GreenKey

Shared learning for improved model accuracy

Like many machine-learned speech recognition engines, GreenKey learns to recognize speech by training from human-generated audio and human-generated transcripts.

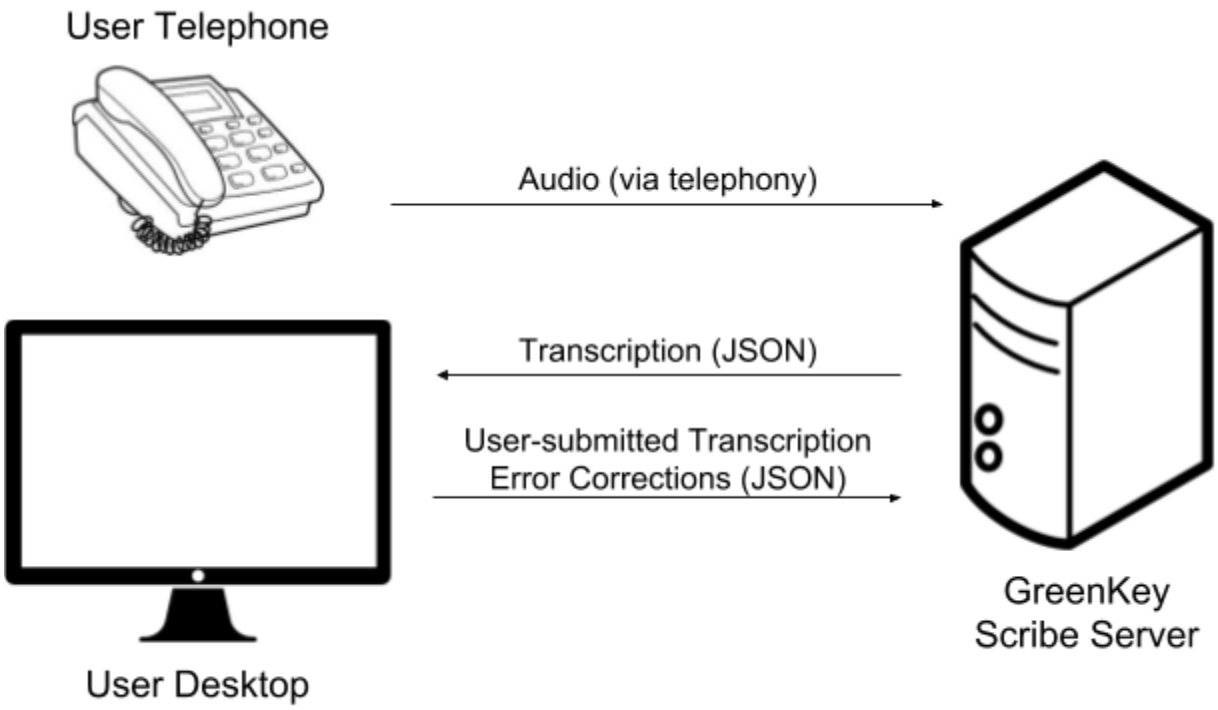
Unlike other engines, GreenKey's base model can learn from customized models without having access to the original raw data.

This ability allows everyone in the GreenKey community to benefit from shared learnings without sacrificing data privacy.

This whitepaper describes the process by which GreenKey's Scribe speech recognition engine learns from user-generated error corrections and shares models in a privacy-safe way.

Model Learning via Error Corrections

GreenKey deploys it's speech recognition engine Scribe on-premise at a customer site or in a GreenKey-hosted cloud environment. As Scribe is used, users can correct mistakes that the engine makes. These "error corrections" are saved for the engine to learn from on-premise.



Architecture diagram of data flows during transcription and error correction.

The error corrections contain two parts:

1) Raw audio data and the corrected transcript
2) Hashed signature of the data

The on-premise training trains from the raw data (1) and creates a record of accuracy improvements using the hashed signature (2). The data in (1) is transient and never leaves customer premises. Having both data types allows a firm's accuracy to improve over time through model training entirely within their premises while also providing a way for data to be shared and tracked for model combination.

```
{'dev': {'efa3b793b5f979d96412c93f1b33b63dc7c98838adf46c8b00b09ac922da1a5011997e24dae9bf71': {'audio_file':
'input-data/dev/TejasSetBTest.sph',
   'audio_file_hash': 'efa3b793b5f979d96412c93f1b33b63dc7c98838',
   'transcript_file': 'input-data/dev/TejasSetBTest.stm',
   'transcript_file_hash': 'adf46c8b00b09ac922da1a5011997e24dae9bf71'}},
 'test': {'0b3cf9544edf0bc72f6bdb1992d06c3c36cfb9c3499030fcfb3721f0a69ea9bea451cc0e8951377f': {'audio_file':
'input-data/test/bgc20170201.sph',
   'audio_file_hash': '0b3cf9544edf0bc72f6bdb1992d06c3c36cfb9c3',
   'transcript_file': 'input-data/test/bgc20170201.stm',
   'transcript_file_hash': '499030fcfb3721f0a69ea9bea451cc0e8951377f'},
  '1121f4f4a538e9779e5502a4b7b626a719b8e5450f19e948fe7fb67abf3d535d23b751bb3329e002': {'audio_file': 'input-
data/test/generated_dollarswap_gain_bdl2.sph',
   'audio_file_hash': '1121f4f4a538e9779e5502a4b7b626a719b8e545',
   'transcript_file': 'input-data/test/generated_dollarswap_gain_bdl2.stm',
   'transcript_file_hash': '0f19e948fe7fb67abf3d535d23b751bb3329e002'},
  '40aa85db33271fe4827d6822fee0cee27efd27ff22d6f09584ab2227593d8fbfb386ba7fe8946cc0': {'audio_file': 'input-
data/test/tradeticketfixedincome.sph',
   'audio_file_hash': '40aa85db33271fe4827d6822fee0cee27efd27ff',
   'transcript_file': 'input-data/test/tradeticketfixedincome.stm',
   'transcript_file_hash': '22d6f09584ab2227593d8fbfb386ba7fe8946cc0'},
  'a751b0680ba1443ef41f86a9dceb020257725ca89ede785fb15efb9953bfed469ce43a6417fbbf54': {'audio_file': 'input-
data/test/generated_dollarswap_gain_c1b.sph',
   'audio_file_hash': 'a751b0680ba1443ef41f86a9dceb020257725ca8',
   'transcript_file': 'input-data/test/generated_dollarswap_gain_c1b.stm',
   'transcript_file_hash': '9ede785fb15efb9953bfed469ce43a6417fbbf54'},
  'e5a90da185148798d3f15dc39cca6a180ea9865cacf2a72db8a7f9db7ee4ca7ab94c6c390c258406': {'audio_file': 'input-
data/test/bgceuroswap20170307.sph',
   'audio_file_hash': 'e5a90da185148798d3f15dc39cca6a180ea9865c',
   'transcript_file': 'input-data/test/bgceuroswap20170307.stm',
   'transcript_file_hash': 'acf2a72db8a7f9db7ee4ca7ab94c6c390c258406'}},
...
```

An example of data described above in (1) that is transient and stays on premise.

```
{'dev': ['efa3b793b5f979d96412c93f1b33b63dc7c98838adf46c8b00b09ac922da1a5011997e24dae9bf71'],
 'test': ['a751b0680ba1443ef41f86a9dceb020257725ca89ede785fb15efb9953bfed469ce43a6417fbbf54',
  '1121f4f4a538e9779e5502a4b7b626a719b8e5450f19e948fe7fb67abf3d535d23b751bb3329e002',
  'e5a90da185148798d3f15dc39cca6a180ea9865cacf2a72db8a7f9db7ee4ca7ab94c6c390c258406',
  '0b3cf9544edf0bc72f6bdb1992d06c3c36cfb9c3499030fcfb3721f0a69ea9bea451cc0e8951377f',
  '40aa85db33271fe4827d6822fee0cee27efd27ff22d6f09584ab2227593d8fbfb386ba7fe8946cc0'],
 'train': ['b3c8efb9f59cb315b5269cd89a2633b5ce47897d3a8993fe0e52bcadc634e577a681640fad8fd658',
  '8be813a361870a4363edc4c883d71a9f71fe175a1cc8ede636442e7b4be7409ffba3354ccd10537d',
  '0917616a8254b47df5ddb5a7c0cd5e32fbe6c7bf9558a0c5d317e797ecb66ecfe501840a5b628ee5',
  '0ee9734e78e97cdb7f5a6d5a6c534e692c5e62af9ecc7515ced0f11faa831935faabc1cb72199a2a',
  '280e58ac48b3d32a16108eca1a91916ef2c43e45a4296b068aac195f5b2dbc866a2bf172fe0fb996',
  '1f6460747d1c67c048a3412c4b0398968db345ad495e200af712d98a4acb74ea4bc8143df4b34311',
  'bbead96d93d662c717d159e0c57a049bc8bb37a05c23b934b47814694b17d42f42791366c73f1ec3',
  '315ea901e54aec248ab541d7d90685bd8c5376d24e6c2e2684e098f0c16f64a860010d510ff4c38f',
  '5d41d9ff1f0f820e8e3f8ff94d23a44572b7658cdbb0f20b8f45b41531cd5e08ba48d173a7e517eb',
  '226918f92dd00a9fd9bd069638ac130a9071e5d7e8ab123c6f3a1ecd71eceb73966bb58b593f399a',
  '27b6bd0f929712695b741ae4392232a094a3b3faf06c030892992808173eff390cd1dae710fb2c67',
  '0c9e022ea5838838bb5d5cede37c8238848b88bb76662ade3a4622e0afd8a4fdb72793227cfa01c8',
  'd71a0b444ea3bec39175bf39296f7197ee50bf5b7b281e1841923ef6ef1c66ebe7756c2101396a2c',
  'cb804d74778dad32cc1cd42981644e2c2e08754aa49572da527f4d53f1f961af167b44f7fb0bb4d5',
  'e56092efcc507e6cdd7783b660b4a3997fd9b7dd12eb2d76eb20e131a9700342392f07b84b479649',
  '32d23412fb18802fc9b0ff492b29dc9171dab8ecde653d76b972907b920c675f040558f27d638bf4',
  '3a0725a999cce16da67e7ede550ec4f36a2beda0958ef84913bf85e7b09fb93c82a78ba95f001c16',
  'dfb1fab4caeca0ec70e68eb6e06518bcf6aa5dd322c7ea2e871bfd56bfb6d981036dfa3e86073544']}
```


An example of data described above in (2) training corpus of hashed files split into “development”, “training”, and “test” sets.

Data retention

By default, GreenKey does not retain any audio or transcription long term. The only data retained is the trained model.

Intra-day recordings and transcriptions are typically stored on a RAM disk. This type of disk stores data in the temporary memory within a server. When the server is restarted, the data is wiped. Data can also be wiped periodically based on any desired purge policy. If desired, data can be sent to a downstream system for long term retention.

Model training process

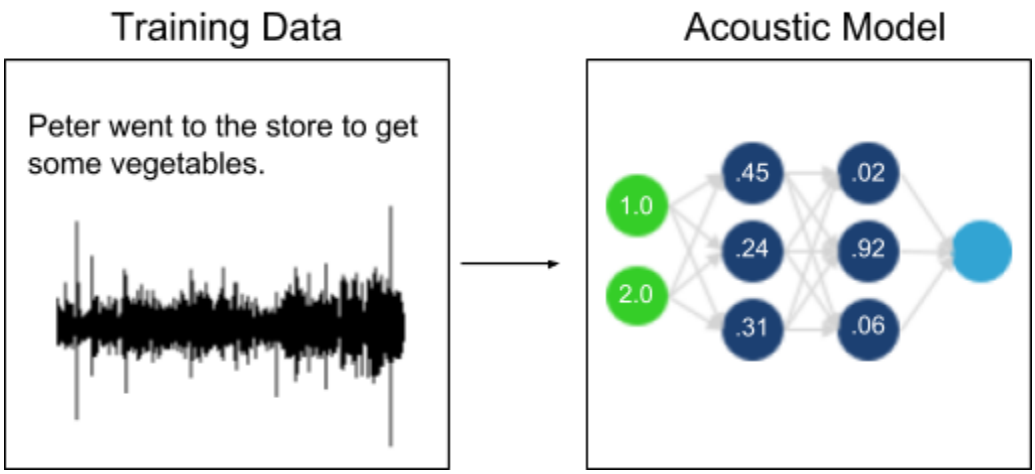
Model training proceeds by extending a GreenKey-provided “base model” using the error corrections as new data. Thus, on-premise model training is never “starting over” and is only additive to the base model provided. The base model can be updated by GreenKey at any time, and model training can then proceed seamlessly from the new “base model”.

When a model is done training, a number of files are generated. These files are all non-reversible representations of the data. The original source data can never be reproduced or identified. The hashed signature of the data is later combined with an accuracy metric to signify that the model improved over time.

GreenKey’s models have two components -- acoustic and language models. Both of these models can be trained, shared, and combined together or independently.

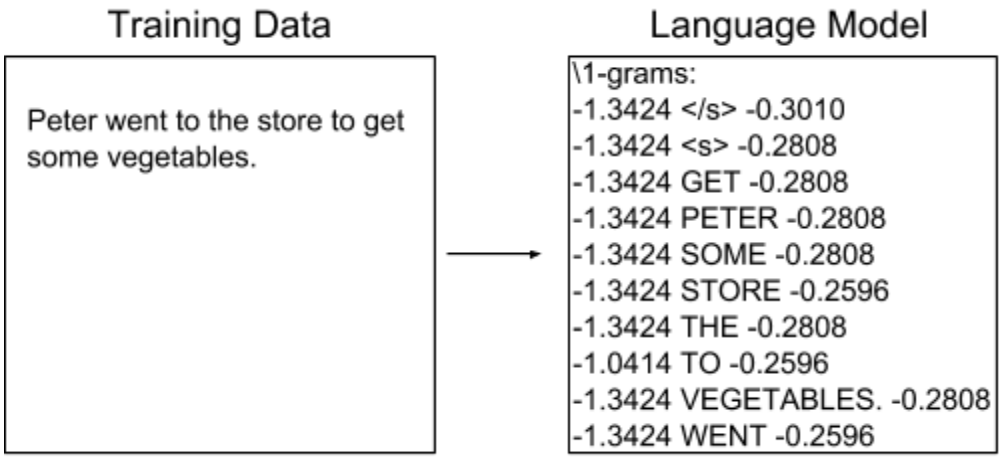
Acoustic Models

The acoustic model contains a large mathematical function that takes audio bytes as an input and outputs potential phonemes. The content of the acoustic model cannot be used to reconstruct the original audio the model was trained on.



Language Models

The language model contains a map of the probability of words and phrases. If GreenKey learns a completely new word that it has never heard before, it will exist inside the language model with a probability provided. The original data cannot be constructed from this model, though the model will contain matches for new words it has learned (like proper nouns).



An example language model based on the training data shown. GreenKey uses standard ARPA formatted n-gram language models. You can try to create a language model yourself to see how your data is transformed here.

A firm can choose to only train the acoustic model and not the language model, or vice versa.

Privacy-safe data verification

Diagnostic information is shared with GreenKey to ensure that a model is improving in accuracy without showing any original data. The hashed signatures and accuracy measurements are a privacy-safe way for GreenKey and a client to assert that a model is indeed improving over time. If new data is introduced that improved or reduced model quality, either party can identify that data.

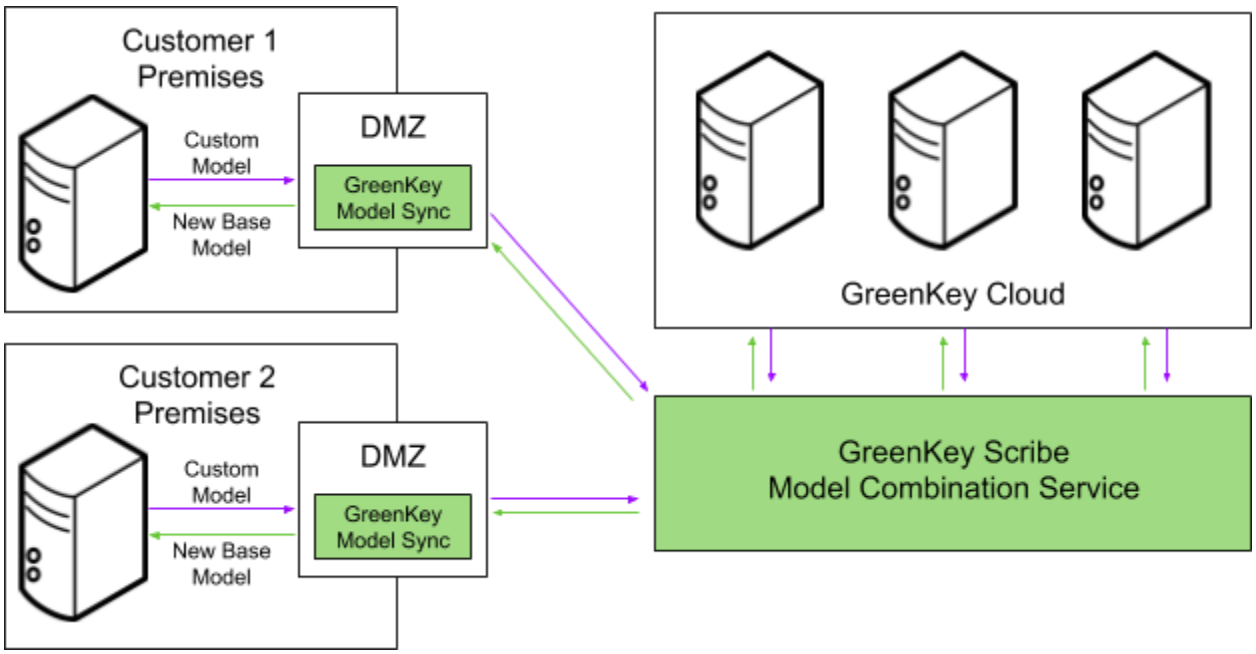

```
'0225c86c34283320a8fd2e25c2633615daafcfc4a033806fcf6aff08a397d6ab98b741265b09e00d': {'num_words': 684,
 'wer': 22.953216374269005},
'026659f9d83cb66ffc33664580ba610bc2244ac089266918daa24f72fcee5356be2b8533f6e56a87': {'num_words': 17,
 'wer': 88.23529411764706},
'0294683d31f1d5d7072d64b5a4d2aa8fd035d287c73cba07b20c9538a2cbbe2dbaf1acad8bab44f7': {'num_words': 11,
 'wer': 63.63636363636363},
'02c7456f2f0ff9c00e3ddd056b6b1dc34351cb4d4b4fc5fd2c422b2c8189966a7e80f37f5ce8a6d5': {'num_words': 29,
 'wer': 44.827586206896555},
'0347af89844389352c66ee090d0ef54fe4834a451bc93bc40d724a23944f30aff8f84bbfb9ec91e4': {'num_words': 12,
 'wer': 50.0},
'03817e2859d67446469072d607e2c96abaff06a5a580854473fc687b752aa92569142f19f2c39796': {'num_words': 8,
 'wer': 75.0},
'03bab42cd32b95d916f81a5e420981b90630faa373ea7f6e294dcd87b2aa552a435c74c52db9774c': {'num_words': 30,
 'wer': 80.0},
'042c2b56cac6f6e07f21d71c93595b61976cc24fe6f7abda75243da9fc0da95bd3a20470a5e3731a': {'num_words': 10,
 'wer': 40.0},
..
```


An example of model accuracy statistics, showing hashed files and the Word Error Rate (wer) that the model achieves on each file.

Model Combination

Because the model has no personal identifying information (PII), the models can be freely shared between the client and GreenKey. GreenKey can then combine the learnings from each model through its proprietary process to improve accuracy for all users.

Models can be manually sent to GreenKey or moved to a DMZ where GreenKey will automatically receive your model, learn from this, and provide you the latest, improved base model.



Architecture diagram of data flows during model combination.

By sharing your model, you benefit from the shared learnings of the community without exposing any confidential information.