2021 International Conference on Intelligent Computing, Automation and Systems (ICICAS) | 978-1-6654-2810-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICAS53977.2021.00050

Research on the Quantity Evaluation of Speech Datasets for Model Training

Sun Li
Institute of Cloud Computing and Big Data
China Academy of Information and Communications Technology
Beijing, China
lisun@caict.ac.cn

Feng Cao
Institute of Cloud Computing and Big Data
China Academy of Information and Communications Technology
Beijing, China
caofeng@caict.ac.cn

Zishan Liu*
Institute of Technology and Standards
China Academy of Information and Communications Technology
Beijing, China
liuzishan@caict.ac.cn
*Corresponding author

***Abstract*—With the maturity of intelligent speech technology and product application, the demand for high-quality speech datasets is increasing. There have been some researchers put effort on the quality evaluation of the structured data, but there are few standards appeared for the speech datasets. By analyzing the construction principle of speech algorithm model and analyzing the construction demand of speech datasets, a unified quality assessment framework for the speech datasets is presented. The framework proposes to evaluate the speech datasets in terms of four dimensions: breadth coverage, anthology distinction, professional field depth and data integrity. By proposing specific speech datasets quality evaluation metrics, calculation methods and evaluation steps, and analyzing the experimental examples and results of speech datasets quality evaluation in vehicle application field, this paper provides a reference basis for evaluating speech datasets quality and promoting datasets construction. Considering the diversified applicability, privacy issues, efficiency and automation requirements of speech datasets construction, some suggestions for the future development of high-quality speech datasets construction are put forward.**

Keywords—artificial intelligence (AI), speech datasets, quality assessment, algorithm, model, intelligent speech

I. INTRODUCTION

In recent years, with the rapid development of artificial intelligence (AI) technology and improvement of the accuracy of speech recognition, intelligent speech has been widely used in the fields of voice assistant, intelligent speaker, intelligent wearable devices and so on. The algorithm construction and model training of the applications highly rely on the quality of the speech datasets. Large scale and high-quality speech datasets are of great significance to the construction of speech recognition system. At present, the dataset quality assessment for the artificial intelligence applications has attracted much research effort, but most of the existing data quality evaluation schemes are aimed at structured data. For the quality evaluation of unstructured speech datasets, there still lack a systematic and complete quality evaluation framework globally.

This paper aims to promote the construction of the speech dataset, and the improvement of dataset quality, by developing a unified speech datasets quality evaluation framework. Firstly, the paper analyses the speech data construction demand for the intelligent speech recognition algorithm model, and points out that the speech dataset construction should consider the general technical performance, application optimization ability and field adaptability, and should be evaluated from the aspects of breadth coverage, anthology distinction, professional field depth and data integrity. This paper proposes the systematic evaluation framework for speech datasets, in terms of four dimensions: breadth coverage, anthology distinction, professional field depth and data integrity, with the specific evaluation metrics for each dimension, the selection reasons, calculation methods and evaluation process. Then we select the speech datasets in vehicular application field for datasets quality evaluation experiment and result analysis, so as to provide reference examples for evaluating the quality and promoting the construction of speech datasets. Finally, considering the diversified applicability, privacy issues, efficiency requirements and automation requirements of speech datasets construction, this paper puts forward some suggestions for the future development of high-quality speech datasets construction.

The remainder of the paper is organized as follows. The related work is presented in Section II. The speech datasets quality evaluation framework is proposed in Section III, with the experiment and analysis of speech datasets quality evaluation presented in Section IV. Section V concludes the paper, and puts forward the development suggestions for the construction of high-quality speech datasets.

II. RELATED WORK ABOUT DATASETS QUALITY

A. Datasets Quality Evaluation

With the development of big data and artificial intelligence, the importance of datasets quality is becoming more and more obvious, attracting research communities, industry area and standardization organizations to participate in the relevant research and construction of datasets.

Professor Xiao Li Meng of Harvard University pointed out that the datasets quality of AI is far more important than the data volume, and the construction of data quality evaluation framework is a key problem to be solved urgently. At present, datasets quality evaluation mainly focuses on the structured data, and have appeared some evaluation model and standards. The research on the overall framework of data quality began in the early 1990s. In [1], Professor Richard Y. Wang of MIT launched the total data quality management (TDQM) and put forward the comprehensive data quality management method, including definition stage, measurement stage, analysis stage and improvement stage. In 2002, Lee et al. put forward the Assessment Information Management Quality (AIMQ) [2], providing the method of data quality evaluation and difference

978-1-6654-2810-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICICAS53977.2021.00050

analysis, and forming the framework of TDQM. Relevant research results have had a far-reaching impact on the development of data quality. The team also proposed a design quality assurance (DQA) method to guide the general principles of data quality metric definition [3]. The method defines three stages: subjective and objective evaluation, subjective and objective evaluation comparison and improvement for data quality assessment. In this method, data quality metrics are mostly defined for specific problems. In China, the information quality research group (IQRG) has established in 2008. In addition, the “basic theory and key technology of data quality management (No. F020204)” was listed as the annual key research project of the National Natural Science Foundation of China (NSFC) in 2011. Professor Tang of Peking University and his research team use the form of six tuples to describe the data quality evaluation model [4], and put forward the method of constructing the model and how to calculate the metrics. Huang proposes a metadata driven data quality evaluation framework [5]. Overall, the data quality research still lacks the systematic frameworks and achievements on the specific speech datasets.

Up to now, data quality related standards include ISO/IEC 25012, ISO/IEC 25024, GB/T 25000.12-2017 and GB/T 25000.24-2017. The data quality evaluation metrics proposed in GB/T 25000.24-2017 mainly include 15 categories, such as data integrity, effectiveness, etc., and each category includes relevant metrics. However, the relevant standards are mainly for the structured datasets, which is difficult to meet the requirements of quality evaluation of the speech dataset, which are unstructured and depended on the specific applications.

B. Speech Recognition Model Training Principle and Datasets Construction

For speech recognition, speech dataset is used as training data to analyze the speech signal and convert it into text sequence, as shown in Fig. 1. Speech recognition algorithm model [6], first extract the feature vector required by the decoder through the signal processing module, converting the speech into a digital sequence or vector that can be recognized, and then find the optimal solution in the decoder according to the extracted feature vector. The acoustic model in the decoder connects the feature of the speech signal with the speech modeling unit of the sentence, training through a large number of speech data to obtain the probability corresponding to each frame and state. The language model outputs the text sequence with the maximum probability. The pronunciation dictionary contains the set of words that can be processed by the system, which is the mapping relation between the modeling unit of acoustic model and language model to form a search state space for the decoder to decode. Finally, the sequence with the highest score is decoded in the network composed of acoustic model, language model and pronunciation dictionary, and the output is considered to be the result of recognition. It can be seen that in the process of model training and application, the important factors such as acoustic environment, language content and knowledge need to be considered. By fully considering the general technical performance, application optimization ability and field adaptability, the speech datasets could be suitable for the algorithm model.

In the field of traditional linguistics, a lot of basic research work has been done on the design and construction of corpus [7], which provides a sufficient basis for the construction and evaluation of speech datasets. In corpus selection, it is necessary to cover the most speech phenomena by constructing the least corpus as possible. The basic algorithm of corpus selection is the greedy algorithm [8]. There are some researches on the construction and evaluation of Chinese speech corpus, however, its purpose is mainly for the analysis of language knowledge and the development of language application. The research is inconsistent with the requirements of algorithm model for speech datasets, and cannot be directly applied to the evaluation of datasets quality.

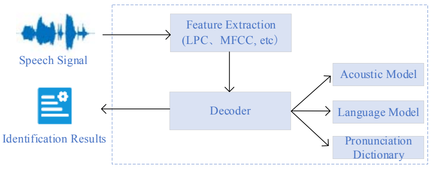
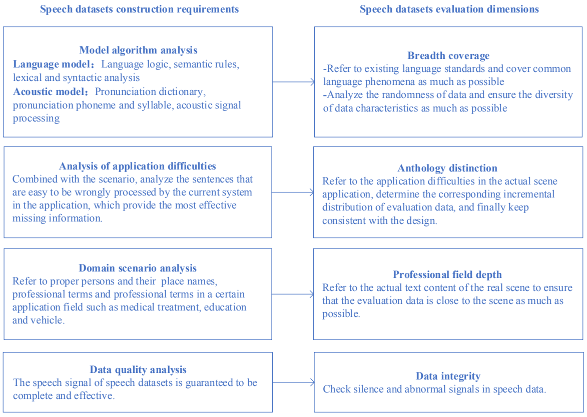


Fig. 1. Speech recognition technology framework.

III. SPEECH DATASETS QUALITY EVALUATION METRIC BASED ON ALGORITHM MODEL

Based on the speech model training principle and corpus construction related demand analysis, aiming at the speech unit and speech knowledge of the speech data itself, combined with the model application scenario, this paper proposes a speech datasets quality evaluation system for algorithm model training, including four dimensions: breadth coverage, anthology distinction, professional field depth and data integrity, as shown in Fig. 2.


Speech datasets construction requirements
Speech datasets evaluation dimensions

Model algorithm analysis
Language model： Language logic, semantic rules, lexical and syntactic analysis
Acoustic model： Pronunciation dictionary, pronunciation phoneme and syllable, acoustic signal processing

Breadth coverage
-Refer to existing language standards and cover common language phenomena as much as possible
-Analyze the randomness of data and ensure the diversity of data characteristics as much as possible

Analysis of application difficulties
Combined with the scenario, analyze the sentences that are easy to be wrongly processed by the current system in the application, which provide the most effective missing information.

Anthology distinction
Refer to the application difficulties in the actual scene application, determine the corresponding incremental distribution of evaluation data, and finally keep consistent with the design.

Domain scenario analysis
Refer to proper persons and their place names, professional terms and professional terms in a certain application field such as medical treatment, education and vehicle.

Professional field depth
Refer to the actual text content of the real scene to ensure that the evaluation data is close to the scene as much as possible.

Data quality analysis
The speech signal of speech datasets is guaranteed to be complete and effective.

Data integrity
Check silence and abnormal signals in speech data.

Fig. 2. Quality evaluation system of speech datasets based on algorithm model.

A. Breadth Coverage

A large-scale speech recognition system needs to face a variety of possible recognition scenarios. If these recognition scenarios do not appear in the training corpus, the recognition effect of the recognizer in this scenario will decline sharply. In order to construct a good recognition system, its training data must cover all possible environments and conditions in the actual application scenario. Therefore, the main evaluation metrics of breadth coverage have two aspects, including the coverage of language and data features.

a) Language coverage: The speech datasets should reflect the real language phenomenon as much as possible by accumulating various pronunciation conditions, pronunciation methods and grammatical structures. Taking Chinese as an example, in terms of structure, part-of-speech in language includes nouns, verbs, adjectives, numerals, quantifiers and pronouns, as well as adverbs, prepositions, conjunctions, auxiliary words, onomatopoeia and interjections. Language coverage mainly evaluates the coverage of basic information related to linguistics in the speech datasets. Its evaluation method is as follows:

$$X = \frac{A \cap B}{B}, \quad (1)$$

where A represents the number of specified data information samples, and B is the total number of reference data samples; $X \in [0, 1]$, The larger the X value, the greater the coverage of the specified data sample.

b) Amount of feature information: The speech datasets should accumulate various data sources so that the model can meet various variability and reflect the robustness of the application. It must cover all possible environments, pronunciation conditions and modes in the actual application scenario, including but not limited to: (1) speaker, (2) recording equipment, (3) transmission channel, (4) environmental noise, (5) field atmosphere, (6) style and emotion, (7) regional accent, (8) specialty and field, etc. Therefore, one of the most important requirements of speech datasets construction is to accumulate diversified data sources, so that the model can have sufficient generalization ability and reflect the robustness of algorithm model recognition.

Entropy can be used as amount of feature information to measure the complexity of specific speech features in the datasets. The more complex the speech features are, the more diverse the different categories of features appear, and the larger the entropy is. Since the speech datasets have many complex features such as multi-dimensional, multi-level, multi relevance and repeatability, the ratio of information entropy and maximum entropy of each type of sample data is used as the amount of feature information to evaluate the speech datasets. For any feature, the calculation method is:

$$I = -\sum_{i=1}^{n} \frac{p(x_i)\, log_2\, p(x_i)}{Y}, \quad (2)$$

where n is the total number of data samples, $p\left(x_i\right)$ represents the probability that the element value of feature X is x_i, and Y represents the maximal entropy. $I \in [0, log_2(n)]$, the higher the value of I, the higher the information entropy provided by the representative feature X.

B. Anthology Distinction

When accumulating speech corpus data sources, we need to consider the real situation of specific application scenarios and accumulate the data that can bring feature incremental value to the algorithm model. For example, in speech recognition scenarios, speech data could include multiple noises and multi-person conversations that should be considered. In the application of speech synthesis, the accumulation of mixed Chinese and English speech data should be considered. And in the voiceprint recognition field, the speech datasets need to include synthetic speech, converted speech and recording, so as to improve the anti-attack ability of the algorithm model.

In order to improve the model recognition accuracy and robustness provided by speech datasets, the annotation and accumulation of error prone data recognized by the current system should be strengthened during speech data accumulation. This means that when selecting data, not only the value of the data itself, but also the incremental value brought by the data to the system should be considered, which is consistent with the feature distribution increment of the application scenario. According to the feature distribution of the speech datasets and the feature distribution of the application field, a distribution consistency evaluation method is proposed to evaluate the feature matching degree. For the speech datasets that are numeric, the Euclidean distance between the features of the speech datasets and the features required by the application scenario is calculated, as follows:

$$S(Data_1, Data_2) = \sum_{i=1}^{n} \omega_i \times \left(Data_1{}^{(i)} - Data_2{}^{(i)}\right)^2. \quad (3)$$

It is assumed that the feature dimension of the datasets is N, $[1, n]$ means that the feature type is numerical. $Data_1^{(i)}$ represents the i^{th} feature of the datasets to be evaluated, $Data_2^{(i)}$ represents the i^{th} feature of the actual application scenario datasets, and ω_i is the weight of the corresponding feature.

For the speech datasets whose feature type is distribution, the distribution consistency evaluation is to calculate the KL-divergence between the features of the speech datasets and the features required by the application scenario, such as frequency distribution, gender distribution, age distribution, etc., as follows:

$$E(Data_1, Data_2) = \sum_{i=n+1}^{N} \omega_i \times \left(\sum_{j=1}^{M^{(i)}} Data_1{}^{(i)}(x_j) \times log\left(\frac{Data_1{}^{(i)}(x_j)}{Data_2{}^{(i)}(x_j)}\right)\right), \quad (4)$$

where $[n + 1, N]$ indicates that the feature type in the datasets is distributed, $Data_1^{(i)}$ represents the i^{th} feature of the datasets to be evaluated, $Data_2^{(i)}$ represents the i^{th} feature of the application scenario datasets, and ω_i is the weight of the corresponding feature. In the distribution feature, each feature is represented by distribution. It is assumed that it has $M^{(i)}$ dimensions, and each dimension is x_j. Finally, the average distribution distance is used as the feature matching degree between the speech datasets to be evaluated and the application scenario datasets.

C. Professional field depth

When constructing *The Modern Chinese corpus of the State Language Commission*, it is divided according to the professional content with reference to the industry attributes, including *humanities and social sciences* (divided into 8 major categories and 30 sub categories), *natural sciences* (including agriculture, industry, medicine, electronics, engineering technology, etc.) and *comprehensive categories* (applied literature and corpus difficult to classify) [11]. Analogizing the training and learning of artificial intelligence model and widely accumulating diversified speech data sources can only help to improve the general recognition ability of artificial intelligence model. However, for the professional background in the application field, such as the adaptability of relevant application models such as law, telecommunications, medical treatment, smart home, finance and so on, it is also necessary to accumulate the greetings, situational dialogue, person name, place name, professional terms and other contents as speech corpus data sources.

It can be seen that the matching between speech data and professional field content has become an important measure of professional field depth. In this paper, TFIDF-COS algorithm is used to quantify the similarity from the perspective of data by abstractly decomposing the word frequency of the text [12]. To evaluate the professional field depth of speech datasets and measure it by content similarity, the specific methods are as follows:

First, calculate the word frequency vector of the speech datasets to be evaluated and preset professional field datasets [13]. The formula to calculate the word frequency vector *u* of speech datasets *X* is as follows:

$$u = \frac{f_n}{f} \cdot \log \frac{N}{p_n}, \tag{5}$$

where, f_n is the number of occurrences of specific word in the datasets, f is the number of occurrences of the most frequent word in the datasets, *N* is the number of articles in the general corpus, and p_n is the number of documents containing the word.

In the second step, based on the cosine similarity algorithm, calculate the similarity of the word frequency vector of the speech datasets to be evaluated and the preset professional field datasets [14]. Assuming that *P* and *Q* are datasets of *n* dimensional vectors, the cosine similarity of *P* and *Q* is calculated as:

$$N = \frac{\sum_{i=1}^{n}(p_i \times Q_i)}{\sqrt{\sum_{i=1}^{n}(p_i)^2 \times \sqrt{\sum_{i=1}^{n}(Q_i)^2}}} \tag{6}$$

Where *P* represents the word frequency vector of the datasets to be evaluated; *Q* represents the word frequency vector of the preset professional field datasets; $N \in [0,1]$, the larger the value of *N*, the closer the two-word frequency vectors are, and the higher the text similarity is.

D. Data integrity

In the process of speech data collection, due to the reasons of equipment, environment and speaker, it will cause problems such as missing signal and incomplete recording, and produce a lot of invalid information. Speech datasets are usually continuous sampling points. Signal anomaly refers to discontinuous speech signal segments affected by sound acquisition equipment or environmental noise; silent segments refer that speech datasets appear as silent signal segments by the problems of speakers and recording equipment, which are invalid segments of speech datasets.

For speech signal quality detection, usually based on voice activity detection (VAD), which obtain the signal invalid segment in the speech datasets [15]. The integrity of speech datasets is evaluated by content effectiveness, as follows:

$$X = 1 - A/B, \tag{7}$$

where *A* represents the invalid data duration in the speech datasets; *B* is the duration of the speech datasets to be evaluated; $X \in [0,1]$, The higher the value of *X*, the higher the validity of the data content.

IV. Experiment and Analysis of Speech Datasets Quality Evaluation

A. Experimental Steps

According to the above research, the quality evaluation of speech datasets is mainly divided into three steps, as shown in Fig. 3, as follows:

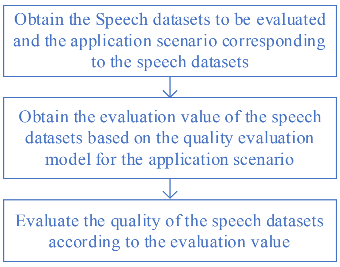

Obtain the Speech datasets to be evaluated and the application scenario corresponding to the speech datasets
Obtain the evaluation value of the speech datasets based on the quality evaluation model for the application scenario
Evaluate the quality of the speech datasets according to the evaluation value

Fig. 3. Voice dataset quality assessment process

B. Experimental preparation

In this paper, the open-source speech datasets AISHEL1(OPENSLR-SLR33) and THCHS-30(OPENSLR-SLR18) are selected for information statistics. Combined with the speech corpus construction requirements [16], the collection standards and texts of speech datasets in the field of vehicle applications are designed. The specific settings include:

a) Content setting: Vehicle driving involves vehicle devices, driving operations, location names, etc. and includes service scenarios such as driving mode selection, fault early warning, telephone service, voice navigation, traffic condition broadcast, etc.

b) Speed setting: Slow (less than 9 words in 3 seconds) accounts for 30%, medium (9-13 words in 3 seconds) accounts for 50%, fast (more than 13 words in 3 seconds) accounts for 20%.

c) Noise setting: Low noise (below 55 dB) accounts for 70%, medium noise (55-60 dB) accounts for 20%, and high noise (61-70 dB) accounts for 10%.

d) Speaker setting: The ratio of male to female is 1:1, and accent standard reach second class B or above in *Putonghua Proficiency Test.*

C. Experimental Result

According to the experimental steps and metrics calculation method, the actual results are as follows:

a) Language coverage assessment: Make statistics on the number of types of initials, tone vowels, soft tones, Er Hua, diphones and trisyllabic in *The Seventh Edition of Modern Chinese Dictionary*, *The List of Required Soft Tone Words for Putonghua Proficiency Test* and *The List of Er Hua words for Putonghua Proficiency Test* as *A*, as well as the number of types of various language phenomena in the speech datasets to be evaluated as *B.* Then obtain the coverage of each speech phenomenon according to the calculation formula, and calculate the average to obtain the overall language coverage, as shown in Table I. It can be observed from the results that in terms of the coverage of language phenomena, AISHEL1 contains richer language pronunciation elements than THCHS-30, which can better reflect the actual Chinese language pronunciation phenomena.

b) Amount of Feature Information Evaluation: Because the open-source datasets only has the characteristic information of the speaker, the individual speaker is taken as the feature to calculate the proportion of different speaker and the ratio of the information entropy and the maximum entropy. Among them, the AISHEL1 covers 400 speakers in total, and 141908 audio output, the distribution is shown in Fig. 4. For 400 speakers, according to the information entropy calculation formula, the maximum information entropy is 17.117, the speaker information entropy is 2.601, and the corresponding amount of feature information is 0.1519. The THCHS-30 covers 60 speakers in total, and 13388 audio outputs. The distribution is shown in Fig. 5. For 60 speakers, according to the information entropy calculation formula, the maximum information entropy is 3.710, the speaker information entropy is 1.743, and the corresponding amount of feature information is 0.1271. Compared with THCHS-30, AISHEL1 provides more speaker information features, more diversified data sources and rich generalization and robustness for model training.

TABLE I. LANGUAGE COVERAGE STATISTICS

	AISHEL1			THCHS-30		
	A	***B***	***cov1 (%)***	***A***	***B***	***cov1(%)***
Initial consonant	21	21	100	21	21	100
Tonal vowel	185	168	99.40	185	168	98.81
Whisper	154	137	80.29	110	137	52.55
Er Hua	1	176	0	1	176	0
syllable	1355	1859	69.77	1208	1859	61.00
Diphone	15057	9853	86.38	8254	9853	62.71
Trisyllabic	187696	67281	75.87	35613	67281	26.16
Language coverage	-	-	73.1	-	-	57.3

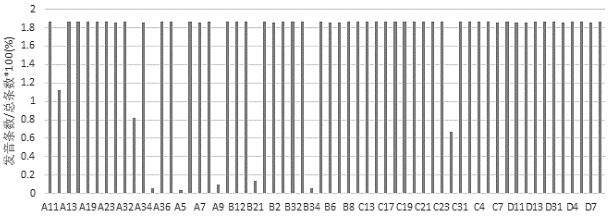


Fig. 4. Distribution statistics of the number of speakers in AISHEL1.

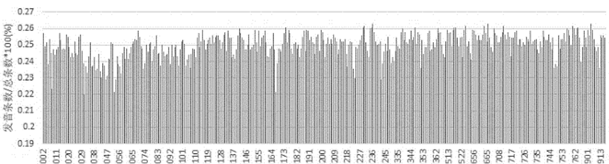


Fig. 5. Distribution statistics of the number of speakers in THCHS-30.

c) Feature matching degree: According to the collection standards in the experimental preparation, the standard reference distribution is formed by specifying the speech speed, noise, speaker and other factors. However, the AISHEL1 and THCHS-30 open-source datasets do not label the above features, so it is impossible to make direct statistics. Fortunately, the signal-to-noise ratio and speech speed can be statistically calculated by VAD algorithm. After calculation, the signal-to-noise ratio distribution feature matching degree of AISHEL1 *D1* is 0.243064205, the speech speed feature matching degree *D2* is 0.15490196 with the mean value of 0.1989. The signal-to-noise ratio distribution feature matching degree of THCHS-30 *D1* is 0.536289422, the speech speed feature matching degree *D2* is 0.15490196 with the mean value of 0.3456. In terms of feature distribution matching degree, THCHS-30 is more in line with the data construction standard of theoretical speech model than AISHEL1.

d) Content similarity: This paper analyzes and construct the word frequency and text content of the reference text (3062 sentences) of the vehicle field speech datasets, as shown in Fig. 6.

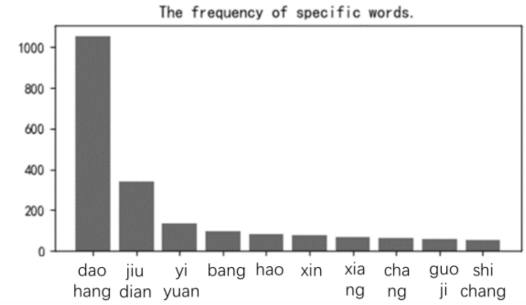


Fig. 6. Reference text analysis of speech datasets in vehicle field

TFIDF-COS algorithm is used for full sample keyword extraction and word frequency analysis. The results are shown in Fig. 7 and Fig. 8. After calculation, the content similarity of AISHEL1 is 3.15%, and that of THCHS-30 is 2.18%. It can be seen that in the vehicle field, the field correlation of AISHEL1 and THCHS-30 are low.

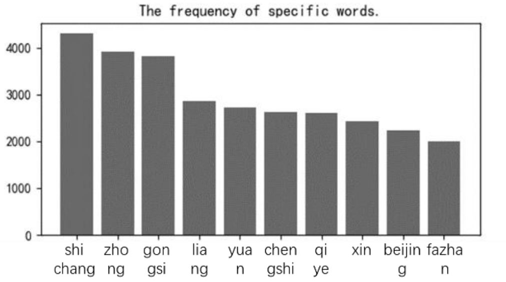


Fig. 7. Word frequency analysis of AISHEL1.

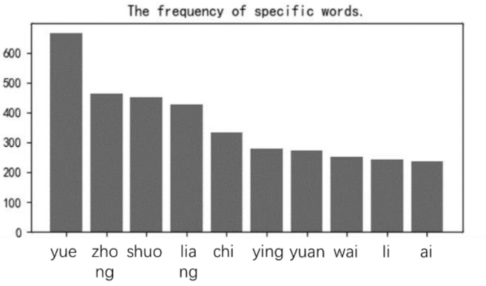


Fig. 8. Word frequency analysis of THCHS-30.

e) Content validity: The VAD algorithm is used to count the number of valid data which removes the silent segment and abnormal speech of the speech datasets, so as to evaluate the content validity. The mean value of speech duration ratio of AISHEL1 is 70.5%, and the standard deviation is 6.4%.

Among them, 24 have clipping, and the clipping ratio is 0.017%. Excluding the silent segment and abnormal signal, the content validity is 0.7049, and the duration ratio distribution is shown in Fig. 9. The mean value of speech duration ratio of THCHS-30 is 71.8%, standard deviation is 5.6%. Among them, 3138 has clipping, and the clipping proportion is 23.4%. Excluding the silent segment and abnormal signal, the content validity is 0.484, and the duration ratio distribution is shown in Fig. 10. From the analysis of content validity of speech signal, it can be concluded that the signal integrity of AISHEL1 is higher than that of THCHS-30.

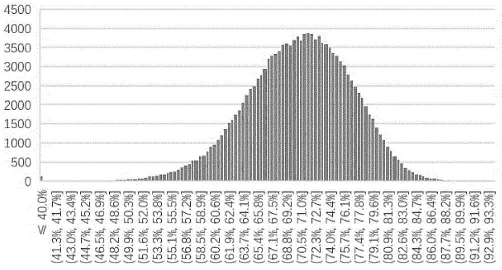


Fig. 9. Speech data duration ratio distribution of AISHEL1.

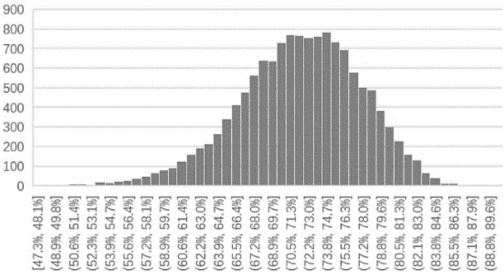


Fig. 10. Speech data duration ratio distribution of THCHS-30.

V. Conclusion

This paper proposes a unified speech datasets evaluation method, including the evaluation dimension and definition of datasets quality, datasets quality evaluation metrics and methods, which provides a reliable basis for evaluating the quality of speech datasets and promoting the construction of datasets. However, in practical applications, there are still some urgent problems to be solved, including the diversified applicability of speech datasets, privacy problems, efficiency and automation requirements and so on. Based on the above analysis, this paper puts forward the following development suggestions for the construction of high-quality speech datasets:

1) Establish a comprehensive speech datasets quality evaluation system: This paper puts forward a number of evaluation metrics and methods for the speech datasets required by the intelligent speech algorithm model. However, it is difficult to build a single evaluation system suitable for all kinds of algorithm models. In the future, the establishment of comprehensive evaluation metrics and standards, evaluation criteria and methods, and the specific requirements should be proposed for the quality of speech datasets for different application scenarios, which is an essential technical element to ensure the quality of speech datasets.

2) Design an adaptive speech datasets quality evaluation framework: For different intelligent speech application scenarios, the training performance of the same speech datasets for the model is quite different. Therefore, how to design a general, efficient and adaptive speech datasets quality evaluation framework and build an application-oriented datasets is an essential part of the construction and development of high-quality speech datasets.

3) Realizing the construction and development of high-quality speech datasets under multi-objective equilibrium: On the one hand, the acquisition and opening of high-quality speech datasets are restricted by data security and privacy protection. whose technologies often affect the quality of datasets. Therefore, in practical application, the balance between datasets quality and privacy protection needs to be considered. On the other hand, the construction of large-scale high-quality speech datasets has high requirements in terms of personnel, environment and tools. Therefore, in practical application, it is necessary to consider the balance between the information increment brought by the datasets as well as the model performance optimization and cost.

References

[1] Wang R Y, Storey V C, Firth C P. A framework for analysis of data quality research[J]. IEEE transactions on knowledge and data engineering, 1995, 7(4): 623-640.

[2] Lee Y W, Strong D M, Kahn B K, et al. AIMQ: a methodology for information quality assessment[J]. Information & management, 2002, 40(2): 133-146.

[3] Pipino L L, Lee Y W, Wang R Y. Data quality assessment[J]. Communications of the ACM, 2002, 45(4): 211-218.

[4] Yang Q Y, Zhao P Y, Yang D Q, et al. Research on data quality evaluation method[J]. Computer engineering and Application, 2004, 40(9): 3-4, 15.

[5] Huang G, Yuan M, Wu X Y, et al. Research on metadata driven data quality evaluation architecture[J]. Computer engineering and Application, 2013, 0(8): 114-119, 181.

[6] Shan Y H, Li J, Wang X R, et al. The generation method of speech recognition training data and the training method of speech recognition model: CN111402865A[P]. 2020.

[7] Zu Y Q. Corpus design of Chinese continuous speech database[J]. Journal of acoustics.1999,(3);236-247.

[8] Wu H, Xu B, Huang T Y. Automatic corpus selection algorithm based on triphone model[J]. Journal of software, 2000, 11(2): 271-276.

[9] Zhuang J L. Research and application of quantitative analysis of data quality[D].2019.

[10] Jin J. Research on the value evaluation of information entropy in the era of big data[D]. Jilin University, 2019.

[11] Liu L Y. Development of modern Chinese Corpus[J]. Language application, 1996(03):3-9.

[12] Gheith M , Aboul-Ela M , Arafa W . Learning Word Graph Representation for Document Classification[C]// 27th Conference for Computer Science, Statistics and Operation Research, Egyptian Computer Society. 2002.

[13] Gou H W, Gou X T. Analysis of word separation and sentence similarity based on word vector[J]. Scientific and technological innovation, 2018(33):55-56.

[14] Gu B, Li J H, Liu K Y. Chinese text clustering based on COSA algorithm[J]. Chinese Journal of information, 2007, 21(6):65-70.

[15] Liu P, Wang Z Y. Multimodal speech endpoint detection[J]. Journal of Tsinghua University (natural science edition), 2005(07):896-899.

[16] Wang T Q, Li A J. Design of continuous Chinese Speech Recognition Corpus[C]. National Conference on modern phonetics. 2003.