Multi-Grained Role Labeling Based on Multi-Modality Information for Real Customer Service Telephone Conversation

Weizhi Ma, Min Zhang*, Yiqun Liu, Shaoping Ma
State Key Lab of Intelligent Technology & Systems; Tsinghua National TNLIST Lab
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
mawz14@mails.tsinghua.edu.cn, {z-m,yiqunliu,msp}@tsinghua.edu.cn

Abstract

Large-scale customer service call records include lots of valuable information for business intelligence. However, the analysis of those records has not utilized in the big data era before. There are two fundamental problems before mining and analyses: 1) The telephone conversation is mixed with words of agents and users which have to be recognized before analysis; 2) The speakers in conversation are not in a pre-defined set. These problems are new challenges which have not been well studied in the previous work. In this paper, we propose a four-phase framework for role labeling in real customer service telephone conversation, with the benefit of integrating multi-modality features, i.e., both low-level acoustic features and semantic-level textual features. Firstly, we conduct ΔBayesian Information Criterion (ΔBIC) based speaker diarization to get two segments clusters from an audio stream. Secondly, the segments are transferred into text in an Automatic Speech Recognition (ASR) phase with a deep learning model DNN-HMM. Thirdly, by integrating acoustic and textual features, dialog level role labeling is proposed to map the two clusters into the agent and the user. Finally, sentence level role correction is designed in order to label results correctly in a fine-grained notion, which reduces the errors made in previous phases. The proposed framework is tested on two real datasets: mobile and bank customer service calls datasets. The precision of dialog level labeling is over 99.0%. On the sentence level, the accuracy of labeling reaches 90.4%, greatly outperforming traditional acoustic features based method which achieves only 78.5% in accuracy.

1 Introduction

Call center plays an important role in customer service of many kinds of companies, such as retailer, bank, mobile service and e-commence. There are some self-service platforms, but currently the quality of automatic processing is not able to meet users' complex requirements. In these cases, users still prefer to give a call to the customer service call center for help.

Customer service calls include many valuable information. We can get the hot topics, the problems about products and other information that customers are concerned about, and they are helpful for improving product quality. On the other hand, call center service satisfaction can be evaluated according to the conversation content. As far as we know, the analysis in customer service is based on human analysis with sampling. However, it is possible to conduct the analyses in an automatical way in the big data era.

Role recognition is a fundamental work of such automatic analysis. While there are still two problems in telephone conversation role recognition: 1) A telephone call is a continuous audio stream which records the mixed information conveyed by users and agents. Therefore, roles have to be separated in conversations. 2) Unlike previous speaker recognition studies, the speakers in this study are not in a pre-defined set. We do not know who will call customer service for help and there may be thousands of agents online to answer the calls. Before analyzing the customer service calls, the two problems have to be addressed to get satisfactory role labeling results.

In this paper, we propose a four-phase framework for role labeling in practical customer service telephone conversation based on acoustic and textual features. Different from most of the previous work which conducted speaker recognition only with acoustic features, we tried to integrate low level acoustic features with high level textual features. Moreover, we designed a text-based post-processing with the help of semantic information in the conversation to reduce the errors accumulated in previous phases. The results indicate that our framework performs better than the single modality work.

We applied this model to two actual customer service telephone conversation role labelings, a mobile service and a bank dialog datasets. The precision of clusters and roles mapping in both datasets in dialog level is over 99.0%. Compared with only using acoustic features, the accuracy of sentence level labeling achieved 90.4%. which increased by 11.9%.

The main contributions of the work are:

- We propose a uniformed role labeling framework which utilizes both low level acoustic features and high level text based features. Most of the previous work only concentrated on making use of acoustic features.

*Corresponding author

- It is a multi-modality role labeling framework including two labeling steps: dialog level role labeling and sentence level correction which are able to reduce the mistakes generated in previous phases.
- The proposed approach is domain-independent and can be successfully applied to real scenarios as the groundwork of large scale data analysis.

The remainder of this paper is organized as follows: we introduce related work in Section 2. The overview of the proposed four-phase framework is shown in Section 3. We give the detailed description of each phase in Section 4. In Section 5, we introduce experimental settings with real customer service dialog datasets and report the comparative results. The conclusions and the outline of future work are drawn in Section 6.

2 Related Work

In previous work, there are several topics which are related to our role labeling work: speaker diarization and role labeling, automatic speech recognition, and multi-modality work.

Speaker Diarization and Role Labeling

Speaker diarization focuses on grouping speech segments according to the speakers in an audio stream [Tranter *et al.*, 2006], which is critical for automatic audio transcription [Tranter *et al.*, 2006], spoken document retrieval [Wang, 2004] and speaker recognition [Zhou *et al.*, 2012]. It has been studied in conversational telephone speech [Zhao and Fan, 2004], broadcast news data [Barras *et al.*, 2006] and other fields [Pardo *et al.*, 2007].

Speaker diarization systems usually include two core parts: speaker segmentation and speaker clustering. Speaker segmentation splits the audio stream into segments. Window-growing-based segmentation [Zhou and Hansen, 2005], fixed-size sliding window segmentation [Malegaonkar *et al.*, 2007] and DISTBIC [Delacourt and Wellekens, 2000] are three popular distance-based segmentation approaches. Then, speaker clustering step groups the segments into speaker clusters, and Hierarchical Agglomerative Clustering (HAC) is usually applied to speaker clustering [Barras *et al.*, 2006]. Each cluster contains the speech segments produced by a speaker. Several ΔBIC based speaker diarization methods are proposed in [Cheng *et al.*, 2010]. Some studies, like [Katharina *et al.*, 2005], [Zhang and Tan, 2008] and [Das, 2011] are aimed at speaker recognition based on acoustic feature.

But different from our work, in most of the previous work, the speakers are in a pre-defined set and the speaker role is indistinct in most of previous work. Therefore, the models cannot be applied to role labeling in practical telephone conversations.

Automatic Speech Recognition

The goal of automatic speech recognition is to handle different speaking styles, channels and environmental conditions as effectively as human does. In the past years, Gaussian Mixture Model (GMM) has remained as the state-of-the-art model to compute probabilities of Hidden Markov Models (HMM) in ASR fields. HMM-GMM based ASR models obtain notable performance with some parameters adjusting methods, such as minimum Bayes risk [Gibson and Hain, 2006] and large margin estimation [Li and Jiang, 2006].

With the development of Deep Neural Network (DNN), which can better combine various features, it is used to replace GMM in ASR field. In previous work, it has shown that DNN-HMM based ASR systems outperforms traditional HMM-GMM based systems in phoneme recognition [Mohamed *et al.*, 2012] and large vocabulary continuous speech recognition task [Seide *et al.*, 2011]. Different work is conducted to find a DNN-HMM ASR model which has better performance and faster training speed, like [Zhou *et al.*, 2012] and [Zhou *et al.*, 2014].

The ASR systems are applied to various fields. ASR is also an important phase in our framework for role labeling which facilitates the transformation from audio to text. After that, we can extract textual features from the text.

Multi-modality Work

As mentioned before, our role labeling work is based on combining low level acoustic features with high level textual features, which is a multi-modality work rather than single modality work that is based on acoustic features.

Many multi-modality studies are applied to various fields. There are several applications of multi-modality features in the phoneme sentiment analysis field. An emotion recognition model using acoustic prosodic information and text semantic labels is proposed [Wu and Liang, 2011]. Text tagged data from twitter and acoustic features are applied in some studies in order to get better speech emotion recognition performance [Hines *et al.*, 2015] . The multi-modality features used in emotion analysis field are analyzed [Cambria *et al.*, 2013]. Furthermore, McAuley et al. integrate images and text features in the recommendation system, which performs better than other systems [McAuley *et al.*, 2015].

Most multi-modality work has better performance than single modality based work, which shows that multi-modality can help the models to become more powerful. To the best of our knowledge, multi-modality has not been applied to role labeling.

3 Multi-modality Role Labeling Framework

In this section, we will introduce our four-phase framework for role labeling in real customer service telephone conversation, which integrates low level acoustic features with high level text content features.

As introduced in Section 1, there are two fundamental challenges in this work: 1) A telephone conversation is a continuous audio stream. Speaker diarization is conducted to get the segments of an audio stream. In practice, the audio segments will be translated into text sentences in ASR phase for further analysis. 2) This study aims at real conversation speaker recognition, e.g the speakers are not pre-defined. Mapping and classification methods which can recognize if the speaker of a sentence is an agent or a user are applied to conduct role labeling to deal with this problem. The flow chart of our framework is drawn in Figure 1. The input and output of each phase are presented in Table 1.

Table 1: The input and output of each phase

Phase Name	Input	Output
Speaker Diarization	Acoustic features	Two audio segments clusters
ASR	Acoustic features	Text content
Dialog Level Role Labeling	Audio segments, textual features	Clusters and roles mapping result
Sentence Level Role Correction	Textual feature	Text segments with labeling

The framework contains four-phase: *Speaker Diarization*, *ASR*, *Dialog Level Role Labeling*, and *Sentence Level Role Correction*. *Speaker Diarization* is designed based on acoustic features extracted from the audio stream, using Mel Frequency Cepstrum Coefficient (MFCC) and ΔBIC algorithm. Filter bank feature is used in *ASR* phase for speech recognition in DNN-HMM model. *Dialog Level Role Labeling* takes both the outputs of *Speaker Diarization* and textual features extracted from *ASR* phase into account and get the primary role labeling results. In the last phase, *Sentence Level Role Correction*, the labeling results are revised in a fine-grained level, which turns out to be helpful in reducing the error accumulated in the previous phases. In next section, we will introduce each phase in detail.

4 Four-phase Model Construction

4.1 Speaker Diarization Based on Acoustic Feature

The first phase designed for role labeling is *Speaker Diarization*, to split the audio stream into segments clusters. We assume that the feature vectors of each segment arise from some probability distribution, so we will try to decide if the segments are in the same distribution, which means the segments are given by the same speaker.

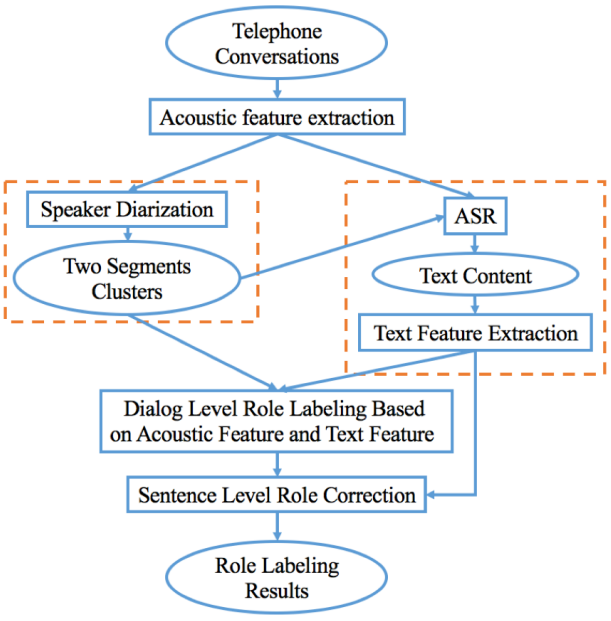

Figure 1: The flow chart of role labeling

Usually, there are only two speakers in a telephone conversation. In service dialog, the two speakers are an agent and a user. After splitting the audio stream into segments, we divide the segments into two clusters in this step based on acoustic features, and they are taken as a prior information in the following phases.

Firstly, a telephone conversation audio stream is splitted into audio segments by silent durations. Apparently, that is just a coarse grained segmentation. Then, a fine-grained segmentation is conducted based on ΔBIC-based algorithm [Cheng *et al.*, 2010]. Given two audio segments represented by feature vectors, $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_n$, the following two hypotheses are evaluated:

$$H_0 : x_1, x_2, ..., x_n, y_1, y_2, ..., y_n \sim N(\mu, \mathbf{\Sigma})$$

$$H_1 : x_1, x_2, ..., x_n \sim N(\mu_{\mathbf{x}}, \mathbf{\Sigma_y}), y_1, y_2, ..., y_n \sim N(\mu_{\mathbf{y}}, \mathbf{\Sigma_y})$$

H_0 means that X and Y are derived from the same multivariate Gaussians distribution, while H_1 means that they are from different distribution. The ΔBIC value can be calculated as the difference between the BIC value of H_0 and H_1 as follows:

$$\Delta BIC_{X,Y} = BIC(H_1, X \cup Y) - BIC(H_0, X \cup Y)$$

The larger the value of ΔBIC is, the less similar the two segments will be. Speaker change point can be located in this way. This segmentation method is window-growing based, in which the segments are sequentially input, and all change points are detected via this method.

At last, hierarchical agglomerative clustering is applied to cluster the segments into two clusters. MFCC feature extracted in audio stream is used in this phase.

4.2 Automatic Speech Recognition with DNN-HMM Model

In the last phase, acoustic features are applied in speaker diarization. The text content of audio stream is very useful in role labeling. For example, if a speaker says: Can I help you? Obviously, this is the agent speaking. Therefore, it is necessary to translate the audio stream into text content. Automatic speech recognition is conducted with the help of the ASR algorithm in this phase. Although both *Speaker Diarization* and *ASR* will introduce some errors, they are still valuable compared with the information they bring for role labeling. Moreover, the errors will be fixed in later phases.

DNN-HMM model is applied to implement automatic speech recognition step , which is one of the state-of-the-art ASR model [Povey *et al.*, 2011] and [Zhou *et al.*, 2012]. Figure 2 is the framework of the DNN-HMM model. The filter bank feature is extracted from the audio stream, which is the input of DNN.

Further more, as shown in Figure 1, the inputs of this phase are the acoustic features of audio stream segments splitted in *Speaker Diarization*, instead of the acoustic features of the whole telephone conversation. The outputs of *ASR* are text segments.

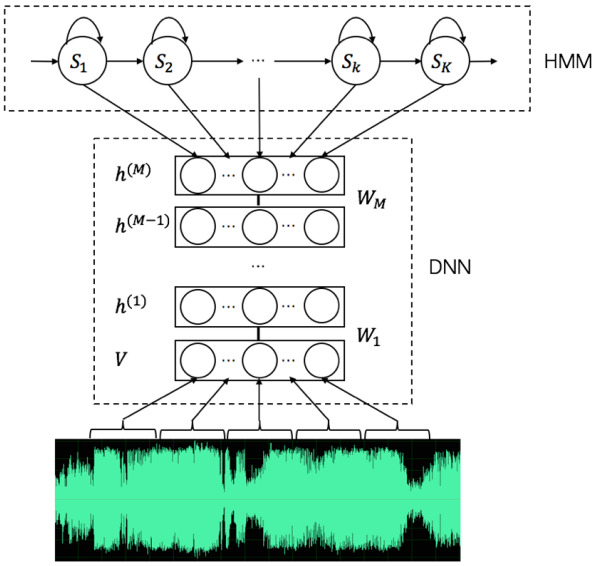

Figure 2: The framework of DNN-HMM

In practice, a GMM model with HMM model is pre-trained using MFCC feature for initialization. Then, DNN model will be used to replace the GMM model, and the input features are changed into filter bank features. After being trained, this DNN-HMM model becomes the final model for automatic speech recognition. The pre-training step is set up for HMM model parameters initialization, which can save the training time of the DNN-HMM model. Admittedly, other *ASR* models can also be applied here to get the text content, while DNN-HMM model is chosen considering its better performance.

4.3 Dialog Level Role Labeling

Low level acoustic features and high level text content features are integrated to conduct a coarse-grained level role labeling in this phase. The two audio segments are constructed according to acoustic features, and the text features are extracted from *ASR* phase. We name this phase coarse-grained level role labeling, because the two clusters are mapped to the two speaker roles, an agent and a user, as dialog level role labeling. The algorithm of textual features extraction and cluster classification for mapping is shown in Algorithm 1. The acoustic feature is remained in dialog segments set, and text feature is applied to construct the mapping relationship between the clusters and speakers.

As shown in the algorithm description, there are 5 steps in dialog level role labeling. Firstly, V is initialized with 0. The length of V is determined by the length of F. Secondly, each dimension of V records the difference between cluster X and Y in word frequency. Notice that in this case, the values of some dimensions can be very large, and it is not excepted in classification. Therefore, in next step, we normalize the value into $\{0,1,2\}$. Then, we use a classifier to calculate the mapping relationship: X to A, Y to U or X to U, Y to A. According to the mapping relationship, we replace the role label l_i in D into l'_i, and get D'. D' is the role labeling result.

Algorithm 1 Dialog Level Role Labeling Algorithm

Definition: D : Dialog segments set; X and Y are the two clusters; A : Agent; U : User; F : A pre-defined feature words set; V : A map records (word, count), saving the vectorized D; R : Classification result, 1 or 0.
Input: $D = \{(t_1,l_1), (t_2,l_2), (t_3,l_3) \ldots (t_n,l_n)\}$, t_i is segment i's text content, l_i is a cluster label, $l_i = X$ or Y.
Initialize: Pre-train a binary-classification $classifier$.
Output: $D' = \{(t_1,l'_1), (t_2,l'_1), (t_3,l'_2) \ldots (t_n,l'_1)\}$, $l'_i = A$ or U

1: Initial map V with F, the feature words are mapped with 0 and stored in V.
2: Traverse the set D and accumulate the frequency of each feature word, and refresh the V. Notice that if the label of the word is Y, we will minus the count of this word in V.
3: Normalized the values of each key in V into 0,1,2. Positive number, 0 and negative number will be replaced by 0,1and 2.
4: Take V as the input of $classifier$, get the classification result.
5: Refresh the labels in D with label U and A according to classification result, get D'.

In Section 5, we will show that different classifiers have been used in this phase, and they all have a good performance.

4.4 Sentence Level Labeling Correction

From the 3 phases above, we will get role labeling results, while that is not enough. Even though the mapping relationship between clusters and roles is perfectly constructed in *Dialog Level Role Labeling* phase, the role labeling results in sentence level might be wrong. Since there are errors made and accumulated in the *Speaker Diarization* and *ASR* phases, we need to correct the role labeling results. The two phases are based on acoustic features, meaning that it is hard to correct it with only acoustic features. In this phase, textual features are used to deal with the mistakes accumulated in previous phases.

A basic assumption in this phase is that most of sentences are labeled correctly, and we will modify the sentences that are highly possible to be falsely labeled. The feature words to vectorize the sentences were strictly selected, and a probability algorithm – logistic regression is applied. The text features used here are also bag-of-words. Different from phase 3, the feature words in this phase are selected according to relative entropy, which is computed with the probability of the words in word set, rather than the frequency of them. The high frequency words perform well in vectorization, while they may be unqualified for sentence level role classification, because they are frequently used by both speakers. On the other hand, the high relative entropy words are usually the typical words that are distinguishable.

Basic symbol notations are defined in Table 2, and the formulations to calculate the relative entropy of word x are defined as following:

$$P_x = \begin{cases} \frac{agent_x}{|a_i|} & \text{if } x \text{ in agent's word set} \\ \frac{1}{|a_i|} & \text{otherwise } x \text{ not in agent's word set} \end{cases}$$

Table 2: Basic Symbols Notation

Symbol	Definition
$agent_x$	The frequency of word x in agent's sentences.
$user_x$	The frequency of word x in user's sentences.
$\|a\|, \|u\|$	The number of sentences in $agent$ set, $user$ set.
$RE_{P,Q,x}$	The relative entropy of word x in $agent$'s set comparing with in $user$'s set.

$$Q_x = \begin{cases} \frac{user_x}{|u_i|} & \text{if } x \text{ in user's word set} \\ \frac{1}{|u_i|} & \text{otherwise } x \text{ not in user's word set} \end{cases}$$

$$RE_{P,Q,x} = P_x * log(\frac{P_x}{Q_x})$$

$$RE_{Q,P,x} = Q_x * log(\frac{Q_x}{P_x})$$

The probability of each word for relative entropy calculating depends on the train set. And the top ten words in $RE_{P,Q,x}$ and $RE_{Q,P,x}$ are used as feature words for sentence vectorization to select the most discriminative features. Some sentences may vectorized into zero vector, and these sentences will not be modified in this phase due to the low correction confidence.

Logistic regression is leveraged as the correction method in this phase. $\mathbf{X}$ represents a vectorized sentence s. The probability of s being said by an agent or a user is calculated by the following formulations:

$$P(s = \text{“}agent\text{”}|\mathbf{X}) = \frac{1}{1 + e^{\theta^T \mathbf{X}}}$$

$$P(s = \text{“}user\text{”}|\mathbf{X}) = 1 - P(y = 0|x)$$

In traditional logistic regression steps, the classification result is determined by whether the possibility is larger than 0.5. But in this phase, only the labeling results which have enough confidence will be revised. The label of a sentence will be modified into “agent” or “user” if and only if the possibility satisfied one of the following two inequations.

$$P(s = \text{“}agent\text{”}|\mathbf{X}) \quad or \quad P(s = \text{“}user\text{”}|\mathbf{X}) > 0.5 + \tau$$

These are the steps we designed for correction, and the outputs of this phase are the final role labeling results. The experiments in Section 5 verify that our probability based correction method helps in getting better performance.

5 Experiments

In Section 3 and 4, we introduced our four-phase model for role labeling. In this section, we will report the experiment results which are conducted in real customer service dialog datasets based on our framework .

5.1 Dataset

The datasets used in our work come from two different fields, a mobile service call center and a bank service call center, and both of them are real customer service telephone conversations. All conversations are in Chinese.

Mobile Service Dialog Dataset (MSDataset) , contains 34 telephone conversations with over 2,000 sentences. This is a normal size for acoustic recognition, like [Malegaonkar *et al.*, 2007] and [Cheng *et al.*, 2010]. We also adopt our method to a much larger dataset, **Bank Service Dialog Dataset** (BKDataset), which contains 85,336 conversations, to test whether our method is effective.

5.2 Evaluation

We chose 3 evaluation metrics to evaluate the labeling results: **AccuD**: the accuracy of dialog level labeling (mapping relationship between clusters and speakers). **AccuS**: the accuracy of sentence level labeling result. **AccuT**: the accuracy of time level labeling result, which considers the percentage of time that the correct labeled sentences occupied. In addition, we used segmentation error rate (SER) to evaluate the speaker diarization result in phase one following previous approaches [Barras *et al.*, 2006], which takes two kinds of errors into account: missed speech (MiS) and false alarm speech (FaS).

We labeled some conversations to evaluate the experiment results. Each sentence in the MSDataset is labeled with speaker role by hand according to audio and text content. Then, we selected 1,000 conversations from the BKDataset based on stratified sampling according to the length of the conversations and labeled the dialog level roles.

5.3 Results

Firstly, compared with human recognition results, the segmentation error rate in *Speaker Diarization* on dataset is 14.28% (9.53% in FaS and 4.75% in MiS). The error rate can be reduced by using better speaker diarization methods, while we did not concentrate on dealing with it in this study.

Dialog Level Role Labeling

As introduced in Section 5.1, there are two dialog datasets, and the BKDataset is larger than the other. Our dialog level role labeling experiment is conducted on BKDataset at first. The 1,000 sampled conversations are used for classification. Each conversation is labeled by two professional annotators. If their opinions are different, they would discuss carefully until an agreement is reached.

We use bag-of-words features for classification. Notice that not all of the words are used to vectorize dialog D, because the vector will be too long if so. We only select the top 20 frequent words in the two clusters separately to construct the bag-of-words vector. As mentioned in Section 4.3, a binary classifier is used for classification. We have adopted different methods: Decision Tree, SVM, Naive Bayes, and so on. In 5-fold cross validation, the performance of Decision tree is the best which **accuD** achieves 99.5%. Moreover, other methods also get more than 97.8% in **accuD**, which indicates that our framework has a steady performance.

We applied the trained Decision Tree classifier to MBDataset. The **accuD** is 97.1%, only 1 of the 34 telephone

conversations get a wrong labeling result, indicating that the dialog level role labeling classifier is domain-independent.

Sentence Level Correction

However, when we concern about sentence level labeling accuracy, the **AccuS** is 87.1 % and the **AccuT** is 88.2% in MB-Dataset (The accuracy reported is calculated in the right recognized sentences). That is due to mistakes made in previous phases. The mistakes cannot be fixed with only dialog level labeling. Logistic regression model is adopted in correction.

The feature extraction steps and correction algorithm are introduced in Section 4.4. The **AccuS**'s variation with the increase of τ (in 5-fold cross validation) is drawn in Figure 3. As illustrated in Figure 3, the blue line is the **AccuS** before correction, and the red line records the **AccuS** after correction with τ. The corrected results have worse performance than before because when τ is low, some modifications are not with enough confidence. There are some sentences even modified into wrong label. With the increase of τ, **AccuS** increases. Specially, at the same time, fewer sentences are checked and modified, for the reason that both possibilities are lower than $0.5+\tau$. So the **AccuS** drops when τ is over 0.4. The accuracy of logistic classification result continuously increases when we use larger τ, and the classification result achieves 100% when τ is larger than 0.47. The best performance of **AccuS** achieves 90.5% when τ equals to 0.39.

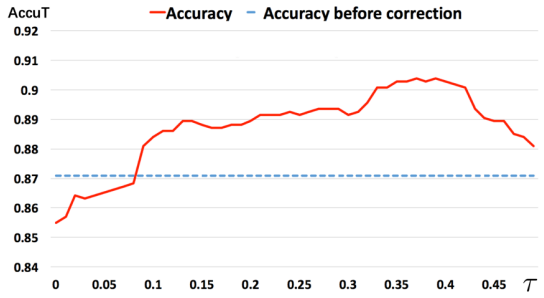

Figure 3: Classification AccuS in MSDataset

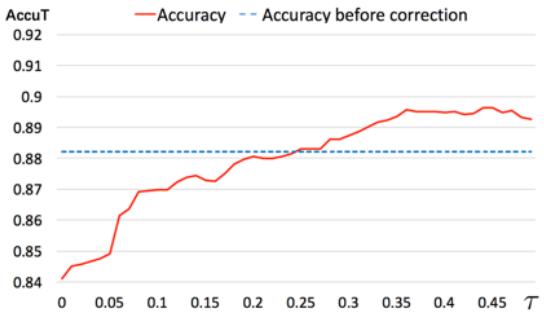

Figure 4: Classification AccuT in MSDataset

In Figure 4, **AccuT** shows similar results with **AccuS** with the increase of τ. **AccuT** is sensitive with the length of correctly labeled segments. The accuracy before correction is higher than **AccuS**, showing that it is easier to label the long segments correctly. And the accuracy dropping less than **AccuT** means that most modified labeling segments are short.

Comparison with Other Methods

In this part, we compared our framework with several other methods. To the best of our knowledge, there is not an appropriate solution. Therefore, all baselines are based on the four-phase framework: 1) Keywords based dialog labeling with acoustic features. 2) Labeling without segment clustering based on text features. 3) Role labeling without correction. The experiment results are presented in Table 3. It is obvious that our acoustic and textual features based framework performs better than other methods that use a single type of feature do. The correction step helps improve the final results.

Table 3: Comparison with other methods in MSDataset

Type	Acoustic Feature	Text Feature	Without Correction	Our Framework
AccuS	78.5%	75.9%	87.1%	90.4%
AccuT	69.5%	82.0%	88.2%	89.6%

6 Conclusions and Future Work

In this paper, we present our work on role labeling in real customer service telephone conversation. This multi-modality work is based on both acoustic and textual features. Differing from previous speaker recognition work, our goal is not mapping the speaker into a pre-defined people group.

We propose a four-phase framework for role labeling: *Speaker Diarization*, *ASR*, *Dialog Level Role Labeling*, and *Sentence Level Role Correction*. *Speaker Diarization* and *ASR* are based on acoustic feature of a telephone conversation, which are the basic steps of role labeling. The clustering results in *Speaker Diarization* and text features extracted in the output of *ASR* are used for *Dialog Level Role Labeling*. With the help of Decision Tree, role mapping accuracy is over 99.0%. In the last phase, logistic regression is applied to labeling result correction based on text features, which improves the performance. The final accuracies in sentence level and time level achieve 90.4% and 89.6%.

To the best of our knowledge, this is the first work that takes both acoustic and textual features into consideration in role labeling. Our multi-grained framework performs better than other methods in realistic datasets based experiments.

Our future work includes two parts: 1) We will try to use better speaker diarization method to minimize the mistakes in segments clustering, which will be helpful in improving the final performance. 2) As mentioned before, this is the fundamental work of customer satisfaction evaluation and customer service evaluation, and we would like to go further in customer service telephone conversations analysis.

Acknowledgments

We thank Ji Cao, Libo Yang for their insightful discussions and help. This work was supported by National Key Basic Research Program (2015CB358700), Natural Science Foundation (61472206, 61073071) of China and joint project of Beijing Sino Voice Technology, Co., Ltd.

References

[Barras *et al.*, 2006] C. Barras, Xuan Zhu, S. Meignier, and J. Gauvain. Multistage speaker diarization of broadcast news. *Audio Speech & Language Processing IEEE Transactions on*, 14(5):1505–1512, 2006.

[Cambria *et al.*, 2013] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, (2):15–21, 2013.

[Cheng *et al.*, 2010] Shih Sian Cheng, Hsin Min Wang, and Hsin Chia Fu. Bic-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. *IEEE Transactions on Audio Speech & Language Processing*, 18(1):141–157, 2010.

[Das, 2011] Amitava Das. Speaker recognition via voice sample based on multiple nearest neighbor classifiers, 2011.

[Delacourt and Wellekens, 2000] Perrine Delacourt and Christian J Wellekens. Distbic: A speaker-based segmentation for audio data indexing. *Speech communication*, 32(1):111–126, 2000.

[Gibson and Hain, 2006] Matthew Gibson and Thomas Hain. Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. In *INTERSPEECH*. Citeseer, 2006.

[Hines *et al.*, 2015] Christopher Hines, Vidhyasaharan Sethu, and Julien Epps. Twitter: A new online source of automatically tagged data for conversational speech emotion recognition. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 9–14. ACM, 2015.

[Katharina *et al.*, 2005] Von Kriegstein Katharina, Kleinschmidt Andreas, Sterzer Philipp, and Giraud Anne-Lise. Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, 3(3):379–382, 2005.

[Li and Jiang, 2006] Xinwei Li and Hui Jiang. Solving large margin hmm estimation via semi-definite programming. In *Proc. of 2006 International Conference on Spoken Language Processing (ICSLP'2006)*, 2006.

[Malegaonkar *et al.*, 2007] Amit S Malegaonkar, Aladdin M Ariyaeeinia, and Perasiriyan Sivakumaran. Efficient speaker change detection using adapted gaussian mixture models. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(6):1859–1869, 2007.

[McAuley *et al.*, 2015] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

[Mohamed *et al.*, 2012] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):14–22, 2012.

[Pardo *et al.*, 2007] J. M. Pardo, X. Anguera, and Chuck Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computers*, 56(9):1212–1224, 2007.

[Povey *et al.*, 2011] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[Seide *et al.*, 2011] Frank Seide, Gang Li, and Dong Yu. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech*, pages 437–440, 2011.

[Tranter *et al.*, 2006] Sue E Tranter, Douglas Reynolds, et al. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, 2006.

[Wang, 2004] Hsin Min 1 Wang. The sovideo mandarin chinese broadcast news retrieval system: Special double issue on chinese spoken language technology. *International Journal of Speech Technology*, (7):189–202, 2004.

[Wu and Liang, 2011] Chung-Hsien Wu and Wei-Bin Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *Affective Computing, IEEE Transactions on*, 2(1):10–21, 2011.

[Zhang and Tan, 2008] Cuiling Zhang and Tiejun Tan. Voice disguise and automatic speaker recognition. *Forensic Science International*, 175(2-3):118–22, 2008.

[Zhao and Fan, 2004] Xiu Zhen Zhao and Xi Yun Fan. Acoustic change detection and segment clustering of two-way telephone conversations. *Journal of Dalian University*, 2004.

[Zhou and Hansen, 2005] Bowen Zhou and John HL Hansen. Efficient audio stream segmentation via the combined t 2 statistic and bayesian information criterion. *Speech and Audio Processing, IEEE Transactions on*, 13(4):467–474, 2005.

[Zhou *et al.*, 2012] Pan Zhou, Lirong Dai, Qingfeng Liu, and Hui Jiang. Combining information from multi-stream features using deep neural network in speech recognition. In *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*, pages 557–561, 2012.

[Zhou *et al.*, 2014] Pan Zhou, Lirong Dai, and Hui Jiang. Sequence training of multiple deep neural networks for better performance and faster training speed. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 5627–5631. IEEE, 2014.