# Individual differences in estimating the importance of speech utterances based on acoustic features

Jiating Liu, Sumio Ohno

Tokyo University of Technology, Hachioji, Tokyo, Japan
g212303237@edu.teu.ac.jp,ohno@stf.teu.ac.jp

**Abstract.** It is observed that there is an amount of casual talk, serving as a warm-up, prior to the commencement of lectures or meetings. When reviewing videos to save time and increase efficiency, attention is directed towards keywords and significant parts within the recording. In the previous study, a method is proposed that determines the importance of each utterance in TED talk videos on YouTube, based on acoustics features. In this study, compared to the previous study, data from different speakers was added. To improve the accuracy of annotations, the importance of the utterances is labeled using the LLM chat systems like ChatGPT, Copilot, Gemini, Perplexity, Claude3, and GropChat. To balance the train data, SMOTE is employed to synthesize new samples for the minority class. As a result, the random forest classifier proves effective for the training set, however, it is observed that the model exhibits low predictive performance for data of different speaker. Furthermore, acoustic features, especially, related to F0 is found to make a significant contribution to the training. In the future, there are two main aspects for improvement. One is that normalization is adopted to process the speech signal, reducing the model's dependence on specific inputs, thereby enhancing the robustness and generality of the training model. Another is that it is a considered to the utilization of other methods of machine learning.

**Keywords:** important decision, acoustic features, linguistic features, random forest classifier

## 1.1 Introduction

The field of speech information processing has achieved remarkable advancements in recent years. As of 2023, speech recognition systems have achieved accuracy rates exceeding 90% [1], and emotion recognition technology has demonstrated precision rates surpassing 85% in studies conducted in 2022 [2].

In verbal communication, not only is linguistic information conveyed through speech, but non-linguistic information is also embedded within acoustic features. The relationship between acoustic features and intelligibility has been investigated in the previous study [3]. Given these advancements, it is indicated that speech recognition and analysis technologies have rapidly evolved, with particular attention being paid to analyses based on acoustic features.

In the previous study [4], a new method for determining the importance of utterances in speech based on acoustic features was proposed. Compared to the previous study, in this paper data from different speakers was added. To improve annotation accuracy, the

importance of the utterances was labeled using several Large Language Model (Hereinafter, it is referred to as LLM) chat systems like ChatGPT, Copilot, Gemini, Perplexity, Claude3, and GropChat. To balance the positive and negative train data, Synthetic Minority Over-sampling Technique (Hereinafter, it is referred to as SMOTE) [5] is employed to synthesize new samples for the minority class.

## 2 The previous study

In the previous study, a system for estimating the importance of speech based on acoustic features was proposed. A model that can classify the importance of utterance for the training data is obtained. It was found that statistical measures of acoustic features related to F0 are utilized as important features for classification. However, upon conducting cross-validation tests and evaluating the test data, it was observed that sufficient accuracy was not achieved, indicating a need for further examination.

Compared to the previous research, five updates have been implemented in the current study. Firstly, the quantity of data has been quadrupled. Secondly, SMOTE has been employed to synthesize new samples for the minority class, thereby balancing the data. Thirdly, testing has been conducted in an open testing format. Fourthly, the number of LLM chat systems for annotation has been increased and their average values have been used to enhance the precision of the annotations. Lastly, in contrast to the previous research, which only relied on confusion matrices for model evaluation, the current experiment incorporates the use of F-values for assessing model performance.

Synthesizing the results of two experiments, it was observed that statistical measures of acoustic features related to F0 are frequently utilized as important features for classification. The use of SMOTE for synthesizing data has proven to be highly effective in handling imbalanced data in random forest classifier. And the use of F-values for model evaluation has enhanced the objectivity and accuracy of the testing process.

## 3 Preliminary Experiment

### 3.1 System design

This system is designed to evaluate the importance of each utterance in TED talk videos using random forest classifier. The system architecture is divided into two main phases: the learning phase and the estimating phase. The learning phase involves data preparation and model training, while the estimating phase focuses on predicting the importance of new utterances. The overall workflow is illustrated in the Fig. 1.

In the learning phase, TED talk videos are first downloaded from YouTube using a downloader, resulting in audio waveform files. Then, these audio waveforms are transcribed and segmented using the Whisper module, generating text transcripts and time-segmented utterances. Next, to improve accuracy of the supervised label, the

importance of the transcribed utterance text is annotated using six different LLM chat systems, including ChatGPT (GPT-3.5), Copilot (GPT-4 Turbo), Gemini (Gemini 1.5 Pro and Flash), Perplexity (GPT-3.5), Claude3 (Claude 3 Sonnet), and GropChat (Model: Llama3-8b-8192). SMOTE is used to generate synthetic samples for the minority class, balancing the training data. Acoustic features of utterances are extracted from the segmented audio using the openSMILE tool, resulting in feature vectors. Finally, a random forest classifier is trained using these feature vectors and the corresponding importance labels to creat a predictive model.

In the estimation phase, acoustic features are extracted from new test audio waveforms using openSMILE, and these acoustic features are inputted into the trained model to predict the importance labels of the new utterances.

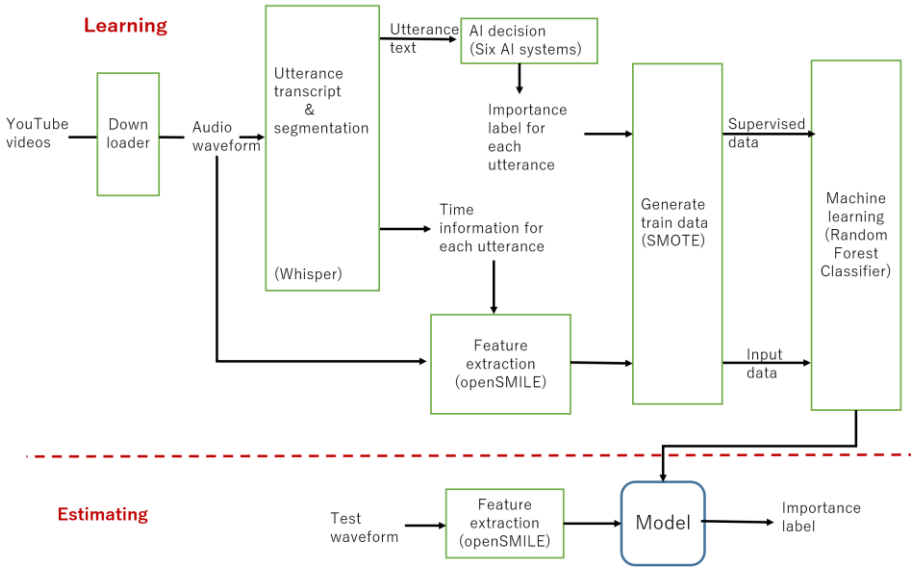


**Fig. 1** Importance estimating system design

## 3.2 Data processing

### 3.2.1 Automatic labeling

To confirm the differences in male and female voices, two male and two female speakers were selected from YouTube TED videos, each exhibiting significant vocal variations. The selected videos are: (1) ″ひとりじゃ円陣組めない″ (You can't make a circle by yourself) [6], (2) ″感動を創造する言葉の伝え方″ (Methods for conveying words that create emotional resonance) [7], (3) "Breakthrough 突破する力！～″ZONE” 人間としての能力を最大限発揮する方法″ (Breakthrough Power! ~ Methods to Maximize the Capabilities of a 'ZONE' Human) [8] and (4) ″あなたは今を選ぶことができる″ (You Can Choose Now) [9]. Each video is approximately

20 minutes long, and the Whisper module's medium model was utilized for transcription.

The transcription timestamps provided by Whisper were used to segment the original audio. Annotation was performed using a LLM chat system on the transcribed text. To ensure the accuracy of importance annotation for the audio, six LLM chat systems were employed: ChatGPT, Copilot, Gemini, Perplexity, Claude3, and GropChat. These LLM chat systems not only provided annotation results but also explained the reasoning behind their annotations, thereby enhancing accuracy. Important texts were labeled as 1, while unimportant texts were labeled as 0. If the average of the six label is greater than or equal to 0.5, it is deemed as important utterances, otherwise, it is considered as unimportant utterances, the processed labels and speech features are utilized as the train input. Due to the limitation on the number of characters per query to the AI, three types of prompts were primarily used:

① YouTube動画に「…」というTEDの講演があります。この講演で伝えたい主旨を簡単にまとめてください？ (There is a TED talk titled "…" on YouTube. Could you please summarize the main point of this speech?)

②〞以下の内容はある講演の発言録です。ID、ファイル名、発話の開始時間、発話の終了時間、発言内容が並んでいます。この講演の主旨は「~」であることを強調しています。主旨に基づいて、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 speech. The ID, filename, start time of the utterance, end time of the utterance, and the content of the statement are listed. The main point of this speech is "~". Based on the main point, please list only the IDs of important utterances.)

③〞以下の内容は上の講演の発話内容の続きです。ID、ファイル名、発話の開始時間、発話の終了時間、発言内容が並んでいます。主旨に基づいて、重要な発話のIDだけを列挙してください。〞(The following content is a continuation of the utterances from the above speech. It lists the ID, filename, start time of the utterance, end time of the utterance, and the content of the statement. Based on the main point, please list only the IDs of important utterances.)

For each original dataset, an imbalance of positive and negative data was observed. Much of the utterances were labeled as 0, with only a minority labeled as 1. To address this issue, the SMOTE was employed to generate new data for the minority class. A comparison between the number of the original data and the data after resampled is shown in Table 1.

**Table 1.** Comparison of important/unimportant utterances in original and SMOTE data

| | Label | Original | Resampled |
|---|---|---|---|
| Male 1 | unimportant | 131 | 131 |
| | important | 34 | 131 |
| Male 2 | unimportant | 100 | 100 |
| | important | 52 | 100 |
| Female 1 | unimportant | 346 | 346 |
| | important | 283 | 346 |
| Female 2 | unimportant | 132 | 132 |
| | important | 100 | 132 |

### 3.2.2 Balancing the imbalanced data

In this study, 15 models were trained using a random forest classifier, utilizing various combinations of four datasets. During the training process, the most suitable hyperparameters were determined and recorded through the implementation of grid search and cross-validation. The parameter settings for the grid search were established as shown in Table 2 below [10].

**Table 2.** Hyperparameter settings

| parameter | range |
|---|---|
| n_estimators | 1~20 |
| criterion | gini, entropy |
| max_depth | 1~4 |
| random_state | 0~100 |
| class_weight | balanced |

In the initial dataset, instances labeled as ‘0’ were found to be in majority, while those labeled as ‘1’ were in minority. Therefore, in the preliminary experiments, the applicability of the SMOTE method was also evaluated [10].

An example is shown using data from one male speaker. In the original dataset, there were 131 utterances labeled as ‘unimportant’ and 34 utterances labeled as ‘important’. After applying SMOTE, the ‘important’ labeled data was synthesized to match the quantity of the ‘unimportant’ labeled data. Subsequently, the balanced total dataset, consisting of 262 utterances, was split in an 8:2 ratio to serve as training and testing data respectively.

As show in Table 3 below, during the training process, the total count of true-negative and true-positive increases with ‘no smote’ is 136, accounting for 95.45% of the total. Upon synthesizing balanced data using the SMOTE method, the total count

of true-negative and true-positive instances increases to 198, still constituting 94.7% of the total, showing there is not much difference in train phase.

During the testing process as shown in Tables 4 and 5 below, the total count of true-negative and true-positive instances with ‘no smote’ is 28, accounting for 85% of the total, with the F-value of 29%. However, upon synthesizing balanced data using the SMOTE method, the total count of true-negative and true-positive instances increased to 46, constituting 87% of the total, with a significantly higher F-value of 89%. This substantial difference underscores the effectiveness of the SMOTE method in synthesizing new samples for the minority class, thereby enhancing the accuracy of the tests. Consequently, the SMOTE method was employed in all subsequent experiments.

**Table 3.** Comparison of results with and without SMOTE in train

| Train | without SMOTE | | with SMOTE | |
|---|---|---|---|---|
| true \ predicted | unimportant label | important label | unimportant label | important label |
| unimportant label | 98 | 5 | 97 | 10 |
| important label | 1 | 28 | 1 | 101 |

**Table 4.** Comparison of results with and without SMOTE in test

| Test | without SMOTE | | with SMOTE | |
|---|---|---|---|---|
| true \ predicted | unimportant label | important label | unimportant label | important label |
| unimportant label | 27 | 1 | 19 | 5 |
| important label | 4 | 1 | 2 | 27 |

**Table 5.** Comparison of F-value with and without SMOTE in test

| Test | without SMOTE | with SMOTE |
|---|---|---|
| F-value | 29% | 89% |

### 3.3 Experiment result

As shown in Table 6, all values represent the F-values obtained from the test results. The darker the frame color is to red, the smaller the number, while the darker it is to blue, the larger the number. The bold numbers with underscores indicate instances where the training and testing data originate from the same individual’s audio, with the original data being partitioned into training and testing data at a ratio of 8:2, the testing is conducted in a ‘closed’ dataset. The numbers only with underscores indicate instances where the training data originates from at least two individuals’ audio, and the testing data overlaps with some of the training data, i.e., the testing is conducted in a ‘closed’. The others indicate instances where the training and testing data are derived

from different individuals' audio, i.e., the testing is conducted in an 'open' dataset. In all the training instances mentioned above, the SMOTE method was utilized to synthesize training samples for the minority class to avoid imbalances between positive and negative data.

In this research, a total of 60 tests were conducted on four datasets using the 15 models. The number of open tests was 28 and the number of closed tests was 32. Regarding the train result, it was observed that 100% of them yielded an F-value greater than 0.65. Regarding the test result, it was observed that 75% of them yielded an F-value of less than 0.5 in the open test, 84% of them yielded an F-value greater than 0.65 in the closed test.

Based on the calculated F-values, it can be concluded that this classification problem achieved good results in terms of model training. The outcomes from the closed test for the model were significantly better compared to the open test, to the extent that the open test could be considered ineffective. This implies that the model is only capable of learning from specific training data, and its predictive performance is low when applied to new data.

**Table 6.** F-value of test result

| Test F-value (test / Train) | M1 | M2 | F1 | F2 |
|---|---|---|---|---|
| M1 | **0.885** | 0.000 | 0.000 | 0.215 |
| M2 | 0.091 | **0.889** | 0.587 | 0.301 |
| F1 | 0.352 | 0.558 | **0.885** | 0.527 |
| F2 | 0.093 | 0.355 | 0.519 | **0.653** |
| M1+M2 | 0.739 | 0.718 | 0.021 | 0.246 |
| M1+F1 | 0.581 | 0.557 | 0.708 | 0.557 |
| M1+F2 | 0.791 | 0.000 | 0.014 | 0.736 |
| M2+F1 | 0.085 | 0.514 | 0.661 | 0.515 |
| M2+F2 | 0.087 | 0.838 | 0.407 | 0.749 |
| F1+F2 | 0.000 | 0.380 | 0.703 | 0.712 |
| M1+M2+F1 | 0.758 | 0.643 | 0.681 | 0.439 |
| M1+M2+F2 | 0.653 | 0.691 | 0.440 | 0.711 |
| M1+F1+F2 | 0.632 | 0.171 | 0.699 | 0.719 |
| M2+F1+F2 | 0.140 | 0.748 | 0.656 | 0.664 |
| M1+M2+F1+F2 | 0.500 | 0.681 | 0.683 | 0.708 |

Tables 7 and 8 show the top 5 acoustic features given as variable importance in the classifiers, a total of four training sessions. The acoustic features contributed to each training are different, suggesting that there are individual differences in estimating the

importance of speech utterance base on acoustic features. Furthermore, it is found that acoustic features related to F0 and MFCC play an important role in training.

**Table 7.** Acoustic features contributing to train of males

| Speaker / Acoustic features | M 1 | M 2 |
|---|---|---|
| Top 1 | F0semitoneFrom27.5Hz_sma3nz_ stddevNorm | logRelF0-H1-H2_sma3nz_amean |
| Top 2 | F0semitoneFrom27.5Hz_sma3nz_ stddevRisingSlope | F1amplitudeLogRelF0_sma3nz_stddevNorm |
| Top 3 | slopeUV0-500_sma3nz_amean | spectralFlux_sma3nz_stddevNorm |
| Top 4 | HNRdBACF_sma3nz_amean | F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2 |
| Top 5 | VoicedSegmentsPerSec | F1amplitudeLogRelF0_sma3nz_amean |

**Table 8.** Acoustic features contributing to train of females

| Speaker / Acoustic features | F 1 | F 2 |
|---|---|---|
| Top 1 | MeanVoicedSegmentLengthSec | HNRdBACF_sma3nz_stddevNorm |
| Top 2 | StddevUnvoicedSegmentLength | F0semitoneFrom27.5Hz_sma3nz_ stddevFallingSlope |
| Top 3 | mfcc3V_sma3nz_stddevNorm | F3bandwidth_sma3nz_amean |
| Top 4 | loudness_sma3_stddevNorm | mfcc2V_sma3nz_stddevNorm |
| Top 5 | mfcc3_sma3_amean | equivalentSoundLevel_dBp |

## 4 Conclusion

It is included two main parts in this paper: data preparation and machine learning in this study. In the data preparation part, utterances are extracted from TED talk videos on YouTube. Six LLM chat systems are utilized to label the utterances. To balance the imbalanced data, SMOTE is employed to balance the data, synthesizing new samples for the minority class. In the machine learning part, the random forest classifier is utilized to train the model.

As a result, it is found that there are individual differences in estimating the importance of speech utterances based on acoustic features. Furthermore, it is found that representative acoustic features that make a significant contribution to the training include acoustic features related to F0 and MFCC.

For future work, there are two main aspects for improvement. One is that to enhance the robustness and generality of the training model, regularization is adopted to process the sound, reducing the model's dependence on specific inputs. Another is that it is a considered to the utilization of other methods of machine learning to improve model accuracy.

## References


1. Evaluating the accuracy of speech-to-text in 2023, https://www.cxtoday.com/speech-analytics/how-accurate-is-speech-to-text-in-2023-assemblyai/, last accessed 2024/6/27.
2. Manohar, K., Logashanmugam, E. (2022). Speech-Based Human Emotion Recognition Using CNN and LSTM Model Approach. In: Bhateja, V., Satapathy, S.C., Travieso-Gonzalez, C.M., Adilakshmi, T. (eds) Smart Intelligent Computing and Applications, Volume 1. Smart Innovation, Systems and Technologies, vol 282. Springer, Singapore.
3. Keishin Saga and Makoto Imura, Analysis of Acoustic Features for Improving the Intelligibility of Spoken Speech, Entertainment Computing Symposium (EC2019/9)
4. Liu, J. and Ohno, S., "A System for Estimating the Importance of Speech Based on Acoustic Features", IWACIII 2023. Communications in Computer and Information Science, vol 1932.
5. SMOTE, https://zhuanlan.zhihu.com/p/671983490, last accessed 2024/6/24.
6. Takashi Yamada's speech in YouTube, https://www.youtube.com/watch?v=oFX8XWcm0EA, last accessed 2024/6/25.
7. Masaki Sato 's speech in YouTube, https://www.youtube.com/watch?v=4jfcE8u9KOM, last accessed 2024/6/25.
8. Keiko Ihara 's speech in YouTube, https://www.youtube.com/watch?v=w50ElZTtzXE, last accessed 2024/6/25.
9. Maho Shono 's speech in YouTube, https://www.youtube.com/watch?v=FEBtLi4a470, last accessed 2024/6/25.
10. Random Forest (Classification) and Hyperparameter Tuning, https://qiita.com/FujiedaTaro/items/61ded4ea5643a6204317, last accessed 2024/6/27.