doi:10.20944/preprints202103.0513.v1

(Article

Automatic Voice Query Service for Multi-Accented Mandarin Speech

Kejing Xiao [1*], and Zhaopeng Qian [2]

[1] Information School, Renmin University of China; xiaokejing@ruc.edu.cn
[2] School of Artificial Intelligence, Beijing Technology and Business University; qianzhaopeng@btbu.edu.cn
* Correspondence: xiaokejing@ruc.edu.cn; Tel.: +86 17801025299

Featured Application: Automatic Voice Query Service of Human and Computer.

Abstract: Automatic Voice Query Service can extremely reduce the artificial cost, which could improve the response efficiency for users. The automatic speech recognition (ASR) is one of the important component in AVQS. However, many dialect areas in China make the AVQS have to response the multi-accented Mandarin users by single acoustic model in ASR. This problem severely limits the accuracy of ASR for multi-accented speech in the AVQS. In this paper, a new framework for AVQS is proposed to improve the accuracy of response. Firstly, the fusion feature including iVector and filterbank acoustic features is used to train the Transformer-CTC model. Secondly, the transformer-CTC model is used to construct the end-to-end ASR. Finally, key words matching algorithm for AVQS based on fuzzy mathematic theory is proposed to further improve the accuracy of response. The results show that the final accuracy in our proposed framework for AVQS arrives at 91.5%. The proposed framework for AVQS can satisfy the service requirement of different areas in mainland of China. This research has a great significance for exploring the application value of artificial intelligence in the real scene.

Keywords: Automatic Voice Query Service; Automatic Speech Recognition; Multi-Accented Mandarin Speech Recognition

1. Introduction

The Telephone/Mobile phone (T/M) Automatic Voice Query Service (AVQS) of is one important application research in the field of intelligent speech communication [1, 2]. Users could fetch their required information by T/M AVQS. The cost of human resource is too high for manual voice query service because the amount of customers is too large, for example recently about 1.597 billion users receive the mobile phone service in China. Therefore, AVQS is an effective solution for decreasing the cost of human resource. In AVQS, the automatic speech recognition (ASR) is one of the key part [3]. However, the accented Mandarin speech makes the ASR become a great challenge in AVQS [4, 5].

Besides, the uncertainty of user townships increases the difficulty of ASR task for Mandarin speech. In China, seven main dialect townships include "Mandarin, Cantonese, Wu, Xiang, Min and Gan and Kejia" [6]. Dialect is the first language (native language) of speaker, while the other speech may be their second language. There is a great difference between the second language pronunciation and native language pronunciation of speaker, such as rhythm and tone variation [7]. This difference makes the speech recognition for second language be a great challenge [8, 9]. Pronunciation of Mandarin is different from the pronunciation of other dialects, therefore the Mandarin speech spoken by dialect speaker is the accented Mandarin speech and the acoustic domain of Mandarin speech does not match the acoustic domain of accented Mandarin speech [10]. Undoubtedly, this mismatch would further improve the difficulty of ASR for AVQS. Several researches were worked to solve the mismatch problem. These methods could be categorized into two main types: dictionary adaption [11-16] and model adaptation [17-22]. Dictionary adaptation focuses on variation of phoneme, for example, expansion of phoneme list and pronunciation vocabulary are used generally. However, the dictionary adaptation may lead to the confusion of lexical. Therefore, the effect of dictionary adaptation is limited. The model adaptation method is used to reduce the confusion of pronunciation by acoustic model based on the phoneme level or state of Hidden Markov Model (HMM) [15, 16]. The model adaptation method focuses acoustic variation, and it requires a large amount of accented speech to train the acoustic model directly.

AVQS is a complicated application field of ASR. In this task, not only multi-accented Mandarin speech requires to be modeled, but also the different severity accented Mandarin speech should be processed. Therefore, our purpose in this paper is to explore how to improve the accuracy of multi-accented Mandarin speech recognition. Recently, the speaker identification feature, such as iVector has been used in the accented speech recognition [23-25], and its experimental results illustrate that the fusion features including speaker identification and acoustic features are very useful to improve the accuracy of accented speech recognition. Moreover, the end-to-end ASR has a great performance due to the development of ASR [26-28], especially the connectionist temporal classifier (CTC) for neural networks is very useful in end-to-end ASR. In this paper, we proposed a novel framework to improve the accuracy of ASR for AVQS. This framework includes three main parts: 1) the fusion features include iVector and filterbank; 2) the end-to-end ASR; 3) the key words matching algorithm based on fuzzy mathematic theory. Especially, the key words matching algorithm is designed according to the mismatch between pronunciations of accented Mandarin speech and pronunciations of standard Mandarin speech.

The contributions in this paper mainly include: 1) exploring the suitable ASR, which has a good robustness for multi-accented Mandarin speech, in the application of AVQS; 2) iVector and filterbank are fused into the fusion features, which are used to train and test the ASR for multi-accented Mandarin speech; 3) the key words matching algorithm is proposed to further improve the response accuracy of AVQS according to the mismatch of accented Mandarin speech and standard Mandarin speech.

The rest of this paper includes that Section 2 introduces the overview of the AVQS framework; Section 3 introduces the end-to-end ASR for AVQS; Section 4 introduces the fusion features including iVector and filterbank features; Section 5 introduces the key words matching algorithm based on the fuzzy mathematics theory; Section 6 includes the experiment setup and results; Section 7 introduces the discussion and conclusion of our research.

2. Framework of AVQS for T/M speech

Two parts in framework of T/M AVQS, shown in figure 1, are query request of T/M voice and query response of T/M voice, respectively. Three steps for query request of T/M voice are as follows: 1) extraction of fusion feature including iVector and acoustic features; 2) ASR; 3) extraction of keywords based on named entity recognition (NER). In addition, two main steps for query response of T/M voice are as follows: 1) fuzzy matching for keywords; 2) answering based Text-To-Speech (TTS).

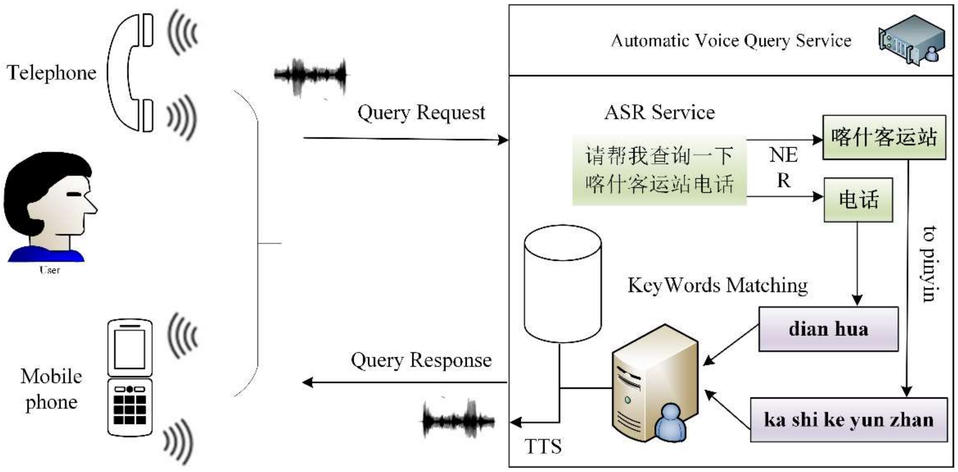

Figure 1. Overview of T/M AVQS framework.

Figure 1 shows the overview of AVQS in T/M environment. Two types of users include "telephone users" and "mobile phone users". After the voice of user is dialing into server, the ASR service would process the query request from user. Then, the NER procedure would extract the keywords from the recognition result. Another procedure would finish the matching process according to the record pre-saved in the database. Finally, the TTS procedure would translate the result selected from database into speech and send the speech to user. The above process is one round of voice query interaction in T/M AVQS.

3. ASR based on Transformer-CTC in AVQS

Transformer is proposed by Ashish Vaswani et al [29], which is a typical seq2seq model performing better than the BiRNN on machine translation. Recently, some researches about ASR based on seq2seq are proposed to improve the accuracy [30, 31]. In this paper, we proposed the Transformer-CTC-based ASR to get the content from T/M speech referring [32]. The ESPNet toolkit [33] is used as the basis for developing ASR.

doi:10.20944/preprints202103.0513.v1

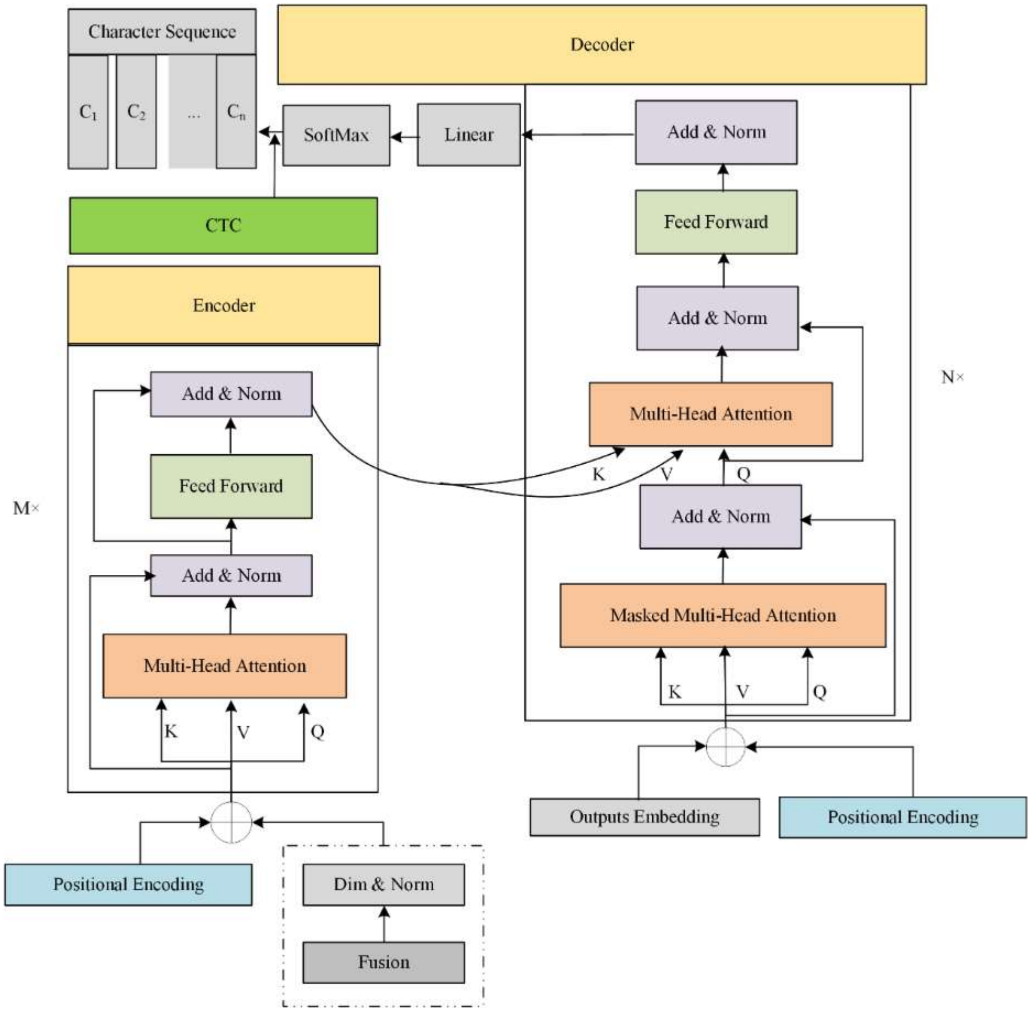

Figure 2. Framework of ASR based on Transformer-CTC

In Figure 2, the fusion features are normalized as input sequence of Transformer Encoder. The output label embedding is as the Transformer Decoder. The Encoder, Decoder and CTC are used to train seq2seq ASR based on Transformer-CTC.

3.1 Encoder Stack

The Encoder of Transformer has M (=6) same neural networks stack. In this stack, the first sublayer of each layer is designed based on "Multi-head Attention", and the second sublayer is a simple full-connected feed-forward neural networks. Surrounded every two sublayers, one keep-residual connection layer [34] and one normalization layer are used [35]. They can be calculated by Equation (1),

$$
\begin{gathered}
h^t = f(\tfrac{g}{\sigma^t} \odot (a^t - \mu^t) + b) \\
\mu^t = \tfrac{1}{H} \textstyle\sum_{i=1}^{H} a_i^t \\
\sigma^t = \sqrt{\tfrac{1}{H} \textstyle\sum_{i=1}^{t} (a_i^t - \mu^t)^2}
\end{gathered}
\tag{1}
$$

where b and g are defined as the bias and gate parameters with same dimension of h^t, respectively. In equation $a^t = W_{hh} h^{t-1} + W_{xh} x^t$, W_{hh} represents the weights of recurrent hidden layer, and W_{xh} represents the weights of line that from input layer to hidden layer. Please note that, the outputs of every sublayer are $\mathrm{LayerNorm}(\mathrm{x} + \mathrm{Sublayer}(\mathrm{x}))$, where $\mathrm{Sublayer}(\mathrm{x})$ is achieved by every sublayer themselves.

3.2 Decoder Stack

Similar to Encoder, the Decoder is designed by stack including N(=6) same neural networks layers. Different from Encoder, the Decoder has three sublayers. The third layer of Decoder processes the outputs of Encoder stack based on multi-head attention mechanism. Every sublayer of Decoder is designed by keep-residual connectionist neural networks. Following each sublayer, normalization is operated. The modification of every sublayer based on self-attention is proposed to prevent the model from paying too much attention to the follow-up positions. Combing with the embedding outputs, the above operations could ensure that all the predictions depends on the previous outputs.

3.3 Attention Mechanism

Attention in Transformer could be described as the map of key, value and query, where the key, value and query are all vectors. The final results are the weighted sum of outputs.

3.3.1 Scaled Dot-Product Attention

The attention mechanism in this paper is designed based on scaled dot-product attention. The input does not only include query and key, whose dimensions are d_k, but also include the value, whose dimension is d_v. The detail process of calculation is shown as equation (2).

$$\mathrm{SDPA}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})\mathrm{V} \tag{2}$$

where SPDA is the scales dot-product attention mechanism; the SPDA is used in Transformer due to its low time complexity and space complexity. Larger the d_k dimension is, the more cost of calculation for dot-product is. Softmax function is used to decrease the gradients extremely. In equation (2), the coefficient $\frac{1}{\sqrt{d_k}}$ is used to counteract this effect.

3.3.2 Multi-Head Attention

Different from single attention mechanism, Transformer processes the linear results of query and key with d_k dimension, and value with d_v dimension.

The multi-head attention could capture the different positions of subspace representation. However, if the model has only one head, the averaging operation would suppress such scattered representation. The multi-head attention could be calculated by equation (3).

$$\mathrm{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Concat}(head_1,..,head_h)W^O \tag{3}$$

where $head_i$ could be obtained by equation (5), $W^O \epsilon \mathbb{R}^{hd_v \times d_{model}}$.

$$head_i = \mathrm{SDPA}\left(\mathrm{Q}W_i^Q, \mathrm{K}W_i^K, \mathrm{V}W_i^V\right) \tag{4}$$

where the parameter matrixes of projection are $W_i^Q \epsilon \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \epsilon \mathbb{R}^{d_{model} \times d_k}$ and $W_i^V \epsilon \mathbb{R}^{d_{model} \times d_v}$, respectively.

3.4 CTC

CTC is a connectionist temporal classification model proposed by Alex Graves et al. [36] in the application of temporal classification for neural networks, can solve the problem that training data in ASR require pre-segmentation and post-processing for label sequences. This requirement constrains the performance of neural networks. The CTC model performs better than HMM because it can predict the corresponding label sequences directly according to the unsegmented input data. Especially, a L-length sequence includes some Chinese characters in CTC, such as $\mathrm{C} = \{c_l \in \mu| = l = 1,2, \ldots, \mathrm{L}\}$. In addition, blank symbol '<b>' is also used in CTC to define the boundary of a word. The set C' with '<b>' can be defined as equation (5).

$$C' = \{< b >, c_1, < b >, c_2, \ldots, c_L, < b >\} = \{c_l' \epsilon \mu \cup \{< \mathrm{b} >\} | l = 1,2, \ldots, L\} \tag{5}$$

where c_l' would be '<b>' if l is an odd number; and c_l' would be Chinese character if l is an even number. The acoustic model could be calculated by CTC, such as equation (6).

$$\mathrm{p(Z|X)} = \prod_{t=1}^{T} p(z_t|z_{1,},\dots,z_{t-1},X) \approx \prod_{t=1}^{T} p(z_t|X) \tag{6}$$

where X denotes the input; z denotes the outputs; T is the amount of frames. Especially, CTC obeys the conditional independent assumption rule, therefore we can obtain that $\mathrm{p(Z|X)} \approx \prod_{t=1}^{T} p(z_t|X)$. Moreover, the length of hidden layer sequence should be less than the length of input sequence. The acoustic model is constructed based on Transformer, and the probability of every state could be calculated by equation (7).

$$p(z_t|X) = Softmax(Linear(Transformer_t(X))) \tag{7}$$

where $Softmax(\cdot)$ is chosen as the active function; $Linear(\cdot)$ denotes the linear layer operation for converting the vectors of hidden layers; $Transformer_t(\cdot)$ catches all of the inputs and outputs vectors of hidden layer at the moment t. The CTC model for character sequence is shown as equation (8).

$$\mathrm{p(C|}Z) = \frac{p(Z|C)p(C)}{p(Z)} = \prod_{t=1}^{T} p(z_t|z_1,\dots,z_{t-1},C)\frac{p(C)}{p(Z)} \approx \prod_{t=1}^{T} p(z_t|z_{t-1},C)\frac{p(C)}{p(Z)} \tag{8}$$

where $p(z_t|z_{t-1},C)$, $p(C)$ and $p(Z)$ are the state transition probability, language model probability for Chinese characters and prior probability of state, respectively. CTC decode the ASR for a sequence of Chinese characters according to $p(C)$ and finite state machine based on language model.

4. Fusion Features for Training ASR

Dehak et al. [37] proposed iVector for speaker identification, which is a milestone in the field of speaker identification. iVector might effectively improve the accuracy of accented speech recognition. Therefore, in this paper, iVector and acoustic features are fused into fusion feature for training ASR of AVQS. The iVector could be obtained by equation (9), and the filterbank is used as the acoustic feature.

$$\mathrm{w} = (I + T^t\Sigma^{-1}N(u)T)^{-1}T^t\Sigma^{-1}\tilde{F}(u) \tag{9}$$

where $N(u)$ denotes a diagonal matrix of $\mathrm{CF} \times \mathrm{CF}$ dimension, whose diagonal blocks are $N_C I(\mathrm{c} = 1,2,\dots,\mathrm{C})$; C represents the components of Gaussian; F represents the dimension of feature space; $\tilde{F}(u)$ is a supervector of $\mathrm{CF} \times 1$ dimension; Σ is a diagonal covariance matrix of $\mathrm{CF} \times \mathrm{CF}$ dimension. T is the total variability matrix [36]. In this paper, the open source toolkit kaldi is used to extract the iVector, and the iVector and acoustic features are composed into the fusion features.

5. Key Words Matching Algorithm for AVQS based on Fuzzy Mathematic Theory

The difference between accented Mandarin speech and standard Mandarin speech leads to the poor accuracy of ASR. And the poor accuracy of ASR further makes the AVQS have a low response accuracy. Therefore, the key words matching algorithm based on fuzzy mathematic theory for AVQS is proposed to further improve the response accuracy of AVQS. The key words matching algorithm is on the basis of pinyin syllable level. After getting the content from T/M speech, the key words could be obtained by named entity recognition (NER). Then the key words would be transformed into pinyin sequence, and the pinyin of key words pre-saved in database is used to match the pinyin sequence obtained by ASR utilizing key words matching algorithm. Finally, the result by matching process would be synthesized into speech and send to user. The whole process could be found in figure 1.

The easily wrong pronunciations of key words are counted statistically, and according to the mapping relationship between easily incorrect pronunciations and correct pronunciations the dictionary is constructed. Finally, the degree of membership could be obtained by equation (10) based on edit distance.

$$\alpha = 1 - \frac{D}{T_{pinyin}} \tag{10}$$

where D denotes the edit distance (also called Levenshtein distance); T represents the total number of characters in one pinyin syllable. During the matching process, the best result would be returned according to the biggest degree of membership.

6. Experiment

doi:10.20944/preprints202103.0513.v1

Several experiments are designed to evaluate the performance of proposed framework for AVQS. The experiments include three parts: 1) comparison of different ASR methods in AVQS testing data; 2) comparison of filterbank and fusion features for AVQS response; 3) evaluation for key words matching algorithm based on fuzzy mathematic theory.

6.1 Experiment Condition Setup

The configuration of parameters for ASR based on Transformer-CTC is shown in Table 1. In this paper, this configure is used in all experiments.

Table 1. Configuration of Parameters for ASR based on Transformer-CTC

Parameter Name	Value
encoder layer	12
encoder unit	2048
decoder layer	6
decoder unit	2048
attention dimension	256
attention head	4
hybrid CTC/attention alpha	0.3
label smoothing weight	0.1
batch size	16
max length of input	512
max length of output	150
transformer learning rate	1.0
transformer attention dropout rate	0.0
CTC beam size	10
CTC weight	0.5
language model weight	0.7
n-gram weight	0.3

6.2 Data Preparation

AIShell-1 (AIShell) speech corpus [38] is chosen in this paper, which is the sub-dataset of AIShell-ASR0009. The speech corpus is recorded by 400 speakers including multi-speaker accented speech. In this paper, the AIShell speech is also used to obtain iVector. The real T/M speech corpus is supported by 114 of China Telecom (114 is a voice service of China Telecom Corporation). In the experiment, the accented Mandarin speech can be separated into 7 areas (standard Mandarin, Cantonese, Wu, Xiang, Min and Gan and Kejia) according to the characteristics of speakers' dialect area. According to the severity of accented Mandarin speech, in this paper, the accented Mandarin speech is split into three severities including "light, medium and heavy". The details of the above speech corpus for training and testing ASR are shown in Table 2.

Table 2. Configuration of Speech Corpus for Training and Testing ASR.

Speech Corpus	Time	Training Amount	Validating Amount	Usage for ASR
AIShell	≈178h	121925 utterances	20000 utterances	Pre-training
T/M Voice Corpus	≈30h	8000 utterances	2000 utterances	Training and Testing

In Table 2, 2000 utterances of speech, which did not participate the training process, were selected from real T/M Voice Corpus to evaluate the performance of the ASR. The voice corpus for training and testing AVQS includes the three severities of accented Mandarin speech, which are "light, medium and heavy", respectively. In addition, the voice corpus also includes all of the 7 dialect-area speech. The details of configuration are shown in Table 3.

Table 3. The Accented Speech from Different Dialect Area for Testing AVQS.

Dialect Area	Accented Level	Validating Amount (utterances)
Mandarin Speech	Standard	200

	light	100
Cantonese	medium	100
	heavy	100
	light	100
Wu	medium	100
	heavy	100
	light	100
Xiang	medium	100
	heavy	100
	light	100
Min	medium	100
	heavy	100
	light	100
Gan	medium	100
	heavy	100
	light	100
Kejia	medium	100
	heavy	100

In Table 3, the standard Mandarin speech has 200 utterances; the others have 300 utterances including light (100 utterances), medium (100 utterances) and heavy (100 utterances).

6.3 Results of AVQS

Character Error Rate (CER), Sentence Error Rate (SER), Key Word Error Rate (KWER) and Response Error Rate (RER) of AVQS (which is optimized by key words matching algorithm). They could obtained by equation (11), equation (12), equation (13) and equation (14), respectively.

$$\mathrm{CER} = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+I+C} \tag{11}$$

$$\mathrm{SER} = \frac{E_S}{T_S} \tag{12}$$

$$\mathrm{KWER} = \frac{E_{kw}}{T_{kw}} \tag{13}$$

$$\mathrm{RER} = 1 - \frac{Acc_S}{T_S} \tag{14}$$

where S denotes the amount of substitutions; D denotes the amount of deletions; I denotes the amount of insertions; C denotes the amount of corrects. This calculation method is different from article [39], because we think that the error rate should not be more than 1. In addition, E_s is the amount of sentences with wrong characters in testing dataset; T_s is the total amount of sentences in testing dataset. E_{kw} denotes the amount of errors of key words; T_{kw} denotes the total amount of key words. Acc_S denotes the amount of correct response sentences of AVQS; T_S denotes the total amount of testing sentences of AVQS..

6.3.1 Accuracy of ASR

Different ASR methods are compared in this experiment, where the filterbank is only used to train the acoustic model of ASR. The results of accented Mandarin speech are shown in Table 4, Table 5 and Table 6 based on T/M testing dataset.

Table 4. Results of Different ASR for Testing Dataset (light accented)

Method	Mandarin	Cantonese	Wu	Xiang	Min	Gan	Kejia	Mean	SER
DNN-HMM	20.67%	32.15%	26.12%	27.15%	30.01%	22.28%	29.66%	26.86%	57.01%

doi:10.20944/preprints202103.0513.v1

bilstm-CTC	15.15%	25.21%	22.05%	25.18%	25.14%	19.11%	25.27%	22.44%	55.15%
Transformer-CTC	11.23%	18.77%	20.11%	20.15%	21.33%	16.25%	20.21%	18.29%	45.20%

Table 4 shows the recognition results of different ASR methods for light multi-accented Mandarin speech. In Table 4, six types of accented Mandarin speech and standard Mandarin speech are tested. For DNN-HMM, the CERs of Cantonese, Min and Kejia accented Mandarin speech are close to each other; the CERs of Wu and Xiang accented Mandarin speech are close to each other; the Gan accented Mandarin speech has the lowest CER; the average CER of DNN-HMM is 26.86%. For bilstm-CTC, the CERs of Cantonese, Xiang, Min and Kejia accented Mandarin speech are close to each other; the CERs of Wu and Gan accented Mandarin speech are close to each other; the average of CER is 22.44%. For Transformer-CTC, the CERs of Wu, Xiang, Min and Kejia are close to each other; the CERs among different accented Mandarin speech have no significant difference. In addition, the SERs of different baseline ASR methods are above 55%, and the SER of Transformer-CTC is the lowest.

Table 5. Results of Different ASR for Testing Dataset (medium accented)

Method	Cantonese	Wu	Xiang	Min	Gan	Kejia	Mean	SER
DNN-HMM	42.17%	43.15%	38.57%	41.51%	50.18%	38.16%	42.29%	76.51%
bilstm-CTC	36.31%	31.75%	33.41%	35.64%	38.35%	35.77%	35.21%	65.35%
Transformer-CTC	25.15%	29.31%	30.18%	27.33%	29.25%	27.21%	28.07%	57.32%

Table 5 shows the results of ASR for medium multi-accented Mandarin speech. For DNN-HMM, the CERs of Cantonese, Wu and Min are close to each other; the CERs of Xiang and Kejia are close to each other; the average of CER is 42.29%. For bilstm-CTC, the CERs of Cantonese, Min, Gan and Kejia are close to each other; the average of CER is 35.21%. For the Transformer-CTC, all of the CERs are close to each other; the average of CER is 28.07%. In addition, the SERs of all ASR for medium accented Mandarin speech arrive above 57%, and the CERs of Gan accented Mandarin speech is the highest.

Table 6. Results of Different ASR for Testing Dataset (heavy accented)

Method	Cantonese	Wu	Xiang	Min	Gan	Kejia	Mean	SER
DNN-HMM	65.31%	52.11%	45.21%	65.72%	73.28%	68.53%	61.69%	85.30%
bilstm-CTC	56.22%	46.19%	47.18%	41.22%	51.95%	50.11%	48.81%	78.15%
Transformer-CTC	34.11%	38.23%	43.27%	40.11%	39.02%	36.33%	38.51%	65.05%

Table 6 shows the recognition results of different ASR for heavy multi-accented Mandarin speech. Please note that, the heavy accented Mandarin speech has very low intelligibility, because the speaker could not pronounce the vowel and consonant correctly. For DNN-HMM, the CERs of Cantonese, Min and Kejia are close to each other, and the CER of Gan is the highest; the average of CER is 61.69%. For bilstm-CTC, the differences among accented Mandarin speech decrease, where the CERs of Wu, Xiang and Min are close to each other and the CERs of Gan and Kejia are close to each other; the average of CER is 48.81%. For the Transformer-CTC, the differences among ASRs further decrease; the average CER is 38.51%. However, the SERs of different ASR for heavy accented speech arrives above 65%.

6.3.2 Results of Fusion Features

In this paper, the fusion feature and filterbank feature are also compared based on testing dataset. During this experiment, only the Transformer-CTC is used to test. The results can be found in Table 7.

Table 7. Results of Trained ASR for Fusion Features

Feature Style	light	medium	heavy
filterbank	18.29%	28.07%	38.51%
fusion	13.10%	20.68%	30.05%

Table 7 shows the results of comparison between the filterbank and fusion features for light, medium and heavy multi-accented Mandarin speech, respectively. Especially, the results in Table 7 are the average CERs of light accented, medium accented and heavy accented speech, respectively. From Table 7, the fusion features can obviously decrease the CERs.

6.3.3 Results of Key Words Matching Algorithm based on Fuzzy Mathematic Theory

In this paper, the key words matching algorithm based on fuzzy mathematic theory are evaluated according to the testing dataset. This experiment is on the basis of trained ASR. The results are shown in Table 8.

Table 8. Error Rate of Different ASR for AVQS.

Average Error Rate	light	medium	heavy
CER	13.10%	20.68%	30.05%
KWER	10.20%	16.50%	22.65%
Response Error Rate (RER)	8.50%	12.63%	18.35%

Table 8 shows the error rate of AVQS, where the KWER represents the error rate of key words in one sentence; the RER represents the error rate of AVQS response after being optimized by the key words matching algorithm based on fuzzy mathematic theory. Obviously, the optimization algorithm can effectively improve the accuracy of AVQS response. The highest accuracy of AVQS arrives at 91.5%, which makes the AVQS satisfy the requirement of the different speakers in China mainland.

7. Discussion and Conclusion

AVQS is an interesting application of artificial intelligence. The powerful AVQS can greatly reduce the cost of human resource, which can improve the efficiency of response. Especially, China has a large amount of people which is a big market. Therefore, it has a great significance to explore how to improve the performance of AVQS for Mandarin speaker.

ASR is one key component of the AVQS. However, some problems of ASR in AVQS should be solved such as the customers with severe accented Mandarin speech. The customers cannot pronounce Mandarin speech standardly enough. In addition, not only the difference between accented Mandarin speech and standard Mandarin speech is too large, but also the differences among different dialect are so large that the different accented Mandarin speech also has a significant difference. The above problems severely limit the performance of AVQS. Therefore, a novel framework is proposed to improve the performance of AVQS, where the framework includes three main parts: fusion features including iVector and acoustic features extraction; the ASR based on Transformet-CTC; key words matching based fuzzy mathematic theory.

The results illustrate that the recognition accuracy of light, medium or heavy accented Mandarin speech has a significant difference. The AVQS need face different speakers with different accented Mandarin speech. Especially, the information technology makes the AVQS more and more popular in the users. However, some users with heavy dialect-accented Mandarin speech extremely limited the response accuracy. In this paper, the end-to-end ASR and the fusion features could effectively improve the accuracy of multi-accented Mandarin speech recognition. Furthermore, we proposed a new key words matching algorithm based on fuzzy mathematic theory according to the statistical characteristics of pronunciation, whose phoneme may occur errors easily. This proposed algorithm can effectively further improve the response accuracy of AVQS. The experimental results show that the highest response accuracy of AVQS arrives at 91.5%. The proposed framework can effectively improve the whole response accuracy of AVQS for light, medium and heavy accented Mandarin speech.

Author Contributions: Conceptualization, Zhaopeng, Qian. and Kejing, Xiao.; methodology, Zhaopeng, Qian.; software,Zhaopeng, Qian.; validation, Zhaopeng, Qian and Kejing, Xiao; formal analysis, Zhaopeng, Qian.; investigation, Keijing, Xiao.; resources, Zhaopeng, Qian.; data curation, Zhaopeng, Qian.; writing—original draft preparation, Zhaopeng, Qian.; writing—review and editing, Kejing, Xiao.; visualization, Zhaopeng, Qian.; supervision, Kejing, Xiao.; project administration, Kejing, Xiao.; funding acquisition, Kejing, Xiao. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Beijing Natural Science Foundation (No. 4202014) and Open Project Program of National Engineering Laboratory for Agri-product Quality Traceability, (No. AQT-2018-YB4).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Not Available.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Song, Y I., Wang,Y Y., Ju,Y C., Seltzer, M., Tashev, I., Acero, A. Voice search of structured media data, 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19-24 April 2009. pp. 3941-3944. DOI: https://doi.org/ 10.1109/ICASSP.2009.4960490
2. Wang Y Y. Voice search - Information access via voice queries. 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), Kyoto, Japan. 9-13 Dec. 2007. pp.123-123. DOI: https://doi.org/10.1109/ASRU.2007.4430095
3. Moreno-Daniel A, Wilpon J, Juang B H. Index-based incremental language model for scalable directory assistance. Speech Communication, 2012, 54(3):351-367. DOI: https://doi.org/10.1016/j.specom.2011.09.006
4. Lee, L., Rose, R. A frequency warping approach to speaker normalization. IEEE Transactions on speech and audio processing, 1998, 6(1): 49-60. DOI: https://doi.org/10.1109/89.650310
5. Sun, S., Yeh, C F., Hwang, M Y., Ostendorf, M., Xie, L. Domain adversarial training for accented speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada, 15-20 April 2018, pp. 4854-4858. DOI: https://doi.org/10.1109/ICASSP.2018.8462663
6. Liu, Y., Fung, P. Partial change accent models for accented Mandarin speech recognition. 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721), St Thomas, VI, USA, 30 Nov.-4 Dec. 2003, DOI:// https://doi.org/10.1109/ASRU.2003.1318413
7. Adams C., Munro R R. In search of the acoustic correlates of stress: fundamental frequency, amplitude, and duration in the connected utterance of some native and non-native speakers of English. Phonetica, 1978, 35(3): 125-156. DOI: https://doi.org/10.1159/000259926
8. Bradlow A R, Bent T. Perceptual adaptation to non-native speech. Cognition, 2008, 106(2): 707-729. DOI: https://doi.org/10.1016/j.cognition.2007.04.005
9. Wilson, E O., Spaulding, T J. (2010). Effects of noise and speech intelligibility on listener comprehension and processing time of Korean accented English, Journal of Speech, Language, and Hearing Research. 53, 1543–1554. DOI: https://doi.org/10.1044/1092-4388(2010/09-0100)
10. Elfeky M., Bastani M., Velez X., Moreno, P., Waters, A. Towards acoustic model unification across dialects. 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13-16 Dec. 2016. pp. 624-628. DOI: https://doi.org/10.1109/SLT.2016.7846328
11. Huang C, Chen T, Chang E. Accent issues in large vocabulary continuous speech recognition. International Journal of Speech Technology, 2004, 7(2): 141-153. DOI: https://doi.org/10.1023/B:IJST.0000017014.52972.1d
12. Arslan, L M., Hansen J H L. A study of temporal features and frequency characteristics in American English foreign accent. The Journal of the Acoustical Society of America, 1997, 102(1): 28-40. DOI: https://doi.org/10.1121/1.419608
13. Ding, G H. Phonetic Confusion Analysis and Robust Phone Set Generation for Shanghai-Accented Mandarin Speech Recognition. 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008), Brisbane, Australia, September 22-26, 2008, pp. 1129-1132
14. Fosler-Lussier E., Amdal I., Kuo H K J. A framework for predicting speech recognition errors. Speech Communication, 2005, 46(2): 153-170. DOI: https://doi.org/10.1016/j.specom.2005.03.003
15. Fosler-Lussier, E. Dynamic Pronunciation Models for Automatic Speech Recognition. Ph.D. dissertation, International Computer Science Institute, Berkeley, CA, USA. 1999
16. Hain, T., Woodland, P C. (1999). Dynamic HMM Selection for Continuous Speech Recognition. Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary, September 5-9, 1999, pp. 1327–1330.
17. Liu, Y., Fung, P. Multi-accent Chinese Speech Recognition. Ninth International Conference on Spoken Language Processing (INTERSPEECH 2006 - ICSLP), Pittsburgh, PA, USA, September 17-21, 2006, pp. 1887-Mon1BuP.8
18. Fung, P., Liu, Y. Effects and modeling of phonetic and acoustic confusions in accented speech. The Journal of the Acoustical Society of America, 2005, 118(5): 3279-3293. DOI: https://doi.org/10.1121/1.2035588
19. Zhang, C., Liu, Y., Xia, Y., Wang, X., Lee, C H. Reliable accent-specific unit generation with discriminative dynamic Gaussian mixture selection for multi-accent Chinese speech recognition. IEEE transactions on audio, speech, and language processing, 2013, 21(10): 2073-2084. DOI: https://doi.org/10.1109/TASL.2013.2265087
20. Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., Yoon, S. (2005). Accent Detection and Speech Recognition for Shanghai-accented Mandarin. Eighth International Conference on Spoken Language Processing (Interspeech'2005 – Eurospeech). Lisbon, Portugal, September 4-8, 2005, pp. 217-220.

21. Vergyri, D., Lamel, L., Gauvain, L. (2010). Automatic Speech Recognition of Multiple Accented English Data. 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), Makuhari, Chiba, Japan, September 26-30. 2010, pp. 1652-1655
22. Fischer, V., Gao, Y., Janke, E. Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer. 9th International Conference on Spoken Language Processing (INTERSPEECH 1998). Sydney, Australia, November 30 - December 4, 1998, pp. 0233
23. DeMarco, A., Cox, S J. Native accent classification via i-vectors and speaker compensation fusion, 14th International Conference on Spoken Language Processing (INTERSPEECH 2013). Lyon, France, August 25-29, 2013, pp. 1472-1476
24. Bahari, M H., Saeidi, R., Hamme, H V., Leeuwen, D.V., Accent Recognition Using I-vector, Gaussian Mean Supervector And Gaussian Posterior Probability Supervector For Spontaneous Telephone Speech, 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, BC, Canada, 26-31 May 2013, pp. 7344-7348, DOI: https://doi.org/10.1109/ICASSP.2013.6639089
25. Chen, M., Yang, Z., Liang, J., Li, Y., Liu, W. Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015). Dresden, Germany, September 6-10, 2015, pp. 3620-3624.
26. Zhou, S., Dong, L., Xu, S., Xu, B. Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin Chinese. 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018). Hyderabad, India, September 2-6 2018, pp. 791-795. https://doi.org/10.21437/Interspeech.2018-1107.
27. Zhou S., Dong L., Xu S., Xu, B. A Comparison of Modeling Units in Sequence-to-Sequence Speech Recognition with the Transformer on Mandarin Chinese. Neural Information Processing. ICONIP 2018. Lecture Notes in Computer Science, 11305. Springer, Cham. https://doi.org/10.1007/978-3-030-04221-9_19
28. Zhou S., Xu S., Xu B. Multilingual end-to-end speech recognition with a single transformer on low-resource languages. arXiv preprint arXiv:1806.05059, 2018.
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. Gomez, A N., Kaiser, L U., Polosukhin, I. Attention is All you Need, Neural Information Processing Systems, Long Beach California, USA, December 4-10 2017, pp. 5998–6008.
30. Dong, L., Xu, S., Xu, B. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15-20 April 2018, pp. 5884-5888. DOI: https://doi.org/10.1109/ICASSP.2018.8462506
31. Karita, S., Soplin, N E Y., Watanabe, S., Delcroix, M., Ogawa, A., Nakatani, T. Improving Transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration, 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019), Graz, Austria, September 15–19, 2019, 1408-1412
32. Watanabe, S., Hori, T., Kim, S., Hershey, J R. Hayashi, T. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8):1240-1253. DOI: https://doi.org/10.1109/JSTSP.2017.2763455
33. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, NEY., Heymann, J., Wiesner, M., Chen, NX., Renduchintala, A., Ochiai, T. ESPnet: End-to-End Speech Processing Toolkit. 19th Annual Conference of International Speech Communication Association (INTERSPEECH), Hyderabad, India, September 2-6 2018. pp. 2207-2211
34. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 27-30, pp. 770–778, 2016.
35. Ba, J L., Kiros, J R., Hinton, G E. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.
36. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, Jun 25-29 2006, pp. 369–376. DOI: https://doi.org/10.1145/1143844.1143891
37. Dehak, N., Kenny, P J., Dehak, R., Dumouchel, P., Ouellet, P. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798. DOI: https://doi.org/10.1109/TASL.2010.2064307
38. Bu, H., Du, J., Na, X., Wu, B. Zheng, H. AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline, 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, South Korea. November 1-3 2017. pp. 2472-7695
39. Klakow, D., Peters, J. Testing the correlation of word error rate and perplexity. Speech Communication, 2002, 38(1/2):19-28. DOI: https://doi.org/10.1016/S0167-6393(01)00041-3