Cross-Domain Matching for Automatic Tag Extraction across Redundant Handwriting and Speech Events

Edward C. Kaiser
Adapx
821 2nd Ave., Suite 1100
Seattle, WA 98104
1-206-428-0732
ed.kaiser@adapx.com

ABSTRACT
In many types of natural human-human interactions people communicate important information redundantly across multiple communication modes, like saying what they handwrite during a presentation or discussion. To detect and benefit from such redundancies a computational understanding system must align the recognition outputs from different perceptual modes like handwriting and speech. Since the recognition domains of each mode differ, researchers refer to tasks like this as *cross-domain matching*. We describe how SHACER (our Speech and HAndwriting reCognizER) currently implements *cross-domain matching*, and compare that to an existing, formally optimal algorithm for this task. Successful alignment and recognition of such multimodal redundancies can be leveraged for automatic tagging of social interactions. These automatically generated tags can benefit retrieval techniques for non-textual documents recorded during computationally perceived social interactions.

Categories and Subject Descriptors
H.5.2 [User Interfaces]: Natural language; Input devices and strategies. I.2.6 [Learning]: Language Acquisition.

General Terms
Algorithms, Human Factors, Languages.

Keywords
Multimodal, Speech, Handwriting. Cross-Domain Matching.

1. INTRODUCTION
High-quality tagging is critical for the success of current retrieval approaches for non-textual data. For example, searching for photo images on the web is currently very dependent on the quality of the textual annotations that are associated with a photo or image. As described in Barthelmess *et al* [6], the problem with tagging non-textual data is that it is tedious. Thus, it is difficult to motivate people to do it. One strategy to overcome this is leveraging common human-human interactions (like the photo sharing discussion shown in Fig. 1) to also serve as a basis for extracting high quality tag annotations [5-7]. Sharing photos with friends and family is a common interaction, which includes multiple communication modalities. The communication is multimodal. Information is presented not only in different modalities (e.g. graphically and orally), but also information across the different modalities is fused and thus understood synergistically. More information is communicated in the integration of modes than can be communicated in any single mode interpreted alone [24, 26].

Presenting information visually (e.g. as text on a screen) while simultaneously presenting the same information orally through speech significantly improves information retention and learning for those viewing multimedia presentations [40]. Terms and

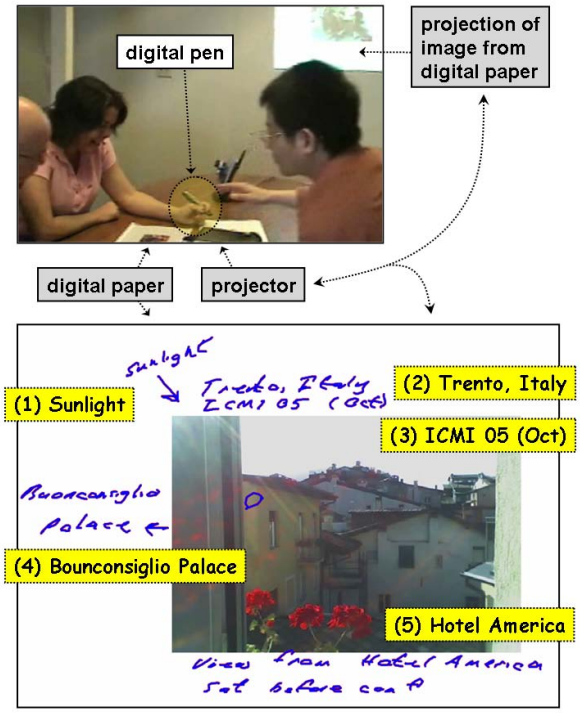

Figure 1: Discussing photos printed on digital paper. As the narrator handwrites annotations they are also projected, so the cross-table participant can view them more easily. Five possible high-value annotation tags are shown. For this photo 93% (all but one) of these handwritten tag words are also spoken.

ICMI '07 Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, November 15, 2007, Nagoya, Japan

information presented redundantly across multiple modes are not only more memorable [4, 28], but also better indexation keys for applying text-based retrieval methods [5, 6], as discussed in the next section.

1.1 Search Techniques and Multimodality

Current online search techniques typically apply to the contents of textual documents. Sometimes they also apply to other non-textual content (like photo images), which have associated textual annotations. Spoken Document Retrieval (SDR) is an emerging search technique where the items to be retrieved are audio-only "documents" — like archived news audio segments [10, 14, 15], or voicemail messages [48]. For audio content, various types of speech recognition technology can be used to automatically transcribe the speech, and these transcriptions can then be used as input to standard text-based search methods.

Aside from audio content, computational perceptual environments are also now capable of recording other recognizable but non-textual input streams. For example tablet PCs can simultaneously record video, audio and inking (e.g. handwriting and sketching) on the tablet surface. This invites the creation of search techniques based on multimodal recognition transcripts. In this regard, distance lecture delivery platforms based on tablet PC technology have been studied [2, 3]. Lectures recorded in this way have multimodal streams that can be analyzed or processed together to produce more accurate transcripts. Our own earlier work has shown that there are significant improvements both in transcription accuracy [25] and retrieval accuracy, based on multimodal integration and subsequent multimodal indexation in the search process [26, 27].

1.2 Cross-Domain Matching as a Basis for Multimodal Tag Processing

In the remainder of this paper we outline the general notion of cross-modal or cross-domain alignment as a basis for processing multimodal streams of information. In particular, we examine how speech and handwriting can be aligned cross modally to support the extraction of useful tag streams. Cross-domain alignment means taking the outputs produced by recognizers in different domains, transforming those outputs into a common description language (like either phonetic sequences or letter sequences), and comparing those descriptions to each other to find alignment matches.

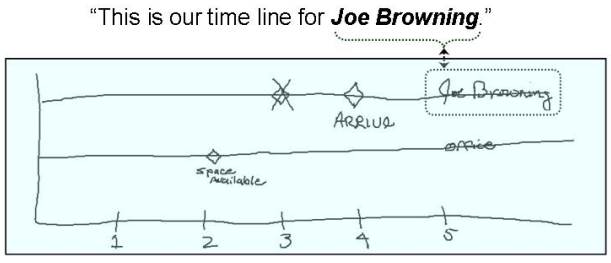

Figure 2: Detection of multimodal redundancy is done by finding closely matched alignments of handwriting and speech. Often times the redundantly spoken words, which accompany a presenter's handwriting, are embedded within a longer utterance. This makes the problem of aligning the two input streams and detecting the redundancy challenging. This example Gantt chart illustrates a handwritten taskline label for a new hire named *Joe Browning*. The accompanying speech is given in quotes.

During multimodal presentations, like lectures or meetings, handwriting that is meant to be public is typically accompanied by redundant speech [27]. Cross modally aligning instances of such redundancies supports a rate of system understanding and recognition that is significantly better than that which is achievable in either mode alone [23, 24, 26]. Our system for processing such alignments to produce multimodal recognitions is called SHACER (pronounced "shaker"), which is an acronym for Speech and HAndwriting reCognizER. SHACER has been tested on sets of recorded meetings in which the participants create Gantt schedule charts (Fig. 2). Handwritten labels in this domain must often be aligned with individual spoken words in a longer accompanying spoken utterance (Fig. 2), which makes the alignment task somewhat similar to word-spotting. SHACER's aim is to unobtrusively observe human-human interactions and leverage multimodal redundancy to learn new words dynamically in context. These new words are often dialogue-critical (like proper names and acronyms), and thus successfully recognizing them amounts to high-value tag extraction from natural human-human interactions [27].

2. CROSS-DOMAIN ALIGNMENT

In order to leverage the occurrence of multimodal redundancy an application like SHACER must first detect it. In order to detect redundancy across different modal domains (like handwriting and speech) the first requirement is a common description language — a language in which the output of both domains can be represented ([12], pgs. 616-17) — like the common phonetic sequence representations shown in Fig. 3. Given a means of transcribing inputs, which originate from different domains and have different error functions (and thus different *transcription costs*), into common language strings, then as Lopresti *et al* [35, 36] have shown, a formal algorithm for returning the optimal

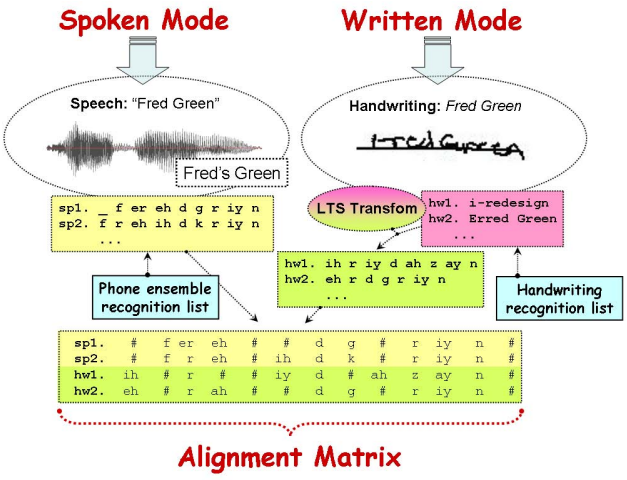

Figure 3. After speech and handwriting streams have been individually recognized, they need to be aligned to check for redundancy. First the handwriting is put through a Letter-To-Sound transform (LTS), which is a transformation of sequences of letters into sequences of phonemes. Then the phone ensemble phone sequences can be aligned with LTS phone sequences.

dynamic programming (DP) distance between them can be described. This distance is called the edit distance or *edit cost*, and whereas DP alignment typically considers only edit costs, Lopresti *et al's* optimal cross-modal approach also includes the *transcription costs* from each differing input domain. Taking both *transcription* and *edit* costs into consideration allows the *edit* score of alignment hypotheses to be weighted by the underlying plausibility of the transcripts being aligned.

For speech and handwriting, SHACER uses both phoneme sequences and letter sequences as common description languages. Fig. 3 illustrates the use phonemes as a common language. The output from a handwriting recognizer (in the form of letter sequences) can be transformed to phoneme sequences using a letter-to-sound (LTS) conversion, and the output from a speech recognizer (in the form of words with underlying phonetic pronunciations) can readily be converted to letter sequences via a statistical sound-to-letter transformation [9]. Often the redundantly spoken words that accompany a handwriting event are embedded in a longer utterance, as shown in Fig. 2. SHACER's approach to detecting redundancy is to align the handwriting and speech recognition outputs (via both common description languages — phoneme sequences and letter strings), identify closely matching segments, and then confirm alignments by extracting word sequences over the temporal bounds defined by the alignment match from deep inside the speech transcription recognizer's output lattice (for more detail see [26]).

2.1 Related Issues in Spoken Document Retrieval

SHACER's alignment-based detection of multimodal redundancy is closely related to the problem of locating query words in a database of documents during Spoken Document Retrieval (SDR). *Documents* in Information Retrieval are objects or computer files, which may have various formats. Text, images, videos, multimedia presentations, and audio recordings can all be considered *documents* in the IR sense of the word [33]. Audio recordings are referred to as Spoken Documents. Sometimes individual utterances within a longer recording are themselves considered as spoken *documents*. In that case the SDR task is to retrieve all the individual spoken utterances in which a single-word search query has occurred and is relevant. This is analogous to what SHACER must accomplish in detecting the location of a spoken redundancy that accompanies a handwriting event. For example, the handwritten *Joe Browning*, in Fig. 2, is effectively a query term. SHACER's alignment of a handwritten term (like *Joe Browning*) with its redundant spoken occurrence in a temporally nearby utterance is parallel to finding a typed-in query word in a database of possible audio documents during SDR.

In SDR, audio input is transformed to a common description language of either word or phone-level representations, which is commonly referred to as a transcription. The individual terms in these transcriptions are then organized into memory structure lists with document-occurrence counts, in a process called indexation. Leath [33] and Saraclar and Sproat [48] both offer comprehensive reviews of current SDR techniques, practices and aims. For spoken documents, which are transcribed at the word-level, indexing and retrieval can basically be implemented within the standard search paradigm, because typed queries are already in the common word-level representation language. Thus audio documents, represented by their automatically recognized transcriptions, can be retrieved by standard query-based web searches. This is the approach taken by the National Science Foundation's *National Gallery of the Spoken Word* project, which uses SpeechFind [15, 58] as an experimental audio index and search engine to make historically significant voice recordings freely available and accessible on the web [33]. Many other systems take this traditional transcription-based approach — like the InforMedia, SpeechBot, THISL, and NPR Online projects [33], as well as some commercial systems, like Nuance's Dragon AudioMining [43] and Virage's AudioLogger [48].

2.1.1 Sub-Word Units in Spoken Document Retrieval

Moreau *et al.* [37] and other SDR researchers [22, 53, 57] point out that the disadvantage of using such traditional text retrieval approaches is that the search vocabulary must be known *a priori*. Out-of-vocabulary (OOV) terms, like important named entities, cannot be recognized by the system. This degrades retrieval performance. Also, the derivation of the complex language models required by the traditional transcription-based approach requires huge amounts of training data, which for constantly growing and changing audio archives may simply be prohibitively expensive to acquire and annotate. Representing audio documents as sequences of sub-word units (like phonemes) avoids this problem. Sub-word unit recognizers can be dramatically smaller than the large vocabulary continuous speech recognizers (LVCSRs) used in the traditional approach. Sub-word unit recognition is independent of vocabulary. Schone *et al.* have even proposed using sub-word unit recognition, because of its vocabulary independence, as a means of searching telephone conversations in any of the worlds languages [50]. They use 119 phones to represent all the phonemes in all of the world's languages. Typed queries are then passed through a letter-to-sound transformer which represents them in a language-appropriate common sub-word unit description language.

The disadvantage of sub-word unit indexing is that phone recognizers typically have much higher error rates than LVCSR systems. For indexing longer documents in large collections this can introduce a lose of discriminatory power, but for collections of utterance length audio documents — like voice mail or teleconference collections — this is less of a problem. For such short-document databases using a vocabulary independent phone recognition system is judged by some researchers in the field to be a very reasonable approach [37].

Hybrid systems that combine word-based and phone-based recognition along with lattice-based indexation and processing are very promising [48] and have been shown to achieve better retrieval results than using either word-based or phone-based systems alone [1, 54, 57].

2.2 SHACER's Alignment Approach

In SHACER handwritten words are not only likely to be OOV proper names or acronyms [27], but because they are OOV they are also likely to be mis-recognized by both handwriting and LVCSR speech recognizers. This compounds SHACER's alignment problem, just as it compounds the accuracy of queries in SDR. Thus, relying on word-level recognition alone to provide the necessary cues for detecting redundancy will not work for SHACER.

SHACER uses a hybrid approach, which combines sub-word unit recognition with word-based recognition, by aligning redundant handwriting and speech in a process that is similar to cutting-edge hybrid SDR systems. SHACER's LVCSR word-level recognizer is run in parallel with an ensemble of phone recognizers. Currently this is a relatively slow, off-line process. However, in the future, as SHACER continues to develop, it could take advantage of better, much faster phone-level recognition. For example in recent research on keyword spotting in informal speech documents, phonetic processing speeds of 0.02 x real-time [53] have been reported.

For optimal cross-domain alignment, as outlined in the algorithm given by Lopresti *et al.* [35], both *transcription* and *edit* costs need to be known. For SHACER, transcription costs are equivalent to the recognition likelihoods provided by speech and handwriting recognition (along with the accompanying letter-to-sound or sound-to-letter transformation costs), while edit costs are based on a distance metric that defines how close two symbols in the common phonetic sequence description language are to each other ([26], Appendix A).

2.2.1 Transcription Costs in SHACER

SHACER's phone recognition is capable of providing transcription cost approximations at the individual phone level, but transcription costs for individual letter recognitions are not currently available from handwriting recognition. Thus, currently, SHACER approximates both phone and letter transcription costs by extrapolating and averaging word level recognition likelihoods.

Determining phone-level transcription costs is complicated by the fact that SHACER's ensemble of phone recognizers uses both phonemes and syllables as sub-word units. Both SDR and name recognition approaches have shown that better phone-level recognition can be achieved by using syllables rather than individual phones as sub-word units [52, 57]. The transformation from syllables to phone sequences within SHACER is trivial because syllables are named by their respective phonetic pronunciation sequences (e.g. cat = "K_AE_T" = "K AE T"); however, recovering individual phoneme likelihoods from syllable-level recognitions is not currently possible within SHACER, so only costs from the individual-phone recognizers within the ensemble are used..

SHACER employs an ensemble approach to phone recognition because it is an effective means of improving underlying phonetic recognition. Recent tests of various phone ensemble configurations, ranging from 5 phonetic recognizers to 1 phonetic recognizer, have been performed. These configurations vary both the number of phonetic recognizers and also the type of phonetic recognizer: type (1) is fast but relatively inaccurate phone recognition (ranging from 6% to 26% phone-level accuracy), and type (2) is slow but more accurate phone recognition (~70% phone-level accuracy). Results show that Gantt-chart label recognition (e.g. Fig. 2) improves in all configurations when type (2) phonetic output is used. Results also show that as expected recognition is faster with fewer phone recognizers, but that label recognition accuracy does not degrade that much relative to the level of speed-up. For the single phone recognizer configuration (with no type (2) recognition), the speed up is on the order of 70% (moving from ~11 x real-time to ~3.3 x real-time), and the corresponding loss in label recognition accuracy is ~15%. For the configuration using a single type (1) and a single type (2) phone recognizer, compared to the original configuration of four type (1) phone recognizers, there was a reduction in processing time from ~11 x real-time to ~3.4 x real-time (a 69% reduction in processing speed), with a loss of label recognition accuracy of about 13%.

2.2.2 Extracting Phonetic Information from Aligned Phone Matrices in SHACER

The reason that using more and differently configured phone recognizers improves accuracy is that the correct phones and phone sequences are more often available in the ensemble alignment matrix. SHACER uses an articulatory-feature based alignment metric, described in the next section, which allows non-identical but similar phones to be aligned with each other. These points are illustrated in Fig. 4, where the frequent differences between the outputs as well as the alignment of non-identical phones are highlighted.

The way that SHACER extracts the phonetic information from an alignment matrix is more complex than just a simple majority

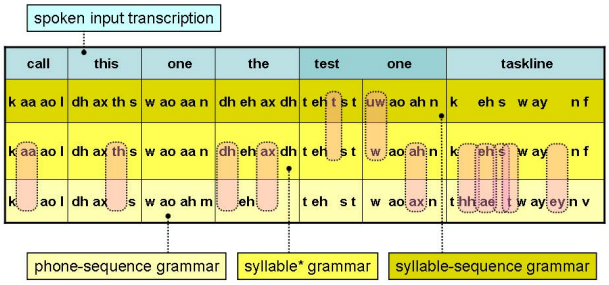

Figure 4: Phone sequence outputs for different ensemble recognizers: (bottom) unconstrained phone-sequence, (middle) unconstrained syllable sequence grammar (the * or *star* means that any syllable can follow any other) , (top) constrained syllable sequence grammar. The differences between outputs are highlighted.

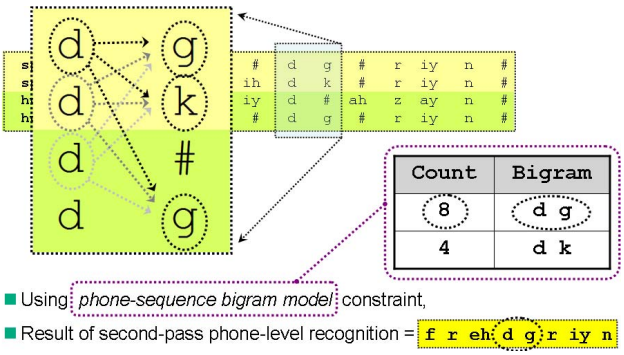

Figure 5: SHACER uses cross-row phone sequence information from the alignment matrix to create a bigram phone sequence model. This model can combine information from both handwriting and speech phone sequences to resolve ambiguous transitions like that shown here from the last phoneme of *Fred* to the first phoneme of *Green* (in the proper name, *Fred Green*). Using the model from the entire matrix to constrain a second-pass phone-level recognition yields the correct pronunciation.

vote at each position. A positional bigram model of phone sequences is extracted from the alignment (Fig. 5). This model is then used to constrain a second-pass, phone-level recognition. Thus, information from the alignment matrix is used like a language model to constrain the likelihood of transitions between terms, which in this case are phonemes. Both the existence of phones in the matrix and their positions relative to each other (both within and across rows) is taken into account by the sequence model. During the second pass phone recognition, information from the alignment-matrix-derived phone sequence model is weighted in relation to the phone-level acoustic scores. This weighting serves to scale the scores of the sequence model in relation to the acoustic model scores, so that scores from one model do not overwhelm the scores from the other model. Thus, this second pass recognition ameliorates the lack of reliable phone-level transcription costs, which currently are unavailable in SHACER.

Another reason for using an ensemble of phone recognizers is that individual phonetic time boundaries must be known. This is critical for SHACER's approach to locating and detecting OOV terms. Using longer sub-word units (like syllables) provides better phone-level recognition accuracy; but, within-syllable phonetic time boundaries are not easily recoverable. There are methods for recovering these phonetic boundaries. For example, Siede *et al.* [51] replace syllable-like units with their equivalent phone sequences by an approach that is similar to techniques for automatic time-alignment of phonemes [16]. For syllable-based phone recognizers within SHACER the within-syllable phone transitions are very roughly estimated by simple interpolation with respect to the syllable start and end times, while for individual phone recognizers the temporal information for phonetic boundaries is fully available. The combined temporal boundary information allows alignment outliers to be discarded, so that more accurate pronunciations of OOV words can be discovered.

2.2.3 Word-Spotting in Cross-Domain Alignment

A sub-task within SDR is to actually word-spot the individual occurrences of query terms in the audio database [1, 17, 57], as opposed to just determining which spoken documents are most relevant to the query. As pointed out above, this word-spotting task is conceptually very close to the task that SHACER must accomplish in aligning handwriting and speech redundancies. Handwriting in SHACER's domain can be considered the equivalent of a query term whose words must be spotted in the surrounding spoken utterances. The end-goal for word-spotting in SDR is to retrieve an audio document or segment for play back in the retrieval interface. The end-goal for word-spotting in SHACER is to dynamically learn the spelling, pronunciation and contextual semantics of a redundantly presented word and enroll it in the system's vocabulary to improve subsequent recognition. There is no dynamic learning of new words involved in SDR.

2.2.3.1 Two Approaches to Word-Spotting

There are two approaches to the word-spotting task within the SDR research community. One approach is the vector space model (VSM), which defines a space in which both documents and queries are represented by vectors. Each vector is composed of term/weight tuples, which can also store positional information. A typical means of assigning the weight or relevance of a term in a document is the *tf-idf (term frequency - inverse document frequency)* weighting scheme [47]. The process of creating the term/weight tables for a given database of documents is called indexing. The VSM approach can used with either transcripts or lattices [48, 57, 59]. If the query keywords are represented as words then the lattices are word-level lattices. If the query keywords are represented as phoneme sequences then the lattices are phone-level lattices. Transforming query keywords to phone sequences is done by a text-to-speech engine; or alternatively, when the query words are spoken, speech recognition automatically provides phonetic pronunciations.

Presently the main current of SDR research is indexing both word and phone lattices together, so that query keywords can then be treated as words when they are in-vocabulary and treated as phone sequences when they are out-of-vocabulary [1, 22, 48, 54]. Retrieval based on VSM searching of word lattices is fast and scalable to large databases. VSM searching of phone lattices is at least an order of magnitude slower [10] than searching word lattices; however, both research systems [55] and commercial systems offer very fast searching based on phone matrices. For example, for a commercial system (e.g. Nexidia [41]) that pre-processes audio to produce searchable phonetic tracks, Cardillo *et al.* [11] report phonetic pre-processing rates of 4 times faster than realtime (4 x realtime) and search rates of 36,000 x realtime (equivalent to searching 10 hours of media recordings in 1 second for significant queries).

The second approach to word-spotting in SDR relies on dynamic programming (DP) matching techniques that don't use vector indexation [19-21]. This approach hypothesizes location slots where query words could exist in the document database, estimates the probability of a slot/query-word match using some sort of probabilistic edit distance measure, and then computes the relevance of a document based on those probabilities [37]. This approach is much slower than VSM techniques due to the computational cost of slot detection and probabilistic distance matching. However, for small databases like finding utterances in which a certain city name was spoken it performs significantly better than VSM approaches, as reported by Moreau *et al.* [38]. It also possible to use this approach to find partial matches to spoken queries, so that users may utter queries that contain extraneous words like “well,” or “let's see” [21]. Another use of this approach is finding repeated words in lectures or other recorded audio [21, 45]. Repeated words in lectures tend to be important, subject-specific words, so this approach could aid in the process of identifying and potentially learning new words, as reported by Park and Glass [46] in recent work at MIT.

2.2.3.2 SHACER’s Approach to Word-Spotting

SHACER uses a dynamic programming matching technique, as opposed to a VSM technique, for word-spotting redundancies across handwriting and speech. Currently SHACER does exhaustive dynamic programming (DP) searches to discover redundancies, but the window of spoken utterances that are examined is relatively small. Currently the five utterances temporally preceding the moment at which the DP search starts are examined. DP matching techniques for small databases, where speed is less of an issue, perform significantly better than vector space modeling techniques [37]. In the future we will experiment with VSM approaches for identifying the particular utterances in

which redundancies are highly likely to be located, and then within those utterances deploy a DP search. SHACER's DP search could potentially be faster by using optimization techniques like those described by Itoh [20, 21].

For discovering repeated spoken words during lectures both Itoh [21] and Park [46] match speech to speech. SDR systems that allow spoken queries, like that of Moreau *et al.* [18, 37, 42] also match speech queries to a spoken database. SHACER's matching task is complicated by having to perform cross-domain matching from handwriting to speech. Work on dynamic programming algorithms specifically for cross-domain matching between handwritten queries and text produced via Optical Character Recognition of scanned documents has been described by Lopresti *et al.* [35, 36]. However, we are not aware of any other system that has implemented cross-domain matching between handwriting and speech as SHACER does. Kurihara *et al.* [32] have developed a system called *SpeechPen* that uses speech recognition to allow note takers to create verbatim transcripts of spoken Japanese instructions or lecture presentations. It allows users to choose from a list dynamically predicted speech recognition alternatives to extend their current note-taking strokes and thus increase the speed of taking verbatim notes. Currently *SpeechPen* does not perform any DP matching between handwriting and speech. Schimke *et al.* [49] have proposed an architecture for collecting time-stamped speech and handwriting, with an aim to integrating them for increased recognition accuracy, but have not to our knowledge reported on an actual implementation.

2.2.4 Summary: SHACER Transcription Costs

In the preceding sections we have shown how SHACER offers various interpretations of the *transcription* costs involved in cross-domain alignment, because for both phoneme transcripts and letter transcripts the low-level sub-unit likelihoods are not typically available from their respective recognizers. This is also the case for the letter-to-sound and sound-to-letter transformations SHACER employs — only word-level transformation costs are exposed, not phoneme or letter-level costs. In the next section we turn to how SHACER handles *edit* costs in the alignment process.

2.3 Gauging Phonetic Distance: Edit Costs

Judging whether a redundancy has occurred across speech and handwriting in SHACER depends on cross-domain alignment. That alignment supports the discovery of other information gleaned from the LVCSR's word lattice using the alignment's temporal boundaries. The alignment also provides a model of phone sequences for constraining second pass phonetic recognition (Fig. 5) as explained earlier.

The first level of edit costs is checking for word-level transcription matches of handwriting letter-string alternatives to terms in either (1) the large vocabulary continuous speech recognizer (LVCSR) transcript or (2) the sound-to-letter transformations of the phone ensemble recognition outputs. If there is an exact match then the redundancy judgment is simplified, and subsequent processing is reduced to exploring alternative pronunciations present in the phone ensemble outputs. If there is no exact word-level transcription matches then the handwriting and speech are phonetically aligned with each other.

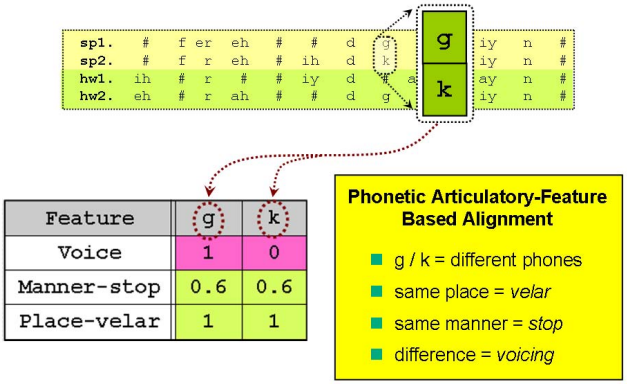

Figure 6: SHACER uses an articulatory-feature based alignment mechanism, which does not insist that phones be spelled the same way in order to match. Thus ***g*** and ***k*** are aligned here, because they are both *velar stops* and differ only in that one is *voiced* while the other is *non-voiced*.

To perform this alignment SHACER uses a phonetic articulatory-feature based edit distance (Fig. 6) as the basis of its alignment technique (based on work by Kondrak [29, 30]). Many researchers in SDR measure phonetic distance by performing speech recognition on a training corpus, and then building a statistical model of the frequency with which one phone is mis-recognized as another phone by the recognizer [1, 21, 39]. The phone-to-phone matrix in which these statistics are stored is called a confusion matrix. The advantage of using a confusion matrix is that it is data driven and recognizer specific. However, the fact that it is recognizer specific is also somewhat of a disadvantage, because if the vocabulary or language model of the recognizer changes then the confusion matrix needs to be recomputed. Since SHACER's goal (as a dynamic learning system) is to be constantly adding to the vocabulary and language model of both speech and handwriting recognizers, then using a recognizer-specific confusion matrix that requires constant re-computation may not be the most effective approach. Kondrak's ALINE approach, based on static articulatory features that are not recognizer specific, out-performs simple Levenshtein edit distance [34] based on the spelling of phone symbols [31], and it also out-performs other articulatory-feature based alignment techniques that use only binary features [8, 44, 56], because of its assignment of saliency weights to the various categories of phonetic features [29, 30]. For example, the manner of articulation (e.g. stop, affricate, fricative, approximate, high/mid/low vowel) of two phones is generally more important in comparing them than considering their respective nasality or roundness, because nasality and roundness are features that only a few phones have. Therefore manner of articulation has a much greater saliency weight.

In recent work Kondrak has highlighted corpus-trained machine-learning approaches to determining phonetic distance, using either paired Hidden Markov Models or Dynamic Bayes Net (DBN) models. Both of these machine-learned distance models out-perform his static, salience-weighted, articulatory-feature based ALINE routine on cognate recognition tasks [31]. However, the drawback of a DBN machine-learning approach, like that of Filali and Bilmes [13], is that it requires a large training corpus. In the future, as larger multimodal handwriting/speech databases

become available, such methods could be tried in SHACER and compared against the performance of its current salience-weighted articulatory-feature based distance measure.

3. DISCUSSION AND CONCLUSIONS

This paper has introduced the potential value of detecting and extracting cross-modal redundancies as a form of automatic tagging. We described a scenario involving the discussion of travel photos, printed on digital paper, annotated in the presence of an audience, that naturally elicits such handwritten and spoken redundancies. In other work we have shown (a) that such tag-events occur naturally as part of many types of human-human interaction, (b) that alignment across different domains is both possible and yields recognition accuracies that are better than what is achievable in either mode alone, (c) that such tag-events serve as natural foci of attention and are thus more memorable than other terms, and (c) that such terms are significantly better index terms for standard search routines.

To accomplish automatic tagging across handwritten and spoken redundancies requires a cross-domain matching algorithm. This paper has compared the approach employed by SHACER to the optimal formal method described by Lopresti *et al* [35], which requires the low-level costs of both *transcriptions* resulting from the recognition processes involved (e.g. handwriting and speech recognition) and *edit-distances* for comparing the common description language sequences generated from the recognition outputs of the different domains.

For SHACER's current implementation, phone-level transcription costs are only partially available and individual letter transcription costs are not available all. Thus SHACER uses various extrapolation and averaging methods to estimate these costs. The transformations of either letters-to-sounds or sounds-to-letters also do not at present yield individual phone or letter transformation costs. However, in all of these cases it is possible to alter the recognition or transformation processes involved to provide these low-level costs, and thus lend support to a formally optimal cross-modal alignment approach. We hope that as multimodal retrieval grows in importance commercial recognizers will expose these needed capabilities.

The articulatory-feature based edit distance measures employed by SHACER already are suitable fom use within an optimal algorithm. However, recent research suggests that corpus-based confusion matrices may ultimately out-perform the salience-weighted, static approach currently employed by SHACER.

There is much work to be done in the area of cross-domain matching. We hope this paper has served to illustrate (1) the potential value in terms of automatic tagging that could result from better cross-domain matching techniques, and (2) the actual feasibility of using such an approach to process natural human-human interactions and make them more available to modern search and retrieval techniques.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

5. REFERENCES

1. Amir, A., A. Efrat, and S. Srinivasan. *Advances in Phonetic Word Spotting*. in *Tenth International Conference on Information and Knowledge Management*. 2001. Atlanta, Georgia.
2. Anderson, R., et al., *Speech, Ink and Slides: The Interaction of Content Channels*, in *Proceedings of the 12th Annual ACM International Conference on Multimedia, ACM Multimedia '04*. 2004. p. 796-803.
3. Anderson, R.J., et al. *A Study of Digital Ink in Lecture Presentation*. in *CHI 2004: The 2004 Conference on Human Factors in Computing Systems*. 2004. Vienna, Austria.
4. Ballard, D.H. and C. Yu. *A Multimodal Learning Interface for Word Acquisition*. in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*. 2003. Hongkong.
5. Barthelmess, P., E. Kaiser, and D. McGee. *Toward Content-aware Multimodal Tagging of Personal Photo Collections*. in *International Conference on Multimodal Interfaces (ICMI '07)*. 2007. Nagoya, Japan.
6. Barthelmess, P., et al. *Collaborative Multimodal Photo Annotation over Digital Paper*. in *Eighth International Conference on Multimodal Interfaces (ICMI'06)*. 2006. Banff, Canada.
7. Barthelmess, P., et al. *Demo: Collaborative Multimodal Photo Annotation over Digital Paper*. in *Eighth International Conference on Multimodal Interfaces (ICMI'06), to appear*. 2006. Banff, Canada.
8. Bates, R., *Speaker Dynamics as a Source of Pronunciation Variability for Continuous Speech Recognition Models*, in *Electrical Engineering*. 2003, University of Washington.
9. Black, A.W. and K.A. Lenzo. *Flite: a small fast run-time synthesis engine*. in *The 4th ISCA Worskop on Speech Synthesis*. 2001. Perthshire, Scotland.
10. Burget, L., et al. *Indexing And Search Methods For Spoken Documents*. in *Proceedings of the Ninth International Conference on Text, Speech and Dialogue, TSD '06*. 2006. Berlin, DE.
11. Cardillo, P.S., M. Clements, and M.S. Miller, *Phonetic Searching vs. LVCSR: How to Find What You Really Want in Audio Archives.* International Journal of Speech Technology, 2004. **5**(1): p. 9-22.
12. Charniak, E. and D. McDermott, *Introduction to Artificial Intelligence*. 1985, Reading, MA.: Addison-Wesley Publishing Company.
13. Filali, K. and J. Bilmes. *A Dynamic Bayesian Framework To Model Context And Memory In Edit Distance Learning: An Application To Pronunciation Classification*. in *Proceedings of the Association for Computational Linguistics (ACL)*. 2005. University of Michigan, Ann Arbor.
14. Garofolo, J., G. Auzanne, and E. Voorhees. *The TREC Spoken Document Retrieval Track: A Success Story*. in *RAIO-2000: ContentBased Multimedia Information Access Conference*. 2000. Paris, France.
15. Hansen, J.H.L., et al. *SPEECHFIND: Spoken Document Retrieval for a National Gallery of the Spoken Word*. in *Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004*. 2004. Espoo, Finland.
16. Hosom, J.-P., *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*, in *Thesis in Computer Science and Engineering*. 2000, Oregon Graduate Institute: Portland, Or. p. 189.
17. Hsieh, Y.-c., et al. *Improved Spoken Document Retrieval With Dynamic Key Term Lexicon And Probabilistic Latent Semantic Analysis (Plsa)*. in *ICASSP*. 2006.
18. Hu, H., et al. *Spoken Query for Web Search and Navigation*. in *Poster Proceedings, Tenth International World-Wide Web Conference*. 2001.
19. Itoh, Y. *A Matching Algorithm Between Arbitrary Sections of Two Speech Data Sets for Speech Retrieval*. in *ICASSP '01*. 2001. Salt Lake City, Utah.

20. Itoh, Y., *Shift Continuous DP: A Fast Matching Algorithm Between Arbitrary Parts Of Two Time-Sequence Data Sets.* Systems and Computers in Japan, 2005. **36**(10): p. 43-53.
21. Itoh, Y., K. Tanaka, and S.-W. Lee, *An Algorithm For Similar Utterance Section Extraction For Managing Spoken Documents.* Multimedia Systems, 2005. **10**(5): p. 432-443.
22. Jones, G.J.F., et al. *Retrieving Spoken Documents by Combining Multiple Index Sources*. in *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1996. Zurich, Switzerland.
23. Kaiser, E.C. *Dynamic New Vocabulary Enrollment through Handwriting and Speech in a Multimodal Scheduling Application*. in *Making Pen-Based Interaction Intelligent and Natural, Papers from the 2004 AAAI Symposium, Technical Report FS-04-06*. 2004. Arlington, VA., USA.
24. Kaiser, E.C. *Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application*. in *Proceedings of the International Conference on Intelligent User Interfaces*. 2005. San Diego, CA.
25. Kaiser, E.C. *Using Redundant Speech and Handwriting for Learning New Vocabulary and Understanding Abbreviations*. in *International Conference on Multimodal Interfaces (ICMI '06)*. 2006. Banff, Canada.
26. Kaiser, E.C., *Leveraging Multimodal Redundancy for Dynamic Learning, with SHACER, a Speech and Handwriting Recognizer*, in *Ph.D. Thesis, Computer Science and Electrical Engineering*. 2007, OHSU: Portland, Oregon. p. 285.
27. Kaiser, E.C., et al. *Multimodal Redundancy Across Handwriting and Speech During Computer Mediated Human-Human Interactions*. in *Conference on Human Factors in Computing Systems, CHI '07*. 2007. San Jose, CA.
28. Karpicke, J. and D.B. Pisoni, *Memory Span and Sequence Learning Using Multimodal Stimulus Patterns: Preliminary Findings in Normal-Hearing Adults*, in *Research on Spoken Language Processing*. 2000, Indiana University.
29. Kondrak, G., *Alignment of Phonetic Sequences*. 1999, Technical report CSRG-402, Department of Computer Science, University of Toronto.
30. Kondrak, G. *A New Algorithm for the Alignment of Phonetic Sequences*. in *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL)*. 2000. Seattle, WA.
31. Kondrak, G. and T. Sherif. *Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification*. in *COLING-ACL*. 2006. Sydney, Australia.
32. Kurihara, K., et al. *Speech Pen: Predictive Handwriting Based on Ambient Multimodal Recognition*. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2006. Montréal, Québec, Canada.
33. Leath, T., *Audient: An Acoustic Search Engine, (http://www.infm.ulst.ac.uk/~ted/exassets/confirmation/Confirmation.pdf, [Viewed March 11, 2007])*. First Year Report, unnumbered, (PhD. program). 2005: University of Ulster.
34. Levenshtein, V.I., *Binary codes capable of correcting spurious insertions and deletions of ones (original in Russian).* Russian Problemy Peredachi Informatsii, 1965. **1**: p. 12-25.
35. Lopresti, D. and G. Wilfong. *Cross-Domain Approximate String Matching*. in *Proceedings of Sixth International Symposium on String Processing and Information Retrieval*. 1999.
36. Lopresti, D. and G. Wilfong. *Cross-domain Searching Using Handwritten Queries, (http://citeseer.ist.psu.edu/520275.html [Viewed March 11, 2007])*. 2000. Amsterdam: Proceedings of Seventh International Workshop on Frontiers in Handwriting Recognition.
37. Moreau, N., S. Jin, and T. Sikora. *Comparison of Different Phone-based Spoken Document Retrieval Methods with Text and Spoken Queries*. in *Interspeech'2005-Eurospeech*. 2005. Lisbon, Portugal.
38. Moreau, N., S. Jin, and T. Sikora. *Comparison of Different Phone-based Spoken Document Retrieval Methods with Text and Spoken Queries*. in *INTERSPEECH*. 2005.
39. Moreau, N., H. Kim, and T. Sikora. *Phonetic Confusion Based Document Expansion for Spoken Document Retrieval*. in *SIGIR*. 2004.
40. Moreno, R. and R.E. Mayer, *Verbal Redundancy in Multimedia Learning: When Reading Helps Listening.* Journal of Educational Psychology, 2002. **94**(1): p. 156-163.
41. Nexidia. *White Papers*. 2006 [cited; Available from: http://www.nexidia.com/technology/whitepapers.html.
42. Ng, K. *Towards Robust Methods For Spoken Document Retrieval*. in *ICSLP '98*. 1998. Sydney, Australia.
43. Nuance. *Dragon AudioMining: http://www.nuance.com/audiomining/sdk/*. [cited.
44. Palmer, D. and M. Ostendorf, *Improving Out-of-Vocabulary Name Resolution.* Computer Speech and Language, 2005. **19**(1): p. 107-128.
45. Park, A. and J.R. Glass. *Towards Unsupervised Pattern Discovery In Speech*. in *Proc. ASRU*. 2005. San Juan, Puerto Rico.
46. Park, A. and J.R. Glass. *Unsupervisedword Acquisition From Speech Using Pattern Discovery*. in *ICASSP '06*. 2006. Toulouse, France.
47. Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information Processing & Management, 1988. **24**(5): p. 513-523.
48. Saraclar, M. and R. Sproat. *Lattice-Based Search for Spoken Utterance Retrieval*. in *HLT/NAACL*. 2004. Boston.
49. Schimke, S., et al. *Integration and Fusion Aspects of Speech and Handwriting Media*. in *in Proceedings of the Ninth International Speech and Computer Conference (SPECOM'2004)*. 2004.
50. Schone, P., et al. *Searching Conversational Telephone Speech in Any of the World's Languages, (http://haircut.jhuapl.edu/publications.html [Viewed January 30, 2007])*. 2005. Mclean, VA: International Conference on Intelligence Analysis.
51. Seide, F., et al. *Vocabulary-Independent Search in Spontaneous Speech*. in *ICASSP '04*. 2004. Montreal, Canada.
52. Sethy, A., S. Narayanan, and S. Parthasarthy. *A Syllable Based Approach for Improved Recognition of Spoken Names*. in *Proceedings of the ISCA Pronunciation Modeling Workshop, (http://www.clsp.jhu.edu/pmla2002/cd/ [Viewed March 11, 2007])*. 2002. Estes Park, CO.
53. Szoke, I., et al. *Comparison of Keyword Spotting Approaches for Informal Continuous Speech*. in *Interspeech'2005 - Eurospeech*. 2005. Lisbon.
54. Yazgan, A. and M. Saraclar. *Hybrid Language Models for Out of Vocabulary Word Detection in Large Vocabulary Conversational Speech Recognition*. in *ICASSP '04*. 2004.
55. Young, S.J., et al. *Acoustic Indexing for Multimedia Retrieval and Browsing*. in *ICASSP '97*. 1997.
56. Yu, C. and D.H. Ballard, *A Computational Model of Embodied Language Learning*. 2003, Computer Science Deptartment, University of Rochester: Rochester, New York.
57. Yu, P., et al., *Vocabulary-Independent Indexing of Spontaneous Speech.* IEEE Transactions on Speech and Audio Processing, 2005. **13**(5): p. 635- 643.
58. Zhou, B. and J.H.L. Hansen. *SpeechFind: an Experimental On-Line Spoken Document Retrieval System for Historical Audio Archives*. in *ICSLP-2002*. 2002. Denver, CO. USA.
59. Zhou, Z., et al. *Towards Spoken Document Retrieval for the Internet: Lattice Indexing For Large Scale Web Search Architectures*. in *Human Language Technology Conference / North American chapter of the Association for Computational Linguistics Annual Meeting*. 2006. New York City.