

A Supplementary Materials for KeSpeech

A.1 Datasheets

These questions were copied from "Datasheets for Datasets"¹.

A.1.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
This dataset involves the most spoken Chinese dialect Mandarin and its 8 subdialects, supporting multiple speech tasks, such as speech recognition, speaker recognition, subdialect identification and their multi-task learning and conditional learning on each other.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
Beike company and Tsinghua University together.
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
Beike company.
- **Any other comments?**
No.

A.1.2 Composition

Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Speech and its transcription.
- **How many instances are there in total (of each type, if appropriate)?**
1,542 hours of audio.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
It contains all possible instances.
- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Raw data, but it presets several tasks by preparing training set, development set and test set.
- **Is there a label or target associated with each instance?** If so, please provide a description.
Speech and its transcription, subdialect type. Speaker and identify number.
- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This

¹<https://arxiv.org/pdf/1803.09010.pdf>

does not include intentionally removed information, but might include, e.g., redacted text.
No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.
N/A.
- **Are there recommended data splits (e.g., training, development/validation, testing)?**
If so, please provide a description of these splits, explaining the rationale behind them.
It presets several tasks by preparing training set, development set and test set.
- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
The labeling of subdialect type is of much subjectivity for some accented utterances that can not be classified definitely into Mandarin (standard Chinese) or subdialect, and each utterance was just labeled by one native people, lacking of quality control.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
Self-contained.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.
No.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.
No.
- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
Yes.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.
No.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.
No.
- **Any other comments?**
No.

A.1.3 Collection Process

As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior section, the answers to questions here may provide information that allow

others to reconstruct the dataset without access to it.

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
Collected by recording with reading style.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?
Via an application.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
N/A.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
Crowdworkers, and were rewarded with virtual currency.
- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
N/A.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
Yes. The legal counsel of our company reviewed it.
- **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.
Yes.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
Via an application.
- **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Yes. Candidates had to sign an agreement for open-sourcing their speech for academic usage.
- **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Yes. Candidates had to sign an agreement for open-sourcing their speech for academic usage.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
Yes. It can be read from the company database.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

- **Any other comments?**

No.

A.1.4 Preprocessing/cleaning/labeling

Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks involving word order.

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.
No.
- **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the “raw” data.
- **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.
- **Any other comments?**

A.1.5 Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

- **Has the dataset been used for any tasks already?** If so, please provide a description.
Yes, as presented in the paper.
- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.
Yes. See the paper.
- **What (other) tasks could the dataset be used for?**
Speech related, such as speech factorization.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No.
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
No.
- **Any other comments?**
No.

A.1.6 Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
No.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
Self-contained website.
- **When will the dataset be distributed?**
Before the conference.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
CC BY-NC-SA.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.
- **Any other comments?**
No.

A.1.7 Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

- **Who is supporting/hosting/maintaining the dataset?**
Beike company.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
Email. tangzhiyuan001@ke.com
- **Is there an erratum?** If so, please provide a link or other access point.
tangzhiyuan001@ke.com
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
No.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.
No.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to

users.

No.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Under the same license.

- **Any other comments?**

No.

A.2 Responsibility Statement

The authors declare that they bear all responsibility for violations of rights related to this dataset.