Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230490

PreSoramimiset: Establishing dataset for Chinese misheard lyrics generation

Zihao Li

Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada

18zl44@queensu.ca

Abstract. Mondegreen is a common phenomenon during conversation and is more obvious during listening to songs. To simulate the impression that the audience will have when they hear a particular piece, the Soramiminet model is introduced in this paper. The model is a combination of wave2vec, and transformer used for generating misheard lyrics in Chinese with various input songs. This article will focus more on the establishment of the dataset for wave2vec model. The criteria and several methods of creating a high-quality dataset are summarized and present in this paper. Additionally, the negative impact of a defective dataset and how to avoid it is discussed. The main limitations and biases of this dataset, and how to address them and future work are explored. The dataset includes a set of audio clips from 21 tracks by 9 different singers in AVI form and a text file for annotation which can be processed by python dataset module. The dataset is relatively biased, since all the annotations are done by the author personally. There is no "correct" labelling, given that the model is generating misheard or "wrong" results.

Key words: Dataset, Transformer, Nature Language Processing, Machine Learning, Artificial Intelligence.

1. Introduction

Mondegreen is coined by American writer Sylvia Wright in 1954[1], recalling a childhood memory of her mother reading a Scottish ballad, and mishearing the words "laid him on the green" as "Lady Mondegreen". In 2002, Oxford English dictionary explained the term as a misunderstood or misinterpreted word or phrase resulting from a mishearing. esp. of the lyrics to a song[2].

The current research on mondegreen mainly focuses on pedagogy and linguistics. In addition, some studies are aimed at solving the negative impact of the mondegreen problem on ASR[3], and the research on the generation of the mondegreen itself is blank. Moreover, the research of misheard lyrics generation is insufficient especially in Chinese. Compared to current ASR modules, the soramiminet is aimed at generating the wrong results that best match the right one. The wave2vec model[4] can be trained to process audio clips and then generate pinyin of misheard lyrics in the clip. Thus, the collection and processing of high-quality audio files is one of the major topics of this project. In addition, this article will also introduce how to obtain suitable mondegreen material in Chinese.

The purpose of this module is not only to achieve mondegreen lyrics generation, but also to establish an acoustic-linguistic model which present the nature of human acoustic-linguistic signal processing.

DOI: 10.54254/2755-2721/4/20230490

This model can assist in testing human communication skills and mimic the audience's understanding of a particular conversation.

2. Literature review

2.1. Shortcoming and negative impacts of some datasets

No matter how complex and complete a model is, it requires high-quality data for training, validation, and testing. A bad dataset will not only lead to technical problems like overfitting and underfitting that will destroy the accuracy of the model, but also may cause deeper impacts on the scientific, social, and ethical aspects. The health systems relying on commercial prediction algorithms which use certain training set for ground truth exhibit significant racial bias. Black patients are potentially receiving 17.7% to 46.5% help as the system predict them to be "sicker" than white patients[5]. Furthermore, according to the study conducted by Stanford University[6], the cutting edge Automated speech recognition systems designed by Amazon, Apple, Google, IBM, and Microsoft are all exhibiting varying degrees of racial disparities. The average word error rate (WER) for black speaker is 0.35. On the other hand, the WER for white speaker is only 0.19. Moreover, the WER for men is typically higher than women. All these five systems are trained and acknowledged as the-state-of-art model.

Data annotation and labelling are another possible source of bad dataset. The quality of training data is critical to the success of supervised machine learning, in which models are automatically generated from labeled training data. Annotators, as humans with different cultural backgrounds and professional skills, will label the data with certain bias and shortcoming. The recruitment machine learning system built by Amazon team was determined to have gender bias[7]. This happened even the gender is not one of the training criteria. The system is trained to be more attracted by candidates who tend to use male engineering verbs. The main reason for the failure of this system is the system is that trained based on the HR teams and engineers with gender gaps in their workforce. As clearly indicated in the

Table **1**, the training details of labeler is not classified in 84% paper. Furthermore, merely half of the labeler was informed with specific definitions or examples before they start annotating.

Table 1. Labeler training details.

	Count	Proportion
Some training details	7 of 45	15.56%
No information	38 of 45	84.44%
Instruction with formal definitions or examples	21 of 45	46.67%
No instruction beyond question test	2 of 45	4.44%
No information	22 of 45	48.89%
Total of applicable papers (involving original human labeling)	45	100%

Note. The table was based on "Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?" by Geiger et al.

The dataset problem is not only reflected in small datasets, but also in benchmark datasets used by major companies. As shown in the research conducted by MIT, several benchmark datasets have labelling errors[8]. Among the evaluated six image dataset, even the best-case MNIST had the 0.15 percent error rate of the total labels. The three text test sets 20news, IMDB, and Amazon have error rates of 1.11, 2.9, 3.9 respectively. The only audio dataset AudioSet has an error rate of 1.35. The benchmark datasets used for training and testing are not as perfect as it supposed to be. A noisy and uncleaned dataset are one of the mean reasons for overfitting and underfitting issues[9]. Additionally, researchers tend to use existing dataset. Only 26.7% study across different disciplines are presenting new human-labeled dataset. Even if some teams mentioned the dataset, 64% of the reports did not specify how many labelers were on the team. 88% of teams did not provide a link to the dataset they were using[10].

DOI: 10.54254/2755-2721/4/20230490

2.2. How to create a high-quality dataset

The first thing one should do before collecting data is setting up objectives. In order to solve the problem effectively, questions about model architecture, training techniques, data acquiring, data labelling, and performance analysis must be considered[11]. Artificial neural networks can be trained using different algorithms. Based on the problem trying to solve, a suitable model can be chosen. The creator should determine whether to apply parallelism or Non-parallelism training techniques[12], and consider what type of data needed to be collected. Additionally, dataset builder needs to figure out what type of data is noisy and should be deleted, and how to measure the success of the model after the training is done.

2.2.1. Data collecting. Data collecting can be achieved by different methods. If the current dataset cannot meet the requirements for the project, the data generation is necessary. There are two main methods to generate new data, which are human creating and synthetic data generation. Crowdsourcing is a potential solution as it involves a large number of participants from different backgrounds. Although crowdsourcing may improve speed, quality, and diversity of the collecting process[13], it should be noted that the data collected through crowdsourcing may be unclean and biased. The researcher should examine all the data before processing it. On the other hand, some neural networks can be used to generate data. For instance, generative adversarial network[14].

Data augmentation is another method to modify data, so it matches to a certain project. An old dataset can be augmented with required external new sources. This technic helps to increase the efficiency of the data collection, since it reduces search requirement for new data. Possible approaches include data wrapping and synthetic over-sampling[15].

2.2.2. Data cleaning. Raw data cannot be directly used to establish a data set, since it may contain noisy data. The dirty data will reduce the accuracy of the model and increase the cost. The possible tools and framework include BoostClean and MLClean.

The BoostClean is able to select an ensemble of error detection and repair combinations using statistical boosting[16]. The system applies conditional repairs, which is a combination of data repairs and prediction repairs. Data repairs is aimed at modifying the data from the training record. Prediction repairs, on the other hand, will set the value to default if the input record is too corrupted. The main idea of boosting is to find the “weak” predictors in each iteration. These “weak” predictors will then be combined into one prediction rule which will be more accurate than all the “weak” ones. As demonstrated by the author, the BoostClean system successfully increased the accuracy of eight different models from Kaggle. Moreover, it improves the AUC of the downstream model by eight to nine percent.

The authors of MLClean[17] unify the three data preprocessing techniques, which are traditional data cleaning, unfairness mitigation, and data sanitization. The traditional data cleaning is based on integrity constraints denial constraint, and functional dependencies need to be satisfied. However, it has security and fairness issues. To deal with fairness issues during data cleaning, the machine learning community programmed unfairness mitigation system. In order to improve model fairness, these techniques typically trade off some model accuracy. The last technique is data sanitization, which aimed at cleaning the poisoned data before the training stage. The main problem with this system is that it cannot distinguish between erroneous data and poisoned data. The MLClean system applies data sanitization as an extreme cleaning and uses unfairness mitigation system to reduce the bias of the processed data. The final combined system yields an average accuracy of 0.71.

2.2.3. Data labelling. Data labelling is required for supervised machine learning model. The labelling process can be differentiated by various data types (e.g., text, image, audio, video). Manual labelling requires trained annotator to spend a lot of time to complete, which lead to increasing demand for labelling assistant software.

In case of image annotating, the labelling process can be divided into four categories according to the different automation degrees[18]. With the lowest automation level, manual labelling requires

DOI: 10.54254/2755-2721/4/20230490

humans to identify the features in the image. The software is aimed at providing clear user interface to support human accomplish the job more efficiently. There are various methods to achieve manual labelling. As the name single user labelling reveals, only one human is annotating the image. Although the labelling process can be proceeded offline, it may contain bias. Collaborative labelling allows a group to finish the annotation together. The same image can be marked by different members and the results are combined. Or it can be labelling by one person first, and then reviewed by another member. This method can improve the accuracy of the data. However, all the possible method requires duplicate work, which reduces the cost efficiency.

Semi-automated labelling allows algorithm to accomplish annotation first, and the results will be reviewed and corrected by human. The labelling time can be reduced. Although the accuracy is slightly inferior to manual labelling, the overall efficiency is relatively high. Automated labelling does not require any human interaction during the process. After the clustering model is trained, the computer will analyze the input and finish the annotating procedure. Automated labelling system produces a better result than manual labeling in some cases, since computer will not be influenced by bias that human may have. Different form semi-automated labelling, interactive labelling allows human to correct and update annotations during iterations. The annotating accuracy will increase as the system learn from its mistake.

3. Dataset creation

3.1. Data collection and preprocessing

Suitable mondegreen materials in Chinese is hard to discover, since the study of this subject is insufficient. The main source of the data is from different social medias and blogs. In view of the fact that misheard lyrics is a more subjective judgement, human review is required after the information is collected. After the manual verification is completed, label the selected clips, record the pinyin corresponding to the misheard lyrics, record the corresponding Chinese misheard and correct lyrics, and metadata such as the singer's name, song title and album name to facilitate the later search for the corresponding audio clips. After obtaining the full-song audio clip, the file format needs to be adjusted to make sure all the clips are in wav format. For instance, several clips in flac and mp3 file were transferred to wav file. All the clips are at 44.1kHz sampling rate, clips with sampling rate lower than 24Khz were replaced. In order to achieve preliminary audio and annotation alignment, the collected full-song clips also need to be edited, and only the clips corresponding to the misheard lyrics are retained. Clips vary in length from three to ten seconds, with most being four to five seconds.

DOI: 10.54254/2755-2721/4/20230490

3.2. Dataset content

0 ruoguotianheizhiqianlaidejiwoyaowaleniyanjing 如果天黑之前来得及我要挖了你的眼睛 如果天黑之前来得及我要忘了你的眼睛 南山南
1 jiangzhendehuibuhuishiwobeiguimixijiaole 讲真的会不会是我被闺蜜洗脚了 讲真的会不会是我被鬼迷心窍了 讲真的
2 yufenfenqiukulicaomushenwotingwennvshizhuyiigeren 雨纷纷秋裤里草木深我听闻女施主一个人 雨纷纷旧故里草木深我听闻你始终一个人 烟花易冷
3 shuangwaidemaquezaidianxianganshangluoshui 爽歪的麻雀在电线杆上裸睡 窗外的麻雀在电线杆上多嘴 七里香
4 nixialeyanwodeaiyichujiuxiangyushui 你瞎了眼我的爱溢出就像雨水 雨下整夜我的爱溢出就像雨水 七里香
5 zaigeiwolianggencongrangwobajiyijianchengbing 再给我两根葱让我把记忆剪成饼 再给我两分钟让我把记忆结成冰 最长的电影
6 bingxiangdebaleinaohaizhonghaizaixuanzhe 冰箱的八类脑海中还在选着 冰上的芭蕾脑海中还在旋转 最长的电影
7 tiandiyouyouguokechongchongchaoqiyouchaoluoenenyuanyuanshengsibaitoujirennengkantou 天地悠悠锅壳重重炒起又炒落恩恩怨怨胜似白头几人能砍头 天地悠悠过客匆匆潮起又潮落恩恩怨怨生死白头几人能看透 潇洒走一回
8 wonengxiangdaozuilangmandeshijiushiheniyiqimaimaidiannao 我能想到的最浪漫的事就是和你一起卖卖电脑 我能想到的最浪漫的事就是和你一起慢慢变老 最浪漫的事
9 weiweixiaoxiaoshihouwotamazhidao 微微笑小时候他妈我知道 微微笑小时候的梦我知道 稻香
10 yankanzheninanguowoliudehuaqumeiyoushuo 眼看着你难过我刘德华却没有说 眼看着你难过挽留的话却没有说 说好不哭
11 tangguoguanlihaoduoyanseweixiaoquegutianle 糖果罐里好多颜色微笑却古天乐 糖果罐裡好多顏色微笑却不甜了 明明就
12 gongqiaoxieposheitamadesheijide 拱桥斜坡谁他妈的谁记得 拱桥斜坡水岸码头谁记得 天涯过客
13 nisanshenmonanrensanshenmonanren 你三婶摸男人三婶摸男人 你算什么男人算什么男人 算什么男人
14 gushidexiaohuangguacongchushengnanianjiupaozhe 故事的小黄瓜从出生那年就泡着 故事的小黄花从出生那年就飘着 晴天
15 woqianzhenideshoujingguozhongmanyadandeshanpo 我牵着你的手经过种满鸭蛋的山坡 我牵着你的手经过种麦芽糖的山坡 麦芽糖
16 weishenmezhengyazinilazhewoshuoniyouxieyouyu 为什么蒸鸭子你拉着我说你有些鱿鱼 为什么这样子你拉着我说你有些犹豫 半岛铁盒
17 wogeinideaizhizhiqiyuanqian 我给你的爱只值七元钱 我给你的爱写在西元前 爱在西元前
18 bujiaonishushusuoyijiaoniyongbufushu 不叫你叔叔所以叫你永不服输 不想你输所以叫你用工读书 听妈妈的话
19 chongqianchongshizheshijielubanliuliankanzhetianmeisizaiyanqian 冲钱充实这世界鲁班六连看着天美死在眼前 从前初识这世间万般流连看着天边似在眼前 起风了
20 xiangzhefengyongbaocaikongyonggandexiangqianzou 想着疯拥抱踩空勇敢的想欠揍 向着风拥抱彩虹勇敢的向前走 你的答案
21 zheshifeiyangyangdegandie 这是沸羊羊的干爹 这是飞翔的感觉 勇敢的心
22 woyongguzhidekudangzuochengxingnang 我用固执的裤裆做成行囊 我用固执的枯藤做成行囊 光明
23 meiyigedanshenderendeikantouxiangaijiubiepashangtou 每一个单身的人得砍头想爱就别怕上头 每一个单身的人得看透想爱就别怕伤痛 单身情歌

Figure 1. Screen shot of labelling in text file format (made by the author).

Annotation: In text formatting with labelling, pinyin for misheard lyrics, misheard lyrics in Chinese, correct lyrics in Chinese, and the name of the track.

Figure 2. Audio files (made by the author).

Audio: High-quality audio clip files in WAV format with sample rate of 44.1KHz from 21 tracks by 9 different singers

DOI: 10.54254/2755-2721/4/20230490

4. Bias and limitations

The labelling bias of this database is relative serious because it applies single user labeling. Given that misheard lyrics is based on subjective opinions, the accuracy and correctness of the annotations are ambiguous. In the labeling process, although most of the adopted misheard lyrics were recognized insights online, some labels are completed by the labeler alone and no one else reviewed it. The language habit of the annotator also affects the accuracy of the dataset annotation to a certain extent.

Sample bias is another aspect. All music pieces in the dataset are in Chinese, partly because the labeler is a native Chinese speaker. Out of total twenty-three clips, fifteen tracks are from the same singer, which may be another source of bias. Additionally, of the nine singers in total, only three of them are women, this might leads to bias too.

One of the main limitations is audio noisy. Since the selected clips are all popular songs, while the singer is singing, the musical instruments and synthesized sounds in the background will greatly affect the model's recognition of the lyrics. Especially in clip 19, the singer sings too fast, and the vocals and background sounds cannot be clearly separated. Clips 7 and 8 are due to the earliest release time of the song and limited by the recording technology at that time, the clarity of the song is slightly worse than other pieces. Due to the limitation of the editing software, the clip clips are accurate to the second at most, resulting in some clips that are slightly shorter or longer than the expected clips of mondegreen. Clips 1, 17, 18 and 22 are mostly affected, which may lead to bad synchronization between audio and annotation during training process. Another limitation is that the sample size of the dataset is too small to train a complete model. Moreover, the low generality of the dataset makes it difficult to train other models.

5. Conclusion

Mishearing is a common linguistic phenomenon that occurs during conversation, and more often when listening to music. Through the method of machine leaning, the mapping of acoustics to linguistics is achieved by inputting audio clips and outputting the corresponding misheard lyrics. The purpose of this project is not only limited to entertainment, but also to explore the reasons for mishearing. This paper reviews the issues associated with building datasets; a flawed dataset will affect the accuracy of the model, furthermore, it also affects the daily lives of real people, such as the Amazon recruitment discrimination incidents. Three aspects: data collection, data cleaning, and data labeling are discussed in relation to how a high-quality dataset is built. Data collection can be done through full collection or data augmentation based on existing data. Two possible data cleaning method, BoostClean and MLClean are introduced. In the data labeling section, this paper focuses on the four different automation levels of image labelling.

The main achievement of this project is the initial establishment of a dataset generated for Mondegreen. The dataset includes 21 music tracks from 9 different singer. Bias and limitations of the datasets are also discussed in this paper.

References

[1] S. Wright, *Get away from me with those Christmas gifts, and other reactions*. New Yor k,: McGraw-Hill, 1957. Accessed: Jul. 22, 2022. [Online]. Available: http://archive.org/details/getawayfrommewit00wrig

[2] “mondegreen, n.,” *OED Online*. Oxford University Press. Accessed: Jul. 22, 2022. [Online]. Available: http://www.oed.com/view/Entry/251801

[3] S. S. Sodhi *et al.*, “Mondegreen: A Post-Processing Solution to Speech Recognition Error Correction for Voice Search Queries,” 2021, Accessed: Jul. 22, 2022. [Online]. Available: https://arxiv.org/abs/2105.09930

[4] “Fine-Tune Wav2Vec2 for English ASR in Hugging Face with 🤗 Transformers.” https://huggingface.co/blog/fine-tune-wav2vec2-english (accessed Jul. 22, 2022).

[5] “Dissecting racial bias in an algorithm used to manage the health of populations.” https://www-

DOI: 10.54254/2755-2721/4/20230490

science-org.proxy.queensu.ca/doi/10.1126/science.aax2342 (accessed Jul. 24, 2022).
[6] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proc. Natl. Acad. Sci.*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020, doi: 10.1073/pnas.1915768117.
[7] "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 10, 2018. Accessed: Jul. 25, 2022. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[8] "Label Errors in Benchmark ML Datasets." https://labelerrors.com/ (accessed Jul. 24, 2022).
[9] "What is Overfitting?," Mar. 06, 2021. https://www.ibm.com/cloud/learn/overfitting (accessed Jul. 24, 2022).
[10] R. S. Geiger *et al.*, "'Garbage In, Garbage Out' Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?," *Quant. Sci. Stud.*, vol. 2, no. 3, pp. 795–827, Nov. 2021, doi: 10.1162/qss_a_00144.
[11] "How to Create a Dataset to Train Your Machine Learning Applications," *Kili Technology*. https://kili-technology.com/blog/create-dataset-for-machine-learning (accessed Jul. 25, 2022).
[12] "Techniques for Training Large Neural Networks," *OpenAI*, Jun. 09, 2022. https://openai.com/blog/techniques-for-training-large-neural-networks/ (accessed Jul. 25, 2022).
[13] R. Buettner, "A Systematic Literature Review of Crowdsourcing Research from a Human Resource Management Perspective," Jan. 2015. doi: 10.13140/2.1.2061.1845.
[14] "Create Data from Random Noise with Generative Adversarial Networks," *Toptal Engineering Blog*. https://www.toptal.com/machine-learning/generative-adversarial-networks (accessed Jul. 25, 2022).
[15] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" arXiv, Nov. 26, 2016. Accessed: Jul. 25, 2022. [Online]. Available: http://arxiv.org/abs/1609.08764
[16] S. Krishnan, M. J. Franklin, K. Goldberg, and E. Wu, "BoostClean: Automated Error Detection and Repair for Machine Learning." arXiv, Nov. 03, 2017. Accessed: Jul. 26, 2022. [Online]. Available: http://arxiv.org/abs/1711.01299
[17] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, and S. E. Whang, "Data Cleaning for Accurate, Fair, and Robust Models: A Big Data - AI Integration Approach." arXiv, Apr. 22, 2019. Accessed: Jul. 26, 2022. [Online]. Available: http://arxiv.org/abs/1904.10761
[18] C. Sager, C. Janiesch, and P. Zschech, "A survey of image labelling for computer vision applications," *J. Bus. Anal.*, vol. 4, no. 2, pp. 91–110, Jul. 2021, doi: 10.1080/2573234X.2021.1908861.