VOICE SEARCH OF STRUCTURED MEDIA DATA

Young-In Song[1]*, Ye-Yi Wang*[2]*, Yun-Cheng Ju*[2]*, Mike Seltzer*[2]*, Ivan Tashev*[2] *and Alex Acero*[2]

[1]Korea University, [2]Microsoft Research

ABSTRACT

This paper addresses the problem of using unstructured queries to search a structured database in voice search applications. By incorporating structural information in music metadata, the end-to-end search error has been reduced by 15% on text queries and up to 11% on spoken queries. Based on that, an HMM sequential rescoring model has reduced the error rate by 28% on text queries and up to 23% on spoken queries compared to the baseline system. Furthermore, a phonetic similarity model has been introduced to compensate speech recognition errors, which has improved the end-to-end search accuracy consistently across different levels of speech recognition accuracy.

Index Terms**—** spoken language understanding; voice search; language model based information retrieval; HMMs; phonetic confusability, music metadata.

1. INTRODUCTION

Voice search [1] is a spoken language understanding (SLU) technology underlying many applications. It accepts users' queries in spoken language and searches for the relevant entries in a database. Directory assistance (DA) [2, 3] is a typical voice search application, where users can use spoken queries to search for business or residential phone listings.

Media data (music, movies, etc.) is pervasive now in people's everyday life. With the ever increasing capacity and reducing cost of storage, it is very common that uses have thousands of music/video entries in their mp3 players or media center PCs. Accessing a music/video title becomes more challenging. Voice search technology can be applied in this scenario to provide a natural and efficient UI for users. In [4], a prototype in-car music search system is presented. In [5], an cognitive load sensitive spoken dialog interface for in-car tasks is reported.

Media (voice) search leverages the metadata associated with the media data. For example, each music entry in an mp3 player comes with the metadata about its title, artists' name, album name, composer/conduct, etc. Different from the DA applications, the SLU needs to search a structured database – the metadata contain records with multiple fields. A user's unstructured utterance may contain descriptions of one or multiple fields, which may not exactly match the entries in the structured database. For example, the query "Boyz II men hard to say goodbye" corresponds to the following structured metadata:

Artists: Boyz II Men
Title: It's so hard to say goodbye to yesterday
Album: Legacy – the greatest hits collection
Genre: R&B/Soul
…

Figure 1. *An record in the structured music metadata.*

A record in the structured metadata like the one in Figure 1 corresponds to an *entity*. An entity may contain *fields*. For example, The *Artist* field of the above example has the *content* "Boyz II Men." An entity does not have to contain all fields. Removal of the *Title* field in Figure 1 would result in a new entity that represents an album instead of a specific song.

A currently deployed system [6] allows users to search music by specifying information about a single field with voice commands in the form of keywords followed by the exact content of the field of an intended entity. User studies have found that users often omit the keywords, and they may not know the exact content of a specific field. Table 1 compares the expected form of queries and the actual queries spoken by users. A more natural/flexible speech interface is thus desirable in this case.

Expected form	What users actually said
Play song all rise	All rise, I guess, from blues
Play song Angel	Sarah, in the arms of an angel
Play album legally blonde	Play legally blonde soundtrack
Play artist Glenn Miller	Glenn Miller, jazz.

Table 1. *The form of queries expected by a deployed system and the actual users' queries.*

A natural interface for music search poses new challenges:

1. *Multi-field queries*. Users often specify more than one field of an entity in a query. Figure 2 shows that more than half of the queries contain information about more than one field. The voice search engine needs to identify which field a word in a query is associated with.

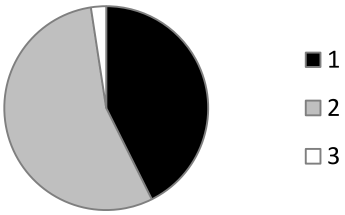

Figure 2. *Distribution of queries containing 1,2, and 3 fields.*

2. *Non-exact match of field contents*: users' specification of a field may not match the content in the metadata exactly. Figure 3 illustrates the percentage of field-wise mismatches between queries and metadata. A robust information retrieval (IR) approach is more suitable than simple pattern matching.
3. *Ambiguity*: Multiple entities may share the same content for a field. Voice search for "Yesterday" may be related to the song "Yesterday" by Beatles or by Leona Lewis, or "Only yesterday" by Carpenters. The problem is complicated by challenge 2 – a word may be shared by different fields of different entities in the metadata. It is important to coordinate the information about multiple fields for disambiguation.

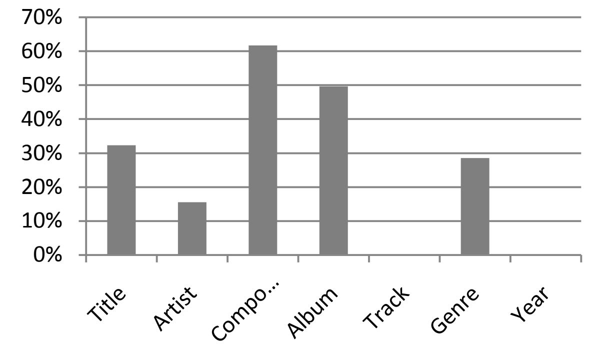

Figure 3. *Percentage of query fields that do not exactly match the content of the corresponding field in the structured data.*

4. *ASR errors*: a recognition error can make an irrelevant entity surface to the top in the search results.

This paper addresses those challenges. Instead of using the traditional structure-agnostic approaches to information retrieval (IR), a field-sensitive model is proposed, and an HMM sequential model is applied subsequently to rescore the hypotheses obtained from the new model (Section 2). The ASR problem is addressed by extending the sequential model to take into account phonetic confusability (Section 3). The different models are evaluated in Section 4, and Section 5 concludes the paper.

2. SEARCH FOR STRUCTURED DATA

2.1. Language Model Based IR

Language model (LM) based IR [7] uses a channel model to find an entity $\hat{E}$ (document) from a document collection given a user's query Q:

$$\begin{aligned}\hat{E} &= \arg\max_E P(E \mid Q) = \arg\max_E P(Q \mid E) P(E) \\ &\approx \arg\max_E P(Q \mid E)\end{aligned} \quad (1)$$

In Eq. (1) every E is assumed equally likely *a priori*. $P(Q|E)$ is modeled generatively with an entity-specific n-gram model that is smoothed with a global background model via linear interpolation:

$$\begin{aligned}P(Q \mid E) &= P(w_1, ..., w_n \mid E) \approx \prod_{i=1}^{n} P(w_i \mid E) \\ &= \prod_{i=1}^{n} \{\lambda \cdot P_{MLE}(w_i \mid E) + (1-\lambda) \cdot P_{MLE}(w_i)\}\end{aligned} \quad (2)$$

LM-based IR is an alternative to other common approaches to IR, like the Tf-Idf weighted vector space model [8]. In a series of experiments on music metadata search, we have found that it had superior performance over other approaches. Therefore, we focus on LM-based structured data search in this paper.

2.2. LM-Based IR for Structured Data

The Baseline Systems

We adopt two different baseline systems. The first (BM1) mimics the deployed system by treating each field as a separate entity. So the exemplar record in Figure 1 produces the following entities: "Boyz II Men" for artist, "It's so hard to say goodbye to yesterday" for song title, "Legacy – the greatest hits collection" for album, and "R&B/Soul" for genre, etc. This baseline system works well only when users specify information about a single field in the query, as expected by the deployed system. The second one (BM2) collapses the structural information and treats the words in each field indifferently. So the structured entity in Figure 1 can be represented as a bag of words: {Boyz, II, Men, It's, so, hard, to, say, goodbye, to, yesterday, R&B, Soul}. In doing so, a multi-field query can be handled.

Interpolation of Field Retrievals

To leverage the structural (field) information in the music metadata, a refined model (FM) based on the interpolation of field specific retrieval models for $P(Q|E)$ is proposed:

$$\begin{aligned}P(Q \mid E) &= P(w_1, ..., w_n \mid E) \approx \prod_{i=1}^{n} P(w_i \mid E) \\ &= \prod_{i=1}^{n} \sum_{F_i} P(w_i, F_i \mid E) = \prod_{i=1}^{n} \sum_{F_i} P(w_i \mid F_i, E) P(F_i \mid E)\end{aligned} \quad (3)$$

Here $P(w_i | F, E)$ is an entity-field specific language model, which is obtained via maximum likelihood estimation (MLE) and smoothed by interpolation with an entity specific model and a global model:

$$P(w_i \mid F, E) = \lambda_f P_{MLE}(w_i \mid F, E) + \lambda_e P_{MLE}(w_i \mid E) + \lambda_c P_{MLE}(w_i) \quad (4)$$

where $\lambda_f + \lambda_e + \lambda_c = 1$. While the interpolation weights can be set using held-out data, we found that the search performance is not very sensitive to their values as long as long as none of the weights is set too close to 0.

An entity independent prior field distribution $P(F)$ is used to derive the entity-specific field distribution by redistributing the probabilistic mass of the fields absent from an entity to the existing fields:

$$P(F = x \mid E) = \begin{cases} \dfrac{P(F = x)}{1.0 - \sum\limits_{\forall y \notin E} P(F = y)} & \text{if } x \in E \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Throughout the paper, the following prior distribution $P(F)$ is used. It incorporates the knowledge about field popularities:

Title	Artist	Album	Composer	Genre	Track	Year
0.3	0.3	0.3	0.025	0.025	0.025	0.025

Sequential Model for Rescoring

The model in Eq. (3) takes into account the structural information in IR ranking. One advantage of the model is that the indexing technique for LM-based IR is applicable for speedy entity search. However, the model is not very accurate in the following aspects:

1. The summation over fields in Eq. (3) does not depict the generative process correctly. Users do not pick a word by considering the contents of all fields. Instead, they think about a field first and then select words to specify its content.
2. The independence assumption of fields at different time is not accurate. Adjacent words are more likely to express the content of the same field. The lack of constraints on field transition in Eq. (3) results in frequent field hopping.

An HMM sequential model (HMM) is introduced to overcome the problems. First the decision rule is modified to search for an entity that gives rise to the highest likelihood of a query Q under the Viterbi field alignment F (an alternative decision rule using

—	m	aa	r	k	—	r	aa	n	s	ax	n	—	ae	n	d	—	ev	m	iy	—	w	ay	n	hh	aw	s	—
SK	SK	SK	SK	SK	—	r	ay	*	z	ih	ng	—	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK	SK

Figure 4. The phonetic alignment between the field content "Mark Ranson and Amy Winehouse" and the recognized word "rising."

summation over all possible alignments is also studied in the experiments for comparison):

$$\begin{aligned}\hat{E} &= \arg\max_E P(E\,|\,Q) = \arg\max_E P(Q\,|\,E)P(E) \\ &\approx \arg\max_E P(Q\,|\,E) = \arg\max_E \max_F P(Q,F\,|\,E)\end{aligned} \tag{6}$$

Here $P(Q,F\,|\,E)$ is modeled by an HMM:

$$P(Q,F\,|\,E) \approx \prod_{i=1}^{n} P(w_i\,|\,F_i,E)P(F_i\,|\,F_{i-1};E) \tag{7}$$

The emission probabilities are modeled in the same way as in Eq. (4). The transition probability is assigned as follows to penalize frequent field hopping:

$$P(F_i\,|\,F_{i-1};E) = \begin{cases} 0.7 & \text{if } F_i = F_{i-1} \\ 0.3 \times P(F_i\,|\,E) & \text{otherwise} \end{cases} \tag{8}$$

where $P(F\,|\,E)$ is the same as in Eq. (5).

The HMM sequential model requires Viterbi decoding, hence it increases the search time by a factor of n^2, where n is the number of fields. To expedite voice search, the sequential model is used to rescore the n-best search results from the previous model in Eq. (3).

3. SEARCH WITH SPOKEN QUERIES

Another advantage of applying generative models for IR lies in the fact that it opens the door for modeling ASR errors according to phonetic confusability. This can be illustrated by an example – the query "Mark Ranson and Amy Winehouse" is recognized as "the rising band Amy Winehouse" mistakenly. As the result, the IR component returns a wrong entity. If the model has the knowledge that "Ranson" and "rising", "and" and "band" are phonetically similar, then the evidence for the correct entity that has both Mark Ranson and Amy Winehouse as the artists would be stronger. Formally, let *R* be the ASR output of a spoken query, then the sequential model with phonetic similarity measure (HMM/PS) can be expressed as

$$P(R,F\,|\,E) = \prod_{i=1}^{n} P(r_i\,|\,F_i,E)P(F_i\,|\,F_{i-1};E) \tag{9}$$

$$\begin{aligned}&P(r\,|\,F;E) = \\ &\quad \lambda_1 P_{MLE}(r\,|\,F;E) + \lambda_2 P_c(r\,|\,F;E) + \lambda_3 P_{MLE}(r\,|\,E) + \lambda_4 P_{MLE}(r)\end{aligned} \tag{10}$$

where $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. Eq. (10) differs from Eq. (4) by the inclusion of $P_c(r\,|\,F,E)$, which is determined by the phonetic similarity between *r* and (part of) the content of *F*. More precisely, the phonetic transcription of a recognized word *r* is aligned to that of the content of a field *F*, and $P_c(r\,|\,F,E)$ is computed according to the operations in the alignment $A(F,r)$ obtained via dynamic programming (DP). In addition to the traditional deletion, insertion, and substitution operations in DP, a skip operation (SK) is introduced. The probability of the operation is 1 if the operation occurs before the first (or after the last) operation involving a phone of *r*, and 0 elsewhere. This is essential since *r* only has to be aligned to a consecutive portion of the content of *F*. Figure 4 shows an exemplar alignment, where "*" on the recognized word side represents a deletion, and "*" on the field content side indicates an insertion. Note that the word boundary "—" is also included to penalize the easy matching of short recognized words like "I". A confusion matrix that contains the probability for each insertion, deletion and substitution operation $P(o)$ is estimated with a data-driven approach from an independent data set. Given the alignment $A(F,r)$ and the $P(o)$ matrix, the phonetic similarity probability is computed as follows:

$$P_c(r\,|\,F,E) \propto \frac{1}{N(F)} \left\{ \prod_{o \in A(F,r)} P(o) \right\}^{\frac{2}{|r|+1}} \tag{11}$$

where $N(F)$ stands for the number of words in the content of *F*, $|r|$ stands for the length of phonetic transcription of *r*.

4. EXPERIEMNTS

4.1. Experiment Setup

The music search query data were collected from 29 subjects. Each subject was instructed to search for their favorite songs with natural speech. The spoken queries were transcribed and the search targets were subsequently labeled by the subjects themselves. 425 queries were collected this way. 409 of them have legitimate corresponding entities in the metadata. The remaining are commands like "Play next song," which are handled by a command & control component instead of voice search. 250 of the data were collected from native English speakers. The metadata of the search targets, together with all other songs in the same albums, were added to a preexisting structured database of about 6,000 entries, resulting in a final dataset of ~11,000 entities. The structured data consist of 75 fields. Among them, only seven fields are searched for by users, including song title (Title), artist name (Artist), composer or conductor (Composer), album title (Album), genre (Genre), track number (Track) and year (Year). Figure 5 shows the distribution of fields being searched for by users. Among them Lyrics and Description are not the actual fields available in the metadata – the metadata-based search is not capable to handle those queries.

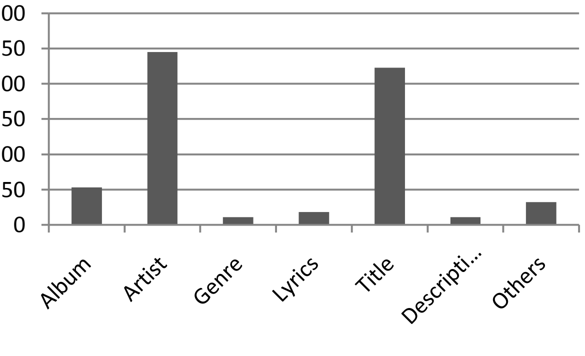

Figure 5. *Percentage of queries containing a specific field.*

4.2. Experimental Results on Text Queries

We first conducted experiments on text queries to study the effectiveness of different models for the search of structured data. For the HMM sequential models, we applied them to rescore the 20-best search results from the field-sensitive IR model (FM). Table 2 shows the n-best (n=1,5,20) search accuracy for the different models.

It is clear that the baseline model BM1, which is based on the assumption that users always specify the information about a single field of a structured entity, is inadequate when natural speech is used. Its accuracy is much lower than the other models.

Model	1 Best	5 Best	20 Best
BM1	57.7%	81.5%	88.8%
BM2	78.8%	89.1%	94.2%
FM	82.0%	91.7%	95.1%
HMM/FB	83.3%	91.9%	95.1%
HMM/Viterbi	84.7%	92.2%	95.1%

Table 2. *N-Best accuracy of music search with text queries.*

Compared to the more reasonable baseline accuracy of BM2, FM, the simple model that takes into account the structural information, has reduced the one-best search error by 15.1%, and the HMM rescoring model with Viterbi decision rule cut the one-best search error rate by 27.8%. The improvement from the HMM rescoring model with decision rule based on all possible field alignments (HMM/FB) has yielded smaller improvements. This confirms our concern about the assumptions underlying the model FM, as discussed in section 2.2.

4.3. Experimental Results on Speech Queries

We have conducted experiments on spoken queries to investigate the robustness of the search algorithm to ASR errors, as well as the effectiveness of the phonetic similarity measure. The experiments were conducted with the subset of the data collected from native English speakers. To study the robustness under different word error rate conditions, different language models were used – some of them have cheating factors. Table 3 lists these language models.

LM	Description
LM1	Trained from the transcriptions of all data.
LM2	Trained from the transcriptions – carrier phrases
LM3	Trained from a subset of metadata + usage patterns
LM4	Trained from a subset of metedata
LM5	Trained from metadata + usage patterns
LM6	Trained from metadata

Table 3. *Language models in the experiments with speech inputs.*

Here carrier phrases like “I want to listen” or “please play” were removed from transcriptions to train LM2. The subset for LM3 and LM4 training contains about 500 metadata entries that cover the entities intended by users. Common usage patterns like “Play *Title* by *Artist*” were introduced in LM3 and LM5 with unified language model [9].

Table 4 shows the word error rates (WERs) and end-to-end search results with different language models and IR models:

LM	WER	BM1	BM2	FM	HMM	HMM/PS
--	0%	65.0%	84.4%	86.0%	88.1%	88.5%
LM1	5.4%	63.4%	81.1%	82.7%	84.4%	84.8%
LM2	14.0%	63.4%	81.5%	83.5%	84.4%	85.6%
LM3	30.0%	60.9%	79.4%	79.8%	81.9%	82.7%
LM4	28.1%	61.3%	78.6%	79.8%	81.9%	82.7%
LM5	25.3%	60.9%	76.1%	78.2%	79.4%	79.8%
LM6	33.2%	57.6%	67.5%	70.0%	70.4%	71.6%

Table 4. *WER and 1-best search accuracy with spoken queries. The first row shows the performance on manual transcriptions.*

Several remarks can be made about the results in Table 4:

1. The end-to-end search results are fairly robust to ASR errors. While the word accuracy has dropped by 33% from manual transcription to the ASR using LM6, the one-best search accuracy has deteriorated by around 19%.
2. Compared to BM2, FM has consistently reduced search error by 2% to 11%, HMM sequential rescoring has reduced the error by 9% to 23%, and HMM rescoring with phonetic similarity measure has reduced the search error rate by 13% to 26%. The HMM/PS model has reduced the error rate by 3% to 8% compared to the sequential rescoring model without phonetic similarity measure. Although the improvements of HMM/PS are not statistically significant, they are consistent across different WER conditions.

5. CONCLUSIONS

The problem of retrieving structured data for voice search applications is investigated in this paper. We have shown that a voice search model restricting users from specifying information about multiple fields in structured data has performed very poorly when users speak naturally. The field sensitive model FM has significantly improved (15% error reduction on text queries, 2%~11% on spoken queries) the retrieval accuracy over the baseline field-agnostic model BM2.The HMM sequential rescoring model has further reduced the search error (27.8% over BM2 on text queries and 13%~26% on spoken queries when phonetic similarity measure is introduced to the model.) Overall the end-to-end search results are relatively robust to ASR errors. For future work, we would extend the work in [10] to train language models aiming at improving search accuracy.

6. REFERENCES

[1] Wang, Y.-Y., et al.: ‘An Introduction to Voice Search’, IEEE Signal Processing Magazine, 2008, 25, (3), pp. 29-38

[2] Yu, D., et al.: ‘Automated Directory Assistance System - from Theory to Practice’. Proc. INTERSPEECH, Antwerp, Belgium. 2007.

[3] Bacchiani, M., et al.: ‘Deploying GOOG-411: Early lessons in data, measurement, and testing’. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, USA, April 2008.

[4] Mann, S., et al.: ‘How to Access Audio Files of Large Data Bases Using In-car Speech Dialogue Systems’. Proc. INTERSPEECH, Antwerp, Belgium. 2007.

[5] Weng, F., et al.: ‘CHAT: A Conversational Helper for Automotive Tasks’. Proc. INTERSPEECH, Pittsburgh, PA, USA, September, 2006.

[6] http://www.syncmyride.com/

[7] Lee, J.H.: ‘Combining multiple evidence from different properties of weighting schemes’. Proc. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1995.

[8] Salton, G.: ‘Introduction to Modern Information Retrieval.’ McGraw-Hill, 1983.

[9] Wang, Y.-Y., et al.: ‘A Unified Context-Free Grammar And N-Gram Model For Spoken Language Processing’. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey. 2000.

[10] Wang, Y.-Y., and Acero, A.: ‘Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy’. Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, US Virgin Islands. 2003.