A System for Estimating the Importance of Speech Based on Acoustic Features

Jiating Liu[1(✉)] and Sumio Ohno[2]

[1] Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology, Hachioji, Tokyo 192-0982, Japan
g212303237@edu.teu.ac.jp

[2] School of Computer Science, Tokyo University of Technology, Hachioji, Tokyo 192-0982, Japan
ohno@stf.teu.ac.jp

Abstract. With the development of AI technology, the accuracy of speech recognition and the range of its use are advancing. With AI-based speech processing, many services such as automatic speech transcription, speech recognition, and speech summarization are now available. In this paper, a method is proposed for determining whether each utterance is important or not based on acoustic information of the utterances. As acoustic features, statistical measures of various acoustic features for each utterance are used. To determine the importance of the utterance, the importance of the transcribed text is labeled using the LLM chat system and trained as a supervised input. In this experiment, TED video speeches on YouTube are used as speech materials. A machine learning model with a random forest classifier is used to determine the importance. As a result, a model that can classify the importance for the training data is obtained. It is found that statistical measures of acoustic features related to the fundamental frequency (F0) of the utterance are frequently used as important features for classification. However, when evaluated on test data, it is found that sufficient accuracy is not achieved, and further examination is necessary.

Keywords: import decision · acoustic features · linguistic features · random forest classifier · labeling

1 Introduction

Automatic meeting minutes tools play an important role in shortening time and reducing burden for businesses. Various services and systems, such as User local audio recording system [1], YOMEL [2], and ACES Meet [3], can be utilized to obtain meeting minutes and summarize the importance of the meeting. The majority of existing services are developed based on linguistic information obtained by the use of speech recognition and AI technology.

In recent years, with the development of AI technology, the accuracy of speech recognition has improved, and it is reported that 80% to 90% [4] of daily conversations can be

B. Xin et al. (Eds.): IWACIII 2023, CCIS 1932, pp. 11–17, 2024.
https://doi.org/10.1007/978-981-99-7593-8_2

transcribed. Acoustic features play an important role in speech information processing. However, statistical measures of various acoustic features for each utterance are used. So a system, based on linguistic information without using acoustic features, utilizing machine learning to analyze acoustic features, can be developed to determine the importance of speech and improve work efficiency. In this paper, we examine a system that distinguishes important utterances based on their acoustic features rather than linguistic information.

The results of this study will be instrumental in enhancing the efficiency of work related to meeting minutes by recording audio information and summarizing the key content of the audio recordings.

The remainder of this paper is organized as follows: Sect. 2 presents an overview of the related works. In Sect. 3 are the explanations of the methodology and preliminary experiment. Finally, we offer the conclusions and future work of this study in Sect. 4.

2 Related Works

In this session, the authors will introduce the related research on the issues of nonverbal communication and the issues on estimation of utterance impression in lecture speech base on acoustic features.

2.1 Acoustic and Linguistic Information Based Chinese Prosodic Boundary Labelling

In Tao's research, a rule-learning approach is proposed for labeling Chinese prosodic boundaries, which are classified into four levels [5]. Acoustic and linguistic features related to prosodic boundaries were extracted from a corpus and used to establish an example database. Comparative experiments were conducted to select the most effective features. The results show that the selected features efficiently characterize boundary features and that the rule-learning approach achieves better prediction accuracy than rule and RNN-based methods while retaining simplicity and understandability.

In contrast to Tao's research, the LLM chat system was utilized to label the importance of each utterance in our research. This approach was found to be simpler and more accurate and objective, as it provides reasons for determination.

2.2 Estimation of Utterance Impression in Lecture Speech Based on Acoustic Features

Tanaka's research estimated speech impressions based on acoustic features [6]. He analyzed the relationship between impressions from uttered speech and acoustic features based on statistical methods such as multiple regression analysis and decision tree. As a result, multiple regression analysis was found to be the most suitable for estimating utterance intentions. In addition, regarding the impression of self-confidence, it was clarified that regardless of the speaker the fluctuation of the mean value of the fundamental frequency within an utterance has greatly contributes to the transmission of the impression.

In Tanaka's research, statistical methods were employed to estimate utterance impressions through the analysis of acoustic features, without the involvement of machine learning. In our study, a new method is proposed to determine whether each utterance is important or not the based on acoustic information of the utterances and Random Forest Classifier was utilized to classify the importance of speech.

3 Methodology and Preliminary Experiment

3.1 Methodology

Features that represent the meaning of speech signals can be extracted from input audio. These features, such as acoustic and linguistic features, can be used to analyze and understand the content of the speech signals, allowing for more accurate and efficient processing of audio information.[6].

In this paper, a new method is proposed for determining whether each utterance is important or not based on acoustic information of the utterances. And a new approach is used to label the linguistic features as supervised data by using LLM chat system,. A machine learning model with a random forest classifier is used to determine the importance (Fig. 1).

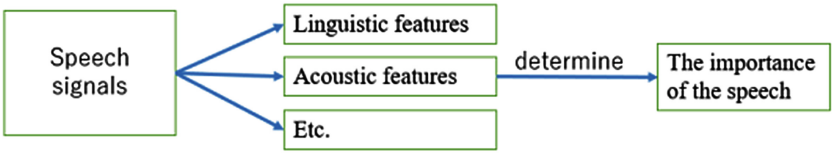

Fig. 1. Proposed method diagram

3.2 System Design

Figure 2 illustrates the flow of the learning process and usage of the proposed system for determining the importance and non-importance for each utterance in a lecture speech. Audio waveform was obtained from YouTube videos. Python's Whisper module was utilized to segment each utterance and obtain start and end time points information. The new method of labelling uses LLM chat system. The LLM chat system, Perplexity AI, was used to determine the importance of each utterance. Importance labels from a linguistic perspective for each utterance were used as supervised data for machine learning. Analyzing audio waveform by using openSMILE, statistic information on acoustic features are obtained, and used as input data for machine learning. For estimating, test waveform segmented into utterance are analyzed by openSMILE and predicted by a model. As a result, it is possible to obtain the estimated results of importance and the which acoustic features play an important role in classification of importance.

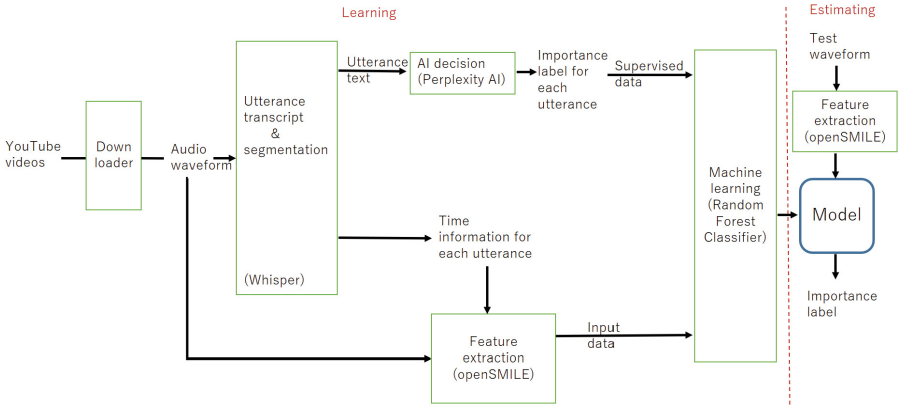

Fig. 2. Importance estimating system diagram

3.3 Data Processing

To ensure experimental accuracy, clear voices were selected for this research. So YouTube videos from Ted × Talks [7] were targeted. Takashi Yamada's speech was chosen due to the need for a voice with a lot of variations for analyzing acoustic features.

There are two methods to segment the targeted, one is module whisper of Python, the other is use the subtitle of YouTube videos. Because of the noise of the original videos, sometimes the subtitle can't be obtained. So module whisper is decided to segment the videos.

The target video, which is approximately 20 min long, was segmented into 165 utterances. Train and test data were split in an 8:2 ratio for application to the Random Forest Classifier. To ensure accuracy and objectivity, the LLM chat system was utilized to label the importance of each utterance.

There are chat AI systems powered by LLM such as ChatGPT, Bing ChatAI, Bard, and Perplexity AI. Perplexity AI was chosen due to its ability to not only judge the importance of each text but also provide the reasoning behind the labeling. The prompt was set to「以下はある講演の発言録です。IDと発言内容が並んでいます。講演の中で重要と考えられる発話とそうでない発話に分類したいです。重要な発話のIDだけを列挙してください。」(Below is a record of utterances from a lecture. The ID and content of the utterances are listed. We would like to classify the utterances into those that are considered important in the lecture and those that are not. Please list only the IDs of the important utterances). The response was shown in Fig. 3. Sentences belonging to the important ID were annotated with an importance of "1", while others were annotated with "0".

3.4 Preliminary Experiment

Experimental Process. Takashi Yamada's speech on YouTube was converted into a wav file and transcribed into text based on Python's Whisper module. The LLM chat system, Perplexity AI, was utilized to determine the importance and determine importance labels for each utterance.

```
PERPLEXITY
The important speech IDs are:
8, where the speaker introduces the topic of his speech
9-11, where the speaker talks about the meaning of "思うは招く" and the
importance of persistence
13-14, where the speaker expresses his desire to find like-minded people
and work together
23-26, where the speaker talks about his company's achievements in space
technology
```

Fig. 3. The label decision by using Perplexity AI Preliminary experiment

The audio was split based on the time of each utterance. Importance labels annotated by LLM for each utterance were used as supervised data for machine learning. Audio waveform was input into acoustic feature extraction based on openSMILE. The acoustic features were then input into machine learning using a Random Forest Classifier to train a model. For testing, the acoustic features of audio waveform were analyzed based on openSMILE and predicted by the model, with the importance label being output.

In the machine learning component, grid search with cross-verification was executed to obtain the most suitable hyper parameters. Hyper parameters were set as [class_weight = balanced, criterion = entropy, max_depth = 4, n_estimators = 12, random_state = 90].

Experimental Result. The Random Forest is a powerful tool for classification problems, but it can require effort to understand the predictions and their context. To facilitate interpretation of the results, the performance of the Random Forest Classifier was evaluated using confusion matrices based on Scikit-Learn.

As shown in Table 1, the training data consists of 132 observations. The values of true-negative and true-positive account for 122, or 92% of the total, indicating that the data is well-trained. However, as shown in Table 2, the test data consists of 33 observations, with true-negative and true-positive values accounting for only 16, or 48% of the total, suggesting that the test data is not well-predicted.

Overall, it was found that the training results were efficient, while the test results were not in this experiment.

Table 1. Result of training

predicted label True label	Unimportant	Important
Unimportant	73	4
Important	7	48

To determine which features contributed to the classification, top 5 in the variable importance of the model was also shown in Fig. 3. This figure revealed that features related to F0 have a significant impact on decision of speech importance. From loudness_sma3_pctlrange0–2, it can also be inferred that when a person speaks about something important, their voice is louder (Fig. 4).

Table 2. Result of test

predicted label True label	Unimportant	Important
Unimportant	13	10
Important	7	3

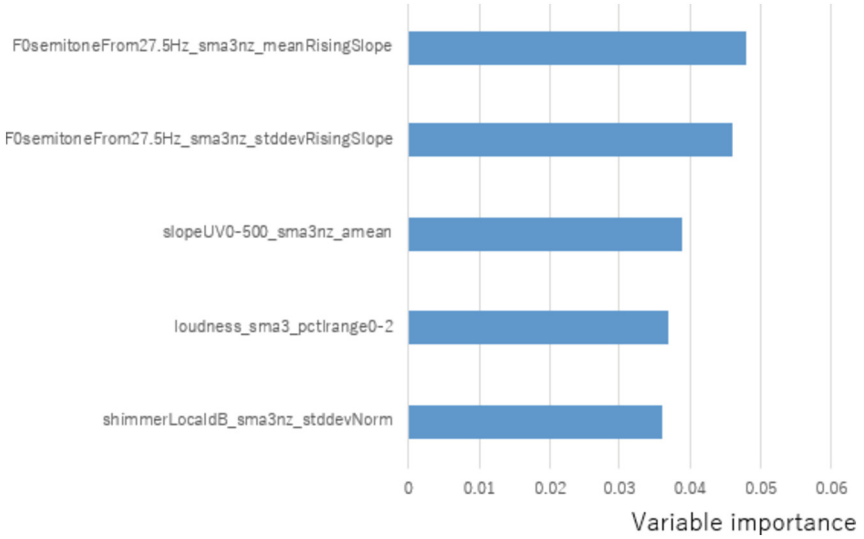

Fig. 4. Acoustic features contributing to classification

4 Conclusion

This paper presents a system for estimating the importance of speech based on acoustic features, designed to assist users in extracting important content from a speech. In this system, there are two main components: data processing and machine learning. Specifically, the LLM chat system was utilized to determine the importance of transcribed text, and audio was classified using the Random Forest Classifier. The results of this study indicate that classifying the importance of a speech using the Random Forest is effective in the training part, but accuracy is not good in the testing part in this trial. Additionally, it was found that acoustic features related to F0 and loudness have a significant impact on speech importance.

In future work, three points for improvement have been identified. Firstly, the linguistic data processing is not rigorous, so a more reliable LLM chat system will be chosen for tagging the importance of each utterance. Secondly, it is necessary to devise the use of acoustic features. Finally, as machine learning is utilized, it will be necessary to add more data to train the model.

References

1. User Local. https://www.userlocal.jp/products/. Accessed 30 June 2023
2. YOMEL. https://ai.yomel.co/gijiroku. Accessed 30 June 2023
3. ACES Meet. https://meet.acesinc.co.jp/. Accessed 30 June 2023
4. selection of automatic AI minutes systems. https://www.aspicjapan.org/asu/article/1742. Accessed 30 June 2023
5. Tao, J.: Acoustic and linguistic information based chinese prosodic boundary labelling. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 489–496. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30120-2_62
6. Tanaka, Y., Ohno, S.: Estimation of utterance impression in lecture voice based on prosodic features. In: 2008 IEICE General Conference Proceedings, no. 1, p. 177 (2008)
7. Takashi Yamada's speech in YouTube. https://www.youtube.com/watch?v=oFX8XWcm0EA&list=PLwvjBjspdgkmOeq6Fd2HtPABCKy99EeRf&index=7/. Accessed 04 Sept 2023