in less colour flicker.

FIGURE 16

Diagram of automatic colourisation system



4.10 Automated programme content creation

In 2018, BBC Four carried out two full days of programming that were entirely selected and scheduled by AI algorithms to optimize to the user demographic. Included in these two days of AI motivated broadcast were portions of content that were generated solely by select AI algorithms. BBC Four leveraged their extensive data archive for training of the algorithm that created the segments that aired with content that was directly generated by AI. Serial stages of learning and generation with the archived content included: 1) scene identification – what it consists of – landscapes, objects, people; text assessment and learning including subtitles of archived programmes, connections between words, topics, themes, content; 3) motion assessment – activity level or energy; 4) amalgamation of learned features and attributes to generate a novel piece of content that aired on BBC Four.

5 Optimization of asset selection – metadata creation

Generation of content metadata for legacy and new content can be extremely time consuming and operationally inefficient. ML/AI algorithms have become quite effective at successfully automating metadata generation in legacy and new content within media asset management workflows.

Broadcasting stations have archives of past broadcast programmes containing video content and other recorded materials. Programme producers often search archived video content and audio files for possible reuse during programme production. The search will be made easier if the video content and audio file are associated with metadata indicating the content information. Image and audio analysis technologies including face recognition, scene text detection and audio feature detection will make it possible to automatically generate metadata describing relevant information and features of the scene.

5.1 Video, audio detection and recognition

The company Prime Focus Technology has partnered with multiple industry affiliates. Their system aims to improve media asset management through implementation of ML/AI algorithms that recognize elements within audio and video content and automatically generate associated metadata. These types of content management automatization systems can have significant impact on operational costs and reduction of asset management and metadata errors.

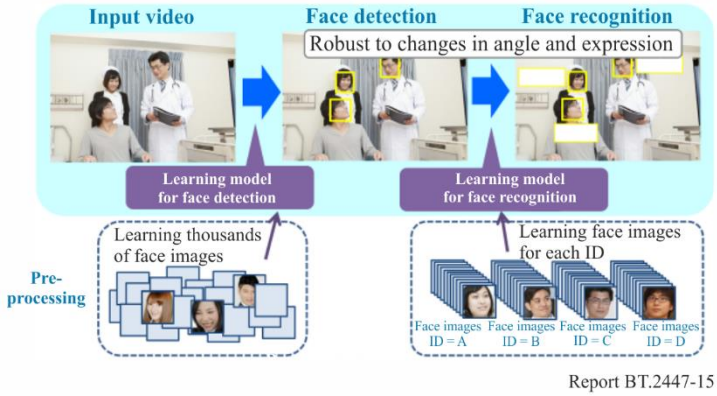
5.2 Face detection and recognition

Face recognition technology can improve operational efficiency when identifying a specific person playing an important role in the video content. However, facial images in television programmes frequently shift due to differences in the varying illumination conditions, orientations and expressions that make it difficult to accurately detect and recognise an individual.

In one example and implementation strategy, improved detection sensitivity and a reduced processing cost for face detection have been achieved by eliminating the influence of different orientations and expressions. This was accomplished by taking into consideration the general positional relationship between the eyes and nose and by designing a detector that utilises a decision-tree structure as shown in Fig. 17. The system can robustly distinguish individual differences by using integrated image features obtained by overlaying multiple small face boundary blocks. This technique works well even when the positions of facial points cannot be obtained accurately [9].

FIGURE 17

Process of face detection and recognition

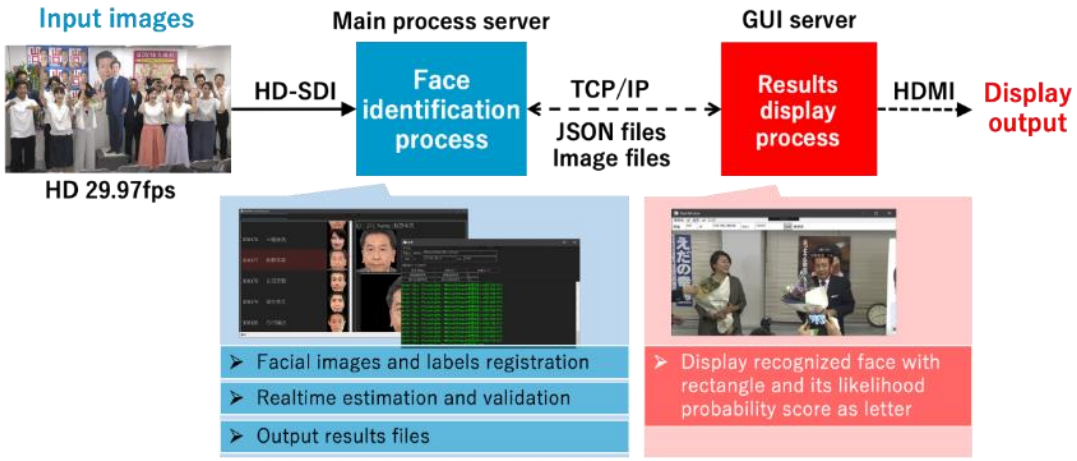


Face detection techniques are usually developed for specific applications and no single technique is applicable to all situations in programme production workflows. Broadcasters have been making efforts to develop suitable methods to meet their needs.

Nippon TV developed a “Deep Neural Network” based (DNN-based) face identification system to identify the candidates of a national election. A library of 2 000 facial images, including the 370 candidates, was used to validate the accuracy of the system and achieved more than 99% accuracy in identifying the candidates. In an election coverage programme, live video of the candidates is relayed to the broadcasting station where the correct profile of each candidate is identified in real time. The system successfully supports this operation with high accuracy and rapidity.

FIGURE 18

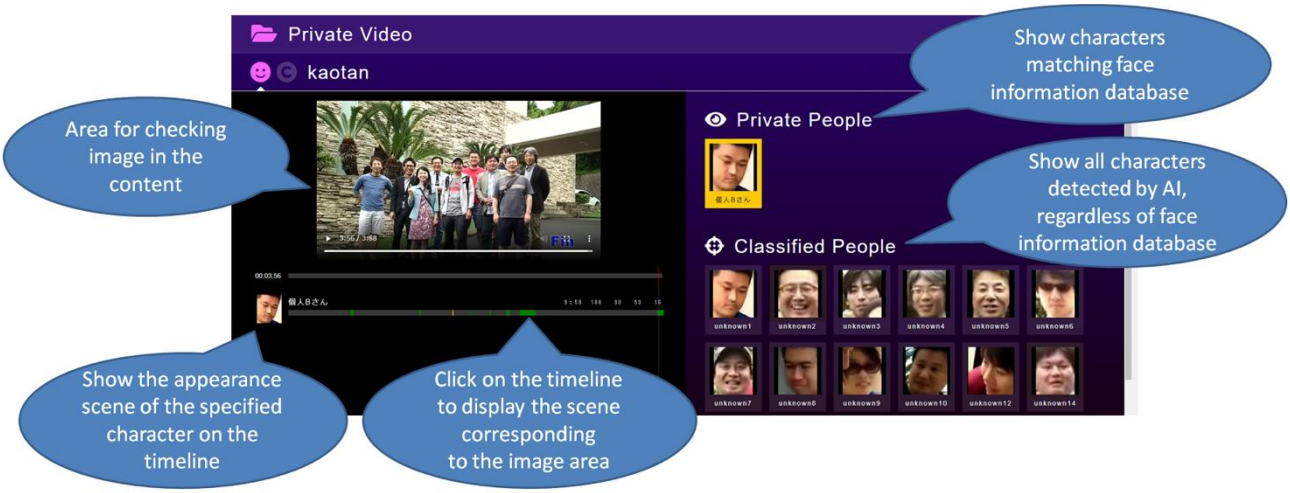
Face identification system



Tokyo Broadcasting System Television, a Japanese commercial broadcaster, has developed a system that identifies performers' appearances in video content on a timeline using a face recognition AI engine (see Fig. 19). It takes a few seconds to find a single person by providing a reference image and no more than a minute to detect all persons with 90% detection accuracy in the content of a 2-hour video. The face recognition performance of this system is less sensitive to appearance changes caused by ageing. For example, it was even possible to search for a specific actor in present content by providing his childhood portrait from over 15 years ago. To take advantage of the robustness against the effects of ageing, this system will be implemented in an archive system that manages a variety of past video material.

FIGURE 19

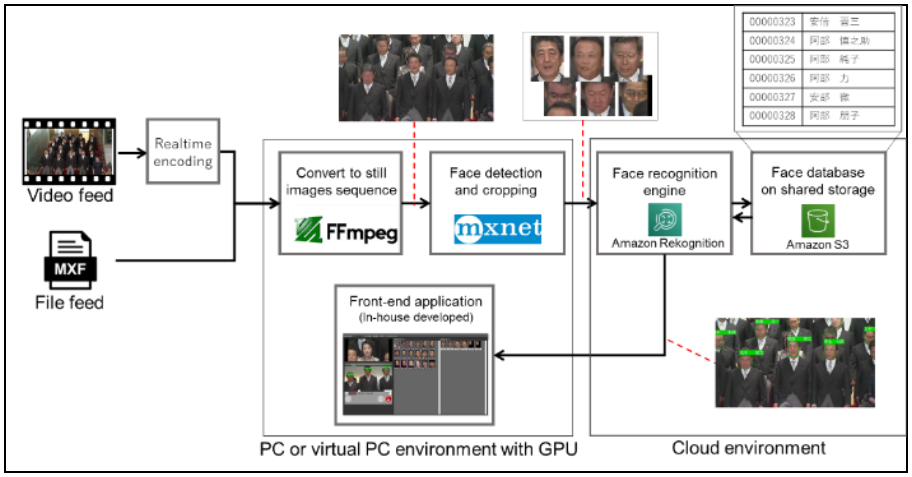
User interface for performer recognition system



Fuji Television has developed a real-time detection system that identifies human faces in video and generates metadata containing their names and the time detected. Figure 20 shows an overview of this system. The system combines a number of open-source software applications and a cloud service to provide the essential functions including image processing and face detection. The system was implemented for a two-week programme production trial where it was able to detect over 20 different human faces in about two seconds and using a library of 8 400 human faces. The system demonstrated a 50% reduction in the time needed to identify individuals in the content over the traditional method.

FIGURE 20

Face detection system



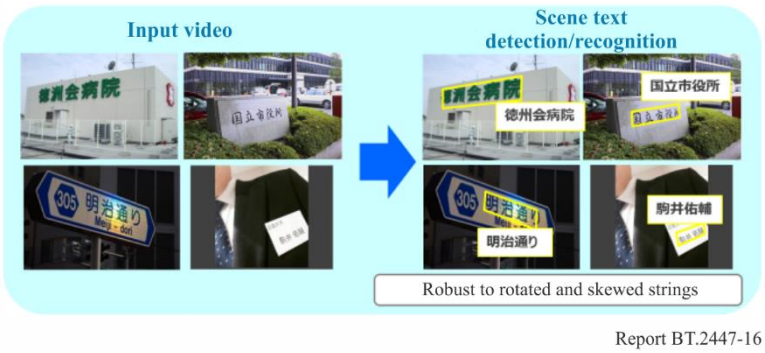
5.3 Detection of text in scenes

Text detection can be used to add metadata about scenes depicted in broadcasts and other production materials. For instance, text that appears on public signs in relevant images can be used to identify the location or address of a building. In this way, an ML/AI algorithm can be used to extract text information from the scene to create metadata that describes the location or building. Attachment of this metadata to the scene enables the scene to be easily located through an effective search algorithm, in turn, enabling it to be located efficiently and used or referenced in many future opportunities.

Text appearing in scenes is often distorted because of skew or rotation. The automatic recognition of scene text is much more difficult than recognising characters in a document. By incorporating the aspect ratio as a feature quantity such that a rotation of the text through any angle can be performed, skewed text can be detected with higher precision (see Fig. 21). Also, by computing features in areas containing multiple texts, a non-contiguous string can be detected [10].

FIGURE 21

Process of scene text detection



Report BT.2447-16

TV Asahi has implemented a system to automatically replace subtitles using text detection technologies. The system uses two features:

1 Text detection – to locate particular text (e.g. a player's name and country) in different patterns of subtitles on international programme feeds, and

2 A replacement module to look up text in an anthroponym English-to-Japanese translation table.

The translated player's name "overwrites" the original English subtitle automatically by evaluating both the accuracy of the character recognition and edit distance (Levenshtein[2] distance) (see Fig. 22). The system has achieved fully automated operations without the need for manual intervention, fulfilling the requirements and reliability required for live programme production and to facilitate an efficient workflow.

FIGURE 22

An example of subtitle replacement

5.4 Object detection and recognition

An application called a real-time indexing system that uses object detection and recognition technologies has been developed for live sports programmes by Nippon TV. The system captures the video stream, detects the player profiles and uniforms, and then learns to recognise "individual players" in real-time using a CNN.

The Nippon TV system has been used in a programme covering a 12-hour relay race to reduce human workload and errors and produce useable metadata. Relevant metadata generation includes times taken for different stages of the relay, names of the athletes and teams, and mileage related to automatically detected objects. This metadata generation is then combined to generate an indexed video stream and can be used to enable a real-time open caption system. Indexed video streams are helpful for editing programmes and use with archival systems to facilitate future use of materials and discovery in search tools.

5.5 Transcription

There is demand for a way to effectively and accurately transcribe speech. This includes complex and live scenarios where content is captured as conversations and interviews. Producers would benefit from ways to rapidly and efficiently evaluate and assess content for use in the programme. Speech recognition is a tool that would greatly improve the workflow process for producers, and its accuracy is critical to realizing this benefit to the transcription process. Figure 23 shows a user interface of the transcription system at NHK [11].

---

[2] The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. The Word Error Rate, WER, is derived from the Levenshtein distance, working at the word level instead of the character level to compare two texts.

FIGURE 23

User interface for speech transcription system

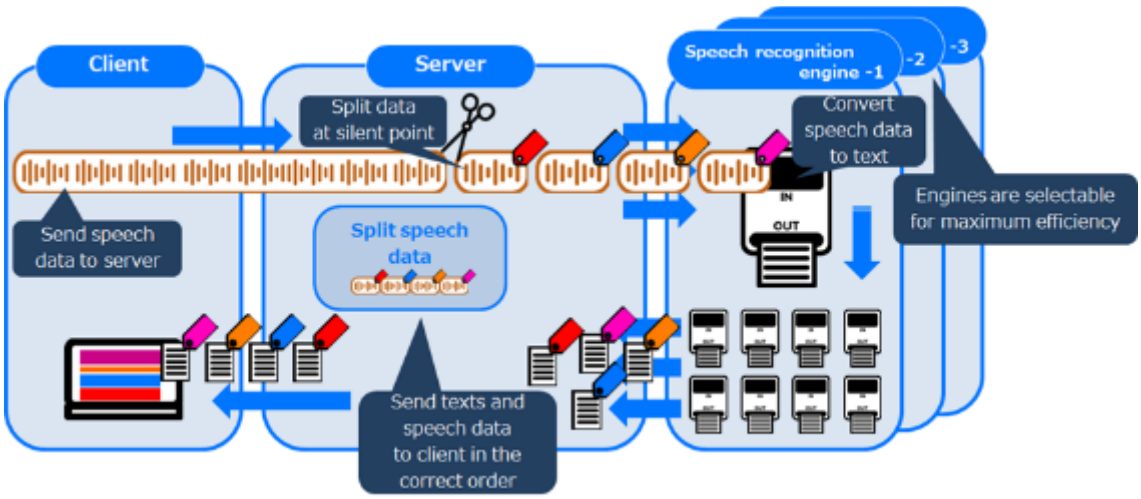
Where speech-to-text is used for programme production transcription, programme makers can tolerate a lower accuracy level than that required for speech-to captioning (see § 4.3). Table 1 shows that transcribed speech content can be understood where the WER is less than 25% meaning the size of the data-set used for an interview is small when compared to the data-set used for live captioning of news programmes.

To achieve this level of WER for transcription, efforts have been made to increase the size of the data-sets for ML and improve DNN learning methods.

Since no one speech recognition technique can cover the entire range of genre and type of conversation, Tokyo Broadcasting System Television has developed a transcription system that supports multiple AI engines for speech recognition. Users then select the best option for their particular use case (see Fig. 24). Specific engines have been found to be better for specific fields. The system also provides functionality to allow users to quickly modify (or correct) the AI transcripts resulting in a system that has enabled a highly efficient workflow, saving about 50% of total work time or about 1 500 hours per month.

FIGURE 24

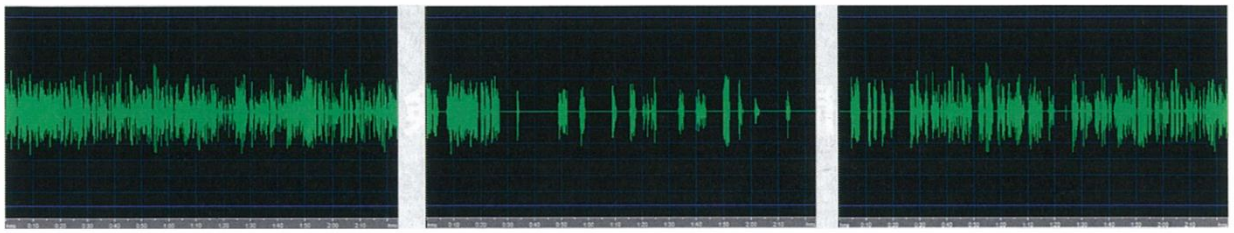
Process of speech recognition system



In some instances, transcription is necessary before editing a TV programme. To improve the success of speech recognition algorithms, it can be helpful to use an AI-driven talker separation algorithm to separate and categorize the dialog of each individual in the programme cast. Inclusion of this additional AI algorithmic step prior to use of a speech recognition algorithm assessing the audio signal, can have notable benefit on the success of the overall automated system. For example, Tokyo Broadcasting System Television has been developing an AI system that separates an audio signal into the signals of each person who is speaking. This system recognises the features of the voice and speech of each speaker using a DNN algorithm before the system attempts to identify the programme dialogue. Figure 25 shows the separation of comments in a TV programme in which a female and a male appeared. A speech recognition system performed better in the condition with the separated audio signals than the condition that assessed the original audio content where the signals remained mixed.

FIGURE 25

Result of separation by AI in the case of two speakers



6 Dynamic product placement and advertising for broadcast

The company Ryff envisions a very different future for the interaction between advertising providers, brands and content. Ryff has developed AI applications that allow dynamic product and brand placement in produced content. Many 3-D objects can be placed in a pre-produced scene or brand skins replaced as contracts expire or timeliness evolves. For example, the same initial content could have a different vehicle/automotive brand appear in the same frames given the demographic or timeslot. This creates a shift in the interaction between the advertiser, content developer, and content distributor. The same piece of content could include, more or less, or, same or different, product and brand placement dependent on demographics or subscription levels.

If successful, this approach will have a notable change on how product placement is handled in creation and distribution. Product placement/brand placement could effectively be decoupled from initial content generation during the production pathway. In an ML/AI inspired product/brand placement workflow, post-production and new processes established during distribution would become key points of product and brand insertion. Interestingly, this can mean that product placement in high value content will have a life cycle that expires. It is an important thought question to consider how content creators may react to this potential malleability in product/brand placement. Brand/product representation has a notable impact on era preservation as a critical part of storytelling. Nonetheless, it could have notable impact on content where advertising drives critical revenue and the time era capture is not critical to the authenticity of the storyline. Moreover, it could have notable implications on the temporal component of advertising such that increased product/brand placement within content could allow reduction in temporal advertising interruptions during content delivery. In many instances, any potential shift will require new considerations for standardization efforts to assure interoperability across demographics and the production and delivery pathway.

7 Content personalization

Content personalization efforts supported by ML/AI algorithms cover a large area of ongoing research. The most relevant efforts to the near-term programme and production pathway include demographically targeted and optimized content for different audiences. This is employed most readily where different geographies are distributing content from other geographies and aiming to improve content relevancy with their audience demographic. It is also used to define what is shown and directed for reality series such as Big Brother, where the human interactions evolve during the course of capture.

User directed storylines are gaining increased interest with the Netflix Black Mirror episode "Bandersnatch". The role of ML/AI algorithms is rich and critical to support any successful learning and user-decision-driven modular storyline efforts in broadcast or otherwise.

Personalization of content features such as dialogue levels and gain driven by a user's sensor captured cognitive effort are also relevant applications of ML/AI algorithms for interface integration into a broadcast ecosystem and personalized user experience.

8 Conclusion

ML/AI algorithms are having significant positive impact on the programme and production workflow. This is apparent by increased operational efficiency in the companies that are employing AI. In addition to numerous workflow gains in efficiency, ML/AI is also enabling a notably improved and relevant audience experience. At this time, there are multiple successful implementations of ML/AI into relevant broadcast programme and production efforts.

A key requirement for the intelligence provided by AI technologies is a large amount of high-quality training data that represents the relation between interrelated events and phenomena for a given task. The development of a training framework utilising real task data is therefore required to build applications with sufficient accuracy for broadcasting.

References

[1] Miyazaki, T., Toriumi, S., Takei, Y., Yamada, I., Goto, J., "Extracting Important Tweets for News Writers using Recurrent Neural Network with Attention Mechanism and Multi-task Learning", 31st Pacific Asia Conference on Language, Information and Computation, #12, 20147, 2017.

[2] Goto, J., Makino, H., Takei, Y., Miyazaki, T., Sumiyoshi, H., "Automatic Manuscript Generation for News Production", ITE Winter Convention, 13B-7, 2017 (in Japanese).

[3] Ito, H., Hagiwara, A., Ichiki, M., Mishima, T., Sato, S., Kobayashi, A., "End-to-end Speech Recognition for Languages with Ideographic Characters", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, PaperID 118, 2017.

[4] Kawahara, T., "Recent Progress of Spontaneous Speech Recognition – Deployment in Parliament and Applications to Lectures", Journal of Multimedia Education Research, Vol. 9, No. 1, S1-S8, 2012 (in Japanese).

[5] Ichiki, M., Shimizu, T., Imai, A., Takagi, T., "Investigation of Simultaneous hearing of Live Commentaries and Automated Audio Descriptions", Autumn Meeting of the Acoustical Society of Japan, 3-5-7, 2017, pp. 1509-1510 (in Japanese).

[6] Kurihara, K., Imai, A., Sumiyoshi, H., Yamanouchi, Y., Seiyama, N., Shimizu, T., Sato, S., Yamada, I., Kumano, T., Takou, R., Miyazaki, T., Ichiki, M., Takagi, T., Ohshima, S., Nishida, K., "Automatic Generation of Audio Descriptions for Sports Program", International Broadcasting Convention [IBC 2017] Conference, 2017.

[7] Kato, N., Kaneko, H., Inoue, S., Simizu, T., Hiruma, N., "Machine Translation to Sign Language with CG-animation", ABU Technical Rview, No. 245, pp. 3-6, 2011.

[8] Kato, N., Miyazaki, T., Inoue, S., Kaneko, H., Hiruma, N., Nagashima, Y., "Development and Evaluation of a Machine Translation System from Japanese Texts to Sign Language CG Animations for Weather News", IEICE Transactions on Information and Systems, Vol. J100-D, No. 2, pp. 217-229, 2017 (in Japanese).

[9] Kawai, Y., Mochizuki, T., Sano, M., "Face Detection and Recognition for TV Programs", IEICE Technical Report, Vol. 117, No. 330, CS2017-68, IE2017-83, 2017, pp. 55-58 (in Japanese).

[10] Endo, R., Kawai, Y., Sumiyoshi, H., Sano, M., "Scene-Text-Detection Method Robust against Orientation and Discontiguous Components of Characters", IEEE Conference on Computer Vision and Pattern Recognition 2017 Workshops, IEEE, 2017, pp. 941-949.

[11] Mishima, T., Ichiki, M., Hagiwara, A., Ito, H., Sato, S., Kobayashi, A., "Experimental Verification of Transcription Interface Using Speech Recognition", ITE Winter Convention, 12C-6, 2017 (in Japanese).