Chinese question speech recognition integrated with domain characteristics

Shengxiang Gao, Dewei Kong, Zhengtao Yu*,
Yang Luo, Jianyi Guo and Yantuan Xian

School of Information Engineering and Automation,
Kunming University of Science and Technology,
Kunming 650500, China
Email: gaoshengxiang.yn@foxmail.com
Email: 179958170@qq.com
Email: ztyu@hotmail.com
Email: 348508816@qq.com
Email: gjade86@hotmail.com
Email: xianyantuan@qq.com
*Corresponding author

Abstract: Aiming at domain adaptation in speech recognition, we propose a speech recognition method for Chinese question sentence based on domain characteristics. Firstly, by virtue of syllable association characteristics implied in domain term, syllable feature sequences of domain terms are used to construct the domain acoustic model. Secondly, in decoding process of domain-specific Chinese question speech recognition, we utilise domain knowledge relationship to optimise and prune the speech decoding network generated by language model, to improve continuous speech recognition. The experiments on tourist domain corpus show that the proposed method gets the accuracy of 80.50% on Chinese question speech recognition and the accuracy of 91.50% on domain term recognition.

Keywords: Chinese question speech recognition; domain characteristic; acoustic model library; domain terms; language model; domain knowledge library.

Reference to this paper should be made as follows: Gao, S., Kong, D., Yu, Z., Luo, Y., Guo, J. and Xian, Y. (2019) 'Chinese question speech recognition integrated with domain characteristics', *Int. J. Computational Science and Engineering*, Vol. 19, No. 3, pp.325–333.

Biographical notes: Shengxiang Gao is currently a Lecturer at School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. She received her PhD degree in Metallurgical Control Engineering in 2016 and MS degree in Pattern Recognition and Intelligent System from Kunming University of Science and Technology in 2005. Her research interests are in speech recognition, nature language processing and machine translation.

Dewei Kong received his MS degree from Kunming University of Science and Technology in 2014. His research interests are in speech recognition.

Zhengtao Yu is currently a Professor and PhD supervisor at the School of Information Engineering and Automation, and he is also the Chairman of Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, China. He received his PhD degree in Computer Application Technology from Beijing Institute of Technology, Beijing, China in 2005. His main research interests are in natural language processing, speech recognition, information retrieval and machine translation.

Yang Luo is currently studying as a MS candidate at Kunming University of Science and Technology, Kunming, China. His research interests are mainly in natural language processing, speech recognition.

Jianyi Guo is currently a Professor at the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. Her main research interests are in natural language processing, speech recognition and information retrieval.

Yantuan Xian is currently studying as a PhD candidate at Kunming University of Science and Technology, Kunming, China. His research interests are mainly in natural language processing, speech recognition, information extraction.

1 Introduction

Speech is one of the most natural and the most important forms of communication. And speech recognition achieves the possibility of natural language exchange between human and computer. With the development of speech recognition technology, Chinese continuous speech recognition technology is maturing. Applied to various domains, it plays an important role in people's lives, such as voice call centre, voice GPS orientation search, voice dialling, voice navigation, voice device control and so on (Huang et al., 2004; Hwang et al., 2009; Sun and Zhao, 2014; Rajavel and Sathidevi, 2015). At present, there are some models which are used for Chinese continuous speech recognition, mainly including hidden Markov model (HMM) and artificial neural network (ANN) model. For raising continuous speech recognition accuracy, depending on these models, many methods have been improved. For example, Dong Yu and Li Deng et al, from Microsoft Asia Research Institute, proposed a hybrid model with context-dependent pre-trained deep neural network and hidden Markov model (DNN-HMM) for large vocabulary speech recognition (LVSR) by directly modelling on senones with deep neural network (Dahl et al., 2011, 2012; Mohamed et al., 2010; Saito et al., 2015). American NEC lab's Ronan Collobert and Jason Weston et al. proposed a framework using ANN to conduct multiple task study in natural language processing (Ronan and Jason, 2008; Yang et al., 2011; Barve and Joshi, 2016). Yang et al. (2012) from Institute of Automation, Chinese Academy of Sciences, proposed a speech recognition decoding arithmetic with induced probability for large vocabulary continuous speech recognition. Y.M. Yang et al. proposed a novel factor-aware training framework, named neural network-based multifactor aware joint training, to improve the recognition accuracy for noise robust speech recognition (Qian et al., 2016a, 2016b). In recent years, there are many related works that use deep neural networks to improve speech recognition effect (Maas et al., 2017; Li et al., 2015). These mentioned measures improved accuracy of Chinese continuous speech recognition in different degrees. On universal domain, these methods have good effect on Chinese continuous speech recognition. But on specific domain, their effects are not good. The reason is that both general acoustic models and common language models are very difficult to describe the difference in speech characteristic sequence of domain terms. In fact, in specific domain, there are a lot of association relationships in pronunciation within domain terms, a lot of domain knowledge in domain text, and these domain characteristics can supervise domain speech recognition effectively. Therefore, this paper discusses how to effectively utilise domain characteristic information to improve domain-specific Chinese question speech recognition. By constructing the domain term acoustics model and the domain knowledge base, we integrate them into the speech recognition framework to improve the performance of speech recognition on specific domain.

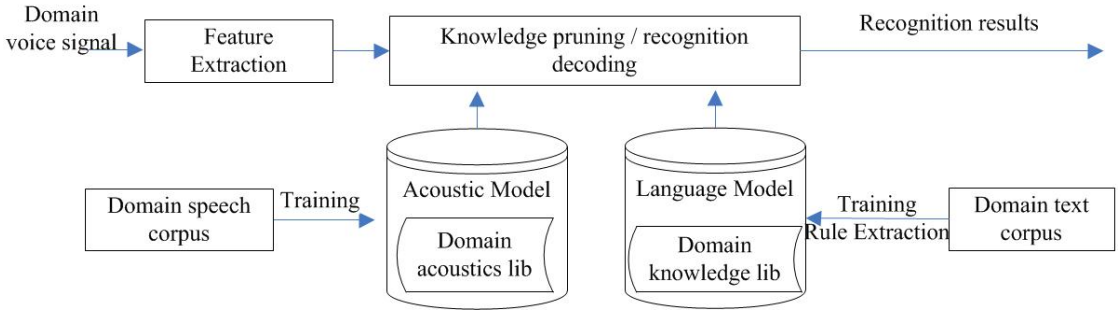
2 Chinese question speech recognition integrated with domain characteristics

2.1 Speech recognition frame integrated with domain characteristic

In speech recognition, building acoustic model is the key step in processing acoustic level. By acoustic modelling, we can transform speech characteristic sequence vector into speech recognition unit. The quality of acoustic model will determine the accuracy of speech recognition to a large extent. In domain-specific question sentences, domain terms and domain vocabularies appears frequently, and glossary of terms is relatively stable. In specified domain term speech, co-articulation phenomenon between syllables is very widespread. In order to retain more information about the associated context pronunciation, in this paper, taking phoneme as an information unit, taking term characteristic vector sequence as a matching template, we build an acoustic library for specific area. In domain-specific voice service system, using this acoustic library, we can effectively improve the recognition results, and also avoid building a large and complex acoustic model. At the language model level, language model is usually used to guide the recognition of linguistic level. For building a practical language model, grammar rules will be very large, expanded paths and search space will also increase sharply. As a result, the decoding efficiency and accuracy will decrease. In domain-specific question sentences, there are some knowledge relationships among entities. Using the knowledge relationships, we can optimise the language model to improve the recognition efficiency. Thus, this paper proposes a pruning idea that uses domain knowledge to optimise decoding search in the language layer.

Based on the idea, this paper constructs an acoustic model library and a knowledge library for specific domain to develop a speech recognition model which adapts to specific fields. The recognition model is shown in Figure 1.

Figure 1 The speech recognition frame integrated with domain characteristic (see online version for colours)



In this domain speech recognition, firstly, we collect speech corpus and text corpus from the tourism domain. Secondly, on the basis of the speech corpus, speech features, including the co-articulation features within domain term, are extracted, the acoustic model is trained, and the domain acoustic library about tourist domain terms is constructed. At the same time, on the basis of the text corpus, domain knowledge features are extracted, the language model is trained and the domain knowledge base is constructed. When a speech signal is inputted as test data, its speech features are extracted, and its acoustic model is constructed. Then they are decoded by being matched with the domain acoustic library, at the same time, the domain knowledge base is used to prune. At the last, the target with the maximum probability is regarded as the recognition result and is outputted.

2.2 The construction of domain acoustic model library

Acoustic model is used to describe acoustic feature of speech. Construction of acoustic model is an important step in Chinese continuous speech recognition. In the process of acoustic modelling, a unit is needed to describe the corresponding relationship between the speech segments and the vector feature sequences. Thus, the base unit selection of acoustic modelling will be the primary problem. In Chinese continuous speech recognition, the choices of base unit are word, initials/finals, syllable, semi-syllable, and phoneme.

In the field of tourism, domain terms and domain-related words occur more frequently, and the term vocabulary is relatively regular. The co-articulation between the syllables is very common. Therefore, when building the tourism domain acoustic model, the selected primitive of speech recognition need to satisfy the following requirements:

1 It can better respond to the phenomenon of articulation between domain terms. The articulation of Chinese continuous speech is very common, but it is more serious for the domain terms which have fixed format and occur frequently. Take tourist attractions as an example, 白云观（White cloud temple）is read as (b ai2 y un2 g u an4) in Chinese, and白元观 (White yuan temple) is read as (b ai2 y u an2 g u an4).
景山公园(Jing-shan park) is read as (j ing3 sh an1 g ong1 y u an2) and金山公园 (Jin-shan park) is read as (j in1 sh an1 g ong1 y u an2). 东方明珠 (Oriental Pearl) is read as (d ong2 f ang2 m ing2 zh u2) and东方明主 (Eastern ming zhu) is read as (d ong2 f ang2 m ing2 zh u3). It is obvious that those pronunciations are regular and related in tone, initials and finals, and their initials/finals combinations. These features need to be responded in primitive selection.

2 The number of primitives should not be too large, especially when constructing context-dependent models, the number of primitives should facilitate the construction of acoustic models. Primitives should benefit speech annotation, speech corpus training and speech recognition. For example, there are more than 1,300 syllables with tone in Chinese, and syllable is obviously not suitable as recognition primitive for questioning speech.

3 It can better respond to phonetic and linguistic knowledge in Chinese speech. Like this, there will be more phonetic and linguistic knowledge to be used to supervise acoustic modeling. Especially, for speech recognition of specific ethnics, the selected primitives should be suitable to the national language features. Only in this way, the primitives will be more beneficial to improve the quality of acoustic model. For examples, the Chinese pronunciations of茶马古道(the ancient tea horse road) are ‘ch a2 m a3 g u3 d a o4’， the Chinese pronunciations of 普达措 (Pudacuo) are ‘p u3 d a2 c uo4’， the Chinese pronunciations of陆良彩色沙林(Colored Sand Forest in Luliang) are ‘l u4 l I ang2 c ai3 s e4 sh a2 l in2’， and the Chinese pronunciations of博白绿珠井(Lv Zhu Jing in Bobai) are ‘b o2 b ai3 l v4 zh u1 j ing3’.

Considering the above factors, this paper chooses initials/finals as the basic element to construct the tourism domain acoustic model. In accordance with the standard set of Chinese initials/finals, which includes 21 initials, 38 finals and four tones, initials can be used with finals, and finals can be used with finals. In this way, it will increase the number of expanded primitives, and will also bring trouble to acoustic tagging. In addition to, when some initials are read as zero tone, for the problem that they cannot be corresponded to initials with zero tone. Therefore,

this paper defines an extended set of initials/finals as the speech recognition primitives. This extended set of initials/finals primitives, using a more regular method, adds six zero-initial primitive elements, so that each syllable has both the initial part and the final part. The six zero-initial primitives are {_a, _e, _o, _I, _v, _u}, respectively. The extended set of initials/finals is shown in Table 1.

After expansion, before and after a final will only be initials or mute. Thus, combining with contextual information, the number of primitives will be significantly less.

Consider the sequence feature of hidden Markov model (HMM), we apply HMM to construct our acoustic model. Each speech primitive can be expressed by three states in HMM. For example, the word of 昆明(Kunming) can be expressed as the following HMM in Figure 2. Among them, the first state and the last state only play the role of connection, no output probability.

In continuous speech, the pronunciation of each phoneme is affected by the co-pronounced before and after, and will show different characteristics. A single-factor model is clearly inappropriate. The context-dependent tri-phoneme model can effectively express the co-articulation regularity between initials and finals of domain terms. Therefore, the context-dependent tri-phoneme model is applied to construct our domain acoustic model library. Its constructing process is shown in Figure 3.

In the above model constructing, a classical continuous probability density HMM is used to model continuous speech in tourism domain. Vocabularies and sentences in the field of tourism are represented by initials/finals strings. The acoustic model is constructed corresponding to the pronouncing strings. For example:

Lable：西双版纳 (Sipsongpanna)

Strings of initials/finals: silence x i1 sh u ang1 b an3 n a4 silence.

Table 1 The initials/finals primitives in tourism domain

Initials (A total of 27)	*Finals (A total of 38)*	*Tones (A total of four)*
b, p, m, f, d, t, n, l,	*eng, en, ei, e, ao, ang, an, ai, a*	*1*
g, k, h, j, q, x,	*ian, ia, i2, i1, I, ou, ong, o, er*	*2*
zh, ch, sh, z, c, s, r,	*iou, iong, ing, in, ie, iao, iang*	*3*
_u, _v, _I, _o, _e, _a	*ueng, uen, uei, uang, uan, uai, ua, u*	*4*
	vn, ve, van, v, uo	

Figure 2 The HMM expression of the Chinese word of 昆明'Kunming' (see online version for colours)

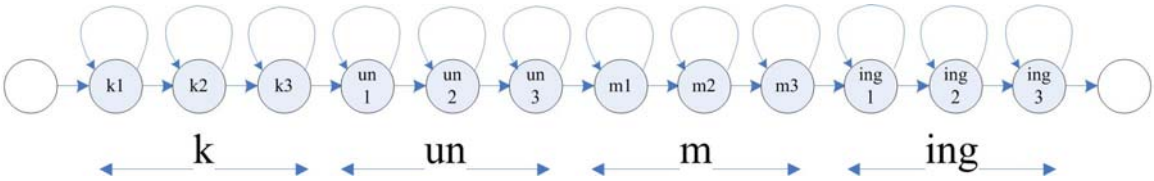
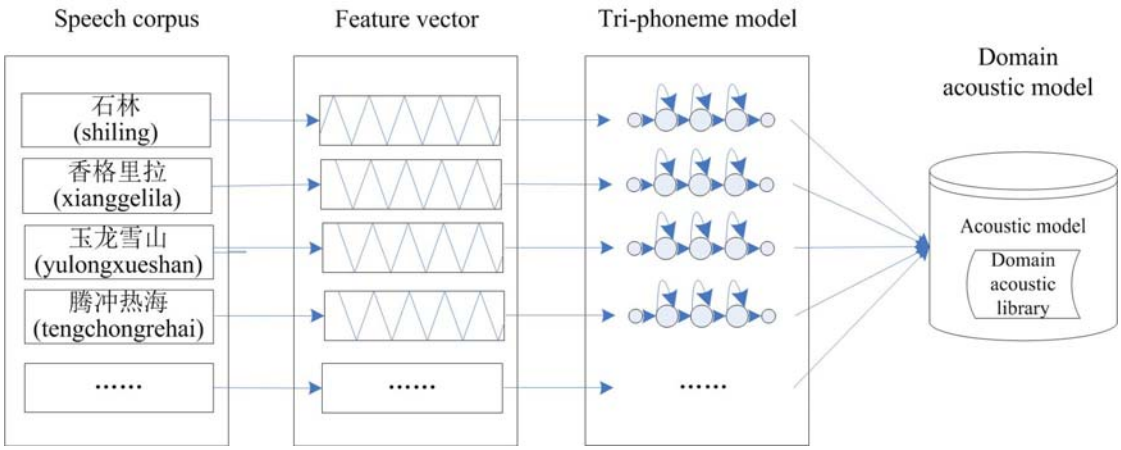
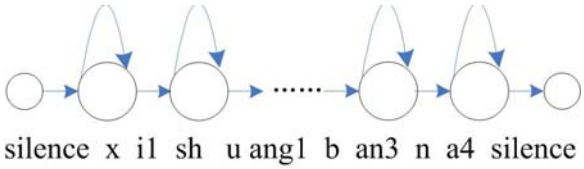


Figure 3 Constructing process of acoustic model library on domain terms (see online version for colours)



The acoustic model for term '西双版纳 (Sipsongpanna)' is constructed as shown in Figure 4.

Figure 4 The constructed acoustic model for term '西双版纳 (Sipsongpanna)' (see online version for colours)



For travel question travel containing, its tri-phoneme model is also constructed in the similar way, for example:

Label: 现在故宫的门票多少钱(How much is the ticket for the Imperial Palace)?

Its strings of initials/finals: silence x i an3 z ai3 g u3 g ong1 de5 m en2 p i ao3 d uo2 sh ao3 q i an2 silence

2.3 A Chinese question speech recognition method integrated with domain acoustic model

In HMM-based continuous speech recognition system, the inputting speech can be seen as a given observation sequence $O=O_1O_2\cdots O_T$. In a trained model $\mu=(A,B,\pi)$, that Viterbi algorithm is used to search the 'optimal' state sequence $Q=q_1q_2\cdots q_T$ so as to explain the observed sequence $O=O_1O_2\cdots O_T$, is the HTK's speech recognition process (Steve and Gunnar, 2009; Ban and Kim, 2015). In the solution process with Viterbi algorithm, it is generally believed that the target is to find the state sequence which can maximise the conditional probability of $P(Q|O,\mu)$. This is expressed as equation (1):

$$\hat{Q}=\arg\max_{Q} P(Q|O,\mu) \tag{1}$$

Viterbi algorithm uses dynamic programming search to find the optimal sequence. Viterbi variable $\delta_t(i)$ is defined as the maximum probability in which at time *t*, HMM goes along a path to reach a state S_i and output an observing sequence $O_1O_2\cdots O_t$. It is expressed as equation (2):

$$\delta_t(i)=\max_{q_1,q_2,\cdots,q_{t-1}} P(q_1,q_2,\cdots,q_t=s_i,O_1O_2\cdots O_t|\mu) \tag{2}$$

Thus, for a given observation sequence $O=O_1O_2\cdots O_T$ and the model $\mu=(A,B,\pi)$, the improved Viterbi recognition algorithm is to find the optimal state sequence in the case of known $P(q_m\cdots q_n|O_j\cdots O_k)$. For the terms in our domain-acoustics library, their transition probabilities are expressed as equation (3):

$$w_i=P\left(q_m\cdots q_n\middle|O_j\cdots O_k,\mu\right)+\lambda\bullet P\left(O_j\cdots O_k\middle|\mu'\right) \tag{3}$$

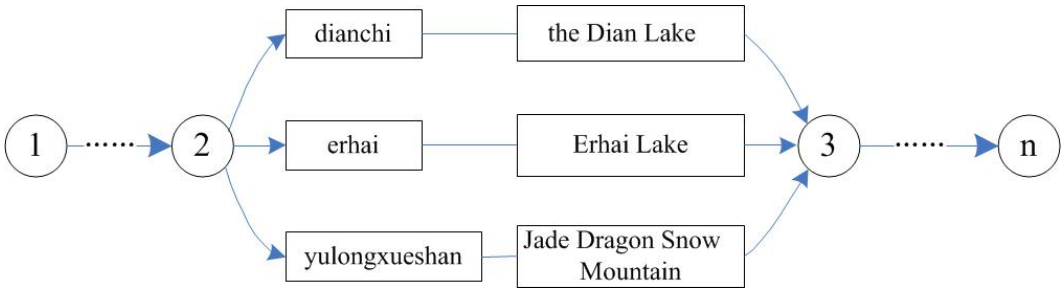
Wherein, μ' is the term model of acoustic library, and $O_j\cdots O_k$ is the observation sequence corresponding to the characteristic sequence of term w_i. Thus, the recognition process can be considered as calculating equation (4):

$$\delta_t(i)=\max_{q_1,q_2,\cdots,q_{t-1}} P(q_1,q_2,\cdots,q_t,O_1O_2\cdots O_t|\mu,\lambda\mu') \tag{4}$$

Wherein, λ is a impact factor. When making speech recognition on a question sentence, the term characteristic sequence in question will impact acoustic layer searching with a large transition probability. Compared with the decoding process from the traditional acoustic model, this method will improve recognition speed and recognition accuracy on domain terms in question sentences. The decoding network is shown as Figure 5.

The speech recognition system is actually a kind of pattern recognition system, including feature extraction, pattern matching, corpus training, the training module text proposes to use domain knowledge to reduce the decoding branch on the search in the language layer, so as to improve the recognition accuracy.

Figure 5 The decoding process of domain-specific question speech recognition by using domain-acoustics library (see online version for colours)



2.4 Pruning strategy guided by domain knowledge

Domain knowledge refers to a collection of concepts, relationships between concepts, and constraints on concepts within a domain. Domain knowledge plays an important supervised role in the decoding optimisation of domain-specific speech recognition.

In a specific domain, terms are often stable. In this paper, The KDML description language of HowNet is chosen to describe the domain concept accurately. When tagging a domain concept, we should determine its first sememe. Secondly, we should find out the other sememe contained by the concept of domain. And finally we determine the relationship between them. For Yunnan tourism domain example, sememes can only be one of scenic, snacks, local specialties, hotels, festivals, customs. The term of 泸沽湖(Lugu Lake) is marked as follows process:

First of all, we give the natural language description of 'Lugu Lake': 'Lugu Lake' is a scenic spot firstly, is a lake, is a legendary-related and plateau-related scenic spot, and located in Ninglang County of Lijiang. The keywords in the natural language description are extracted as the initial label of 'Lugu Lake', such as scenic spot, lake, legend, plateau, Ninglang County. And then the corresponding exactly domain sememe related with scenic spot, lake, legend, plateau, Ninglang County, one by one is determined. Eventually the DEF label of term 'Lugu Lake' is: DEF={scene|景区：{lake|湖泊：RelateTo={legend|传说}{ plateau|高原}， belong={NingLangXian|宁▪县}}}. Here, scene|景区，lake|湖泊，legend|传说，plateau|高原，Ning Lang Xian|宁▪县are domain sememes of Yunnan tourism.

In domain-specific continuous speech recognition, domain entities in a sentence often have particular contacts with each other. Such contacts can be considered as the syntactic and semantic relationships between entities under specified fields, which we called domain knowledge relationship. In some specific areas, domain knowledge is relative centralised and stable. For example, about scenic consultation in tourism field, the questions usually are: "What are the characteristics of Yulong Snow Mountain in Lijiang?", "What are the characteristics of Jade Water Village Scenic in Lijiang?", "Where is the Primitive Forest Park in Xishuangbanna,Yunnan?", "How much is the ticket of Atami SPA in Tengchong?" and so on. If you use a grammar model to describe the syntax rules of certain questions, the word graph (Meng et al., 2011) of the question "Does it really fun about Dali Erhai Lake?" is shown in Figure 6.

In the word graph, each arc represents a word, different nodes are connected by arcs, and thereby they form a network from the beginning node to the end node. A complete path constitutes a complete sentence. All possible paths in a word graph are all possible sentences in the grammar rules. The question speech recognition process is the very process searching the best path from the word graph.

In larger scale search space, we can use the semantic relationships between domain entities to make search pruning. In Figure 6, the decoder will search 27 paths. For example, the decoding path "how to go to the Dian Lake in Lijiang?", the question is grammatically right, but does not meet tourism domain knowledge, the reason is that the Dian Lake is in Kunming not in Lijiang. When the grammar model generates its decoding network, because the fact in tourism domain knowledge is that the Dian Lake locates in Kunming, Erhai Lake locates in Dali, and Jade Dragon Snow Mountain locates in Lijiang, Erhai Lake, the Dian Lake will first be cut, thereby the search space is reduced. The pruned word graph is shown in Figure 7.

Figure 6 The word graph model sample of tourism field questions (see online version for colours)

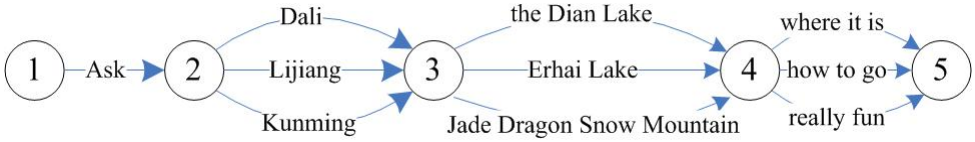
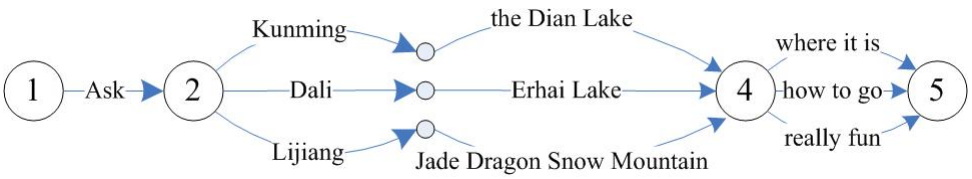


Figure 7 The pruned word graph model sample of tourism field questions (see online version for colours)



3 Experimental results and analysis

3.1 Experiment preparation

In experiments, we construct a HTK-based speech recognition system which is integrated with domain characteristic (Steve and Gunnar, 2009). In order to effectively verify the proposed method, we design three experiments of domain term recognition and question recognition which is integrated domain characteristics. The experiments of domain terminology recognition use a total of 5,000 term vocabulary, and are divided into the experiment of weighting factor selection and the experiment of what effect the vocabulary size has on recognition accuracy in domain term recognition. Chinese question speech recognition integrated domain features uses a total of 1,200 tourism domain questions. The speech corpus is recorded by us, and the parameters are set as 16 b, 16 kHz and moderate speak speed. The training set consists of the speech corpus, which is recorded by 100 men and 50 women about duration of 60 hours. The test set consists of the speech corpus, which is recorded by 25 male and 20 female about duration of five hours. A 39-dimensional mel-frequency cepstrum coefficient (MFCC) is used as acoustic feature parameter, including energy components, and their first-order and second-order differences. Speech features are extracted by HCopy in HTK tool. HMM model (hidden Markov model) is used to build the acoustic model. Both training and testing of the model are based on HTK3.3 toolkit.

To construct a domain acoustic library, we collected and collated 5,000 tourism terminologies from the internet. Let 100 men and 50 female recorded the terms into speech corpus. Taking phoneme as marking unit, using the HTK tool of Cambridge University to label the speech corpus and to extract speech feature, we constructed phonemic acoustic models for the terms. As a result, all the phonemic acoustic models were integrated into a domain acoustics library. In tagging speech corpus, we take initials/finals as annotation primitive. And HTK tool developed by Cambridge University is used to label the speech corpus. The speech annotation for tourism domain terms is as follows:

石林 [石林] shilin

九乡 [九乡] jiuxiang

滇池 [滇池] dianchi

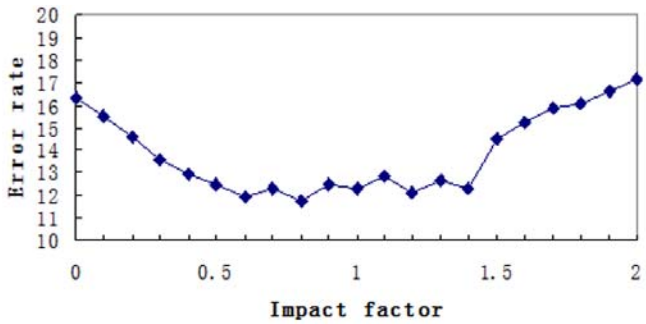
To build a domain knowledge base, we collected the related questions in Yunnan tourism industry by manual, and acquired a domain knowledge base of 1,200 sentences by K-means clustering. Their speech corpus was recorded by 100 male and 50 female, and the speech was annotated with HTK tool of Cambridge University. In tagging domain knowledge, we take KDML description language of HowNet.

3.2 Experiments on speech recognition of domain terminologies

3.2.1 Weighting factor selection in domain terminology recognition

Acoustic model and domain-acoustics library have a certain impact on domain terminology recognition accuracy. In this section, this experiment aims at select the appropriate impact factor to make the acoustic model and the domain-acoustics library better integrated, so as to achieve a better recognition results. In order to better perform the changes of recognition accuracy, this experiment is conducted on the dataset of selected 2000 vocabularies. Among them, about a half is domain terminologies and the other is not. The result is shown in Figure 8.

Figure 8 The relationship between impact factor and recognition error rate (see online version for colours)



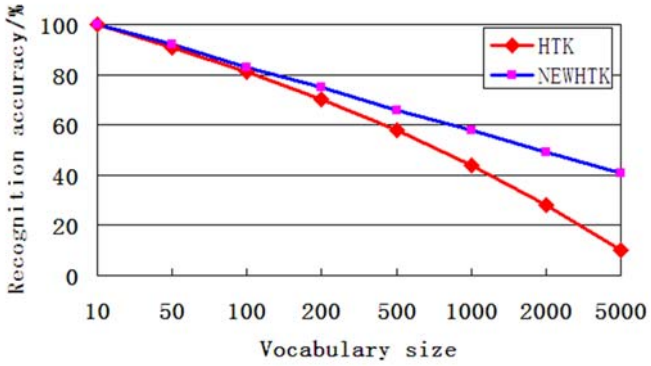
It can be seen from Figure 8, when $\lambda = 0$, the acoustic model does not contain the domain-acoustics library, and the recognition error rate is the same as the general system. With λ increases from 0 to 1.4, the recognition error rate decreases significantly, and the lowest is to 11.5%. When compared with the general system, the recognition rate drops by 4.6%. With λ continually increasing, however, the recognition error rate no longer decreases, but increases.

In essence, the domain-acoustics library is to increase the shift probability between current phoneme and next phoneme, thereby to increase the ability of distinguishing glossary terms. On the datasets which contain a certain percentage of terms, when λ given and unchanged, this effect can better distinguish the phoneme association between before phoneme and after phoneme. However, excessive increase of the weight λ, it will interfere with the probability computation of normal words, and result in the error rate raising.

3.2.2 The impact of vocabulary size on recognition accuracy

In recognition of domain terminologies, domain-acoustics library's effect on raising recognition accuracy need to be verified on a large of data sets. This paper uses several data sets which contains 10, 50, 100, 200, 500 and other number terms. In each dataset, its domain terminologies have about a half of its total vocabulary. The result is shown in Figure 9.

Figure 9 The vocabulary size's impact on recognition accuracy (see online version for colours)



In Figure 9, NewHTK is the recognition accuracy integrated with the domain-acoustics library. As can be seen from Figure 9, the acoustic model integrated with the domain-acoustics library has a better effect on recognition of domain terminologies. When the words number is up to 5,000, the recognition accuracy is still higher than that of the general model by about 30%. Meanwhile, with the increase of words, the recognition accuracy declines. When the number of words exceeds 1,000, the recognition accuracy will decrease rapidly.

After analysis, the main reasons for this trend are two types: (1) The same or similar pronunciation among domain terms. For example: lijiang (漓江) and lijiang (丽江), lujiang (庐江) and lujiang (怒江), gudong (古洞) and gugong (故宫) and so on. (2) The same or similar pronunciation between domain terms and common words. For example: dazhaosi(大昭寺) and dachaoshi(大超市), xitou(溪头) and xitou(洗头), mohe(漠河) and mohe(墨盒) and so on.

3.2.3 *Experiment on Chinese question speech recognition integrated with domain characteristic*

In order to analyse the effect of domain-acoustics library and domain-knowledge library on Chinese continuous speech recognition, we use three different methods to test the real-time rate and the accuracy of continuous speech recognition of 200 Chinese questions from tourism domain. The results are shown in Table 2.

Table 2 The test results of Chinese question recognition in tourism domain

	Results		
	Universal system	*ANN model*	*Integrated with domain characteristic*
Average real-time rate	1.45	1.42	1.39
recognition accuracy	75.50%	78.90%	80.50%

As can be seen from Table 2, compared with the general HMM model and the ANN model, our continuous speech recognition method, owing to using domain-acoustics library and domain-knowledge library to decoding and pruning, has a certain increase in recognition speed and recognition accuracy. The improvement in recognition accuracy benefits from the domain-acoustics library. Meanwhile, the recognition speed has only less improving. This also shows that using domain-knowledge to make pruning, it will have a certain actual effect on reducing searching space, but it is not obvious. On the point, it needs further study and improvement.

4 Conclusions

This paper proposes an integrated domain-characteristic Chinese question speech recognition method. This method uses domain term feature sequence to build the domain acoustic model, and uses domain knowledge to optimise decoding process of Chinese question speech recognition. The preliminary test results in the tourism questions speech recognition show that this method has a good effect on the specified domain speech recognition. This indicates that the pronouncing correlations within domain terms and domain knowledge have a positive contribution to domain question speech recognition. However, the stability of the acoustic modelling method based on domain term characteristic sequence still need to be further verified on other domains. In addition, how to semantic knowledge optimise continuous Chinese speech decoding more effectively, also need a more in-depth research.

Acknowledgements

This paper is supported by National Nature Science Foundation (Grant Nos. 61472168, 61175068), Key Project of Yunnan Provincial Nature Science Foundation (Grant No. 2013FA030), and Yunnan Provincial Talent Fund for Science and Technology Innovation (Grant No. 2014HE001).

References

Ban, S.M. and Kim, H.S. (2015) 'Weight-space viterbi decoding based spectral subtraction for reverberant speech recognition', *IEEE Signal Processing Letters*, Vol. 22, No. 9, pp1424–1428.

Barve, A. and Joshi, B.K. (2016) 'Fast parallel lexical analysis on multi-core machines', *International Journal of High Performance Computing and Networking*, Vol. 9, No. 3, pp.250–257.

Dahl, G.E., Yu, D. and Deng, L. (2012) 'Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition', *IEEE Transactions on Audio, Speech and Language Processing*, January, Vol. 20, No. 1, pp.30–42.

Dahl, G.E., Yu, D., Deng, L. et al. (2011) 'Large vocabulary continuous speech recognition with context-dependent DBN-HMMs', in *Proceedings of 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague Congress Ctr, Prague, Czech Republic, 22–27 May, pp.4688–4691.

Huang, C., Shi, Y., Zhou, J.L. et al. (2004) 'Segmental tonal modeling for phone set design in Mandarin LVCSR', in *Proceedings of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '04*, Montreal, Canada, 17–21 May 2004, pp.901–904.

Hwang, M.Y., Peng, G. and Ostendorf, M. (2009) 'Building a highly accurate mandarin speech recognizer with language-independent technologies and language-dependent modules', *IEEE Transaction on Audio, Speech and Language Processing*, Vol. 17, No. 7, pp.1153–1262, September.

Li, X.G., Yang, Y.N., Pang, Z.H. and Wu, X.H. (2015) ' A comparative study on selecting acoustic modeling units in deep neural networks based large vocabulary Chinese speech recognition', *Neurocomputing*, Vol. 170, pp.251–256, December.

Maas, A.L., Qi, P., Xie, Z., Hannun, A.Y., Lengerich, C.T., Jurafsky, D. and Ng, A.Y. (2017) 'Building DNN acoustic models for large vocabulary speech recognition', *Computer Speech and Language*, Vol. 41, pp.195–213.

Meng, J.T., Wu, J. and Wang, Z.Y. (2011) 'A speech recognition strategy for specific domain based on wordgraph decoder', *Voice Technology*, Vol. 51, No. 9, pp.1288–1292.

Mohamed, A.R., Yu, D. and Deng, L. (2010) 'Investigation of full-sequence training of deep belief networks for speech recognition', in *Proceedings of 11th Annual Conference of the International-Speech-Communication-Association*, Makuhari, Japan, 26–30 September, pp.2846–2849.

Qian, Y.M., Tan, T. and Yu, D. (2016a) 'Neural network based multi-factor aware joint training for robust speech recognition', *IEEE-ACM Transaction on Audio, Speech and Language Processing*, Vol. 24, No. 12, pp.2231–2240, December.

Qian, Y.M., Bi, M.X., Tan, T. and Yu, K. (2016b) 'Very deep convolutional neural networks for noise robust speech recognition', *IEEE-ACM Transaction on Audio, Speech and Language Processing*, Vol. 24, No. 12, pp.2263–2276, December.

Rajavel, R. and Sathidevi, P.S. (2015) 'Optimum integration weight for decision fusion audio-visual speech recognition', *International Journal of Computational Science and Engineering*, Vol. 10, No. 1, pp.145–154.

Ronan, C. and Jason, W. (2008) 'A unified architecture for natural language processing: deep neural networks with multitask learning', in *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, New York, USA, pp.160–167.

Saito, R., Kuboyama, T. and Yasuda, H. (2015) 'User behaviour modelling by abstracting low-level window transition logs', *International Journal of Computational Science and Engineering*, Vol. 11, No. 3, pp.249–258.

Steve, Y. and Gunnar, E. (2009) *HTK Book*, Cambridge University Engineering Department, Cambridge, UK.

Sun, X. and Zhao, Y.X. (2014) 'Integrated exemplar-based template matching and statistical modeling for continuous speech recognition', *EURASIP Journal on Audio Speech and Music Processing*, Vol. 2014, No. 1 pp.28–37.

Yang, G., Qiao, J.F. and Bo, Y.C. (2011) 'Research on artificial neural networks with spatial architecture based on span connection and lateral inhibition mechanism', *International Journal of Computational Science and Engineering*, Vol. 6, Nos. 1–2, pp.86–95.

Yang, Z.L., Liu, W.J. and Chao, H. (2012) 'Integrating induced probability into decoding for large vocabulary continuous speech recognition', *Acta Acustica*, Vol. 37, No. 2, pp209–217, March.