## 4. DEVELOPMENT

Our development stage consists of three steps: (1) training emotion-recognition models, (2) training service-problem-identification models, and (3) implementing the *SPICE* system in the company’s call centers. We describe the three steps in turn below.

### 4.1 Training Emotion-Recognition Models

Company A provided proprietary data—a sample of service call records—for training our models. Company A recorded and saved its service calls in a Waveform (\*.wav) format (sampling rate: 8000Hz; bit depth: 16 bit). The sample includes 2,400 calls randomly selected from service calls in the year right before the start of our research project. We randomly choose 1,000 of them to train and test our emotion recognition model and the rest of them for our service problem identification model.

***Pre-processing.*** Upon obtaining the sample calls, we went through data pre-processing steps as described in the System Design section, including noise reduction, chopping a call into a series of consecutive speech segments, speaker recognition, speech-to-text conversion, and text preprocess. The data fed into the training of emotion-recognition models are customers’ speech segments.



***Labeling speech segments.*** We formed a research teamincluding CSRs of Company A and research assistants of our research project to label speech segments by tagging each segment whether or not the speaker has a negative emotion. Training (of team members) is necessary to make sure all team members use the same standard to determine negative emotions. During this training stage, all team members worked on the same batch of service calls (*N*=100) and discussed inconsistent labeling. Experienced QA managers of the company and coauthors of the paper led the discussion to make sure the team’s understanding of negative emotions is consistent across team members and with QA managers’. After the training, the team members formed two groups, one for labeling and the other for double check. Because we would train two types of models—acoustic models and linguistic models—for emotion recognition, the labelers assessed each speech segment and gave it two labels. One is “acoustic label”—indicating whether or not the labeler can identify negative emotions from the speaker’s sound per se; the other is “linguistic label”—indicating whether or not the labeler can identify negative emotions from the words used by the speaker.

***Training and testing.*** We trained the acoustic models based on acoustic features and acoustic labels. After the labeling work, we separate the segments into three parts: train set, validation set and test set. We make sure that the segments of one call only appear in the same dataset. Since the data is highly imbalanced (with much more non-negative segments than negative ones), we take a undersampling method (Chawla 2009) to train model. In this way, we randomly select 2,000 “negative segments” (i.e., segments with negative emotions as indicated by the acoustic labels) and 2,000 “non-negative segments” (i.e., without negative emotions as indicated by the acoustic labels) from the training set. We arranged the training and testing data linguistic models in a similar way. Acoustic models and linguistic models share the same validation set and test set for model comparison. In validation and testing, the label is a combination of “acoustic label” and “linguistic label” (i.e., either “acoustic label” or “linguistic label” is negative, the label is negative). The next section presents the results.

Table \*\* Segment numbers of train, validation and test set



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Train set for acoustic models | Train set for linguistic models | Validation set | Test set |
| Negative segments # | 2000 | 2000 | 635 | 322 |
| Non-negative segments # | 2000 | 2000 | 5718 | 1659 |

### 4.2 Training Service-Problem-Identification Models

### As discussed earlier, the goal of training service-problem-identification models is to estimate the parameters in call’s service problem likelihood— (see the discussion above on eq. #)—so that the sequence of a call’s negative emotions can predict the call’s service problems. To do so, we need training data with labels of “having a service problem” (yes or no) for each call.

***Labeling service problems***. We used two sources of information provided by the company to label service problems. First, the company conducts routine operations of QA evaluations. QA specialists listen to sampled service calls, evaluating the calls using a list of QA questions (e.g., whether the CSR knows the key to solve the problem). Second, the company routinely surveys customers, asking them to answer a list of questions regarding the services they received (e.g., whether the CSR has a good service attitude). As discussed in the Background part, these two methods have their merits in assessing service quality for company and we combine them to reflect service quality more completely. We selected items from these two sources to measure four dimensions of the widely used SERVQUAL scale (reliability, responsiveness, assurance, and empathy). Prior research suggests that the tangibles dimension (e.g., equipment, employee appearance, etc.) of the SERVQUAL scale is not relevant to service quality in service calls (Hsieh et al. 2011). Table XXX lists the items we selected from the two sources.[[1]](#footnote-4) We labeled a service call to have a problem on one dimension if at least one item under that dimension is “no”.

Table XXX. SERVQUAL Dimensions and Items

|  |  |
| --- | --- |
| Dimensions | Items |
| Reliability | The CSR knows the key to solve the problem (yes/no). |
| The CSR provides problem solution properly (yes/no). |
| The CSR takes service note correctly (yes/no). |
| The CSR’s service note has no error (yes/no). |
| Responsiveness | The CSR has a good service attitude (yes/no). |
| The CSR provides service actively (yes/no). |
| The CSR gives prompt service (yes/no). |
| Assurance | The CSR uses service words properly (yes/no). |
| The CSR is professional (yes/no). |
| The CSR protects customer privacy (yes/no). |
| Empathy | The CSR understands the customer’s needs (yes/no). |
| The CSR provides service patiently (yes/no). |
| The CSR’s voice has proper tone, manner and speed (yes/no). |

***Training and testing.*** The training data includes 1,400 service calls. This is a balanced sample in that half of them have no service problem. We use the training data to determine a call’s service problem likelihood—i.e., . The testing data consist of 2,704 calls, which happened during one week after the system implementation. As mentioned earlier, the company evaluate 2 percent (about 400 calls) of all the calls each day. All the calls in our test set had been inspected by the QA specialists and through customer surveys (thus having information for the needed SERVQUAL items as shown in Figure 3). Section 5 presents the model evaluation results.

**4.3 Implementation**

We implemented the *SPICE* system in call centers of Company A. The system went live in September 2017. All the five modules as shown in Figure XXX—pre-processing, feature extraction, emotion recognition, service problem identification, and user interfaces—are at work.

[report some system technical specifications somewhere, e.g. latency, tools used, programming language, etc] @@问问饶鹏，写几句描述系统架构@@ Our UI design has gone through an iterative process after the implementation of the system.

As part of the implementation, we worked with the company to redesign its service QA procedure in call centers: the *SPICE* system constantly process service call records, sort calls in terms of their that can predict the likelihood of service problems, and show the results on the QA-UI. The QA specialists can look at the results (the left panel of Figure XXX), where calls with the highest likelihood of having service problems appear on the top, select certain calls to listen, and then report issues and problems, if any, to their supervisor. This is an on-going process; that is, after a call is finished, it goes into a queue awaiting the system to process. The processing time is short so that the *SPICE* system identifies the likelihood of its service problems and reports to the QA specialists in real time.

We also worked with the company to help CSRs in call centers to get familiar with the CSR-UI. During each service call, the CSR-UI visualizes customer emotions over time (Figure XXX). It serves as a virtual memory of customer emotions during the service encounter. This UI is a built-in feature of the call center system and CSRs have the discretion to open it. We have obtained feedback from CSRs who used the feature, telling us that they found it useful as it helped them easily notice the evolutionary path of customer emotion during the phone call, so that they can choose service tactics accordingly. A call center manager commented that this is useful particularly when CSRs are tired and get less attentive to customer emotion. Overall, call center managers including the director are satisfied with the system.

1. The original items (used by both the QA team and customer surveys) are in Chinese. We translated them to English, as shown in Table 3. A call center manager of the company verified the correctness of the translation. [↑](#footnote-ref-4)