Key Information Expansion Applied in Spoken Document Classification based on Lattice

Lei Zhang
Information and Communication Engineering College, Harbin Engineering University, Harbin, China
zhanglei@hrbeu.edu.cn

Zhuo Zhang Xue-zhi Xiang
Information and Communication Engineering College, Harbin Engineering University, Harbin, China
Email: {zhangzhuo, xiangxuezhi}@hrbeu.edu.cn

***Abstract*—Traditionally, query words or key words in spoken document classification are generated by manual. In this paper, based on CHI-square, *TFIDF* and maximum poster probability (MPP) features, a new hybrid feature for key information extraction is proposed. It can combine the advantages of these three features, and the weight of each word in hybrid feature can be further integrated into the classification system. Here, the weights of key words can reveal the relationship between words and topic to some extent. Furthermore, when the query words or key words are not enough, key information expansion part based on focus score can be added to dig the latent information about the topic. In the key information expansion part, not only the documents with key words occurring but also the other documents with no key word participate into the expansion procedure. Additionally, in the classification system, document length as prior information is adopted when no query is found. The whole classification system is based on lattice, which has more information than 1-best result in speech recognition system. Among CHI-square, *TFIDF* and MPP, the system performance of MPP is a little worse than the others. CHI-square is a little better than *TFIDF* when the key words number is increasing. Among these feature, hybrid feature can almost obtain the best performance under the same condition. Combined with document length information, the classification system performance is further enhanced, especially for less key information condition. Experiments show that when the system is combined weight and document length information, hybrid feature can obtain the best performance with a MAP of 0.7817 under 50 key words. When key information is not enough, key information expansion can improve the system performance when only 1, 5, 10 key words here. In the proposed key information expansion approach, since the focus factor is introduced to adjust the effect of documents with no key words, some empty words can be avoided to some extent, and the number of expansion words can be under control.**

***Index Terms*—hybrid feature, key information extraction, document length, spoken document classification, lattice**

I. INTRODUCTION

Nowadays, text-based searching engine has been applied widely, and many technologies such as automatic summarization [1], semantic extraction [2] based on text documents are studied in detail. However, how to integrate these approaches into spoken document processing is still a challenge. Most studies of Spoken Document Retrieval (SDR) use the speech recognizer to generate approximate transcripts and just to apply the text-based information retrieval techniques directly [3]. But for broadcasts and conversation data, the low recognition rate can worsen the performance of classification system. Lattice can reduce the impact of the error rate to some extent by providing multiple hypothesis [4-9]. We have shown the improvement of lattice for spoken document classification system in [10]. Further, in that work, spoken document length is firstly applied in spoken document classification.

However，in most classification systems, the queries of each topic are assigned by manual, and these queries play the same roles during classification. In fact, for each topic, some queries may be more important than others. Here, we combine the key information extraction into the classification system. From the extraction stage, different query words may have different weight, which can be considered during classification. Especially when the queries are not enough, a special approach is proposed to expend the queries based on focus scores.

There are many approaches to extract the key information from text documents, such as document frequency (DF), χ^2 statistics (CHI-square), term strength (TS), mutual information (MI), and information gain (IG). In [11], the performance of these features is compared in text categorization. [12] gives the comparison among IG, CHI and the maximum posterior probability (MPP) measure. Analysis in [11,12] reveals that DF, IG and CHI scores of a term are strongly correlated, and MPP can get better performance in topic identification. So here, CHI and MPP are selected as the baseline approaches to extract key information. Additionally, *TFIDF* as the most common feature in text-based retrieval is also adopted here.

In fact, different feature can reflect different aspect of the key information. In order to combining these aspects together, a hybrid feature is investigated. Different from

doi:10.4304/jcp.6.5.923-930

the queries obtained by manual, the queries extracted automatically have distinct weights. That means even for the same topic, different query word can play different role in the classification.

The whole paper is arranged as follow: section II will introduce the whole system framework and some pre-processing procedure, and section III gives some details about the spoken document classification system with document length as prior information. In section IV, some key information extraction approaches are introduced, based on these features, a new hybrid feature is proposed. Furthermore, a new key information expansion approach based on factor score is proposed in section V. Under some conditions, query words or query information are not ensured to be enough, this approach can be applied to mining the latent information related to the topic. Lastly, the experiments and results are conducted in section VI to evaluate the system performance, and conclusion is given in section VII.

II. SYSTEM FRAMEWORK

The whole system has three parts, as shown in Fig. 1. The first one named off-line part converts the speech signal into lattice. The second part is for classification based on lattice. The last part is for key information extraction and expansion. In the whole framework, the information extraction and expansion are combined into the spoken document classification system.

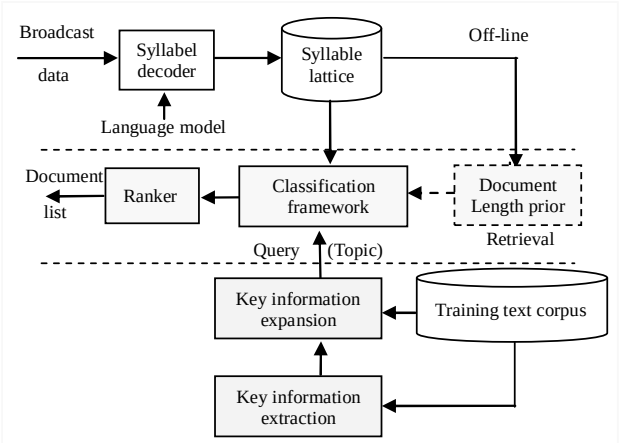

Figure. 1 System architecture

In the first part, HMM model is built by syllable. Since there exist more than 80,000 commonly used words and more than 10,000 commonly used characters in Mandarin Chinese, it is hard to construct the recognition model based on words or characters. Furthermore, all characters are monosyllabic, and for Chinese, there are many homophones, then the total number of phonologically allowed syllables with tone is only 1345 [13]. So in our system, the recognition model is built on syllable. Combined with language model smoothed by modified Katz, the system can output the syllable lattice instead of syllable sequence known as 1-best result. Fig. 2 gives the example of lattice of 'ren2min2fa3yuan4'.

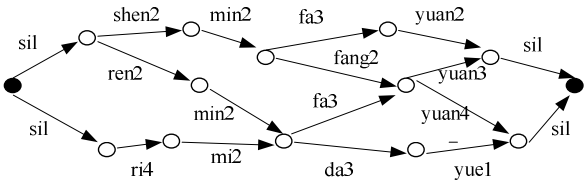

Figure 2. Structure of lattice 'ren2min2fa3yuan4'

As shown in Fig.2, lattice is composed by arcs and nodes. From the start node, there are many paths to the end node. In fig.2, only the corresponding labels are listed for arc, which represents that form start node to the end node of this arc, the decoded result maybe 'shen2' or others labels, further the probability of this case can be denoted by acoustic probability, language probability. These two probabilities and the information of start and end node of this arc are also attached to this arc. For each node *n*, there is time information.

From start node to end node of lattice, there are many paths, and each one corresponds with one recognition result. 1-best result is only one path in the lattice with the largest poster probability of whole sentence. However in many applications as spoken document classification, the most important unit is not sentence, but words. So during speech recognition phase, we want to keep more information than 1-best result to be handled in next phase.

The second part is the classification system based on lattice. According to different queries, the relevance between lattice and queries are computed. Since the multi-path in lattice, expectation calculation during the classification is need. Furthermore, the document length is combined in the classification as the prior probability when no query is in the document.

The last part in fig. 1 is key information extraction and key information expansion. Here, the queries of different topic are generated automatically and a hybrid feature is proposed to combine the advantage of CHI-square, *TFIDF* and MPP. Then the weight of query in hybrid feature can be combined in classification system. As for key information expansion part, it can be adopted under some conditions. When the key information is not enough to represent to topic information, it can be used to obtain the latent information to topic. The more details of this part are introduced in section IV and section V.

III. SPOKEN DOCUMENT CLASSIFICATION WITH DOCUMENT LENGTH

Given a query $\mathbf{q}^t = (s_1^t s_2^t ... s_m^t)$ belonging to a topic *t*, where s_i^t is the *i*-th syllable in query. Then the relation between spoken document *D* and query $\mathbf{q}^t$ can be defined by the probability $P(D|\mathbf{q}^t)$ in (1).

$$P(D|\mathbf{q}^t) = P(\mathbf{q}^t|D)P(D)/P(\mathbf{q}^t) = \frac{P(s^t{}_1 s^t{}_2 ... s^t{}_m|D)P(D)}{P(\mathbf{q}^t)} \quad (1)$$

Supposing the syllables in the query are independent each other, then (1) can be turned into (2).

$$P(D \mid \mathbf{q}^t) \approx \prod_{w^t} \left(\frac{P(w^t \mid D)}{P(w^t)}\right)^{c(w^t, \mathbf{q}^t)} P(D) \quad (2)$$

Where $c(w^t, \mathbf{q}^t)$ is the count of syllable *w* in query $\mathbf{q}$ belonging to topic *t*. Further, $P(w^t)$ is always considered as uniform distribution if there is no limited applying region, and can be dropped for document ranking purpose.

$P(w^t \mid D)$ is the probability of syllable *w* occurring in document *D* under topic *t*, which can be estimated by maximum likelihood algorithm. As for the priors probability of documents, $P(D)$ is equal for every document since there is no prior information for it. So in traditional classification system, $P(D)$ is also dropped in (1) and (2). That means whether the query word occurring in document or not, $P(D)$ has no effect for classification.

In text retrieval method, there is a view that document prior probabilities depend on document length. Many researches tried to establish a connection between the likelihood of relevance and document length. The results in [14] confirm that the prior probability is proportional to document length. That means, the longer documents span more topics and are more likely to be relevant with the query, although no query has been seen in the document. So in our classification system, the document length information is combined in the whole classification framework. Additionally, considering the size and computing speed of lattice, each speech document is divided into *M* segments, and (2) can be turned into:

$$P(D \mid \boldsymbol{q}^t) \approx \begin{cases} \prod_{w} \sum_{k=1}^{M} P(w^t \mid D_k)^{c(w^t, \boldsymbol{q}^t)} & \text{any query} \\ \prod_{w} \sum_{k=1}^{M} P(w^t \mid D_k)^{c(w^t, \boldsymbol{q}^t)} \times \dfrac{\sum_{k=1}^{M} E[\mid D_k \mid]}{L} & \text{no query} \end{cases} \quad (3)$$

Where D_k is represented as lattice of *k*-th segment in spoken document *D*. *L* is the whole length of all documents. Since there are many candidates in lattice, the expectation must be considered here.

In (3), the document length is combined in classification. When there is no query occurring in lattice, the longer document is more likely to relate to the topic. But if there are some queries happening in document, the effect of document length can be ignored.

In order to handle the zero probability, $P(w^t \mid D_k)$ is computed by Jelinek-Mercer (JM) method [15], in which an interpretation of the maximum likelihood model with the priori probability of syllable *w* is adopted.

$$P(w^t \mid D_k) = \frac{\lambda E[c(w^t \mid D_k)]}{E[\mid D_k \mid]} + (1-\lambda) P(w^t \mid \overline{D}) \quad (4)$$

$\overline{D}$ is the text document set with similar topic as spoken document set D. If $\overline{D}$ is huge enough, $P(w^t \mid \overline{D})$ can reflect the priori probability of syllable *w* in topic *t*. When the syllable *w* is seldom occurring in lattice, then this part of priori probability will play more role in (4). Finding a proper parameter λ, a good balance between the real condition and the priori information can be found.

The first term in (4) can reflect the real effect of the syllable *w* in spoken document D_k. Since for lattice, there are many paths and in each path, the syllable will happen at certain probability, the expectation of the count must be considered, which can be expressed as $E[c(w^t \mid D_k)]$. It is calculated as:

$$E[c(w^t \mid D_k)] = \sum_{\boldsymbol{p}} c(w^t \mid \boldsymbol{p}) P(\boldsymbol{p} \mid \boldsymbol{O}_k) \quad (5)$$

Here, $\boldsymbol{p}$ is one of the hypothesized path in lattice D_k, $\boldsymbol{O}_k$ is the corresponding utterance, and $c(w^t \mid \boldsymbol{p})$ is the number of syllable *w* occurring in $\boldsymbol{p}$. Similarly, the expected document length $E[\mid D_k \mid]$ is computed as (6).

$$E[\mid D_k \mid] = \sum_{\boldsymbol{p}} \mid \boldsymbol{p} \mid P(\boldsymbol{p} \mid \boldsymbol{O}_k) \quad (6)$$

Where $\mid \boldsymbol{p} \mid$ denotes the number of syllable in path $\boldsymbol{p}$.

In (5) and (6), $P(\boldsymbol{p} \mid \boldsymbol{O}_k)$ contains the effect of both acoustic model and language model during speech recognition system.

$$P(\boldsymbol{p} \mid \boldsymbol{O}_k) \approx \frac{P(\boldsymbol{p})^{\xi} P(\boldsymbol{O}_k \mid \boldsymbol{p})}{\sum_{k} P(\boldsymbol{p})^{\xi} P(\boldsymbol{O}_k \mid \boldsymbol{p})} \quad (7)$$

In (7), the normalization factor ξ is used to adjust the balance of effect between acoustic model and language model.

IV. KEY INFORMATION EXTRACTION

A. Other Features

Here, three features as CHI-square, MPP and *TFIDF* are selected to extract the key information related to topic.

Supposing that the relationship between term *m* and topic *t* is independent, CHI-square test can be acted as (8).

$$weight_{CHI} = \frac{n(k_{11}k_{00} - k_{10}k_{01})^2}{(k_{11}+k_{10})(k_{01}+k_{00})(k_{11}+k_{01})(k_{10}+k_{00})} \quad (8)$$

Let e_t =0 denote the document is not in topic *t*, and e_t =1 is the other condition, that is the document

occurring in topic *t*. Similarly, e_m =0 or 1 means the document does or not contain term *m*. Then according to the value of e_t and e_m , $k_{e_t e_m}$ denotes the number of document under all conditions. n is the sum of k_{11} , $k_{10}, k_{01},$and k_{00}.

For MPP, *t* also means topic, and the larger $P(t \mid m)$ is, the more important the term *m* is. But if *m* does not occur in topic *t*, the probability will be zero. So here, $P_{map}(t \mid m)$ is adopted as follows, which can handle the zero probability.

$$weight_{mpp} = P_{map}(t \mid m) = \frac{N_{m|t} + \alpha_1 N_t P_{map}(t)}{N_m + \alpha_1 N_t} \quad (9)$$

where $N_{m|t}$ is the occurring number of term *m* in topic *t*, and N_t is the number of distinct topics. N_m is the number of term *m* in training corpus. Furthermore, $P_{map}(t)$ is as

$$P_{map}(t) = \frac{N_{d|t} + \alpha_2}{N_d + \alpha_2 N_t} \quad (10)$$

Similarly, the $N_{d|t}$ is the number of document in topic *t*, and N_d is the whole number of document.

TFIDF is a kind of widely used weight in text-document retrieval. The value $TFIDF(m,t)$ is calculated as

$$weight_{tf} = TFIDF(m,t) = N_{m|t} \times IDF(m) \quad (11)$$

The inverse document frequency $IDF(m)$ can be calculated as:

$$IDF(m) = \log(N / W(m)) \quad (12)$$

where *N* is the total number of documents, $W(m)$ is the whole number of document containing term *m*. (12) is a little different with traditional *TFIDF*. Here, $W(m)$ can also present the character of inverse information. If this value is large, that means this term also happens not only in topic *t*, but also other classes. Thus, the effect of this term should be weakened.

B. Hybrid Feature

The features above are from different aspect to reveal the relationship between term *m* and topic *t*. *TFIDF* can reflect the term's distinction ability to other kinds of documents. CHI-square feature can measure the lack of independence between the term and the topic, and MPP measure can combine the feature selection into topic classification. Fig. 3 gives the top ten key words about law corpus.

CHI-square	*TFIDF*	MPP
lao2dong4	lao2dong4	lao2dong4
fa3lv4	cao3an4	fa3lv4
cao3an4	gui1ding4	cao3an4
gui1ding4	fan4zui4	gui1ding4
zhong1guo2	fa3lv4	fan4zui4
fan4zui4	chu3fa2	bu4men2
chu3fa2	**wu1ran3**	chu3fa2
bu4men2	**bei3jing1shi4**	zhi2fa3
zhi2fa3	zhi2fa3	**bei3jing1shi4**
wu1ran3	bu4men2	**an4jian4**

Figure 3 Key information comparison of three approaches

All terms in Fig. 3 are sorted by descending order. It can be seen that not only the order of some terms is different, but also some terms are distinct from each other, as bold ones show. In hybrid feature, we hope to combine all conditions together to select the most possible terms to represent the topic. Only that kind of terms which are important in all methods, they will play an important role in hybrid feature. The simplest approach is to add each column in fig. 3 together, and then we can choose the first top ones.

Fig. 4 gives the histogram of these three features. Here, *x*-axis is the region of amplified weights with 10^4, and *y*-axis is the number of the weight in corresponding region. From this figure, it can be seen that these three features are distributed in different regions, which means the weights of CHI-square, *TFIDF* and MPP are different. Since the ranges of values in different methods are distinct, here, a linear function is selected to map the different weight into the same region. It can also avoid the large weight weakening the effect of small weight in hybrid feature. The weights of all three approaches are mapped into the region from 1 to 10 by a linear function as (13). * represents one kind of feature from CHI-square, *TFIDF* and MPP approaches.

$$weight_*^{'} = a_* \times weight_* + b_* \quad (13)$$

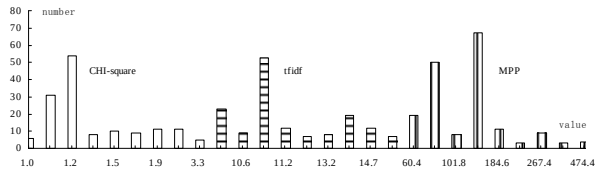

Figure 4 The histogram of log probabilities

In (13), a_* is the slope and b_* is the shift. For CHI-square, *TFIDF* and MPP, these two parameters are different. Selecting proper a_* and b_*, all the weights are in the same region for different approaches. Then the means of weights of these three features can be obtained as the weights in hybrid feature. After these processing, the top *N* weights in hybrid feature must be the top in all these three features.

$$weight_w = \frac{1}{3}(weight'_{tfidf} + weight'_{CHI} + weight'_{MPP}) \quad (14)$$

For this weight, it can reflect the relation with topic. The larger of this weight is, the closer relationship with topic exits. So the weight in hybrid feature can be further combined in the classification framework. (3) can be turned into (15).

$$P(D \mid \boldsymbol{q}^t) \approx \begin{cases} \sum_w weight_w \times \log(\bar{P}) & \text{any query} \\ \sum_w weight_w \times \log(\bar{P}) \times \frac{\sum_{k=1}^{M} E[\mid D_k \mid]}{L} & \text{no query} \end{cases} \quad (15)$$

Where $\sum_{k=1}^{M} P(w^t \mid D_k)^{c(w^t, q^t)}$ is represented as $\bar{P}$. The weight should be normalized as the same magnitude as $\log(\bar{P})$.

V. KEY INFORMATION EXPANSION

Under some conditions, if the key information is not enough, it can be expanded. There is a view that if the words occurring in the same document with key information, there are some kinds of relation with the key information. Here, this view is called related key information extraction. In [16], this view is applied into the intelligent expansion of query of searching related scientific articles. Here, it shows the shortcoming of this method. Further more, a new idea in key information expansion approach based on focus factor is proposed.

A. Related key information extraction

Supposing there are N documents in topic t, W_i is the word set of D_i document, and $i \in [1, N]$. As proposed in IV, the key information set K has been extracted by hybrid feature. Additionally, each key word in set K is represented as k_j, $j \in [1, M]$.

If word w is not in key information set K, and the number of documents in which the word w and any key word k_j both happen is largest, then the word w is the most nearest key word. It can be expressed as (16).

$$w = \underset{w \in W_i; i \in [1,N]}{\operatorname{argmax}} \mid \{D_i \in A_w \mid w \in W_i, w \notin K; K \cap W_i \neq \phi\} \mid \quad (16)$$

Here, A_w is the document set that satisfies the condition, and $\mid\ \mid$ is the number of the document in set.

Since the word w is selected with maximum number of A_w in (16), the real number of expansion word is hard to control. For example, if the maximum number of the document set A_w is 5, the all the words with this number should be selected at the same time and the exact number can be known before.

TABLE I The expansion result of topic law and topic sport with one key word

Topic	Key word	Expansion word
Law	lao2dong4	er4shi2/gui1ding4
sport	bi3sai4	zhong1guo2/zuo2tian1

Table I gives the results of the two topics of law and sport. Under the condition of one key word each topic, there are two expanding words. Since the empty word, which has little meaning to topic, often happen in each document including the document which key words also happen, it is more possible to be selected as the expanding word. Like the word ‘er4shi2’ in table 1, it has little relation with topic law. In normal condition, this kind of empty word should be filtered by a stop word list before. But for the complex applying system, it is hard to construct an all-sided stop word to cover all conditions. If this kind of empty words is used in classification system, it can add the confusion among different topic, and weaken the performance of classification system.

In the idea above, only the document with key words are considered in information expansion. In fact, the documents with no key words also can play some roles during expansion, since they are in the same topic. So in next sub-section, the non-key words documents are also considered in expansion and the weight is less than that of the documents with key word.

B. Key information expansion based on focus fatcor

According to the key words occurring in the document or not, the documents in topic *t* can be classed into tow set as definition 1.

Definition 1: given the key words set *K*, the one with key words occurring is called ‘related document set’ *A*, and the other one is called ‘the other document set’ *B*.

Definition 2: In set *A*, the number of documents in which the word *w* occurring is represented as NMRD (the Number of the Most Relevant Document). Similarly, the number of documents with word *w* in set *B* is called NLRD (the Number of Low Relevant Document).

For word *w*, here, $D(t, w)$ called focus score reflects the influence in key information expansion for topic *t*. it is shown as (17).

$$D(t, w) = \mathrm{NMRD}(t, w) + \mathrm{FF}(t, w) \times \mathrm{NLRD}(t, w) \quad (17)$$

In order to present the difference effect between document set *A* and document set *B*, focus factor $\mathrm{FF}(t, w)$ is introduced here. In (17), it can be seen that for NMFD, the weight equals to 1. It means that for the document set *A*, the effect in focus score is 100% percent. Comparing with the document set *A*, the effect of the document set *B* should be weaken, which is adjust by the focus factor. So the focus factor $\mathrm{FF}(t, w)$ should be less than 1. Here, it is as (18).

$$\mathrm{FF}(t, w) = [\mathrm{NMRD}(t, w) + \mathrm{NLRD}(t, w)] / N(t) \quad (18)$$

Where $N(t)$ is the whole number of documents in topic *t*, it equals $|A|+|B|$. Since the sum of $\mathrm{NMRD}(t,w)$ and $\mathrm{NLRD}(t,w)$ is less than $N(t)$, it can keep $\mathrm{FF}(t,w)$ less than 1.

Combined (17) and (18), each word *w* will have a focus score. A list will be obtained by sorting these focus scores as descending. Then the top *N* can be selected as the expansion information. Table II gives the top 10 expansion words by focus score in topic law with one key word.

From table II, it can be seen that NMRD of 'er4shi2' and 'gui1ding4' are both 3, the maximum documents number in (15). So in table I, these two words are selected as the expansion words. Considering the effect of NLRD, since these NLRD of two words are different, the final focus scores are also distinct. According to table II, for the focus score are normally different each other, the number of expansion word can be chosen arbitrarily. Here, if the number of expansion words is 3, then 'er4shi2' can be dropped off. Table III gives the corresponding result for topic law and sport.

TABLE II focus score in topic law with one key word

Word *w*	NMRD	NLRD	Focus score
gui1ding4	3	8	4.173333
fa3lv4	2	10	3.6
guo2jia1	1	12	3.08
er4shi2	3	0	3
dan1wei4	2	7	2.84
gong1zuo4	1	11	2.76
wen4ti2	2	6	2.64
ren2yuan2	2	5	2.466667
shou1qu3	2	2	2.106667
Lao2dong4zhe3	2	1	2.04

TABLE III The expansion result based on focus score in topic law and sport with one key word

topic	Key word	Expansion words
law	lao2dong4	fa3lv4/guo2jia1/gui1ding4
sport	bi3sai4	xuan3shou3/guan4jun1/ti3yu4

Compared with table I, it can be drawn that by adding 'the other document set' influence, some empty words can be filtered. Furthermore, the number of expansion words is under control.

VI. Experiments and Results

There are speech corpus and text corpus in the experiments. For speech corpus, it include '863' corpus and broadcast corpus. The '863' corpus is reading style pronunciation and consists of 90,821 utterances from 156 speakers. The broadcast corpus for classification in our experiments is from the programs of radio, which include the conversation programs, news and so on. Furthermore, broadcast corpus is classified into 6 topics as national defense, sport, countryside, law, economy and politics. It includes 5924 utterances from different kinds of broadcast programs. For text corpus, the similar topics with speech are selected.

In the speech recognition system, HTK is employed to train acoustic and language models based on "863" corpus. Another part of broadcast corpus which do not belong to any topic is applied as adapt corpus. Maximum likelihood linear regression and maximum a posterior approach are used in adaptation. Acoustic model of this system is context-dependent tri-phone model, whose topology is left-to-right with jump. Every model with five states is jointed as syllable model according to dictionary. Language model is syllable based bi-gram model, and Katz approach is adopted as the smoothing algorithm.

Here, the evaluation is depended on mean average precision (MAP) as follow:

$$\mathrm{MAP} = 1/N_1 \sum_{i=1}^{N_1} (1/R_i) \sum_{j=1}^{R_i} (j/r_{i,j}) \tag{19}$$

Where N_1 denotes the total number of queries, R_i is the total number of documents relevant to the *i*-th query, and $r_{i,j}$ the position of the *j*-th relevant document in the ranked list output by the retrieval method for the query $\mathbf{q}_i$. The software used for the evaluation is the trec_eval.7.0 provided by TREC, which will output the values of MAP based on the results of the spoken document retrieval system and the standard answers.

There are several experiments to test different part of our system. The first one is the determination of parameter ξ in (7). The second one is to evaluate the performance of different feature, as CHI-square, *TFIDF*, MPP and new proposed hybrid feature. Additionally, the document length and the weight in (13) combined in classification system as (14) are tested in this part.

A. Selection of ξ

As analysis above, proper value of ξ will affect the performance of system. Fig. 5 gives the classification results of hybrid feature with 20 query words. It can be seen that when ξ is 5, the system performance is better. In the following experiments, the normalization factor ξ is fixed at 5.

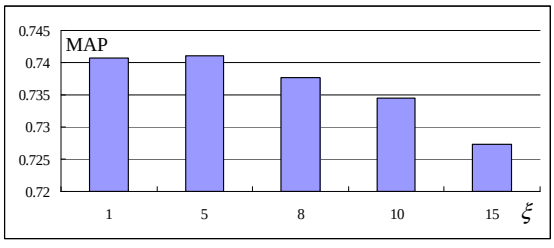

Figure 5 The effect of ξ on retrieval results when 20 query word with hybrid feature

B.Classification results with different feature

Table IV is based on the classification system without the document length and weight effect in (15).

From Table IV, it can be seen that no matter what kind of feature is adopted, with the increasing of the key words number, MAP is increasing too. If there is enough

information in query, keeping increasing the query word has a little influence on MAP. It can be seen in table IV, that compared with the performance of 20 key words, the performance of classification system only increase about 0.02 when key word number is added to 50 for CHI-square, *TFIDF* and MPP. For hybrid feature, it is only about 0.03 enhancements. Furthermore, when the number of key word is only 1, in hybrid feature, there is no more selection in the key word list. In this experiment, the key word is the same as that in CHI-square. So the performance of hybrid feature and CHI-square is the same.

Table IV MAP of different features without document length and weight

	CHI	TFIDF	MPP	hybrid feature
1 key word	0.1982	0.2240	0.2290	0.1982
5 key words	0.4551	0.5125	0.4966	0.5082
10 key words	0.6365	0.6358	0.6280	0.6551
20 key words	0.7317	0.7314	0.6928	0.7410
30 key words	0.7338	0.7544	0.7126	0.7544
50 key words	0.7535	0.7546	0.7166	0.7706

Among CHI-square, *TFIDF* and MPP, the system performance of MPP is a little worse than those of the others. CHI-square is a little better than *TFIDF* when the key words number is increasing. Since hybrid feature can combine the advantage of each one, especially when the number of key words is large, it can get the best result with a MAP of 0.7706 under 50 key words.

Additionally, Fig. 6 gives the difference of MAP between the system with and without the document length information in hybrid feature.

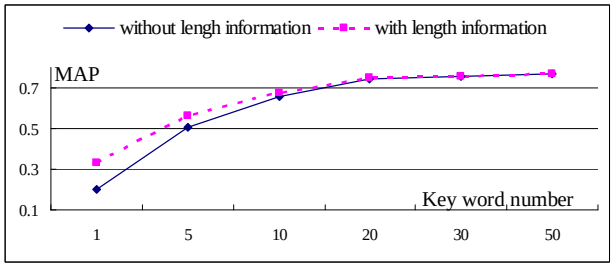

Figure 6 The effect of document length information in hybrid feature

It can be seen that with the increasing of key words number, the effect of document length information is weaken. But for less query words, the influence of document length is notable.

The reason is that in (3), only when the query is not found in the spoken document, the document length information can be merged into the classification system. When the number of key words is smaller, the spoken document is more possible without any key word, and then the document length information can actually take effect.

Lastly, for the weight effect in (15), Table V gives the difference of MAP between the system with and without weight information. When the number of key words is less than 30, the system performances are similar when the weight is added or not. That is, in classification system, the effect of (3) or (15) has no obvious difference. When the number of key words is increasing, the difference is shown. With 50 key words, the system performance can be enhanced to 0.7817 with weight and length information in classification. It is the best result in all experiments.

TableV MAP with weight information for hybrid feature

Hybrid feature	30 key words	50 key words
no weight, with length	0.7579	0.7715
with weight, with length	0.7595	0.7817

C. Classification with key information expansion

Table VI gives results of spoken document classification system when key words number are 1,5, and 10.

TABLE VI MAP with document length of hybrid feature and information expansion

Number of key words	R =0	R =1	R =2	R =3
1	0.3311	0.4213	0.4587	0.4756
5	0.5649	0.5667	0.5761	0.6042
10	0.6767	0.6781	0.6783	0.6870

In this table, hybrid feature is adopted and document length is considered. R means the number of expansion words. When R=0, it is the same as dash line in fig. 6. From table VI, it can be seen that when R is increased, the system performance can be enhanced. When the key word set is small, the system performance can be enhanced obviously. Only with one key word, MAP can be increased from 0.3311 to 0.4765 with 3 expansion words. But for 10 key words, the enhancement is only about 0.0003 in MAP.

The reason maybe that in the proposed approach of key word expansion based on factor score, expansion words will not happen in key words set. Then the less of key words are, the more relation of expansion word with topic. With the increasing of key words set, especially when the key words are enough to reflect the topic information, the expansion words will not play little role on classification.

VII. CONCLUSION

In this paper, we proposed hybrid feature in key information extraction and applied it into spoken document classification system. Further, the weight of hybrid feature can be combined into classification system, thus different query may play different role for the same topic. Combined with the document length information, especially when query words are few, the system performance can be enhanced. The best performance can be achieved as a MAP of 0.7817 for hybrid feature, when the document length information and weight are both combined in classification under 50 key words. In some conditions, there is not enough key information. Key information expansion can be applied further. In the proposed key information expansion approach, both the ‘related document set’ and ‘the other document set’ are considered during expansion. Focus factor is introduced here to weaken the influence of ‘the other document set’.

According to the focus score, some kinds of empty words can be filtered to some extent. Additionally, the number of expansion words can be easy to control. When there is not enough information in query or key words, the expansion words can play a good role to supply the latent information about the topic.

ACKNOWLEDGMENT

It is a project supported by National Natural Science Foundation of China #60702053, also supported by young teacher supporting plan by Heilongjiang province in China #1155G17 and young teacher supporting plan by Harbin Engineering University.

REFERENCES

[1] Y-L Chang and J-T Chien, "Latent Dirichlet learning for document summarization." in *Proc. ICASSP*, Taipei, Taiwan, 19-24 April 2009, pp.1689-1692.
[2] J.T. Malin, and D.R. Throop, "Basic Concepts and Distinctions for an Aerospace Ontology of Functions, Entities and Problems". *Aerospace Conf.* Big Sky, MT, 3-10 March 2007, pp.1-18.
[3] B. Chen, H-M Wang,, and L-S Lee. "Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics." in *Proc. ICASSP*, Istanbul, Turkey , vol 3, 5-9 June 2000 pp.1771-1774.
[4] C-H Meng, H-Y Lee and L-S Lee, "Improved lattice-based spoken document retrieval by directly learning from the evaluation measures," in *Proc. ICASSP*, Taipei, Taiwan 19-24 April 2009 pp.4893 – 4896.
[5] C-L Huang and C-H Wu, "Spoken Document Retrieval Using Multilevel Knowledge and Semantic Verification", *IEEE Trans. Audio, Speech, and Language Process.*, vol 15, Issue 8, pp.2551-2560. Nov. 2007.
[6] T. Mertens and D. Schneider, "Efficient subword lattice retrieval for German spoken term detection." in *Proc. ICASSP*, Taipei, Taiwan, 19-24 April 2009 pp.4885-4888.
[7] R-Q Huang and J.H.L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora." *IEEE Trans. Audio, Speech, and Language Process.*, vol 14, Issue 3, pp.907-919, May 2006.
[8] B. Chen, "Latent topic modelling of word co-occurence information for spoken document retrieval." in *Proc. ICASSP*, Taipei, Taiwan, 19-24 April 2009 pp.3961-3964.
[9] H. Lin, A. Stupakov, and J. Bilmes, "Improving multi-lattice alignment based spoken keyword spotting," in *Proc. ICASSP*, Taipei, Taiwan, 19-24 April 2009 pp.4877-4880.
[10] Lei Zhang, Yunxia Gao, Xuezhi Xang and Dong Lu, "A new syllable-lattice based approach for Mandarin spoken document retrieval," in *Wireless Communications & Signal Processing*, 2009.
[11] Y-M Yang, and J. O. Pedersen. "A comparative study on feature selection in text categorization." in *Proc.ICML-14*, 1997. pp.12-420.
[12] Timothy, J. Hazen, Fred Richardson, and Anna Margolis. "Topic identification from audio recordings using word and phone recognition lattices." In *ASRU* 2007, pp.659-664.
[13] B. Chen, H.M. Wang, and L.S. Lee, "A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents," *ACM Trans. Asian Lang. Inform. Process.*, vol. 3, no. 2, pp.128-145, 2004.
[14] R. Blanco, and A. Barreiro, "Probabilistic document length priors for language models," in *Proc.ECIR-08*, UnitedKingdom, March 2008, pp.394-405.
[15] C-X Zhai, and J Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Information Systems*, vol. 22, no. 2, pp.179-214,2004.
[16] Zhou Qing Zheng Zeqi . "An Intelligent Query Expansion of Searching Related Scientific Articles**",** *Computer Engineering and Applications*, 2004, (12): 48-51.

Lei Zhang Harbin, China, 1973. Received Master degree and Doctor in computer applying field in 2000 and 2004 in Harbin Institute of Technology, Harbin, China. Her main research fields are about spoken document classification and retrieval, robust speech recognition, speaker recognition and other fields about speech signal processing.

Currently, she is professor in Information and Communication College in Harbin Engineering University. She has published more than 20 papers in journal and international meetings, and also published a book about speech signal processing by QingHua University Press.

Dr. Zhang is the member of IEEE.

Zhuo Zhang Changchun, China, 1981. Receive Bachelor degree in signal processing in Harbin Engineering University in 2008. Currently, he is Master candidate in Harbin Engineering University.

Xue-zhi Xiang Harbin, China, 1979. Received Master degree and Doctor in computer applying field in 2004 and 2008 in Harbin Engineering University, Harbin, China. His main research field is about signal and information processing,

Currently, he is the associated professor in Information and Communication College in Harbin Engineering University.

Dr. Xiang is the member of IEEE.