A framework for Single-Channel Two-Speaker Telephone Customer Service Speaker Verification

Hao Liang, Yuanzhe Cai, Jianzong Wang, Jing Xiao
Ping An Technology (Shenzhen) Co.,Ltd, China
Email: {lianghao894, caiyuanzhe259, wangjianzong347, xiaojing661}@pingan.com.cn

***Index Terms*—Speaker Verification, Telephone Customer Service**

***Abstract*—Speaker verification (SV) is the identification of a person from biometrics characteristics of voices (e.g., speech waves, voice pitch, speaking style, etc.) SV technique become ubiquitous, and are widely used in the various areas (e.g., mobile apps login, telephone banking, etc.) Hence, it would be beneficial if Ping An telephone customer service (TCS) can provide the personalized high quality service to the customer with this SV identification. For example, Ping An TCS are able to recommend the most related credit card, deposit plan, or insurance plan to the customer based on their history records. This paper addresses the problem of applying the more accuracy and efficient SV technique in real world telephone customer service.**

Current SV technique requires single-channel speech, but TCS records co-channel speech which means customer and service's voices are mixed together. Directly using the mixed voice will receive a very low accuracy. Hance, we propose a SV framework for this co-channel speech. We first transcribe customer's audio to the set of short texts with the time stamp by automatic speech recognition (ASR), and then customer's short texts are identified by our proposed short text classification approach. In the end, the customer's audio can be easily extracted and concatenated together. Meanwhile, to improve the SV accuracy, a highway long short-term memory recurrent neural network model (HLSTM) [1] are used for automatic speech recognition. The use of a HLSTM-UBM framework in the speaker verification pipeline is attractive as it can integrate phonetic information to help us to discriminate between speakers even in text-independent speaker verification task and show promising results. We present our framework along with extensive experimental analysis that indicates superiority of our approach as compared to other methods.

I. Introduction

Speaker Verification (SV), also called voice recognition, is the process of accepting or rejecting a person from the given personal utterance. Most of the applicable services in which banking system, financial business and security control use of voice to confirm the identity of a speaker for verification purpose. For example, Ping An insurance department utilizes Speaker Verification technique within phone conversation between customers and client services to authenticate their identification.

The current system in Ping An speaker verification uses GMM-based i-vector framework. Although it has achieved great success in some real-world applications (e.g., Happy Ping An APP [1], Ping An Zhi Bird APP [2],) this technique has not been widely used in Ping An TCS (i.e. 95511) system.

There are three main problems in current SV model:

- Co-channel speech: the current system used in Ping An SV task requires single track audio which is limited to certain situations. For instance, the Ping An telephone service platform only records co-channel speech. The SV accuracy for these mix-audio with customer and service are extremely low (error ratio is more than 64%.) Some tentative methods have been tried for handling co-channel speech, i.e. since at the beginning client service always says the same sentence (e.g., "Hello, May I help you?",) we tried to extract customers first utterance speech segments for verification. However, since the first utterance of customers are usually quite short (the average length is 4.4s, with 8.3 words,) the accuracy is still low (error ratio is about 13.2%.) Therefore, how to extract the clean and long customer voice segments is a vital problem in this paper.
- Ping An TCS requires a very short response time (less than 1s) for speaker recognition. The quicker to identify the user's information, the better services Ping An service can provide to customer.
- The high accuracy SV algorithm is always expected in Ping An TCS.

To solve the above problems, we developed a SV system with customer/service identification front-end to extract customer speech segments from the co-channel audio data. Specifically, the system first uses automatic speech recognition (ASR) technique to transcribe mix-audio to text. Then, applies text classification approach to assign the customer label to these text. In the end, according to the text content and related time stamps (from ASR result,) the identification front-end can easily concatenate these customer's homogenous audio segments together.

For the high accuracy aspect, it is known that the current unsupervised GMM-based framework is conceptually an unsupervised way to discriminate speakers from given speech utterance. Specifically, the UBM is trained to cluster MFCCs into unsupervised clusters [2]. This fact decreases the verification performance when two speakers pronounce the same phone from the relative point (a supervised tied triphone state.) Therefore we propose to utilize a HLSTM-based i-vector framework instead of the GMM-based framework.

Contributions: The contributions of this paper are:

[1] http://tech.Pingan.com/product/sass_b_happy.shtml

[2] http://www.zhi-niao.com/

- In this paper, we discuss the co-channel problem for current telephone customer service. We argue that why current single channel SV approach cannot be directly used in TCS speaker recognition task.
- We present a novel SV framework for this co-channel TCS. Apart from speaker's voice biometrics, the content of speaker's dialogue has also been used for identification. The long and clean customer's voice has been extracted from the the co-channel TCS. We also analyze the characters of customer service's dialogs. More accurate, efficient and scalable short text classification has been apply in our framework.
- Extensive experimental analysis is performed on multiple, diverse data sets to show how the proposed algorithm provide more accurate and efficient results than other approaches. Meanwhile, this SV framework has been used in Ping An TCS.

The rest of the paper is organized as follows: In Section V we introduce the related work. In Section II we describe the diarization front-end for managing co-channel speech and the HLSTM-UBM model that we used for our speaker recognition system. We summarize the experiment setup and report experiment results in Section III and conclusions are in Section IV.

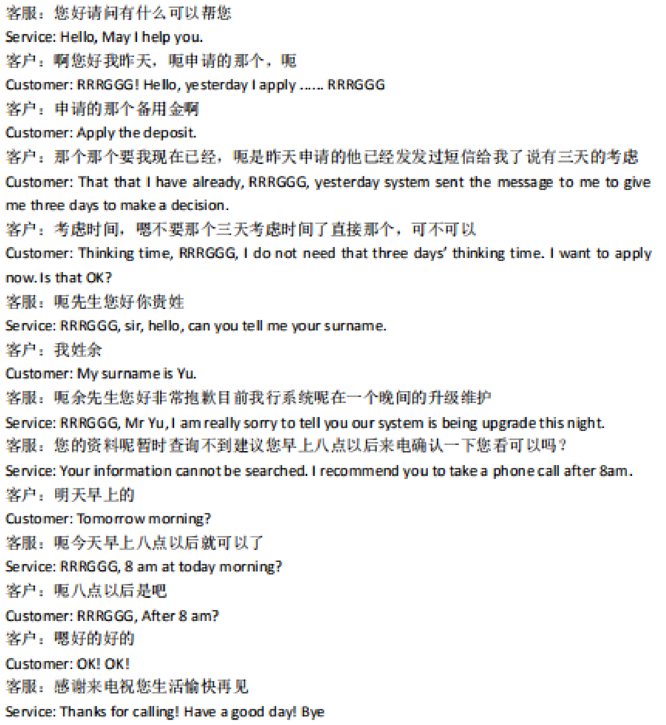
II. Co-channel Speech SRE Framework

A. Motivation

As we know, the quality of the user's audio is very important for the speaker verification. Hence, we need to design a good method to extract the customer's voice from the mix audio. The current Ping An TCS has the following characteristics. First, as the customer service, only two speakers (customer/service staff) are in the system. The binary classification approaches can be applied to identify the customer's audio. Second, the contents of speakers are much different between customer and service. Since our audio come from Ping An Bank (http://bank.Pingan.com/) customer service, the similar dialogues, such as "apply the deposit", "apply the credit" and etc, always appear in the audio set. The similar services questions, such as "What is your surname?", "May I help you?", etc and the similar customers answers, such as "My surname is ....", "apply the deposit" repeat for multiple times. Therefore, it is possible to use the "keyword matching" (e.g., "My" for customer, "help" for service, etc) to assign the "customer" and "service" label to these audio. Thus, our framework extracts the text from the audio by ASR model and use the text classification approach to identify the customer's content and then splices all the customer's audio together.

B. Overview

The proposed system is an ensemble of three modules (See in Figure 2,) that is: (i) an ASR module for generating recognized text from input mixed audio. (ii) a Customer/Service identification module for classifying speakers and identify the customer's audio files, so that we can get the cleaned one-person single track audio file as input to the recognition module. (iii) a speaker recognition module, specifically a text-independent speaker verification, with the input of split audio of customer, verifying the identity of the speaker. The detail information is shown in the following sections.



客服：您好请问有什么可以帮您

Service: Hello, May I help you.

客户：啊您好我昨天，呃申请的那个，呃

Customer: RRRGGG! Hello, yesterday I apply ...... RRRGGG

客户：申请的那个备用金啊

Customer: Apply the deposit.

客户：那个那个要我现在已经，呃是昨天申请的他已经发发过短信给我了说有三天的考虑

Customer: That that I have already, RRRGGG, yesterday system sent the message to me to give me three days to make a decision.

客户：考虑时间，嗯不要那个三天考虑时间了直接那个，可不可以

Customer: Thinking time, RRRGGG, I do not need that three days' thinking time. I want to apply now. Is that OK?

客服：呃先生您好你贵姓

Service: RRRGGG, sir, hello, can you tell me your surname.

客户：我姓余

Customer: My surname is Yu.

客服：呃余先生您好非常抱歉目前我行系统呢在一个晚间的升级维护

Service: RRRGGG, Mr Yu, I am really sorry to tell you our system is being upgrade this night.

客服：您的资料呢暂时查询不到建议您早上八点以后来电确认一下您看可以吗？

Service: Your information cannot be searched. I recommend you to take a phone call after 8am.

客户：明天早上的

Customer: Tomorrow morning?

客服：呃今天早上八点以后就可以了

Service: RRRGGG, 8 am at today morning?

客户：呃八点以后是吧

Customer: RRRGGG, After 8 am?

客户：嗯好的好的

Customer: OK! OK!

客服：感谢来电祝您生活愉快再见

Service: Thanks for calling! Have a good day! Bye

Fig. 1. a Sample for Pingan Telephone Customer Service (The customer information has already been removed.)

C. ASR Module

The ASR model helps to transcribe co-channel speech to conversational text. The co-channel audio data is often received as a non-overlapped mono audio file although there are two speakers within the audio file. And the audio can be segmented into audio data attributed to separate speakers.

As we wish to send short text to the Customer/Service Identification module module, we first apply Sequential Gaussian Mixture Model based Voice Activity Detection (SGMM-VAD) algorithm, described in [3], to split the mixed audio into audio segments. The SGMM-VAD comprises two Gaussian components which respectively describe the speech and non-speech log-power distributions, helps to detect audio parts from noisy speech with high signal noise ratio.

The ASR model is then able to transcribe each audio segments into short text. Specifically, we use Latency-controlled Bidirectional Highway LSTM RNN model (LC-BHLSTM) [1] as the acoustic model of our ASR module, to establish the relationship between the input audio signal and the phonemes or other linguistic units that make up speech. Figure 2 shows an sample of ASR result.

For ASR model setup, the front-end feature consists of 80-dimension log Mel Filterbank (FBANK) with a frame length of 25ms. 3-dimension pitch feature (including probability of

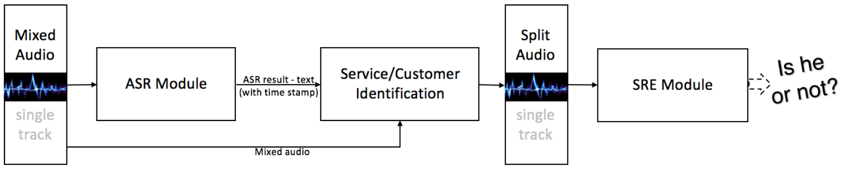


Fig. 2. SV Framework for designed Co-Channel TCS

Voicing (POV) features) is appended to create 83-dimension feature vector. The LC-BHLSTM has 5 layers, with 1024 memory cells together with a 512-node projection layer at each layers output. The output targets are 10k dimensional context-dependent tied triphone states (a.k.a. senones,) of which were determined by the HMM-GMM training stages.

D. Customer/Service Identification

The customer/service identification module uses the supervised learning classification approach, to identify the customer and service label from the transcribed text. In the training stage, we manually marked the "customer" and "service" label to build the training set. After training, we can use this text classification model to automatically assign the "customer" and "service" label to the given text segments.

Different from the traditional text/documents, the texts extracted from telephone customer service are very short (the average text length is only about 8.3.) Figure 1 shows the sample of Ping An Telephone Customer Service. Each line is the segment of text [3], e.g., "Hello, May I help you?" etc. Therefore, to assign the customer/service label to such short text is a real challenge work.

The motivation of our designed method comes from the search engine process, where nowadays if users want to get an idea of something, they input some queries (key words) into these search engine (e.g., google.com, bing.com) to obtain the related information. Like this procedure, the training short-text has been imported into a search engine; the rest short-texts, as the query, are used to search the relevant texts. The label for the text can be easily voted by the searched texts which has the manually marked label.

Figure 3 shows the framework of customer/service classification. First, system builds the search engine [4] by these training text (extract Chinese word segmentation, the build the reserved index on these text, etc.) Second, the new short text, as the query of the search engine, retrieves the top K related texts from these training texts. Third, system use the k-nearest-neighbor (KNN) [5] to vote the label to the input text based on the retrieved results. The short-text's query word can be set the different various weight [6] (e.g., Boolean, TF, TF.IDF.) In experiment section, the accuracy of TF-KNN achieves the better accuracy then the other approach. Leveraging the well studied IR technique, the proposed approach is extremely scale for the large data and easily add the new training text by updating the index.

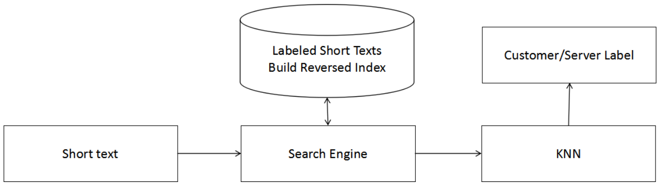


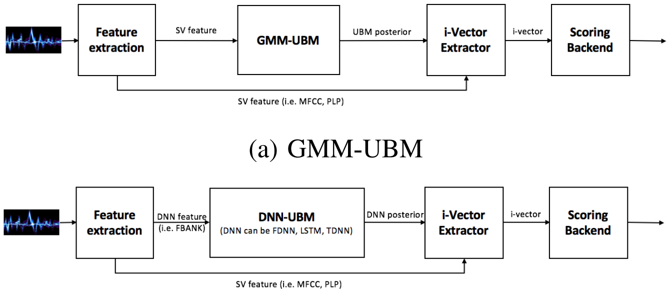
Fig. 3. Customer/Server Classification for Telephone Customer Service System

E. SV Module

The framework of many voice recognition applications in use is a GMM-based i-vector system (See Figure 4(a)). It can be decomposed into four sequential stages: (i) the feature extraction, the collection of sufficient statistics, the extraction of i-vectors and a scoring criterion backend. The feature extraction is a process converts the speech waveform into characteristic parameter, which retains useful speaker information from given speech signal and filters unwanted information such as noise. Parameters like Mel-Frequency Cepstrum Coefficient (MFCC,) Linear Predictive Cepstral Coefficient (LPCC) and perceptual linear prediction (PLP) are often used in feature extraction stage, follow by voice activity detection (VAD.) (ii) The collection of sufficient statistics is a process to calculate the zero-order, first-order, second-order Baum-Welch statistics from a sequence of feature vector from the first stage. These statistics are highly dimensional, which is generated from a large GMM, called Universal Background Model (UBM.) (iii) The extraction of i-vectors in the third stage is a technique that converts the high dimensional statistics into a single low dimension feature vector, which carries only the discriminative characteristic information apart from the other speakers. (iv) Once i-vectors are extracted, a scoring criterion backend is used to make decision that whether accept or reject the person as the request identity. There are three commonly used scoring criterions for voice recognition, cosine distance similarity, linear discriminant analysis (LDA) and probabilistic linear discriminant analysis (PLDA.)

In the DNN-UBM i-vector framework (See Figure 4(b)), the feature extraction stage stays the same as the GMM-UBM i-vector framework. The different part is that the UBM

[3] ASR client identifies each segment by persons talking pause. If a person pause for more than 1s, ASR server will generate a segment.



(a) GMM-UBM

(b) DNN-UBM

Fig. 4. I-Vector based Speaker Verification Framework

now utilizes senone states to evaluate the acoustic feature space. In the experiment section, we have tested various kinds of deep neural network to find the best prototype model in our DNN-UBM framework. We show that the Highway LSTM based Universal Background Model (HLSTM-UBM) outperforms the other type of neural network based UBM with comprehensive comparison in accuracy and response time.

III. Experiment

A. Data Set

We evaluated our system on two data sets: the HKUST Mandarin Telephone Speech corpus and the PAEC financial phone call data set collected by Ping An Technology (Shenzhen) Co.,Ltd China.

- HKUST corpus: the HKUST corpus [7] is a very large scale Mandarin speech corpus which comprises around 200 hours of telephone recordings, with over 2100 Mandarin speakers and including 1206 ten-minute phone call conversations, recorded at an 8k sampling rate, where the two callers do not know each other in advance. Most of the phone calls are recorded from relatively quiet environments, such as home, office.
  All channel A wav data are used to train UBM. We split out all channel B wav data to five 30-second sub-recordings and abandon the rest.
- PAEC corpus: The PAEC data set is a building financial phone call conversations between customer and service staff in the field of banking, insurance and other financial business. Phone calls are recorded from variational environment in 8k sampling rate. For customers, many are using mobile phones or land-line phones, while for Ping An service staffs they use lapel microphones. The data set we used to train and test our system consists of 900 speakers with 40 hours multi-calls for each in training set and 5 hours uni-call for each in test set.
  Meanwhile, to test the accuracy of customer/service identification module, 900 customer/service dialogues (text) are transcribed from these audio by ASR model and 32,445 segments (short text) are extracted from these dialogues. Each segment is manually assigned “customer” or “service” label.

B. Experiment Setup

All experiments are performed on a Intel Core 8 processor Linux machine, 64 GB memory. All these experiments are run 10 times and then calculate the average score.

For SV system performance evaluation, we choose the cosine similarity distance scoring [8]. With the ground true label (target / non-target speaker,) we can calculate the False Acceptance Rate (FAR) and False Rejection Rate (FRR) of the test data with certain threshold theta. Clearly, the FAR and FRR depend on the threshold. The Equal Error Rate (EER) index is applied to evaluate speaker verification system.

C. Experiment Results

Our first experiment shows the performance of various kinds of DNN-UBM versus GMM-UBM baseline speaker verification system. Table I shows that the proposed HLSTM-UBM based SV system performs better than GMM based SV system in both HKUST corpus and PAEC real-world corpus which the test data are 30s clean single track client waveforms.

The experiment setup for various text-independent speaker verification pipeline can be described as follow.

- GMM-UBM baseline: the feature front-end is 40-dimension MFCC vector with delta and delta-delta. VAD method is applied after feature extraction step. We set the UBM 2048 components and for i-vector dimension is 600-dimension. At the end, a cosine similarity scoring back-end is used to evaluate the system performance.
- DNN-UBM: we compared various deep neural networks within DNN-UBM applied in text-independent speaker verification system. The DNN feature is generated as 40-dimension FilterBank vector, VAD method is applied as the same. For DNN-UBM components, we set all neural network model output (number of senone states)to 3000 (this is depended on the last GMM-HMM decision tree clusters, various from different training set, so it might be slightly different to 3000, i.e. for HKUST corpus the number is 2798, for Ping An corpus the number is 2825, there is no harm with a closed difference, we just want to choose closed to the 2048 of GMM-UBM baseline components). The rest of i-vector and scoring back-end stay the same as above.
  1) MN-UBM: the MN abbreviation indicates the maxout network, with introduced p-norm generalized maxout units as described in [9]. The test MN has 5 hidden layers, and the p-norm input dimension is 3000, p-norm output dimension is 300.
  2) TDNN-UBM: the time-delay neural network used to test consists of 7 layers, for different layers, the splice indexes are represented as “-2,-1,0,1,2 -1,2 -3,3 -7,2 -3,3 0 0”. The hidden layers have an input dimension of 300 and an output dimension 3000.
  3) HLSTM-UBM: the highway lstm recurrent neural network has the same structure as LC-HBLSTM described in the ASR module section. It has 5 layers, with 1024 memory cells together with a 512-node projection layer at

each layers output. The output targets are 3k dimensional context-dependent tied triphone states.

4) BLSTM-UBM: the BLSTM components are similar to HLSTM-UBM, it has 5 hidden layers, each layer consists of 1024 memory cells (512 for forward and 512 for backward) with a 512-node projection layer.

As one should know, the HKUST corpus is well designed for ASR task, which should not be the best corpus to verify speaker recognition/verification due to the lack of speaker channel difference, that is, in HKUST we do not have multi-calls from one speaker. We guess that is why in both GMM-UBM/DNN-UBM case the EERs stay quite closed.

Table I shows the EER and response time comparison from GMM-UBM baseline and various DNN-UBM i-vector based SV system. From the results we find that the DNN-UBM has an average of 15% relative improvement from GMM-UBM baseline. BLSTM-UBM achieves the bast equal error rate of 4.56% in Ping An test set, while TDNN-UBM (4.83%) and HLSTM-UBM (4.69%) performs quite closed.

TABLE I
EER(%), RESPONSE TIME BETWEEN GMM-UBM AND DNN-UBM

	HKUST data set	Ping An data set	response time
GMM-UBM	0.1116%	5.63%	408ms
MN-UBM	0.1115%	5.27%	637ms
TDNN-UBM	0.1115%	4.83%	593ms
HLSTM-UBM	**0.1115%**	**4.69%**	**585ms**
BLSTM-UBM	0.1115%	4.56%	1359ms

However, since the response time [4] is very important in Ping An Telephone Customer Service. Our second experiment focuses on the response time of GMM-UBM and various DNN-UBM. From table I, we find that when apply the same test data, e.g. one 30s clean single track audio, BLSTM-UBM (1359ms) performs more than 3 times compared to the other type of SV systems. HLSTM-UBM's response time is a slower than GMM-UBM system (408ms for GMM-UBM and 637ms for HLSTM-UBM.) However, since both response time is less than 1s, HLSTM-UBM with much higher accuracy can be used in the real telephone service system.

Experiment 3 analyze the performance of short text classification. F_1 [10], which is the harmonic average of the precision p and recall r, is used to evaluate this binary (service/customer) classification.

We conducted experiments on PAEC data set, which has manually been labeled customer/service. We evaluate three schemes to set the weight for the query (short-text in test data): Boolean, TF, TF.IDF [6]. Meanwhile, various k value has also been tested in this experiment. The classification accuracy (F1 score) are reported in Figure 5. From the results, we have the following observation. First, all the methods receive the very high accuracy (more than 0.93.) Since the dialog between various customer and service are much similar in TCS, "word matching" technique will receive the high accuracy. Second, as we know, k value is an important character for KNN algorithm.

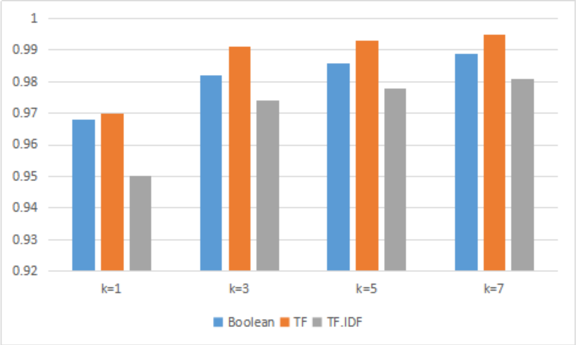


Fig. 5. F1 score for short text classification

Using $k = 1$, our approach gets a poor result (0.93) since the very less votes has been considered in the experiments. However, the accuracy for k in (3, 5, 7) are much similar, since three voting is enough for KNN classification algorithm. Third, TF-KNN and Boolean-KNN receives much similar accuracy, since the same terms in each query only appears one to two times (Remember TF is the term frequency.) Both TF-KNN and Boolean-KNN get the higher accuracy than TF.IDF-KNN. IDF is low when this term (a normal term) appears to the various short text. However, a normal term does not mean this word is not important for the classification. For example, "apply" appears for the various short texts, but it is easy to identify the customers by this term. In sum, with the high accuracy, KNN-TF (k=3) is used in the following experiments.

Experiment 4 compares different methods for co-channel case. In this experiment, three methods has been compared.

- Baseline 1: mixed audio, which including customer and client service in one single channel.
- Baseline 2: the first customer utterance (average 4.4s in PAEC) extracted from the mixed audio in Baseline 1. In PAEC corpus, the first utterance is always belonged to the client service, so that we applied voice activity detection (VAD) method and cut the second utterance as the customers first utterance segment.
- Our Approach: the concatenated customer speech from our SV framework. The mixed_* represents how long our input mixed audio is, which contains customer and client service co-channel speech. For example, mixed_60 means that the mixed audio is about 60s.

The experimental results are shown in Figure 6. We have the following observation: (i) It is clear that baseline 1 obtains the lowest accuracy (err is 64.2%) since for the mixed data the SV system cannot judge the data stand for which speaker is. (ii) Baseline 2 shows the short utterance effects on speaker recognition. As the utterances get shorter (average 4.4s), results deteriorate (err is equal to 13.2%.) (iii) Our approach (mixed_60s[5]) performs much better than that the other two baseline, since we concatenate more clean target speaker data, making the audio length increased, the result

[4] The response time is the testing time, rather than training time.

[5] The average length of concatenated customer speech segments is around 39.6s in PAEC data set.

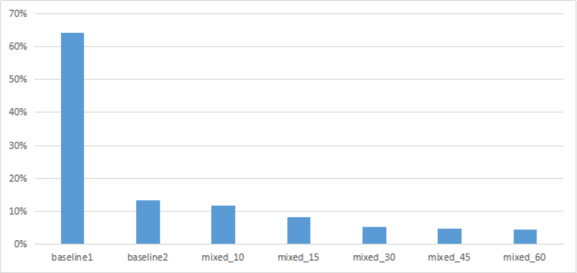


Fig. 6. ERR(%) for Three Approaches

improve dramatically, which closed to the original cleaned single track performance (compare to table I.)

Since in Ping An's financial business, different departments have various need for balancing accuracy with response time, We also list the effect of various length of the input mixed co-channel audio in Figure 6. From mix_10 to mix_30, the err ratio decreases dramatically (from 11.8% to 5.4%.) However, from mix_30 to mix_60, the err ratio decreases slowly (from 5.4% to 4.5%). Therefore, 30s mixed voice are used in our real telephone customer service.

IV. Conclusion

In this paper, we introduce a feasible speaker recognition system with short text classification technique for co-channel situation. This system takes mix speaker audio as input, the audio file is transcribed into text file with time stamp by ASR. Applying short text classification method, homogenous speech segments can be effectively identified so the audio file is segmented into homogenous audio segments, which can be concatenated into a target speaker, say customer. We show the proposed system can work on co-channel audios and the performance is closed to the clean single track target audios with no more than 15% relative equal error rate. We also show that improvement by replacing GMM-UBM with HLSTM-UBM, which achieves about 14.3% improvement in the real-world financial phone call data set.

V. Related Work

We categorize existing work related to our study into two main categories: speaker verification, speaker diarization.

Speaker Diarization: Speaker diarization has emerged as an increasingly important and dedicated domain of speech research. This technique has been used in broadcast news (BN) and conference meetings (CM) to identify the speaker according to their biometric features. The current state-of-the-art speaker diarization system fits into two categories: the bottom-up and the top-down methods. The bottom-up approaches [11], [12], [13], using hierarchical clustering (HC), training a number of clusters which aims at successively merging and reducing the number of clusters until only one remains for each speaker. In contrast, the top-down approaches [14], [15], [16] first cluster the audio stream with a single speaker model and successively add new cluster to it until all the speakers are accounted for. Top-down approaches have performed well against the broader filed of other bottom-up entries. Our system also needs the speaker diarization technique to identify the customer's audio, but telephone customer service is different from BN and CM. First, the above-mentioned clustering approaches cannot be applied in our TCS framework. We assume that the two clusters has been identified from the audio, but we still do not know which cluster is customer's audio. Therefore, the binary classification approach has been applied in TCS. Second, Second, the state-of-the-art speaker diarization system mainly uses the speaker's biometric features, such as audio waves, voice pitch, speaking style, etc, but our framework mainly use the speaker's content information, which is totally different from the former work. By the way, the biometric features cannot be used, since it is very difficult to identify some rules or models for these biometric features since these audio from various customers and services. In the other words, we cannot use the audio waves, voice pitch and speaking style to distinguish the customer and service. However, the speaker's content can be used because of the difference speaking characters between customer and service.

Speaker Verification: So far, the GMM-UBM i-vector based model [17] is still the dominant model used in modern speaker verification systems. With the development of deep neural network models in speech recognition, people start to explore using DNN in speaker verification framework. There are two dominant approaches that are the most widely adopted in research for combining deep neural network into speaker recognition. One is to extract deep neural network bottleneck feature from SV feature (i.e. MFCC). As reported in [18], one important factor that leads to superior performance improvement in distant-talking recognition is utilizing bottleneck feature to restrain channel-mismatch reverberant conditions. The other is to replace GMM posterior with DNN posterior during feature modeling, so called DNN-UBM [19], [20], [21]. This method takes the advantage of deep neural network model ability to directly model phonetic content, rather than an arbitrary acoustic space, achieves a 50% relative improvement compared to the traditional GMM-UBM method which can well perform than the first approach even more. Therefore, our SV framework uses the DNN-UBM approach, test with the various kinds of DNN models to optimize the verification system performance.

References

[1] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory rnns for distant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5755–5759, 2016.

[2] M. Mclaren, D. Castan, L. Ferrer, and A. Lawson, "On the issue of calibration in dnn-based speaker recognition systems," in *INTERSPEECH*, pp. 1825–1829, 2016.

[3] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Noise power estimation based on a sequential gaussian mixture model," in *International Congress on Image and Signal Processing*, pp. 2362–2365, 2011.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the 7th International Conference on World Wide Web*, 1998.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[6] G. Salton and C. Buckley, *Term-weighting approaches in automatic text retrieval*. Pergamon Press, Inc., 1988.

[7] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, *HKUST/MTS: A Very Large Scale Mandarin Telephone Speech Corpus*. Springer Berlin Heidelberg, 2006.

[8] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010.

[9] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 215–219, 2014.

[10] X. Li, Y. Y. Wang, and A. Acero, "Learning query intent from regularized click graphs," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 339–346, 2008.

[11] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: Icsi rt06s meetings evaluation system," in *International Workshop on Machine Learning for Multimodal Interaction*, pp. 346–358, 2006.

[12] T. Nguyen, H. Sun, S. Zhao, S. Khine, H. D. Tran, T. Ma, B. Ma, E. S. Chng, and H. Li, "The iir-ntu speaker diarization systems for rt 2009," in *Rt*, 2009.

[13] T. L. Nwe, H. Sun, B. Ma, and H. Li, "Speaker clustering and cluster purification methods for rt07 and rt09 evaluation meeting data," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 2, pp. 461–473, 2012.

[14] C. Fredouille, "Technical improvements of the e-hmm based speaker diarization system for meeting records," in *International Conference on Machine Learning for Multimodal Interaction*, pp. 359–370, 2006.

[15] S. Bozonnet, E. Nicholas, and C. Fredouille, "The lia-eurecom rt'09 speaker diarization system : enhancements in speaker modelling and cluster purification," *ICASSP 2010, IEEE international conference on acoustics, speech and signal processing*, pp. 4958–4961, 2010.

[16] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The lia-eurecom rt'09 speaker diarization system," in *RT 2009, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, USA*, pp. 4958 – 4961, 2009.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[18] T. Yamada, L. Wang, and A. Kai, "Improvement of distant-talking speaker identification using bottleneck features of dnn.," in *Interspeech*, pp. 3661–3664, 2013.

[19] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Automatic Speech Recognition and Understanding*, pp. 92–97, 2016.

[20] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4814–4818, IEEE, 2015.

[21] M. M. Saleem and J. H. Hansen, "A discriminative unsupervised method for speaker recognition using deep learning," in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*, pp. 1–5, IEEE, 2016.