arXiv:2403.02938v1 [cs.CL] 5 Mar 2024

AIx Speed: Playback Speed Optimization Using Listening Comprehension of Speech Recognition Models

Kazuki Kawamura
The University of Tokyo, Tokyo, Japan
Sony CSL Kyoto, Kyoto, Japan
kwmr@acm.org

Jun Rekimoto
The University of Tokyo, Tokyo, Japan
Sony CSL Kyoto, Kyoto, Japan
rekimoto@acm.org

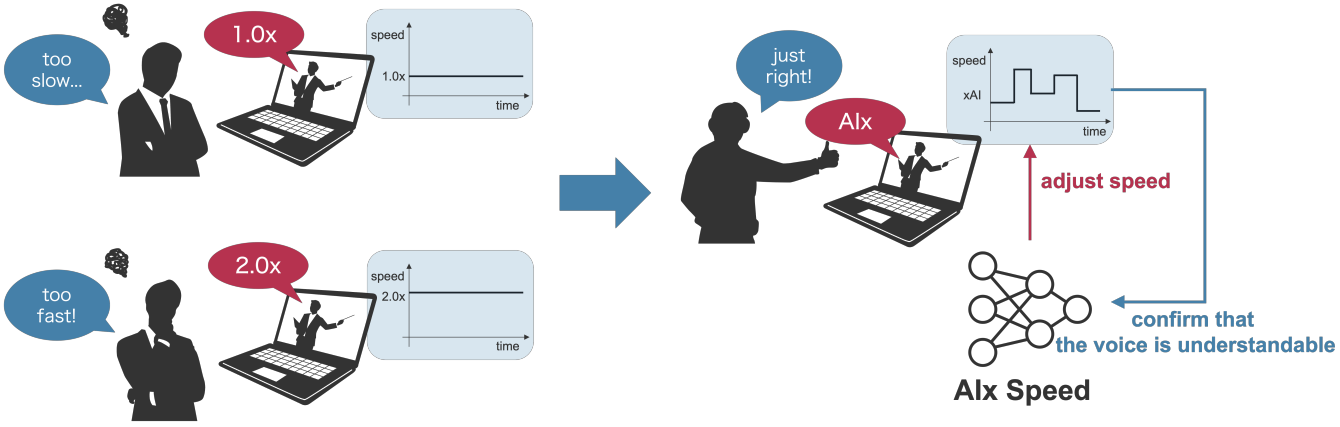


Figure 1: AIx Speed optimizes the playback speed of a video in units as small as phonemes. The system maximizes the speed of audio to the extent that humans can hear it. Users can watch videos at a comfortable speed without having to adjust the playback speed.

ABSTRACT

Since humans can listen to audio and watch videos at faster speeds than actually observed, we often listen to or watch these pieces of content at higher playback speeds to increase the time efficiency of content comprehension. To further utilize this capability, systems that automatically adjust the playback speed according to the user's condition and the type of content to assist in more efficient comprehension of time-series content have been developed. However, there is still room for these systems to further extend human speed-listening ability by generating speech with playback speed optimized for even finer time units and providing it to humans. In this study, we determine whether humans can hear the optimized speech and propose a system that automatically adjusts playback speed at units as small as phonemes while ensuring speech intelligibility. The system uses the speech recognizer score as a proxy for how well a human can hear a certain unit of speech and maximizes the speech playback speed to the extent that a human can hear. This method can be used to produce fast but intelligible speech. In the evaluation experiment, we compared the speech played back at a constant fast speed and the flexibly speed-up speech generated by the proposed method in a blind test and confirmed that the proposed method produced speech that was easier to listen to.

KEYWORDS

video information, playback speed, deep neural network, speech recognition, self-supervised learning

ACM Reference Format:
Kazuki Kawamura and Jun Rekimoto. 2023. AIx Speed: Playback Speed Optimization Using Listening Comprehension of Speech Recognition Models. In *Augmented Humans Conference (AHs '23), March 12–14, 2023, Glasgow, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3582700.3582722

1 INTRODUCTION

With the widespread use of video distribution services, people are increasingly watching videos for various purposes, including information gathering, learning, and entertainment. Humans are capable of understanding naturally observed phenomena at rates faster than the original. Therefore, when watching video or listening to audio, we often increase the playback speed of the content to understand more content in a shorter amount of time. Existing research has shown that when watching videos for learning, under certain conditions, differences in video playback speed do not affect learning effectiveness and may even improve performance [16, 21, 22]. It has also been reported that 30% to 80% of users prefer to watch dramas at high speed, although this varies from country to country [6].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AHs '23, March 12–14, 2023, Glasgow, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9984-5/23/03…$15.00
https://doi.org/10.1145/3582700.3582722

There are many advantages to listening to video and audio at high speeds. As a result, many methods have been proposed to automatically adjust the playback speed according to the structure of the content and the condition of the user, to further enhance the human ability to listen at high speeds. They are widely studied as video summarization [1] and audio summarization [29]. There are two basic strategies to these methods. The first is to scan the user's intentions and behavior and leave only the parts that need to be viewed or adjust the playback speed in proportion to the user's concentration level [13, 15]. The second is to speed up unnecessary parts of the content (i.e., parts that do not contain speech) or slow down parts that contain speech [9, 14, 26, 32]. However, these methods do not explicitly model whether the resulting speech at different playback speeds is intelligible to humans, so it is unclear whether a wide range of speech types can be made intelligible to users. In addition, these systems vary the playback speed for each large chunk of speech, which leaves room for adjustment regarding playback speed for smaller units of time.

Therefore, we propose AIx Speed, a system that adjusts audiovisual output speed while maintaining intelligibility by measuring speech intelligibility after playback speed is increased. As shown in Fig. 1, this system flexibly optimizes the playback speed in a video at the phoneme level. By utilizing the listening ability of a neural network–based speech recognition model, which is said to rival human performance [31], the system simultaneously maximizes the video playback speed and the speech recognition rate after changing the playback speed. In this paper, the validity of using speech recognizers as a proxy for evaluating human listening performance was tested through the correlation of changes in human and speech listening performance when speed is increased. We also whether speech where the playback speed is controlled at the phoneme level, as generated by the proposed method, or speech that is played at a constant speed and a high rate is easier for humans to listen to. The results showed that the utterances generated by the proposed method were easier for humans to understand. Furthermore, the experimental results confirmed that the speech of non-native speakers can be transformed into speech that is easier to understand for native speakers by speeding up the speech with AIx Speed. In summary, the proposed method not only supports the improvement of human speed–listening ability, but also improves the intelligibility of speech by generating speech with adjustable playback speeds that consider the balance between playback speed and speech intelligibility.

The contributions of this paper are summarized as follows.

- Demonstrates that speech recognizers can be a substitute for human listening performance assessment.
- Proposes a method to increase playback speed while maintaining speech intelligibility at the phoneme level.
- Improves the intelligibility of speech for non-native speakers by optimizing speech rate at the phoneme level.

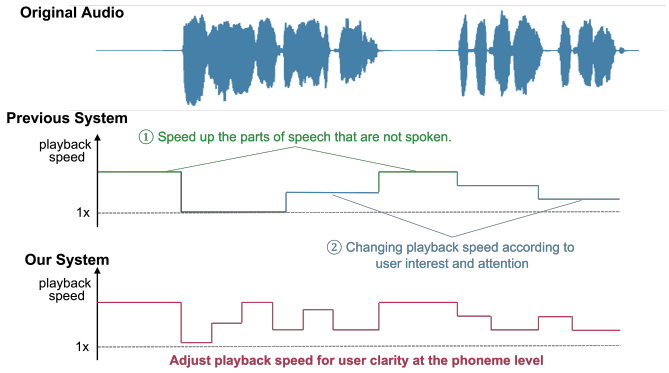
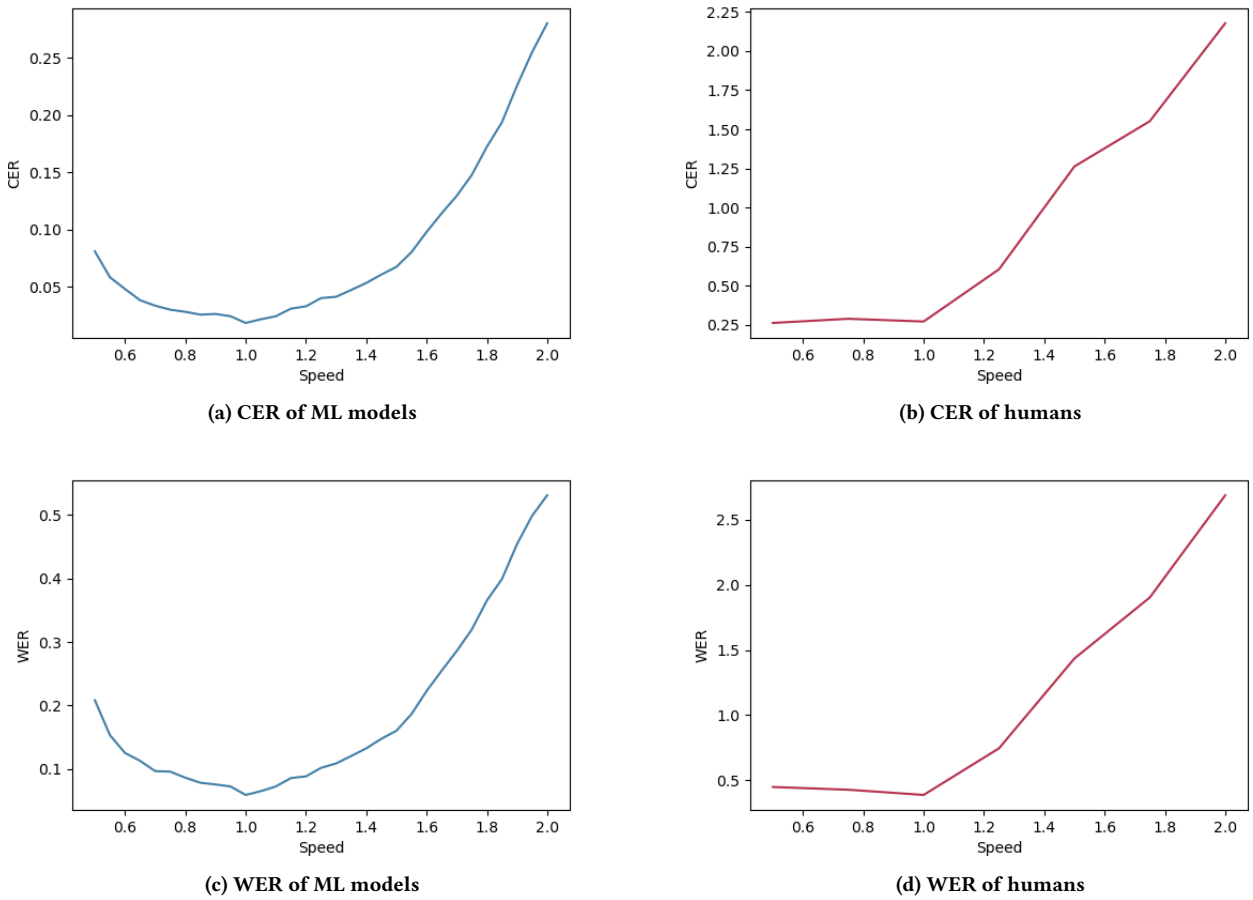


Figure 2: Differences between existing methods and ours. Existing methods speed up the playback speed of silent or unimportant parts of the audio source and slow down the playback speed of important parts of the audio source or when the user is not concentrating on them. Our method, on the other hand, adjusts the playback speed of the sound source at finer intervals according to the audibility of the speaker's speech in the sound source.

2 RELATED WORK

There are two main methods for adjusting video playback speed to improve the time efficiency of video viewing. The first is to remove unnecessary portions by focusing on the content, and the second is to retain the necessary portions based on user interaction. The former removes portions that do not have audio, which is accomplished using systems such as CinemaGazer [15], or portions of sports games that are not highlights of the game [13]. The latter has been studied extensively, especially in the field of human–computer interaction (HCI), and adjusts the playback speed based on the user's behavior. For example, SmartPlayer [4] learns the optimal playback speed based on a user's past viewing history. There are also technologies that allow a user to make a rough selection in advance of what AIx Speed considers important and then fast-forward the rest of the video [9, 14]. Others monitor the user's movements and adjust the playback speed according to the user's level of concentration [26] and comprehension [23, 32]. These technologies have the advantage of tracking the optimal playback speed for each user, but they cannot reflect important factors such as the intelligibility of the conversation in the video and its changes in the playback speed.

These methods do not explicitly model whether the resulting speech is intelligible to humans when the playback speed is varied. Therefore, speeding up the playback of various types of audio and video while making the audio understandable to the user is still an open problem. In addition, these systems vary the playback speed for large chunks of speech, which leaves room for adjustment regarding the playback speed for smaller units of time. In this respect, the proposed system can adjust the playback speed in finer phoneme units and can also generate speech that is easier for the user to understand. The differences between the proposed system and the existing systems are shown in Fig. 2.



(a) CER of ML models

(b) CER of humans

(c) WER of ML models

(d) WER of humans

Figure 3: Comparison of playback speed and listening comprehension in machine learning (ML)-based speech recognizers and humans. Transitions in human and speech recognition model transcription performance at speeds greater than 1x are correlated by a coefficient of 0.99.

3 PILOT SURVEY

The purpose of this study is to automate the maximization of speed to the extent that speech is understandable. To this end, we hypothesize that as the speed increases, the recognition performance of both humans and speech recognizers will decrease in the same way. If this hypothesis is correct, we can evaluate how well a human can hear when playback speed is increased using a speech recognizer instead of a human. Several attempts have been made to evaluate human hearing with speech recognizers in this way. For example, it has been shown that the results of human mean opinion score (MOS) listening tests correlate with the results of the speech recognition-based MOS estimation method introduced in [11], and in [7], the understanding and comprehension scores of a listener with simulated age-related hearing loss were highly correlated speech recognition-based system. A similar hypothesis has been used in speech learning support research to evaluate a learner's speech ability based on speech recognition performance [27]. In other words, if a speech recognizer can recognize speech, it judges that the person speaks well. However, the relationship between the machine learning model and the ability to understand human speech when the playback speed is varied, which is the focus of this study, has not been evaluated. Therefore, we first investigated whether this hypothesis is true.

This study compared human listening performance and speech recognition performance for speech at 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2.0x playback speeds. To measure human listening performance, speech data of English sentences were prepared at each playback speed, and the subjects were asked to transcribe the data. The target English sentences were selected from LibriSpeech [24], a large English speech corpus created for the development and evaluation of automatic speech recognition systems. This corpus contains over 1,000 hours of audiobook readings and transcriptions. The participants were 140 English speakers who lived in the United States and had graduated from a US high school. All participants were recruited from the Amazon Mechanical Turk and were compensated for their time. Each subject was given 15 English sentences that had been randomly sped up by 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, or 2.0x and asked to transcribe them. The speech recognizer

transcription data was collected by inputting 15 English sentences at each playback speed into a Wav2Vec2-based speech recognition model, similar to the human performance evaluation [2].

Figure 3 shows a graph of the change in listening performance of the human and machine learning models when the playback speed was changed. In these figures, the horizontal axis is the playback speed, and the vertical axis is the recognition performance. Recognition performance was evaluated using character error rate (CER) and word error rate (WER), which are commonly used in speech recognition (the lower these values are, the better). The WER was calculated as follows:

$$\frac{(\#\text{ of inserted words} + \#\text{ of replaced words} + \#\text{ of deleted words})}{(\#\text{ of correct words})},$$

and CER was calculated as

$$\frac{(\#\text{ of inserted characters} + \#\text{ of replaced characters} + \#\text{ of deleted characters})}{(\#\text{ of correct characters})}.$$

For both the human and machine learning models, listening performance decreased as playback speed increased from 1.0x. For playback speeds greater than 1.0x, the correlation coefficient between the change in listening performance for the human and machine learning models was 0.9977. On the other hand, when the playback speed was slowed down from 1.0x, the machine learning model showed a decrease in recognition performance, but the decline barely observable for humans. In particular, while the performance of WER decreased slowly, the performance of CER showed almost no decrease. This indicates that human recognition performance on a character-by-character basis does not drop nearly as much when listening to slowed speech. Thus, when the playback speed increased, the recognition performance of the human and machine learning models decreased similarly, but they exhibited different behavior when the playback speed was decreased. However, since this study focuses on increasing the playback speed of speech, the difference in behavior when the playback speed is slower than 1.0x has no effect. Therefore, by taking advantage of the fact that listening performance decreases as playback speed is increased for humans and speech recognition models alike, we replaced human listening performance with speech recognition performance to develop the desired system.

4 AIX SPEED

AIx Speed increases the speed as much as possible, as long as the user can understand it. This system allows users to watch videos in a time-efficient manner without having to adjust the playback speed for each video. In addition, the system can automatically improve intelligibility by adjusting the speech speed to accommodate non-native speakers who are not proficient in the target language. The working process of AIx Speed is illustrated in Fig. 4. The system first extracts the human voice from the target video. Next, it splits this voice into specified equal intervals. Next, using each segmented voice as input, the system calculates the optimal playback speed for each segment voice, taking into account the characteristics of the voice as a whole. Finally, the system changes each voice to the specified playback speed and combines them into a single voice. At the same time, the combined voice is recognized by speech recognition to confirm that the resulting single voice is understandable. This system consists of two mechanisms, as shown in Fig. 5. One is a playback speed adjuster (left), and the other is a speech recognizer (right). The former is used to maximize the

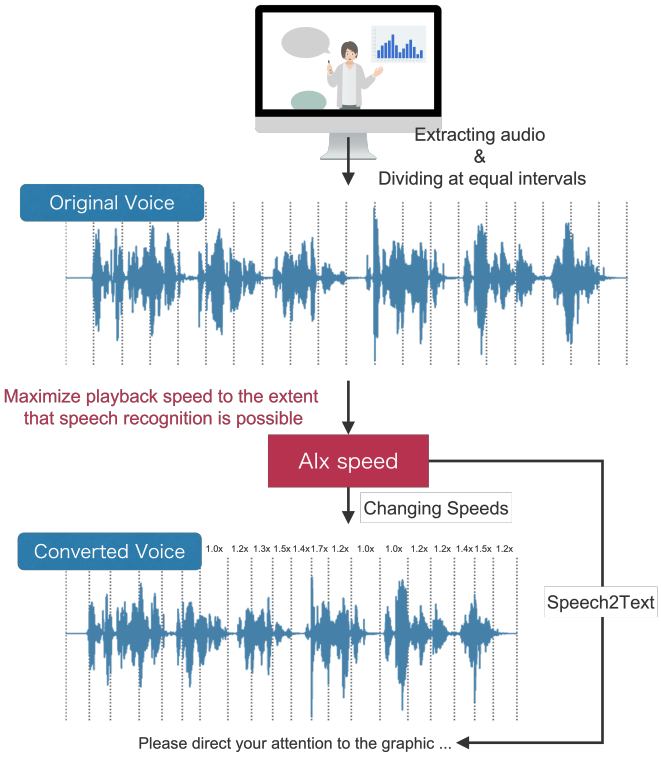


Figure 4: The work process of AIx Speed: ① Extract human voices from the target video. ② Divide the voices into specified equal intervals. ③ Calculate the optimal playback speed for each divided voice. ④ Change each voice to the fixed playback speed and synthesize it into one voice. ⑤ Perform speech recognition on the synthesized voice to confirm that the resulting voice is understandable.

playback speed of the input speech, while the latter is used to evaluate how understandable the input speech is. By training these two models simultaneously, it is possible to generate speech that plays back as fast as possible within the comprehension range. The following subsections describe these two key features.

4.1 Playback speed adjuster

In this study, we use Wav2Vec2, a self-supervised neural network designed for speech signal processing systems, to optimize the playback speed. The playback speed controller is divided into a feature extractor layer and a linear layer, as shown in Fig. 5 (left). They are trained by pre-training through self-supervised representation learning on unlabeled speech data and regression learning, which outputs the playback speed based on the features after the representation learning. The pre-training method for the feature extractor layer is similar to masked language modeling, as exemplified by bidirectional encoder representations from transformers (BERT) [5] in natural language processing, where a portion of the input is masked and the corresponding utterance features are estimated from the remaining input. In this way, the model can learn

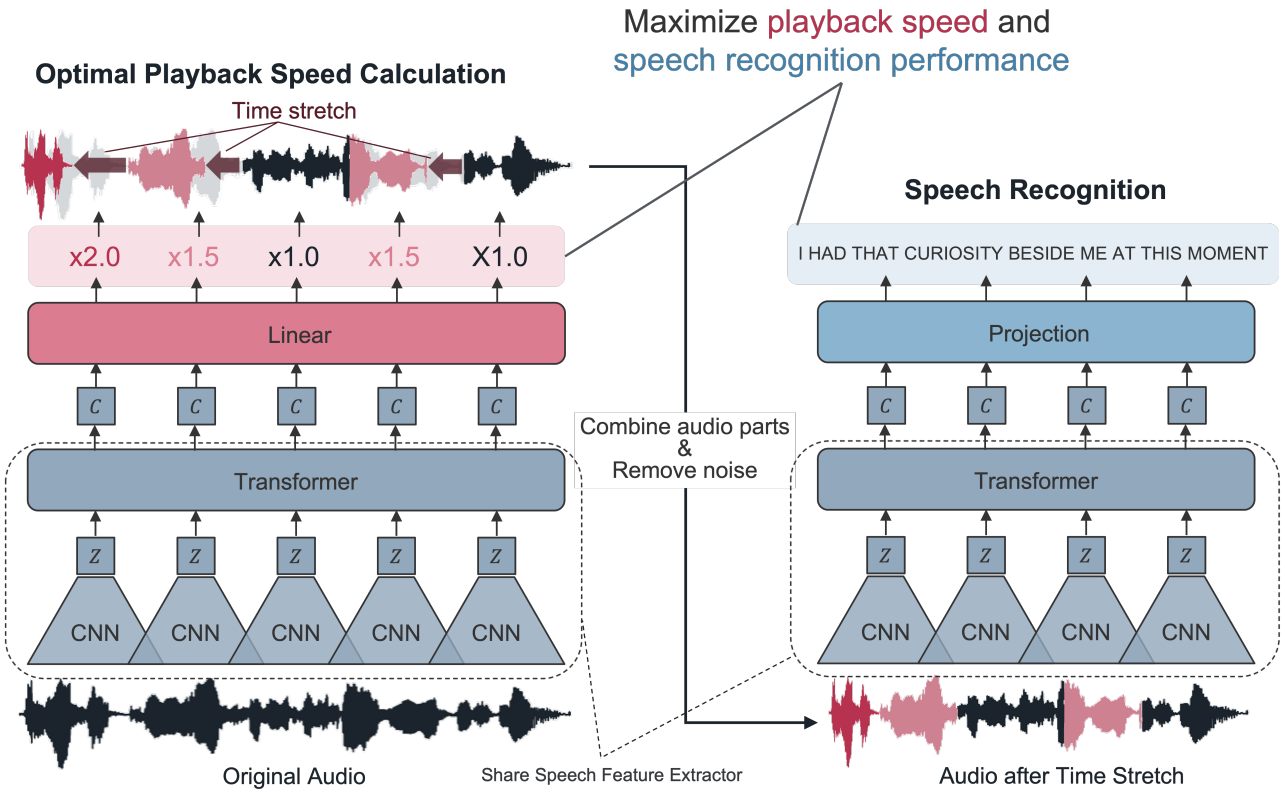


Figure 5: AIx Speed architecture: Our system simultaneously optimizes the playback speed regulator (left) and the speech recognizer (right). By doing so, we can maximize the playback speed to the extent that the model can recognize.

good-quality features of the target language to which the rate of utterance should be adapted. Typically, these self-supervised learners are used to tackle tasks such as speech recognition and speaker identification by pre-training and then fine-tuning with a small amount of label data. For example, in speech recognition, we have added a projection layer and a connectionist temporal classification (CTC) layer [8] to the output of self-supervised learners, such as Wav2Vec2 and HuBERT [10], to enable transcription from speech waveforms. Similar to these methods, we pre-train on unlabeled speech data and then connect and train a linear layer that outputs rates. In general speech processing tasks, such as speech recognition [18] and speaker identification [3, 12], the difference from the class label is minimized as an error function. On the other hand, the goal of the playback speed adjuster is to maximize the speed. Therefore, we designed the following function whose value decreases as the speed increases.

$$\mathrm{Loss}_{\mathrm{speed}} = \left(\frac{1}{10}\right)^{\frac{1}{S}\sum_{s=1}^{S} r_s},$$

where r_t is the playback speed for each segment obtained using Wav2Vec2 and a linear layer with audio X as input as follows:

$$C = c_{[1:T]} = \mathrm{Wav2Vec2}\,(X)\,,$$
$$R = r_{[1:S]} = \mathrm{Linear}\,(C)\,.$$

4.2 Speech recognizer

The speech recognizer transcribes the speech converted to the playback speed obtained by the playback speed adjuster (Fig. 5 (right)). When speech parts with different playback speeds are combined, noise is generated in the speech data and the sound quality is degraded. We use voice separation technology [19] to extract only the speaker's voice and reduce the effect of noise. The resulting speech is then fed into a speech recognizer for speech recognition. The speech recognition process consists of the extraction of acoustic features from the speech waveform, estimation of the classes of acoustic features for each frame, and generation of hypotheses from the sequence of class probabilities. Since the speech recognizer can partially share the neural network with the playback speed adjuster, the overall network size can be reduced. As shown in Fig. 5, the dotted speech feature extractor is shared. The speech features obtained by this mechanism are used as input to generate text in the projection layer. In this process, the speech recognizer is trained to minimize CTC loss, as in normal speech recognition. In summary, the entire model is trained to minimize the following error function, which is a combination of this error function and the error function of the playback speed adjuster.

$$\mathrm{Loss} = \mathrm{Loss}_{\mathrm{speed}} + \lambda \mathrm{Loss}_{\mathrm{ctc}}.$$

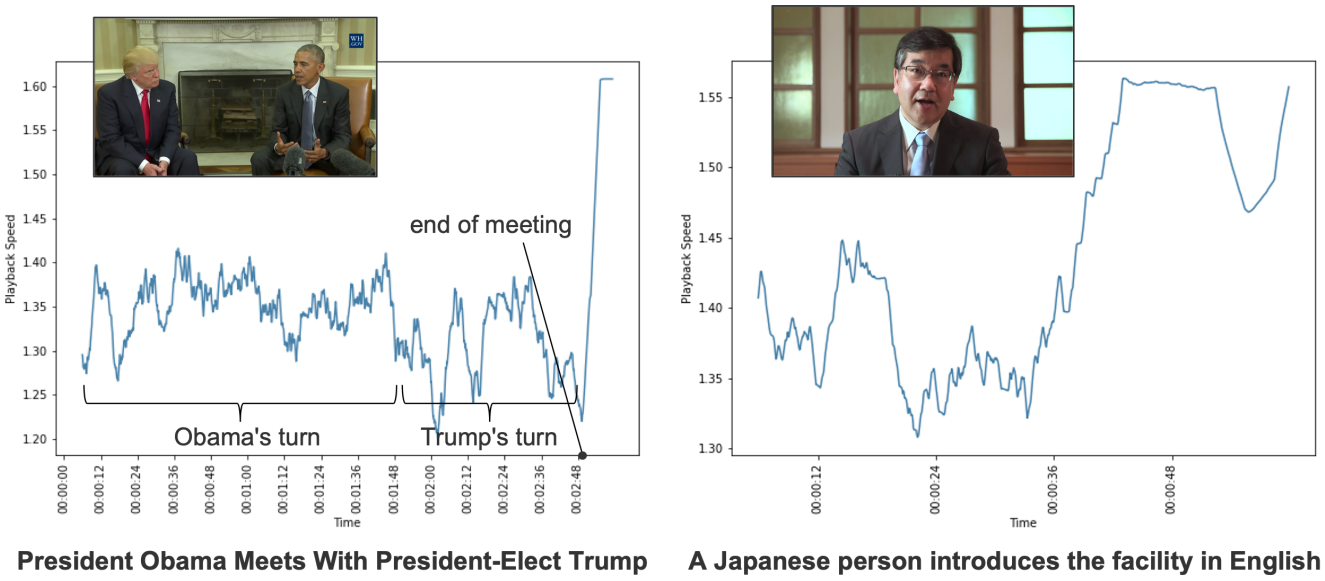


Figure 6: Application example. This graph shows how the playback speed changed when AIx Speed was applied to two videos on YouTube. As you can see, the playback speed changes flexibly according to the speaker's speech.

Here, λ is a hyperparameter that adjusts the importance of the playback speed calculation and speech recognition.

5 PROTOTYPE

As a prototype of AIx Speed, an application that optimizes audio playback speed using English as the target language has been implemented. This section describes the implementation and usage of the application.

5.1 Implementation

Wav2Vec2 was used for the prototype's shared utterance learning model (the utterance learning part shared by the speech recognizer and the playback speed adjuster). For pre-training, LibriSpeech [24] was used as the dataset, train-clean-360 as the training data, and dev-clean as the validation data. The dataset consist of clean speech data in LibriSpeech and were partitioned using the data partitioning method proposed in LibriSpeech for training/validation. The pre-training did not require the corresponding transcribed text, only the speech data. We then trained a speech recognizer and a playback speed adjuster using two sets of speech data, including the transcribed text. One was LibriSpeech's train-clean-100, and the other was the English speech database read by Japanese students (UME-ERJ) [20]. The latter is a dataset of English spoken by non-native Japanese speakers, with 202 speakers (100 males and 102 females) reading simple English sentences. The playback speed adjuster and the speech recognizer were trained separately. First, the speech recognizer was trained using LibriSpeech and UME-ERJ, and then the playback speed adjuster was trained using the same data with fixed weights for the speech recognizer. All speech files used for training were normalized to a sampling frequency of 16 kHz. The Wav2Vec2 used the initial parameters implemented in PyTorch [25], and the final layers of both the playback rate adjuster and the speech recognizer were 768-dimensional linear layers. The hyperparameter λ of the error function was set to 10^{-7}, and training was performed for 5 epochs with a batch size of 32 using the AdamW [17] optimization algorithm.

5.2 Usage of the application

Examples of using AIx Speed are shown in Fig. 6. This is an example of the prototype applied to a video uploaded to YouTube. The horizontal axis is the playback time, and the vertical axis represents the playback speed output by the model for each playback time. The first is an example of speeding up a dialogue, movie, or lecture. Two speakers appear in the video, and the optimal playback speed can be set for each speaker. Of particular interest is that the two speakers in the video speak at different speeds, so the average playback speed for the two speakers is different. It can also be seen that the playback speed increases drastically from the moment when the dialog between the two speakers ends and there is no more speech from the person. Although we did not intend to design this feature, we can see that our system, like conventional playback speed controllers, can speed up the playback speed in the parts where there is no speech. The second example is speeding up the speech of non-native speakers to make it easier to understand. Since the speech of non-native speakers is often slower than that of native speakers, moderately speeding up the speech makes it easier to understand. The change in playback speed shows that the overall playback speed is faster than the speech between native speakers in the first video. This indicates that the model can speed up the speech more because the non-native speakers' speech is slower than that of the native speakers.

Table 1: Performance comparison of models

(a) LibriSpeech

Model	Avg. Speed	CER	WER
Wav2Vec2 (1.00x)	1.00	**4.38**	**12.57**
Wav2Vec2 (1.30x)	1.30	5.61	14.58
Wav2Vec2 (1.50x)	1.50	6.83	17.19
AIx Speed	1.30	5.21	12.96

(b) UME-ERJ

Model	Avg. Speed	CER	WER
Wav2Vec2 (1.00x)	1.00	26.74	55.02
Wav2Vec2 (1.29x)	1.29	33.71	63.90
Wav2Vec2 (1.50x)	1.50	36.93	66.51
AIx Speed	1.29	**26.45**	**53.13**

6 EVALUATION

6.1 Technical evaluation

To demonstrate that the proposed method can optimize playback speed while maintaining content understanding, we compared the CER and WER values at the AIx Speed–modified speech playback speed to those at a constant playback speed. We compared the CER and WER with the speech playback speed modified by the AIx Speed to the CER and WER when the speech was simply played at a constant speed. The speeds of the comparison targets were 1.0x, 1.5x and the average speed times the playback speed of AIx Speed. A standard Wav2Vec2 based speech recognition model, which was the speech recognizer used in our method, was used to compute the CER and WER for comparison. The performance of the models is shown in Table. 1 (a) and 1 (b) for LibriSpeech and UME-ERJ, respectively. AIx Speed produces speech 1.30 times faster on average for LibriSpeech and 1.29 times faster on average for UME-ERJ. Both results show that the playback speed optimized by AIx Speed has lower values for both CER and WER than the average constant speech speed at that playback speed. From these results, it can be said that the proposed model maximizes the playback speed while guaranteeing the recognition performance. In addition, for UME-ERJ, the speech generated at AIx Speed shows better recognition performance in terms of WER than at 1.0x playback speed. Therefore, it is also suggested that the proposed method can be used to convert the speech of non-native speakers into more understandable speech.

6.2 User evaluation

User experiments were conducted to confirm that the generated speech was understandable. The quality of the speech generated by the proposed method was compared with that of the speech played at a constant speed, at the average playback speed of the speech. The quality was evaluated using the mean opinion score, which is commonly used in speech synthesis research [28, 30]. This measure rates speech quality on a five-point scale from 1 (poor) to 5 (excellent). Participants were 50 US residents who used English on a daily basis. 20 sentences were extracted from each of the LibriSpeech and UME-ERJ datasets, and half were converted to speech with the proposed speed-up, while the other half were converted to speech with a constant speed-up. Participants were given a total of 40 sentences of audio and assisted to rate the quality of the audio. Figure 7 shows the quality of the generated speech at baseline and AIx Speed, LThe quality of LibreSpeech and UME-ERJ were 0.5 and 0.8 points higher at speeds generated by the proposed method, respectively.

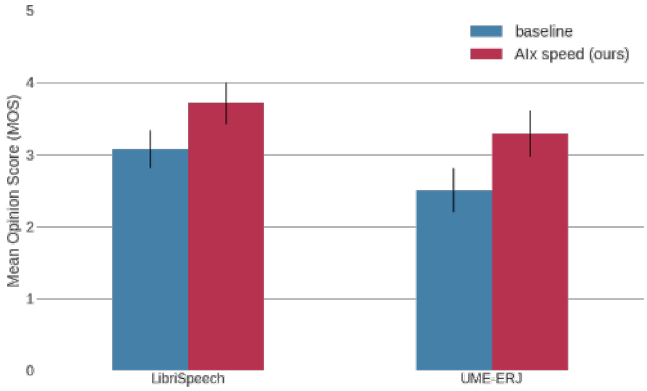


Figure 7: Voice quality at baseline (constant playback speed increase throughout) and AIx Speed (flexible playback speed increase)

7 DISCUSSION

7.1 Evaluation results

The technical evaluation shows that the proposed method can produce speech that is easier to understand than that produced by simply increasing the playback speed in terms of speech recognition performance. The user evaluation also shows that the proposed method can produce speech that is easier to understand for real users. These results show that the proposed method can produce speech with a high playback speed within a range that is easy for users to understand. This allows users to watch videos at a reasonable speed without having to adjust the playback speed for each video. However, the improvement in MOS values by using the proposed method is by no means sufficient. In the current model, the average conversion to a faster playback speed is about 1.3 times, but it is a future task to investigate whether it is possible to make this even faster. In fact, many video playback services implement 1.5x and 2.0x playback speeds, and some people watch dramas and lectures at such speeds. Therefore, we expect that it will be possible to convert up to this speed and make the audio easy to understand. In addition, since each user has a different preferred playback speed, personalizing the model so that it plays at the optimal playback speed for users is also a future issue.

7.2 Listening comprehension and speech recognition

As discussed in the preliminary research chapter, it can be seen that there is a relationship between speech recognition performance and

human transcription ability. Thus, we expect to build automated systems for various tasks and evaluations by replacing human speech comprehension ability with machine learning models, as in this system. On the other hand, speech recognition performance based on playback speed does not perfectly match human speech comprehension. In other words, as the playback speed increases, the dictation performance decreases in both cases, but the performance values are not exactly the same. Thus, we anticipate that by training speech recognition models to match these relationships as closely as possible, it will be possible to use them more generally as alternatives to humans. Distillation, a technique that learns to approximate an output that matches existing results, will be the technical key.

7.3 Adjustment of non-native speakers' speech

Several suggestions can be made as to how increasing the playback speed by the proposed system improves the intelligibility of speech for non-native speakers. One of them is that when non-native speakers read English manuscripts, they may find it easier to understand if they speak naturally (slower from a native speaker's point of view) and then artificially speed up their speech, rather than forcing them to speak quickly like a native speaker. In fact, this study also began with the realization that it is easier to listen to a video of a non-native speaker speaking his or her native language when it is played at a high speed.

8 CONCLUSION

This paper presents a system that applies a speech recognition model to automatically and flexibly adjust the playback speed of video and audio within the range of human comprehension. By using this system, users can consume audiovisual content at optimal speeds without having to manually adjust the playback speed. Experiments have also confirmed that the system makes it easier for users to understand the speech of non-native speakers. In the future, we expect the system to be used in a variety of applications, such as video distribution services and language learning tools.

ACKNOWLEDGMENTS

This work was supported by JST Moonshot R&D Grant Number JPMJMS2012, JST CREST Grant Number JPMJCR17A3, and The Univesity of Tokyo Human Augmentation Research Initiative.

REFERENCES

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proc. IEEE* 109, 11 (2021), 1838–1863.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*.

[3] Zx Bai and Xiao-Lei Zhang. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks* (2021).

[4] Kai-Yin Cheng, Sheng-Jie Luo, Bing-Yu Chen, and Hao-Hua Chu. 2009. SmartPlayer: User-Centric Video Fast-Forwarding. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[6] Songshuang Duan and Xiaoqian Chen. 2019. Why College Students Watch Streaming Drama at Higher Playback Speed: The Uses and Gratifications Perspective. In *International Joint Conference on Information, Media and Engineering*.

[7] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, Julien Tardieu, Cynthia Magnen, Pascal Gaillard, Xavier Aumont, and Christian Füllgrabe. 2017. Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research* 60, 9 (2017), 2394–2405.

[8] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Nets. In *ICML '06: Proceedings of the International Conference on Machine Learning*.

[9] Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proc. ACM Conference on Human Factors in Computing Systems*.

[10] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. (2021).

[11] Wenyu Jiang and H. Schulzrinne. 2002. Speech recognition performance as an effective perceived quality predictor. In *IEEE 2002 Tenth IEEE International Workshop on Quality of Service (Cat. No.02EX564)*. 269–275.

[12] Muhammad Mohsin Kabir, Muhammad Firoz Mridha, Jungpil Shin, Israt Jahan, and Abu Quwsar Ohi. 2021. A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. *IEEE Access* 9 (2021), 79236–79263.

[13] Shunya Kawamura, Tsukasa Fukusato, Tatsunori Hirai, and Shigeo Morishima. 2014. Efficient Video Viewing System for Racquet Sports with Automatic Summarization Focusing on Rally Scenes. In *ACM SIGGRAPH 2014 Posters*.

[14] Seita Kayukawa, Keita Higuchi, Ryo Yonetani, Masanori Nakamura, Yoichi Sato, and Shigeo Morishima. 2018. Dynamic Object Scanning: Object-Based Elastic Timeline for Quickly Browsing First-Person Videos. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*.

[15] Kazutaka Kurihara. 2011. CinemaGazer: A System for Watching Video at Very High Speed. In *Proc. of the Workshop on Advanced Visual Interfaces AVI*.

[16] David Lang, Guanling Chen, Kathy Mirzaei, and Andreas Paepcke. 2020. Is Faster Better? A Study of Video Playback Speed. In *Proc. of the Tenth International Conference on Learning Analytics & Knowledge*.

[17] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[18] Mishaim Malik, Muhammad Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: A survey. *Multimedia Tools and Applications* (2021).

[19] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. Librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*. Citeseer.

[20] Nobuaki Minematsu, Yoshihiro Tomiyama, Kei Yoshimoto, Katsumasa Shimizu, Seiichi Nakagawa, Masatake Dantsuji, and Shozo Makino. 2002. English Speech Database Read by Japanese Learners for CALL System Development.. In *LREC*. Citeseer.

[21] Dillon H. Murphy, Kara M. Hoover, Karina Agadzhanyan, Jesse C. Kuehn, and Alan D. Castel. 2022. Learning in double time: The effect of lecture video speed on immediate and delayed comprehension. *Applied Cognitive Psychology* 36, 1 (2022), 69–82.

[22] Toru Nagahama and Yusuke Morita. 2017. Effect Analysis of Playback Speed for Lecture Video Including Instructor Images. *International Journal for Educational Media and Technology* (2017).

[23] Naoto Nishida, Hinako Nozaki, and Buntarou Shizuki. 2022. Laugh at Your Own Pace: Basic Performance Evaluation of Language Learning Assistance by Adjustment of Video Playback Speeds Based on Laughter Detection. In *Proc. of the Ninth ACM Conference on Learning @ Scale*.

[24] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*.

[26] Sunghyun Song, Jeong-ki Hong, Ian Oakley, Jun Dong Cho, and Andrea Bianchi. 2015. Automatically Adjusting the Speed of E-Learning Videos. In *Proc. of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*.

[27] Cristian Tejedor-García, Valentín Cardeñoso-Payo, and David Escudero-Mancebo. 2021. Automatic Speech Recognition (ASR) Systems Applied to Pronunciation Assessment of L2 Spanish for Japanese Speakers. *Applied Sciences* 11, 15 (2021).

[28] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray

Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *Arxiv*.

[29] Aneesh Vartakavi, Amanmeet Garg, and Zafar Rafii. 2021. Audio Summarization for Podcasts. In *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 431–435.

[30] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *INTERSPEECH*. 4006–4010.

[31] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving Human Parity in Conversational Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[32] Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. 2020. WithYou: Automated Adaptive Speech Tutoring With Context-Dependent Speech Recognition. In *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems*.