A Robust Fusion Method for Multilingual Spoken Document Retrieval Systems Employing Tiered Resources

Murat Akbacak, John H.L. Hansen

Center for Robust Speech Systems, Department of Electrical Engineering
University of Texas at Dallas, Richardson, TX, 75083, U.S.A
Web: http://crss.utdallas.edu
{murat.akbacak,john.hansen}@utdallas.edu

10.21437/Interspeech.2006-357

Abstract

In this study, we present two novel fusion approaches to merge subword and word based retrieval methods within a multilingual spoken document retrieval (SDR) system. Considering the fact that more than 6000 languages are spoken in the world today, resources (e.g., text and audio data, pronunciation lexicon) needed to develop Automatic Speech Recognition (ASR) systems for such a range of languages (accordingly the performances of these ASR systems) can be considered within a tiered structure. Even for resource-rich languages, some applications (e.g., historical digital archives) contain acoustical/lexical variations among time which presents challenges to build effective up-to-date audio indexing and retrieval systems. Within this concept, we focus on creating *robust* multilingual SDR systems employing both word-based and subword-based retrieval methods. Our proposed algorithms employ an OOV-word detection module to generate hybrid transcripts/lattices. In our Dynamic Fusion (DF) approach, hybrid transcripts/lattices are used to assign dynamic fusion weights to each subsystem. In our Hybrid Fusion (HF) approach, queries are searched through hybrid lattices. We evaluated our proposed algorithms in a proper name retrieval task within the Spanish Broadcast News domain, and spoken document retrieval task using our historical speech archive NGSW corpus [1], where the proposed algorithms yield improvements over traditional fusion methods.
Index Terms: audio indexing, subword, retrieval, multilingual.

1. Introduction

Multilingual information search in audio recordings (e.g., audio broadcasts, archives from digital libraries, audio content on the internet, etc.) is expanding at an increasing rate as more audio data becomes available in different languages. In many cases, there are only a few or no linguists in high-interest languages leading to a considerable shortfall in transcription and/or lexicon creation efforts within the task of ASR development. In addition to this, data sparseness is another research problem since the available training material (e.g., audio and text) needed to train ASR systems, generally speaking, is limited and diverse (e.g., different recording conditions, speaking styles, accents/dialects, etc.). All these factors, in what we refer to as tiered resources, result in different operating tiers for multilingual SDR systems. Therefore, developing algorithms that yield robust SDR performance, independent of the tier the system is operating at, becomes crucial within the context of portability and rapid-transition to resource-limited target languages.

Even for resource-rich languages (e.g., English), searchable audio content is sometimes so diverse (e.g., audio content on the internet, or spoken archives from digital libraries [1]) due to acoustical/lexical variations across time, that it is difficult to build effective up-to-date SDR systems. System tuning could help to maintain SDR performance at desired levels, on the other hand this is a costly (time, labor, money) solution. Finding automatic ways of maintaining system performance at the desired level becomes a practical concern across languages and time periods.

In this paper, we focus on rapid transition to resource-limited target languages within the context of multilingual audio indexing and retrieval. Our goal therefore, is to develop robust retrieval methods for new target languages. Different tiers might result from changing lexicon coverage [1] (poor coverage at the beginning, with an evolving lexicon as more resources are employed to obtain better coverage), or data sparseness or mismatch during acoustic model training.

It is a well known fact that during recognition, shorter units (e.g., monophones) are more robust to errors and word variants than longer units (e.g., words), but longer units capture more discriminative information and are less susceptible to false matches during retrieval. In order to move towards solutions that address the problem of misrecognition (both in-vocabulary word and out-of-vocabulary word errors) during SDR, previous studies have employed fusion methods [2, 3, 4, 5, 6] to recover from recognition errors during retrieval. In these methods, fusion weights are optimized for specific tasks. More importantly, these methods assume a homogeneous audio collection where the ASR system achieves similar performance in each audio document. Results from these studies show small improvements over word-based-only retrieval approaches as the number of documents increases.

In our algorithms, in order to use fusion methods more effectively for changing tiers within SDR, for each utterance we first run an OOV-word and misrecognition detection module. Based on the output of this module, the first fusion algorithm, Dynamic Fusion (DF) approach, decides how to merge subword and word based retrieval scores dynamically. The idea is to automatically adapt the fusion weights according to the present tier (lexicon coverage or acoustic model accuracy in this paper) for which the retrieval system is operating. The second algorithm, Hybrid Fusion

This work was funded by grants from the U.S. Air Force Research Laboratory, Rome NY, under contract number F30602-03-0110, and by the University of Texas at Dallas under Project EMMITT.

[1] Lexical variation over time (e.g., 1890's to present) for resource-rich languages are out of the scope of this paper [1].

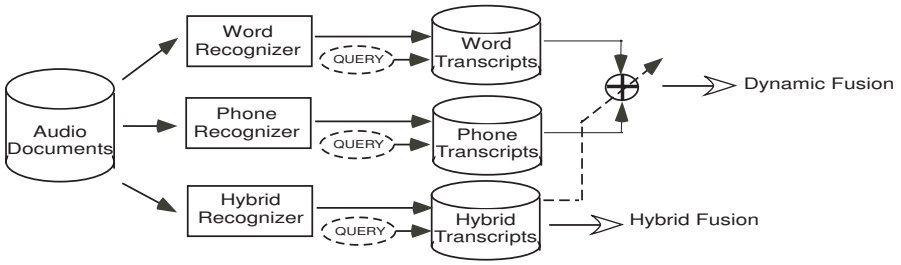

Figure 1: An overview of proposed algorithms employing subword and word based retrieval systems.

(HF) approach, searches query words through the hybrid lattice. These methods could be considered as dynamic back-off strategies where subword-based retrieval scores are merged with word-based retrieval scores according to the performance level of the word-based recognizer. In other words, subword-based retrieval is employed more heavily than usual for utterances where the word-based ASR system is performing poorly.

In Sec. 2, we present the baseline system, and the proposed retrieval algorithms employing Dynamic Fusion (DF) and Hybrid Fusion (HF). Recognition results and evaluation of the proposed retrieval algorithms for two different tasks are presented in Sec. 3. Discussion and future work are presented in Sec. 5, with conclusions in Sec. 6.

2. System Architecture

2.1. Baseline System

In our baseline system, we employ conventional fusion methods where subword and word based retrieval system outputs are merged with fixed weights for in-vocabulary query words. For OOV query words, only subword-based retrieval method is employed. In the current system, only phonemes are used as subword units. In addition to 1-best recognition hypothesis, we also use N-best transcripts and lattices for our word-based retrieval and phoneme-based retrieval systems, respectively. In our word-based retrieval engine, we use a modified version of the MG [7] retrieval system. In this version, the *tfidf* weighting scheme is replaced with Okapi as explained in [1]. Stop-word removal and stemming are applied to the resulting ASR transcripts. For phoneme-based retrieval, we use the system developed in our previous study [8] where Finite State Transducers (FST) are used to index phone lattices, and to retrieve query words using confusion-embedded pronunciations.

In addition to using subword and word based retrieval methods in a fusion approach, one can employ query expansion (QE) to address the problem of misrecognized/OOV word retrieval. Assuming the lexicon coverage is ranging from high rates to somewhat common rates, then QE will not be a valid solution for this problem, and an alternative solution is needed. In addition to this, proper names (which are most likely not in the lexicon) are more discriminative than common words during retrieval.

2.2. Proposed Algorithms

We employ a hybrid-recognition based OOV detection module which is similar to the one proposed in [9]. Different from the word-based recognizer used in the baseline system, we use a generic word model (i.e., every OOV word is mapped to the generic word model) which allows arbitrary phoneme sequences during recognition. In our word lexicon and word-based language model, we use *UNK* tag for the generic word model. N-gram word language model treats *UNK* just like any other word in the lexicon. During decoding, at the end of every word hypothesis, we allow transitions into generic word model *UNK*, and within the generic word model *UNK*, the recognizer switches to the monophone language model and considers the phoneme set as the active lexicon. The module output is then used in the following algorithms.

2.2.1. Dynamic Fusion (DF)

In this method, we assign dynamic weights to phoneme-based and word-based retrieval scores for each utterance. First, we calculate the OOV detection/misrecognition probability (e.g., probability of having OOV-word or misrecognition in the utterance), and use this probability to decide when/how much to employ subword-based retrieval in addition to word based retrieval. For example, when we search for OOV words in a given set of utterances/documents, we employ subword based retrieval more in utterances where it is more likely to have an OOV word. To calculate the probability of having OOV-word or misrecognition in a given utterance/document, we take the ratio of the number of occurences of the generic word *UNK* over the total number of words in the utterance assuming that OOV/misrecognition detection module is performing at a reasonable level.

2.2.2. Hybrid Fusion (HF)

In this approach, we perform retrieval through hybrid recognition lattices generated by OOV detection module. The motivation behind this algorithm can be explained with the following example. When the query word is OOV word, it is not the best approach to search for the query word in a monophone-only lattice since phoneme-based search is not discriminative enough to yield the desired performance. This will increase the recall rate but will have a negative impact on precision. In this method, we search for OOV words in parts of the lattice where the hybrid recognizer generates monophone sequences assuming that those parts of the lattice will correspond to in vocabulary words. In other words, OOV words are being searched in a smaller document space which is assumed to correspond to misrecognized words. For the case of searching in-vocabulary query words, we can use this algorithm in a fusion scheme whenever we want to back-off to subword-based retrieval (e.g., for a small number of returned hits using word-based retrieval).

3. Experimental Results

We evaluated our proposed algorithms in two different tasks. First, we used a proper name retrieval task in the Spanish Broadcast News (Spn-BN) corpus. Second, we used spoken document retrieval task in National Gallery of Spoken Word (NGSW) historical speech archive [1].

3.1. Spanish Broadcast News

In this task, we used Latin-American Spanish (LAS) as the target language, and focused on proper name retrieval within a broadcast news domain. It is important to note that sufficient resources clearly exist for Spanish based ASR development. While other languages (e.g., Dari, Pashto, Somalian, etc.) are possible, we selected Spanish to be able to intentionally limit the available resources to see what performance can be achieved as further data/resources are available. In other words, using a language such as LAS allows us to select the tier level of resources. In our experiments, we intentionally restrict the following resources: lexicon coverage in spoken documents as well as in query words. We denote OOV rate in spoken documents and OOV rate in queries as OOV_{doc} and OOV_{query} respectively.

We trained microphone-speech models (Spn-Mic) and broadcast-news models (Spn-BN) from Latino40 corpus [10] and Spanish Broadcast News speech corpus [10]. We applied a bootstrapping approach to train microphone speech models for Spanish by using English microphone models via a phone mapping during initial alignment, and then iteratively perform alignment and training steps with updated Spanish acoustic models as explained in [8]. Next, Spanish microphone models (Spn-Mic) are used to initially align the Spanish Broadcast News speech corpus, and then train Spanish Broadcast News models (Spn-BN) using 20 hours of speech corpus. Speech corpus used during retrieval experiments was kept seperate.

Different lexicons were created for evaluation purposes: L_{45K}, L_{50K}, L_{51K}. L_{45K} was obtained from Callhome Spanish lexicon [10], and L_{50K} was created with an additional most frequently occuring 5K words from Spn-BN corpus. L_{51K} was created to contain all query words that are used in retrieval experiments. OOV_{doc} and OOV_{query} values for these lexicons are shown in the Table 1. N-gram (N=3) language models at the monophone and word level were trained using Spanish Newswire Text corpus [10] consisting of 5 Million words. Bigram and trigrams occuring less than 4 times are pruned during N-gram counting. Sentences having high OOV rates (in our experiments sentences with more than 40% OOV rate) are also discarded in our language model training to prevent spelling errors, as well as high unigram probability for generic word *UNK*.

	L_{45K}	L_{50K}	L_{51K}
OOV_{doc}	3.80%	1.15%	0.98%
OOV_{query}	100%	100%	0.00%

Table 1: *OOV_{doc} and OOV_{query} for different Spanish lexicons used in proper name retrieval experiments.*

During recognition, we apply single-class MLLR adaptation. We report recognition results in terms of PER (Phone Error Rate), and SSF-WER (Stemmed and Stop-word-filtered WER) as illustrated in Table 2. Rather than WER results, to be consistent with our retrieval engine, here we report SSF-WER since our word-based retrieval engine, MG, removes stop words and applies Porter stemming to the resulting transcripts, which is very common in text retrieval applications. We report recognition performance in terms of oracle performance for different N-best sizes since we perform lattice-based search during retrieval.

		$N=1$	$N=20$	$N=100$	$N=500$
(a)		24.39	22.10	19.64	17.46
(b)	L_{45K}	33.03	30.45	28.72	28.20
(b)	L_{50K}	29.58	26.62	25.23	23.91
(b)	L_{51K}	29.31	25.92	24.82	23.65

Table 2: *Oracle error rates of (a) monophone and (b) word based ASR in Spn-BN corpus.*

Number of Documents	5000
Average Length of Documents	9 seconds
Average # of Words per. Documents	11 words
Number of Queries	100
Average Length of Queries	6 phonemes
Number of Relevant Documents	100
Average Relevant Documents per. Query	1 doc

Table 3: *Description of document and query sets in Spanish BN.*

Table 3 describes the spoken document and query sets used in our evaluation. The test queries were designed to simulate a known item retrieval task. For each query, there is only one document considered relevant for the purposes of this evaluation. While other documents may have some relevance to the query, only the document it was designed to retrieve was scored as a correct retrieval. To reflect the nature of this task, we used Inverse Average Inverse Rank (IAIR) as a performance criteria. One characteristic of the IAIR is that it rewards correct documents near the top more than documents in the middle or towards the end of the rankings:

$$IAIR = \frac{1}{\sum_{i=1} rank_i^{-1}}$$

where $rank_i$ is the rank of document i.

	Baseline	Dynamic Fusion	Hybrid Fusion
L_{45K}	1.69	1.53	1.48
L_{50K}	1.69	1.49	1.45
L_{51K}	1.68	1.48	1.45

Table 4: Inverse average inverse rank (IAIR) for proper name retrieval task within Spn-BN domain for different lexicons.

Proper name retrieval results are shown in Table 4. Proposed methods (DF approach and HF approach) perform better than the baseline system employing fusion approach with fixed weights. Another observation is that for changing lexicon sizes, proposed methods yield more robust and consistent performance than the baseline. In other words, when lexicons with better coverage are used, baseline system peformance does not change much (e.g., L_{45K}). On the other hand, the proposed methods benefit from better lexicon coverages. This is mostly due to the fact that employing lexicons with better coverages does not guarantee less errorful ASR word transcripts, especially when less frequent words are under consideration as in our case where we try to retrieve proper names.

3.2. National Gallery of Spoken Word

Here, we perform experiments to evaluate our proposed retrieval algorithms in an English speech corpus called National Gallery of the Spoken Word (NGSW) containing spoken word collections spanning the 20th century with as much as 60,000 hours of audio archives. Using English Broadcast News corpus (Hub4'96), we trained F-condition specific acoustic models as well as an acoustic model via multi-style training. To initially evaluate recognition performance, 3.8 hours of sample audio data from the past 6 decades in NGSW is used as the test data.

Number of Documents	956
Average Length of Documents	14 seconds
Average # of Words per. Documents	30 words
Number of Queries	25
Average Length of Queries	4.6 words
Number of Relevant Documents	324
Average Relevant Documents per. Query	13 docs

Table 5: *Description of document and query sets in NGSW corpus.*

	$N=1$	$N=20$	$N=100$	$N=500$
PER	29.19	26.12	22.95	21.16
SSF-WER	49.32	47.21	45.47	43.29

Table 6: *Oracle error rates of monophone and word based ASR in NGSW corpus (PER and SSF-WER respectively).*

To ensure sufficient documents from the perspective of IR, the transcript from each recognition segment is treated as a single document. Table 5 describes the spoken document and query sets used in our evaluation. The 25 test queries were designed by an independent human researcher, based on human transcripts of this test collection, and human relevance assessments were made based on the audio content of the corresponding segments. In Table 6 shows recognition performance of the system in terms of PER and SSF-WER for changing N-best sizes. Table 7 shows retrieval results in terms of average precision. It is seen that the proposed algorithms perform slightly better than the baseline system. Performance improvement obtained in this task for DF approach retrieval algorithm is not as substantial as the improvement obtained in Spanish proper name retrieval task. This is partly because the document lengths are longer in this setup making DF approach less effective. On the other hand, HF approach yields absolute performance improvement of 1.4% over baseline system. These results show that the HF approach is more robust to changes in document length than the DF approach is.

	Word-based-only	Baseline	DF	HF
Avg. Prec.	39.42	40.84	40.52	42.27

Table 7: Average Precision values for SDR experiments in NGSW.

4. Discussion and Future Work

In our proper name retrieval task, we focused mostly on OOV word retrieval rather than misrecognized word retrieval. Although different lexicons are employed during recognition, two different OOV_{query} values (either 0 or 100) were used. In the future, we will evaluate the current system within a wider range of OOV_{doc} and OOV_{query} values. In our experiments we have used an OOV word detection module which is a preliminary system. In the future, we will consider using different parameter settings, and observe the effect of OOV detection on the retrieval performance. We will also focus on exploring ways to obtain more homogeneous lattice portions to benefit from Hybrid Fusion. In other words, within each hybrid lattice, the word hypothesis and subword hypothesis should be kept separate. Query expansion method will be integrated into the current experimental setup. Rather than comparing QE with proposed approaches, we will consider the additional improvement obtained using QE on top of current system.

5. Conclusions

In this study, we focused on audio indexing and retrieval having tiered resources (e.g., lexicon coverage, acoustic model accuracy, etc.). We proposed two methods to obtain more robust retrieval performance for systems employing recognition systems operating at changing performance levels due to tiered structure. The first algorithm, Dynamic Fusion (DF), employed a hybrid recognizer to calculate OOV-and-misrecognition-detection probabilities to assign dynamic weights to each subsystem (subword and word based retrieval). In the second algorithm, Hybrid Fusion (HF), we used hybrid lattices, and performed searches through these lattices. Experimental results showed that both DF and HF approaches perform better than the baseline system employing traditonal fusion methods with fixed fusion weights in both proper name retrieval and spoken document retrieval tasks.

6. References

[1] J.H.L. Hansen, et al, "Speechfind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word", IEEE Trans. Speech and Audio Proc., vol. 13, no. 5, Sept. 2005.

[2] M.G. Brown, J.T. Foote, G.J.F. Jones, K.S. Jones, and S.J. Young, "Open-vocabulary speech indexing for voice and video mail retrieval", In Proc. ACM Multimedia '96, pages 307316, Boston, November, 1996.

[3] K. Ng, V.W. Zue, "Subword-based Approaches for Spoken Document Retrieval", Speech Communication, vol. 32, no. 3, pp. 157-186, October 2000.

[4] D.A. James, S.J. Young, "A Fast Lattice-Based Approach to Vocabulary Independent Word spotting", in Proc. IEEE Conf. Acoustics, Speech, and Signal Processing (ICASSP), pp. 1029-1032, Istanbul, Turkey, 2000.

[5] M. Saraclar, R. Sproat, "Lattice-based search for Spoken Utterance Retrieval", in Proc. HLT-NAACL Conference, Boston, 2004.

[6] C. Allauzen, et al., "General Indexation of Weighted Automata - Application to Spoken Utterance Retrieval", in Proc. HLT-NAACL Conference, Boston, 2004.

[7] I.H. Witten, A. Moffat and T.C. Bell, "Managing Gigabytes: Compressing and Indexing Documents and Images", *Morgan Kaufmann Publishing,* Second Edition, 1999

[8] M. Akbacak, J.H.L. Hansen, "Spoken Proper Name Retrieval in Audio Streams for Limited-Resource Languages via Lattice Based Search Using Hybrid Representations", IEEE ICASSP-06, Toulouse, France, 2006.

[9] T.J. Hazen, I. Bazzi, "A comparison and combination of methods for OOV word detection and word confidence scoring", In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, May, 2001.

[10] Linguistic Data Consortium, http://www.ldc.upenn.edu