Speech Retrieval in Unknown Languages: a Pilot Study*

Xiaodan Zhuang# **Jui Ting Huang**# **Mark Hasegawa-Johnson**
Beckman Institute, Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, U.S.A.
{xzhuang2,jhuang29,jhasegaw}@uiuc.edu

Abstract

Most cross-lingual speech retrieval assumes intensive knowledge about all involved languages. However, such resource may not exist for some less popular languages. Some applications call for speech retrieval in unknown languages. In this work, we leverage on a quasi-language-independent subword recognizer trained on multiple languages, to obtain an abstracted representation of speech data in an unknown language. Language-independent query expansion is achieved either by allowing a wide lattice output for an audio query, or by taking advantage of distinctive features in speech articulation to propose subwords most similar to the given subwords in a query. We propose using a retrieval model based on finite state machines for fuzzy matching of speech sound patterns, and further for speech retrieval. A pilot study of speech retrieval in unknown languages is presented, using English, Spanish and Russian as training languages, and Croatian as the unknown target language.

1 Introduction

Dramatic increase in recorded speech media calls for efficient retrieval of audio files. Accessing speech media of a foreign language is a particularly important and challenging task, often referred to as cross-lingual speech retrieval or cross-lingual spoken document retrieval.

Previous work on cross-lingual speech retrieval mostly leverages on intensive knowledge about all the languages involved. Most reported work investigates retrieval in a target language, in response to audio or text queries given in a different source language (Meng et al., 2000; Virga and Khudanpur, 2003). Usually, the speech media in the target language, and the audio queries in the source language, are converted to speech recognition transcripts using large-vocabulary automatic speech recognizers (LVASR) trained for the target language and the source language respectively. The text queries, or transcribed audio queries, are translated to the target language. Text retrieval techniques are applied to retrieve speech, by retrieving the corresponding LVASR transcription in the target language. In such systems, a large-vocabulary speech recognizer trained on the target language is essential, which requires the existence of a dictionary and labeled acoustic training data in that language.

LVASR currently do not exist for most of the 6000 languages on Earth. In some situations, knowledge about the target language is limited, and definitely not sufficient to enable training LVASR. Imagine an audio database in a target language unknown to a user, who needs to retrieve spoken content relevant to some audible query in this unknown language. For example, the user knows how the name "Obama" is pronounced in the target language, and wants to retrieve all spoken documents that contain the query word, from a database in this unknown language. A linguist might find himself/herself in this scenario when he or she tries to collect a large number of utterances containing some particular

*This research is funded by NSF grants 0534106 and 0703624. The authors would like to thank Su-Youn Yoon for inspiring discussion. #The student authors contribute equally.

phrases in an unknown language. Similarly, an information analyst might wish to leverage on speech retrieval in unknown languages to organize critical information before engaging linguistic experts for finer analysis. We refer to such retrieval tasks as *speech retrieval in unknown languages*, in which little knowledge about the target language is assumed.

A human linguist attempting to manually perform speech retrieval in an unknown language would necessarily map the perceived speech (both database and query) into some cognitive abstraction or schema, representing, perhaps, the phonetic distinctions that he or she has been trained to hear. Matching and retrieval of speech would then be performed based on such an abstraction. Two cognitive processes, assimilation and accommodation, take place when human brains are to process new information (Bernstein et al., 2007), such as speech in an unknown language. In accommodation, the internal stored knowledge adapts to new information with which it is confronted. In assimilation, the new information, e.g., speech in an unknown language, is mapped to previously stored information, e.g., subwords (phones) as defined by knowledge about the languages known to the listener.

This paper models speech retrieval in unknown languages using a machine learning model of phonetic assimilation. A quasi-language-independent subword recognizer is trained to capture salient subwords and their acoustic distribution in multiple languages. This recognizer is applied on an unknown language, therefore mapping segments of the unknown speech to subwords in the known languages. Through this machine cognitive process, the database and queries in the unknown language are represented as sequences of quasi-language-independent subwords. Speech retrieval is performed based on such representation. Figure 1 illustrates that speech retrieval in an unknown language can be modeled as a special case of assimilation.

This task differs from the more widely studied known-language speech retrieval task, in that no linguistic knowledge of the target language is assumed. We can only leverage on knowledge that can be applied by assimilation to the multiple known languages. Therefore, this task is more like a cross-lingual sound pattern retrieval task, leveraged on quasi-language-independent subwords, rather than

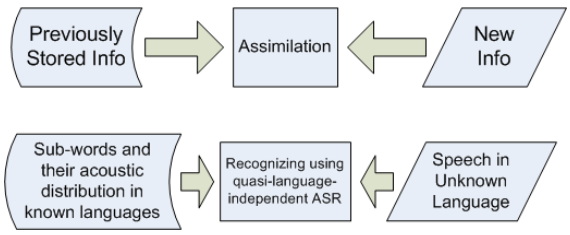

Figure 1: Automatic speech retrieval in an unknown language (below) is modeled as a special case of the cognitive process called assimilation (above).

a translated spoken word/phrase retrieval task using target language LVASR transcripts, as in most cross-lingual speech retrieval systems. The quasi-language-independent subword recognizer is trained on speech data other than the target language, and therefore generates much noisier recognition results, owing to potential mismatch between acoustic distributions, lack of dictionary and lack of a word-level language model.

To manage the extra difficulty, we adopt a subword lattice representation to encode a wide hypothesis space of recognized speech in the target language. Language-independent query expansion is achieved either by allowing a wide lattice output for an audio query, or by taking advantage of distinctive features in speech articulation to propose quasi-language-independent subwords most similar to the given subwords in a query. Finite state machines (FSM) constructed from the speech lattices are used to allow for fuzzy matching of speech sound patterns, and further for retrieval in unknown languages.

We carry out a pilot study of speech retrieval in unknown languages, using English, Spanish and Russian as training languages, and Croatian as the unknown target language. To explain the effect of additional knowledge about the target language, we demonstrate the improvements in retrieval performance that result by incrementally making available subword sequence models and acoustic models for the target language.

2 Quasi-Language-Independent subword Models

2.1 Deriving a subword set

Based on the assumption that an audible phrase in an unknown language can be represented as a sequence

of subwords, the question is to find an appropriate set of subword symbols. Schultz and Waibel (2001) reported that a global unit set for the source languages based on International Phonetic Alphabet (IPA) symbols outperforms language-dependent phonetic units in cross-lingual word recognition tasks, whereas language-dependent phonetic units are better models for multilingual word recognition (in which the target language is also one of the source languages). A multilingual task might benefit from partitioning the feature space according to language identity, i.e., to have different subsets of models aiming at different languages. By contrast, a cross-lingual task calls for one consistent set of models with language-independent properties in order to maximize portability into the new language.

To capture the necessary distinctions between different phones across languages, we first pool together individual phone inventories for source languages, each of which has its phones tagged with a language identity, and then performed bottom-up clustering on the phone pool based on pairwise similarity between their acoustic models. Each cluster represents one distinct language-independent subword symbol. Since this set is still derived from multiple languages, we refer to these subword units as *quasi-language-independent subwords*. A quasi-language-independent subword set is derived by the following steps:

First, we encode all speech in the known languages using a language-dependent phone set. Each symbol in this set is defined by the phone identity and the language identity. One single-Gaussian three-state left-to-right HMM is trained for each of these subword units.

Second, similarity between the language-dependent phones is estimated by the approximated KL divergence between corresponding acoustic models. As shown in (Vihola et al., 2002), KL divergence between single-Gaussian left-to-right HMMs can be approximated in closed form by Equation 1,

$$KLD(U,V) = \sum_{i=1}^{S} r_i \sum_{j=1}^{S} a_{ij}^U \log\left(a_{ij}^U / a_{ij}^V\right) \quad (1)$$

$$+ \sum_{i=1}^{S} r_i I\left(b_i^U : b_i^V\right), \quad (2)$$

where a_{ij} is the transition probability to hidden state j, and b_i and r_i are the observation distribution and steady-state probability for hidden state i. For single-Gaussian distribution, $I\left(b_i^U : b_i^V\right)$ can be approximated by,

$$\begin{aligned} I(b_i^U : b_i^V) &= \frac{1}{2}\Big[\log\frac{\left|\Sigma_i^V\right|}{\left|\Sigma_i^U\right|} \\ &+ tr\left(\Sigma_i^U\left(\left(\Sigma_i^V\right)^{-1} - \left(\Sigma_i^U\right)^{-1}\right)\right) \\ &+ tr\left(\left(\Sigma_i^V\right)^{-1}\left(\mu_i^U - \mu_i^V\right)\left(\mu_i^U - \mu_i^V\right)^T\right)\Big]. \end{aligned}$$

Third, we use the Affinity Propagation algorithm (Frey and Dueck, 2007) to conduct pairwise clustering of phones based on the approximated KL divergence between acoustic models. The tendency for a data point (a phone) to be an exemplar of a cluster is controlled by the preference value assigned to that phone. The preference of a phone i is set as follows to favor frequent phones to be cluster centers:

$$p(i) = k\log(C_i), \quad (3)$$

where C_i is the count of the phone i, and k is a normalization term to control the total number of clusters. To discourage subwords from the same language to join a same cluster, pairwise distance between them are offset by an additional amount, comparable to the maximum pairwise distance between the models.

The resultant subword set is supposed to capture quasi-language-independent phonetic information, and each subword unit has relatively distinctive acoustic distribution. These subwords are encoded using the corresponding cluster exemplars as surrogates.

2.2 Recognizing subwords

An automatic speech recognition (ASR) system (Jelinek, 1998) serves to recognize both queries and speech database, with acoustic models for the language-independent subwords derived from the known languages as described in section 2.1. The front-end features extracted from the speech data are 39-dimensional features including 12 Perceptual Linear Prediction (PLP) coefficients and their energy, as well as the first-order and second order regression coefficients.

We create context-dependent models for each subword, using the same strategy for building context-dependent triphone models in LVASR (Woodland et al., 1994). A "triphone" is a subword with its context defined as its immediate preceding and following subwords. Each triphone is represented by a continuous three-state left-to-right Hidden Markov Model (HMM). Additionally, there is a one-state HMM for silence, two three-state HMMs for noise and unknown sound respectively. The number of Gaussian mixtures (9 to 21 Gaussians) is optimized according to a development set consisting of speech in the known languages. A standard tree-based state tying technique is adopted for parameter sharing between subwords with similar contexts.

The "language model" (LM), or more precisely subword sequence model, should generalize from the known languages to the unknown language. Our trial experiments showed that unigram statistics of subwords and their triphones is more transferable across languages than N-gram statistics. We also assume that infrequent triphones are less likely to be salient units that would carry the properties of the unknown language. Thus, we select the top frequent triphones and map the rest of the triphones to their center phones, forming a mixed vocabulary of frequent triphones and context-independent subwords. The frequencies of these vocabulary entries are used to estimate an unigram LM in the ASR system. Triphones in the ASR output are mapped back to its center subwords before the retrieval stage.

3 Speech Retrieval through Subword Indexing

In many cross-lingual speech retrieval systems, the speech media are processed by a large-vocabulary automatic speech recognizer (LVASR), which has access to vocabulary, dictionary, word language model and acoustic models for the target language. With all these resources, state-of-the-art speech recognition could give reasonable hypothesized word transcript, enabling direct application of text retrieval techniques. However, this is not the case in speech retrieval in unknown languages. Moreover, without the higher level linguistic knowledge, such as a word dictionary, this task aims to *find speech patterns that sound similar*, as approximated by sequences of quasi-language-independent subwords. Therefore, the sequential information in the hypothesized subwords is critical.

To deal with the significant noise in the subword recognition output, and to emphasize the sequential information, we use the recognizer to obtain subword lattices instead of one-best hypotheses. These lattices can be represented as weighted automata, which are compact representations of a large number of alternative subword sequences, each associated with a weight indicating the uncertainty of the data. Therefore, indexing speech in unknown language can be achieved by indexing the corresponding weighted automata with quasi-language-independent subwords associated with the state transitions.

We adopt the weighted automata indexation algorithm reported in (Allauzen et al., 2004), which is optimal for searching subword sequences, as it takes time linear in the sum of the query size and the number of speech media entries where it appears. The automata indexation algorithm also preserves the sequential information, which is crutial for this task. We leverage on two kinds of knowledge for query expansion, namely empirical phone confusion and knowledge-based phone confusion. An illustration of our speech retrieval system is presented in Figure 2. We detail the indexing approaching as well as query expansion and retrieval in this section.

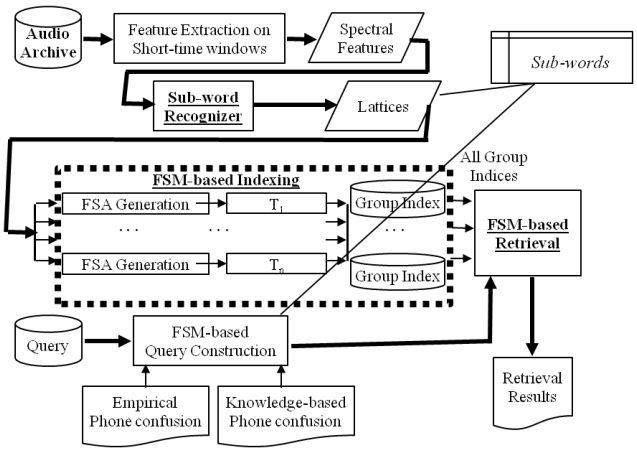

Figure 2: Framework of speech retrieval through subword indexing

3.1 Subword Finite State Machines as Speech Indices

We construct a full index that can be used to search for a query within all the speech utterances $u_i, i \in 1, ..., n$. In particular, this is achieved by constructing a weighted finite-state transducer T, mapping each query x to the set of speech utterances where it appears. Each returned speech utterance u is assigned a score, which is the negative log of the expected count of the query x in utterance u.

The subword lattice for speech utterance u_i can be represented as a weighted finite state automata A_i, whose path weights correspond to the joint probability of the observed speech and the hypothesized subword sequence. To get an automata whose path weights correspond to desired negative log of posterior probabilities, we simply need to apply a general weight-pushing algorithm to A_i in the log semiring, resulting in an automata B_i. In this automata B_i, the probability of a given string x is the sum of the probability of all paths that contains x.

The key point of constructing the index transducer T_i for uttereance u_i is to introduce new paths that enable matching between a query and any portions of the original paths, while properly normalizing the path weights. This is achieved by factor selection in (Allauzen et al., 2004). First, null output is introduced to each transition in the automata, converting the automata into a transducer. Second, a new transition is introduced from a new unique initial state to each existing state, with null input and output. The weight associated with this transition is the negative log of the forward probability. Similarly, a new transition is created from each state to a new unique final state, with null input and output as the label i of the current utterance u_i. The assicated weight is the negative log of the backward probability. General finite state machine optimization operations (Allauzen et al., 2007) of weighted ϵ-removal, determinization and minimization over the log semiring can be applied to the resulting transducer. As shown in (Allauzen et al., 2004), the path with input of string x and output of label i has a weight corresponding to the negative log of the expected count of x in utterance u_i.

To optimize the retrieval time, we divide all utterances into a few groups. Within each group, the utterance index transducers are unioned and determinized to get one single index transducer for the group. It is then feasible to expedite retrieval by processing each group index transducer in a parallel fashion.

3.2 Query Expansion

While sequential information is important, exact string match is very unplausible in this challenging task, even when subword lattices encode many alternative recognition hypotheses. Language-independent query expansion is therefore critical for success in retrieval. We carry out query expansion either by allowing a wide lattice output for an audio query, or by taking advantage of distinctive features in speech articulation to propose quasi-language-independent subwords most similar to the given subwords in a query.

In particular, for a spoken query, ASR will generate a subword lattice instead of a one-best subword sequence hypothesis. With the lattice, the audio query is encoded by the best hypothesis from ASR and its empirical phone confusion. The lattice can then be represented as a finite-state automata.

However, when the query is given as a target language subword sequence, we can no longer use the recognizer to obtain an expanded query. Furthermore, some target language subwords may not even exist in the quasi-language-independent subword set in the recognizer. In this case, knowledge-based phone confusion is engaged via the use of a set of distinctive features $F_j, j \in 1, ..., M$ for human speech (Chomsky and Halle, 1968), including labial, alveolar, post-alveolar, retroflex, voiced, aspirated, front, back, etc.

We estimate similarity from phone a to phone b, or more precisely, substitution tendency as in Equation 4,

$$DFsim(a, b) = \log \frac{N_{ab}}{N_a} \tag{4}$$

where

$$N_{ab} = \sum_{j=1}^{M} (F_j^a \times F_j^b = 1),$$

$$N_a = \sum_{j=1}^{M} (F_j^a \neq 0).$$

The target subword sequence is first mapped to the derived subword set, by locating the identical or nearest member phone in the clustering and then adopting the surrogate for that cluster. This converted sequence of derived subwords is further expanded by adding the most likely alternative quasi-language-independent subwords, parallel to each original subword. Transitions to these alternative subwords are associated with the corresponding substitution tendency based on distinctive features.

3.3 Search

An expanded query, either obtained from an audio query or a subword sequence query, is represented as a weighted finite state automata. Searching this query in the utterances is achieved by composing the query automata with the index transducer. This results in another finite state transducer, which is further processed by projection on output, removal of ϵ arcs and determinization. The output is a list of retrieved speech utterances, each with the expected count of the query.

Apparently, the precision and recall of the retrieval results vary with the width of the subword lattices used for indexing as well as how much the query is expanded. We control the width of the subword lattices via the number of tokens and the maximum probability decrease allowed for each step in the Viterbi decoding. The extend to which a subword sequence query is expanded is determined by the lowest allowed similarity between the original phone and an alternative phone. These parameters are set empirically.

4 Experiments

4.1 Dataset

The known language pool should cover as many language families as possible so that the derived subwords could better approximate language independence. However, as a pilot study, this paper reports experiments using only languages within the Indo-European family. Table 1 summarizes the size of speech data from each language. Croatian is used as the unknown target language, and the other three languages are the known languages used for deriving and training the quasi-language-independent subword models. We extracted 80% of all speakers per language for training, and 10% as a development set.

Language	ID	Hours	Spks	Style
Croatian	hrv	21.3	201	Read+answers
English	hub	13.6	406	Broadcast
Spanish	spa	14.6	120	Read+answers
Russian	rus	2.5	63	Read+answers

Table 1: Summary for data: language ID, total length, number of speakers and speaking style for each language.

4.2 Settings

The speech retrieval task aims to find speech utterances that contain a particular query. We use two kinds of queries: 1) subword sequence queries, transcribed as a sequence of phonetic symbols in the target language; 2) audio queries, each being an audio segment of the speech query in the target language.

Since we aim to match speech patterns that sound like each other, the queries used in this experiment are relatively short, about 3 to 5 syllables. This adds to the challenge in that very limited redundant information is available for query-utterance matching. There are totally 40 subword sequences and 40 audio queries, each occurs in between 18 and 38 utterances out of a set of 576 utterances.

In addition to a cross-lingual retrieval system built using only the known languages, we incrementally augment resource on the target language to build more knowledgeable systems.

AM0LM0: Both the acoustic model (AM) and the language model (LM) are quasi-language-independent, trained using data in multiple known languages. This happens when no transcribed speech data or a defined phone set exist for the target language. Essentially the system has no direct knowledge about the target language.

AM0LMt: This setting examines the performance gap due to the acoustic model mismatch by using a quasi-language-independent AM, but a target language LM. Suppose that a word dictionary with phonetic transcription and possibly some text data from the target language are available, for training a target language subword LM. To find the mapping between target triphones and language-independent source AMs, linguistic knowledge and phonetic symbol notation are the only information

we can use. First, we map each of target monophones to source phone symbols: Any source cluster that contains a phonetic symbol with the same notation as the target phonetic symbol becomes a surrogate symbol for that target phone. If a target phone is unseen to the known languages, the most similar phone will be chosen first. The similarity is based on the distinctive features, as discussed in Section 3.2. Second, the target triphones are converted to possible source triphones for which acoustic models exist. Each target triphone not modeled in the source language AM is replaced with the corresponding diphone (subword pair) if it exists, otherwise the center phone.

AMtLM0: This setting examines the performance gap due to the language model mismatch by using a quasi-language-independent source LM, but a target language AM. For the source triphones and monophones that do not exist in the target AM, they are mapped to target AMs in a way similar as described above.

AMtLMt: Both AM and LM are trained for the target language. This setting provides an upper bound of the performance for different settings.

4.3 Metrics

We evaluate the performance for both subword recognition and speech retrieval, measured as follows.

Recognition Accuracy: The ground truth is encoded using subwords in the target language while the recognition output is encoded using quasi-language-independent subwords in Section 2. To measure the recognition accuracy, we label each quasi-language-independent subword cluster using the most frequent target language subword that appears in that cluster. The hypothesis subword sequence is then compared against the groundtruth using a dynamic-programming-based string alignment procedure. The recognition accuracy is defined as $REC - ACC = \frac{H-I}{N} \times 100\%$, where H, I, and N are the numbers of correct labels, insertion errors and groundtruth labels respectively.

Retrieval Precision: The retrieval performance is measured using Mean Average Precision ($IR - MAP$), defined as the mean of the Average Precision (AP) for a set of different queries x. Mean Average Precision ($IR - MAP$) can be defined in Equation 5. n is the number of ordered retrieved utterances and R is the total number of relevant utterances. f_i is an indicator function whether the i^{th} retrieved utterance does contain the query. Precision p_m for top m retrieved utterances can be calculated as $p_m = \frac{1}{m}\sum_{k=1}^{m} f(k)$.

$$IR - MAP = \frac{1}{Q}\sum_{x=1}^{Q} AP(x),$$
$$AP(x) = \frac{1}{R(x)}\sum_{i=1}^{n(x)} f_i(x)p_i(x). \quad (5)$$

We use $IR - MAP_A$ and $IR - MAP_S$ to denote the retrieval MAP for audio queries and subword sequence queries respectively.

4.4 Results

Table 2 presents a few examples of the derived quasi-language-independent subwords. As discussed in Section 2, these subwords are obtained by bottom-up clustering of all the language-dependent IPA phones in the multiple known languages. The same IPA symbol across languages may lie in the same cluster, e.g., /z/ in Cluster 1, or different clusters, e.g., /j/ in Cluster 3 and 4. Although symbols within the same language are discouraged to be in one cluster, it still desirably happens for highly similar pairs, e.g., /ɨ/$_{rus}$ and /j/$_{rus}$ in Cluster 4.

Cluster ID	Surrogate	Other phone members
1	/z/$_{hub}$	/z/$_{spa}$, /z/$_{rus}$, /zʲ/$_{rus}$
2	/tʃʲ/$_{rus}$	/tʃ/$_{hub}$, /tʃ/$_{spa}$
3	/j/$_{hub}$	/j/$_{spa}$
4	/iː/$_{hub}$	/ɨ/$_{rus}$, /j/$_{rus}$

Table 2: Examples of quasi-language-independent subwords, as clusters of source language IPAs.

Table 3 compares the subword recognition and retrieval performance for the quasi-language-independent subwords and IPA phones. We can

Setting	$REC - ACC$	$IR - MAP_A$	$IR - MAP_S$
IPA	37.18%	17.90%	31.40%
AM0LM0	42.52%	23.24%	32.62%

Table 3: Performance of quasi-languange-independent subword and IPA.

Setting	AMtLMt	AMtLM0	AM0LMt	AM0LM0
$REC - ACC$	73.45%	67.29%	49.88%	42.52%
$IR - MAP_A$	58.82%	52.38%	28.32%	23.24%
$IR - MAP_S$	76.96%	51.86%	34.95%	32.62%

Table 4: Performance of subword recognition and speech retrieval.

see that on the unknown language Croatian, the derived quasi-language-independent subwords outperform the IPA symbol set in both phone recognition and retrieval using two kinds of queries.

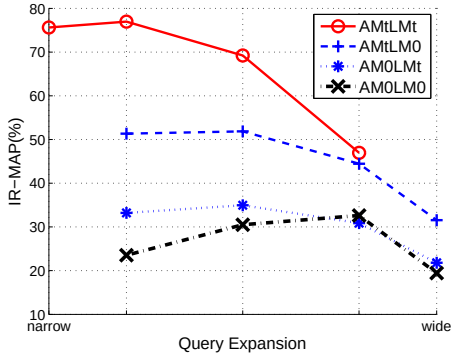

Figure 3: Speech retrieval performance for subword sequence queries

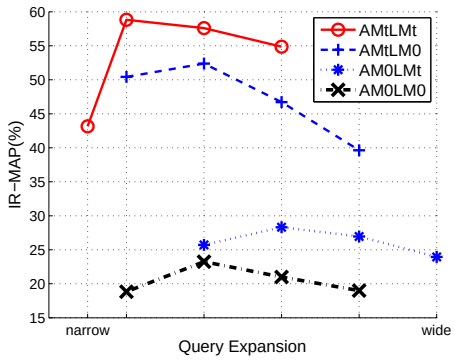

Figure 4: Speech retrieval performance for audio queries

Table 4 presents the subword recognition accuracy and retrieval performance with optimal query width. Figure 3 and Figure 4 presents speech retrieval performance at varying query widths for subword sequence queries and audio queries respectively. It is shown that speech retrieval in completely unknown language achieves MAP of 23.24% and 32.62% while the system trained using the most available knowledge about the target language reaches MAP of 58.82% and 76.96%, for audio queries and subword sequence queries respectively. We also demonstrate access to phone frequency (**AM0LMt**) and acoustic data (**AMtLM0**) both boosts retrieval performance, and the effect is roughly additive (**AMtLMt**).

5 Conclusion and Discussion

In this work, we present a speech retrieval approach in unknown languages. This approach leverages on speech recognition based on quasi-language-independent subword models derived from multiple known languages, and finite state machine based fuzzy speech pattern matching and retrieval. Our experiments use Croatian as the unknown language and English, Russian and Spanish as the known languages. Results show that the derived subwords outperform the IPA symbols, and access to the subword language model and acoustic models in the unknown language explains the gap between this challenging task and retrieval with knowledge about the target language.

The proposed retrieval approach on unknown languages can be viewed as a machine learning model of phonetic assimiliation, in which the segments in an unknown language are mapped to language-independent subwords learned from the multiple known languages. However, another important cognitive process, i.e., accomodation, is not yet modeled. We believe the capability to create new subwords unseen in the known languages would lead to improved performance. In particular, speech segments that are hypothesized by the quasi-language-independent subword recognizer with very low confidence scores can be clustered to form these new subwords, accomodating to the unknown language.

The approach in this work can be readily scaled up to much larger speech corpora. In particular, larger corpora would make it more practical to implement the accomodation process discussed above. Besides, that would also enable online adaptation of the model parameters of the quasi-language-independent subword recognizer. Both are believed to promise reduced gap between retrieval performance in a known language and an unknown language, and are potential future work beyond this paper.

References

C. Allauzen, M. Mohri, and M. Saraclar. 2004. General indexation of weighted automata – application to spoken utterance retrieval. In *Proc. HLT-NAACL*.

C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M.Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *Proc. CIAA*.

Bernstein, Penner, Clarke-Stewart, and Roy. 2007. *Psychology*. Houghton Mifflin Company.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315:972–976.

Frederick Jelinek. 1998. *Statistical Methods for Speech Recognition*. The MIT Press.

Helen Meng, Berlin Chen, Erika Grams, Sanjeev Khudanpur, Wai-Kit Lo, Gina-Anne Levow, Douglas Oard, Patrick Schone, Karen Tang, Hsin-Min Wang, and Jian Qiang Wang. 2000. Mandarin-english information (MEI): Investigating translingual speech retrieval. http://www.clsp.jhu.edu/ws2000/final_reports/mei/ws00mei.pdf.

Tanja Schultz and Alex Waibel. 2001. Language independent and language adaptive acoustic modeling for speech recognition. *Speech Communication*, 35:31–51.

M. Vihola, M. Harju, P. Salmela, J. Suontausta, and J. Savela. 2002. Two dissimilarity measures for hmms and their application in phoneme model clustering. In *Proc. ICASSP*, volume 1, pages I–933 – I–936.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in crosslingual information retrieval. In *Proc. ACL 2003 workshop MLNER*.

P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young. 1994. Large vocabulary continuous speech recognition using HTK. In *Proc. ICASSP*, volume 2, pages II/125–II/128.