Operational Assessment of Keyword Search on Oral History

Elizabeth Salesky, Jessica Ray, and Wade Shen
MIT Lincoln Laboratory
Lexington, Massachusetts, USA
{elizabeth.salesky, jessica.ray, swade}@ll.mit.edu

Abstract

This project assesses the resources necessary to make oral history searchable by means of automatic speech recognition (ASR). There are many inherent challenges in applying ASR to conversational speech: smaller training set sizes and varying demographics, among others. We assess the impact of dataset size, word error rate and term-weighted value on human search capability through an information retrieval task on Mechanical Turk. We use English oral history data collected by StoryCorps, a national organization that provides all people with the opportunity to record, share and preserve their stories, and control for a variety of demographics including age, gender, birthplace, and dialect on four different training set sizes. We show comparable search performance using a standard speech recognition system as with hand-transcribed data, which is promising for increased accessibility of conversational speech and oral history archives.

Keywords: Keyword Search, Oral History, Mechanical Turk

1. Introduction

Oral history has been specifically recorded for the purpose of making history available to future generations. While digitizing such archives has recently gained momentum, these recordings remain largely inaccessible without the ability to index and retrieve relevant items. Though some projects, like the Shoah Visual History Foundation (VHF) have the resources to transcribe their collections by hand (Byrne et al., 2004), this is typically not the case. While there have been many recent advances in the performance of automatic speech recognition (ASR) on spontaneous conversational speech both in and out of this domain, it remains to quantify the resources and techniques necessary for effective search performance.

StoryCorps has recorded more than 56,000 interviews with over 90,000 participants to date, made available through the American Folklife Center within the Library of Congress and selectively online (Dave Isay, 2015). Search is currently only available for metadata, making it difficult to find interviews discussing particular topics. We leverage standard HMM/GMM-based ASR systems and keyword search (KWS) to make the content more accessible.

Training usable ASR systems requires transcribed data: hundreds to thousands of hours for English. Most oral history archives are smaller than this by an order of magnitude. We use recordings from the StoryCorps Griot Initiative, which contains 164 hours of audio, and compare results across varying dataset sizes. We compare standard automatic metrics for assessing search performance with actual human search capability as measured on Amazon Mechanical Turk (MTurk) in an information retrieval (IR) task. On the human task, we compare MTurk worker (Turker) performance on generated transcripts against reference transcripts and a metadata-only search condition. We assess accessibility by measuring the number of searches and time taken compared to the bail out point for an average subject.

Conversational speech, and oral history in particular, typically covers a wide range of demographics which can heavily influence the capabilities of ASR. We control for our demographics in data partitioning and describe performance on gender and two dialects, griot and non-griot. A griot [ɡɹiˈo] is a West-African storyteller, responsible for community history. StoryCorps' Griot Initiative aims to "preserve and present with dignity the life stories of African Americans" (StoryCorps, 2013).

The remainder of this paper is organized as follows: the next section reviews prior work on speech-based information retrieval and oral history. Section 3 discusses our dataset and training methodology. Section 4 evaluates our results. The paper concludes with the implications of this work for oral history collections.

2. Previous Work

The goal of spoken term detection (STD) is to detect the presence of a term, a sequence of consecutive words, in spoken language. The performance of these algorithms depends on many variables. Previous work on STD has largely been in the domain of large vocabulary continuous speech recognition (LVCSR) systems applied to large broadcast news, roundtable meetings, or telephone speech corpora, as in the NIST '06 STD Evaluation Task (Fiscus et al., 2007). Subsequent work has expanded the state-of-the-art through vocabulary-independent keyword search (Mamou et al., 2007), (Parada et al., 2009), and low-resource language evaluations (NIS, 2013), (IARPA, 2011), (Shen et al., 2009), with designs to make better use of less data, and minimize the effects of variability in conversational speech.

Primary work on ASR and keyword search for oral history has been done by the MALACH Project. The goal of the MALACH Project has been to dramatically improve access to large multilingual spoken word collections through ASR and IR techniques (Byrne et al., 2004), (Psutka et al., 2010). Work on this project addressed the approaches in acoustic modeling and adaptation for oral history data (Psutka et al., 2003) and the difficulties in language modeling for a domain where most stories recounted are depen-

This work is sponsored by the Office of the Director of National Intelligence (ODNI) under Air Force Contract FA8721-05-C-0002, IARPA-BAA-11-02. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

dent on named-entities. With ~116,000 hours of Holocaust testimonials in the VHF archives and the resources to hand transcribe and annotate ~10,000, their primary concern has not been lack of data but minimizing ASR error and OOV rates on highly inflectional languages (Psutka et al., 2002). Very few oral history collections are of this size, or have the resources to transcribe significant proportions of their collections (Ashenfelder, 2013). Here, we assess IR performance given more typical data resources, and measure training data size necessary for optimal results.

3. Automatic Speech Recognition

For a range of ASR and keyword search performance, we trained four different acoustic and language models on varying amounts of data. Below, we detail the StoryCorps dataset used for training and the methods used.

3.1. StoryCorps Dataset

The StoryCorps dataset consists of cleanly recorded conversations between an interviewer and interviewee, typically friends or family. Recordings are classified as either griot or non-griot. Each recording is typically 30-45 minutes in length, with 164 hours of audio in total. Segment-based transcriptions were provided for each channel in the conversations. A filtering procedure was applied to remove any problematic audio/transcription pairings. Nine hours of audio were found to have transcriptions that did not align with the audio and were thrown out. The remaining 155 hours were partitioned to create development, evaluation, and training sets.

3.2. Data Partitioning

Four training sets $\{T_1, T_{10}, T_{50}, T_{100}\}$, and a development set D_5, and evaluation set E_{50}, were greedily partitioned from the full dataset to optimally cover selected speaker demographics. The training sets contain {1, 10, 50, 100} hours of training data, respectively. The dev set contains five hours and the eval set contains 50 hours.

Five speaker demographic categories were considered during partitioning, weighted by importance to ensure each was best covered. The five categories, in descending order of importance, were gender, dialect, age, birth state of the interviewee, and the state in which the interview was performed. There was an 11:10 ratio of females to males. 60% of the data was marked griot and 40% non-griot. Interviewees spanned ages 16 to 108, with each age range in {16-46, 47-77, 78-108} represented by 23, 43, and 192 interviewees, respectively. Interview sites were spread across the Northeast, Mid-Atlantic, Midwest, and South, and birth states were spread across the country. E_{50} was filled first to ensure best demographic coverage, followed by D_5. For training, T_{100} was filled first, and T_{50} was partitioned from the data in T_{100}. T_{10} was partitioned from T_{50}, and T_1 was partitioned from T_{10}. Thus, $T_1 \subset T_{10} \subset T_{50} \subset T_{100}$. Files for the same speaker were not separated, so speakers are not split across the six training, dev, and eval sets.

3.3. Acoustic and Language Model

Four acoustic model (AM) and language model (LM) pairings $\{M_1, M_{10}, M_{50}, M_{100}\}$ were trained using the four training sets $\{T_1, T_{10}, T_{50}, T_{100}\}$, respectively.

First, training data was resegmented on silence and non-speech regions of at least 0.5s in length. The feature set consisted of 13 perceptual linear prediction coefficients along with first, second, and third order deltas, resulting in a 52-dimensional feature vector. HTK (Young et al., 1997) was used to train hidden Markov models (HMMs) using state-clustered cross-word triphones with 32 Gaussians per state. 300, 1400, 3000, and 5500 state clusters were used for $\{T_1, T_{10}, T_{50}, T_{100}\}$ HMMs, respectively. All models were discriminatively trained with two iterations of the minimum phone error criterion, although gains were not seen in every model.

Trigram and 4-gram language models were generated independently for each training set from the corresponding transcriptions. The MITLM Toolkit was used for LM training (Hsu and Glass, 2008).

3.4. Decoding and Keyword Search

Decoding and keyword search parameters were tuned on D_5 to optimize both the word error rate and term-weighted value metrics (see section 4.1.).

Decoding on D_5 was performed with each of the four models, starting with a trigram LM with a fixed LM weight of 15 and word insertion penalty of 0. Models were individually optimized using N-best list optimization and lattice rescoring with 4-gram LMs. Speaker adaptation using constrained maximum likelihood linear regression (CMLLR) followed by MLLR was applied to the rescored output, and then N-best list optimization and lattice rescoring were applied again. The resulting parameters for each model were used to decode on E_{50}.

Keyword search with each of the four models was optimized on D_5 also. First, word lattices from decoding were indexed offline into a searchable database. Unigrams, bigrams, and trigrams ('records') were indexed to enable faster and more accurate searching (at the expense of database size). Identical records within 0.5s of each other were merged into a single record and their posteriors summed. Score normalization as defined in (Miller et al., 2007) was applied to search results, and only results with normalized scores above 0.5 were returned. For phrases, a word had to start no more than 0.5s after the end of the previous word to be considered. These values gave the highest term-weighted value on D_5 and were used for keyword search on E_{50}.

4. Evaluation

4.1. Automatic Metrics

Two different evaluation metrics were used to judge ASR and keyword search performance. ASR performance was judged using the standard word error rate (WER) metric, and keyword performance was judged using term-weighted value (TWV) as defined in (Fiscus et al., 2007). In our case, decreases in WER corresponded to increases in TWV, but this is not always true, which is why both metrics are considered.

TWV is a weighted average of false alarms and misses used to 'measure the usefulness of a system to a user' (Fiscus et al., 2007). It is measured at a specific posterior threshold, θ, which is a global detection threshold assigned to all hits returned from a system (the system considers all hits above the threshold to be correct detections, and all hits below to be spurious detections). TWV is calculated via the following equation:

$$TWV(\theta) = 1 - average\{P_{miss}(kw, \theta) + \beta \cdot P_{FA}(kw, \theta)\} \quad (1)$$

where β is a weight that takes the prior probability of a term and the relative weights of misses and false alarms into account. We use the same setting as in the IARPA Babel (IARPA, 2011) program, where the cost of a false alarm is one tenth the cost of a miss.

The highest possible TWV is 1, which corresponds to a system with perfect recall and perfect precision. A TWV of 0 signifies a system returns nothing. Negative values for TWV are also possible with a high number of false alarms. The most common realizations of TWV are actual TWV (ATWV) and maximum TWV (MTWV). ATWV is calculated using a given θ, while MTWV is calculated using the optimal θ for a set of returned hits. As Turkers were only presented results above $\theta = 0.5$, we report only ATWV. An ATWV of 0.3 is considered to be the minimum acceptable baseline for useful system performance, as in Babel (IARPA, 2011). Score normalization as previously desribed was applied to optimize per-keyword thresholds before assigning the global detection threshold, θ.

4.2. Results from Automatic Metrics

To evaluate ASR performance, WER was calculated for each model on the full E_{50} set, and also the dialect and gender subsets of E_{50}, as seen in Table 1. For keyword performance, a set of approximately 7k single and multiword keywords were randomly selected from E_{50}. Multiword phrases contained up to four words. English stop words were excluded from the keyword list. ATWV values for each of the models can be seen in Figure 1. As expected, model performance dramatically improves as the amount of training data increases. The next section will evaluate how this correlates with human performance.

Data	Overall	Griot	Non-Griot	Female	Male
M_1	76.7	77.9	74.6	76.1	77.5
M_{10}	56.1	57.2	54.0	57.1	54.9
M_{50}	42.9	44.2	40.4	43.3	42.4
M_{100}	38.4	40.0	35.5	39.0	37.8

Table 1: *Word Error Rate (WER) on* E_{50}

4.3. Mechanical Turk

To assess human search capability for each of our data partitions, we crowd-sourced experiments on Amazon Mechanical Turk. Turkers were presented with an interface capable of either keyword or metadata search or both, and asked to answer content questions. Turkers were paid a competitive rate equating to an hourly minimum wage. Each 'Human Intelligence Task' (HIT) was limited to three questions to not overwhelm Turkers. The first question was a control, allowing the Turker to familiarize themself with the UI, the results of which are not included in our analysis.

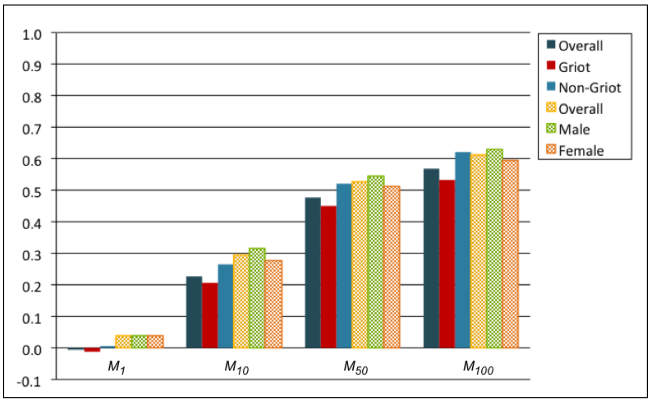

Figure 1: *ATWV for 7k KWs by Demographics, on* E_{50}

We required that Turkers had an approval rate of 95%. We had 271 unique workers complete ~800 HITs. The ratio of approved to rejected HITs was 6:1. HITs were rejected if answered implausibly fast, the answers were incomplete, or they were clearly unrelated to the questions. Each question was answered by at least 5 Turkers to reduce variability, and all questions were asked in each position.

From the UI, participants could listen to the audio files, see metadata, and read full one-best file transcripts when keyword search was available. For systems with KWS, search results returned indexed segments for each matching file containing the query in context, with links to play the audio from that point. We compared KWS results on databases generated from $\{M_1, M_{10}, M_{50}, M_{100}\}$ to M_{Ref}, generated from reference transcripts. We used a system with no search capabilities as our baseline.

We tested all conditions with and without additional metadata search to assess how helpful metadata annotation is to oral history accessibility, with the exception of the baseline; without the aid of keyword or metadata search, it would be unreasonable to ask Turkers to answer content questions from a body of 50 hours of audio.

We measure responses in terms of recall and precision. Here, partial precision corresponds to incomplete answers, for example, finding the correct subject or file but not the answer. Partial recall finds some but not all of the correct subjects or files. One third of questions specifically tested keyword recall, requiring Turkers to find multiple hits: ex) "How many interviewees have degrees in social work?"

4.4. Results from Mechanical Turk

Metadata search has an inconsistent impact on systems. Given additional metadata search, Turkers were in all cases more likely to find an answer, but for $\{M_1, M_{10}, M_{100}\}$, overall accuracy decreased as they needed to sort through more total results. With metadata search capability, Turkers performed 143% more searches, ⅓ of which were metadata searches. The types of keyword searches performed, and accordingly the number of hits returned by the system, also changed. With metadata search, 28% of the top 50 keyword searches made were for named entities, but only 16% without. With worse ASR performance on named entities, the ratio of the number of returned hits for Turker searches with metadata and without is 2:3. As well, without metadata search, performance on M_{50} and M_{100} approaches M_{Ref} accuracy, as can be see in Figure 2, but this is not the case with metadata. Here performance is remarkably similar across $\{M_{10}, M_{50}, M_{100}\}$, and significantly worse than our reference model. Further, Turkers listen to 6.5× more audio segments. While more could be done to improve automatic metrics on all models, it is not clear that this would improve human capability; however, by returning fewer spurious hits, the impact may be felt in decreased search time.

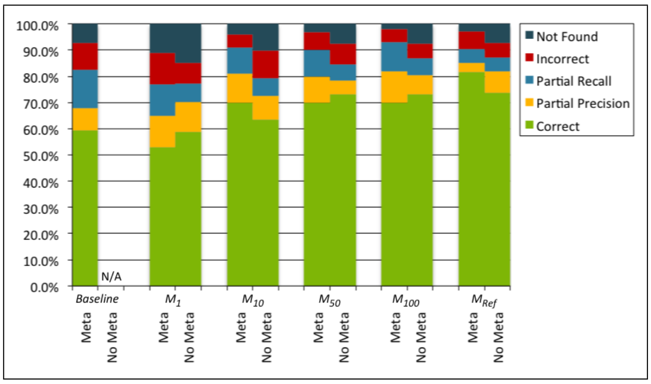

Figure 2: *MTurk Results with & without Metadata Search*

A key difference across systems is the time taken to complete the task. Table 2 shows the average time taken by Turkers to complete the IR task. The first bar in Figure 2 shows that 60% of our questions may be answered by using metadata search to limit the field of interviews, and listening for answers. However, this takes considerably more time than audio search.

Search	Base.	M_1	M_{10}	M_{50}	M_{100}	M_{Ref}
KW only	N/A	9.8	15.5	12.1	14.0	10.5
KW+meta	16.9	17.4	12.9	10.4	15.4	8.8

Table 2: *Avg. Minutes to Complete MTurk Task*

How long are people willing to spend on a task relative to the quality of their results and transcripts? The baseline condition, with metadata search only, required an average 46 searches to complete a task, while M_1 and M_{10} required just half as many, with 23 and 27. With increasingly better transcripts and higher TWV values, M_{50}, M_{100} and the reference case M_{Ref} all required the same number of average searches: 11. These systems return significantly more hits, yet as noted in Table 2, people take more time per search. When given transcripts of higher quality, people are more willing to read the context for answers; as noted through UI logs, the majority of this extra time is spent reading transcripts. In systems with smaller models, Turkers are presented with results of mixed quality and may assume the smaller number of returned hits are the only ones. Further, they may be unwilling to read lower quality transcripts.

These relative task times suggest subjects may bail out faster on lower quality systems. To measure this, we looked only at questions on which Turkers reported not being able to find an answer. Here, they spent on average 1.6× more time performing 2.4× more searches. By dataset, Turkers gave up after {14.6, 20.2, 16.2, 30.6, 18.7} minutes on each of $\{M_1, M_{10}, M_{50}, M_{100}, M_{Ref}\}$, respectively. This is in line with our partial recall statistics in Figure 2 above; when Turkers could locate the file containing the answer, but not the answer itself, they spent on average 1.9× more time doing 3.0× more searches. It may be the case that, given higher quality transcripts and results for other questions, Turkers attribute lack of results to the dataset rather than their inability to find them. It is important to note that it is still possible to nearly match search performance on ground truth before average bail out time, even without extensive metadata access and only 50 hours of in-domain training data for an automatic system.

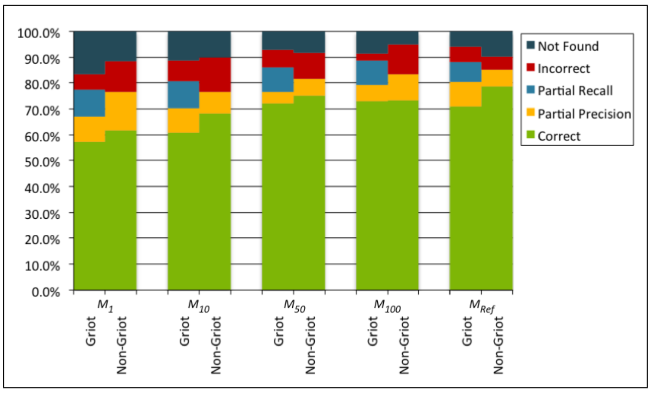

Figure 3: *MTurk Results by Dialect (no metadata search)*

Evaluating by demographics, Figure 4 shows slightly higher ATWV values for males vs. females, which follows from higher WER for female speech. We see in Figure 3 that answers contained in griot interviews have consistently lower accuracy, consistent with the ATWV values in Figure 4. A slightly higher number of griot interviews results in more returned hits from this subset. Access to a larger pool of results explains why people find more answers in the griot data, though with higher rates of partial recall (not all found) and partial precision (file located, but imprecise answer). Surprisingly, demographics do not significantly affect overall human accuracy, which is encouraging for typically diverse oral history data.

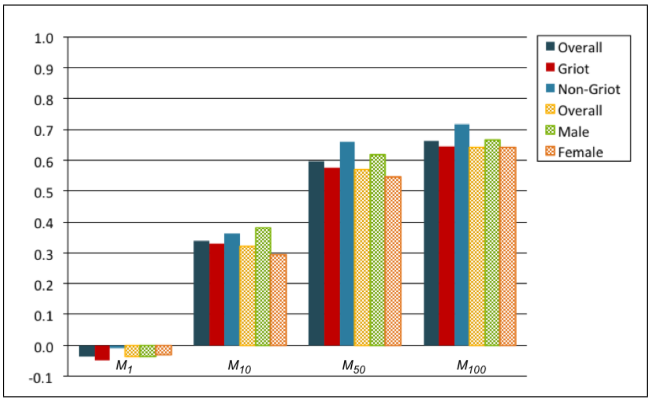

Figure 4: *ATWV for Turker Keywords by Demographic, calculated on* E_{50}

In another set of experiments, we explored using out-of-domain training data for decoding in-domain StoryCorps data. We trained HMM/GMM-based acoustic and language models using both Wall Street Journal (WSJ) and Fisher data. Our best in-domain result with a StoryCorps AM and LM, M_{100}, gives a WER of 38.4 (see Table 1), while with both WSJ models, it is 68.1. When using a WSJ AM with a StoryCorps LM, the WER can be dropped to 55.5, similar to M_{10}. Fisher is still conversational speech, though a different domain, and a Fisher AM and LM yields a much better 49.5, closer to M_{50}. A Fisher AM with an in-domain LM improves WER to 46.4, but surprisingly, a StoryCorps AM and Fisher LM is better still, with 43.0. This may be because of the channel difference; StoryCorps data is microphone, while Fisher is telephone. We would expect that a combined acoustic model may do better, and would like to run a human evaluation with such a model in the future. While these results are not particularly surprising, they suggest that in the likely case of insufficient in-domain training data, reasonable ASR performance can still be achieved by leveraging out-domain resources, particularly if the language model is closer to the target domain. Oral history rarely has significant labeled training data, and so these results, combined with previous sections demonstrating that reasonable ASR performance can still yield favorable information retrieval results, are promising for oral history accessibility.

5. Conclusion

In this paper, we apply automatic speech recognition and spoken term detection to make the StoryCorps oral history archives searchable. We compare word error rate, term-weighted value, and human search capability as measured through Mechanical Turk, across multiple conditions and a variety of demographics. We find that in a human evaluation, the search accuracy seen on ground truth reference transcripts can be approximated using only 50 hours of in-domain training data. Additional metadata access did not significantly impact search capability, though it lengthened search time. Further, when training on publicly available, out-of-domain corpora, we can achieve reasonable performance. This is certainly encouraging for oral history archives, where it is rare to have transcripts or extensive metadata; our results indicate automatic methods are a reasonable approach. We hope these results will promote increased accessibility for oral history in the future.

6. Acknowledgment

The authors would like to thank StoryCorps for the use of their data, and all of our Mechanical Turk workers.

7. Bibliographical References

Ashenfelder, M. (2013). Every voice matters: Storycorps and digital preservation at the library of congress. http://blogs.loc.gov/digitalpreservation/2013/08/every-voice-matters-storycorps-and-digital-preservation-at-the-library-of-congress.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., et al. (2004). Automatic recognition of spontaneous speech for access to multilingual oral history archives. *Speech and Audio Processing, IEEE Transactions on*, 12(4):420–435.

Fiscus, J. G., Ajot, J., Garofolo, J. S., and Doddingtion, G. (2007). Results of the 2006 spoken term detection evaluation. In *Proc. SIGIR*, volume 7, pages 51–57. Citeseer.

Hsu, B.-J. P. and Glass, J. R. (2008). Iterative language model estimation: efficient data structure & algorithms. In *INTERSPEECH*, pages 841–844.

IARPA. (2011). Iarpa babel program - broad agency announcement (baa). http://www.iarpa.gov/Programs/ia/Babel/solicitation_babel.html.

Mamou, J., Ramabhadran, B., and Siohan, O. (2007). Vocabulary independent spoken term detection. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 615–622. ACM.

Miller, D. R., Kleber, M., Kao, C.-L., Kimball, O., Colthurst, T., Lowe, S. A., Schwartz, R. M., and Gish, H. (2007). Rapid and accurate spoken term detection. In *Eighth Annual Conference of the International Speech Communication Association*.

(2013). Nist open keyword search 2013 (openkws13) evaluation. http://www.nist.gov/itl/iad/mig/openkws13.cfm.

Parada, C., Sethy, A., and Ramabhadran, B. (2009). Query-by-example spoken term detection for oov terms. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 404–409. IEEE.

Psutka, J., Ircing, P., Psutka, J. V., Radová, V., Byrne, W. J., Hajič, J., Gustman, S., and Ramabhadran, B. (2002). Automatic transcription of czech language oral history in the malach project: Resources and initial experiments. In *Text, Speech and Dialogue*, pages 253–260. Springer.

Psutka, J., Ircing, P., Psutka, J. V., Radová, V., Byrne, W. J., Hajic, J., Mírovskỳ, J., and Gustman, S. (2003). Large vocabulary asr for spontaneous czech in the malach project. In *INTERSPEECH*.

Psutka, J., Švec, J., Psutka, J. V., Vaněk, J., Pražák, A., and Šmídl, L. (2010). Fast phonetic/lexical searching in the archives of the czech holocaust testimonies: advancing towards the malach project visions. In *Text, Speech and Dialogue*, pages 385–391. Springer.

Shen, W., White, C. M., and Hazen, T. J. (2009). A comparison of query-by-example methods for spoken term detection. Technical report, DTIC Document.

StoryCorps. (2013). Storycorps griot initiative. http://storycorps.org/griot.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (1997). *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge.

8. Language Resource References

Dave Isay. (2015). *StoryCorps collection*. Archive of Folk Culture, American Folklife Center, Library of Congress, Washington, D.C., distributed via StoryCorps, The StoryCorps Collection, ISLRN N/A.