Indexing Presentations using Multiple Media Streams

A Thesis
Presented to
The Academic Faculty

by

Ravikrishna Ruddarraju

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2006

Indexing Presentations using Multiple Media Streams

Approved by:

Professor Irfan Essa, Committee Chair

Professor Irfan Essa, Adviser

Professor JamesRehg
(College of Computing)

Professor Biing-Hwang Juang

Date Approved ______________________

To my parents,

Satyanarayana Raju, Satyavati Devi

ACKNOWLEDGEMENTS

There are so many that have shaped this endeavor in so many ways, it is impossible to thank all of them. I apologize for any omissions, and I hope you understand you are in my mind though not in this section.

First and foremost I would like to thank my parents for taking the courage to send their 18 year old child to United States. While I pursued my education at Tech, I know they missed having their only child by their side. They taught me to have a forward thinking mind, a humble heart, and the courage it takes to do something different.

It says without going, there are several faculty members that inspired my pursuit of new technologies and science behind them. In his words - ”DSP is in my heart” - Ron Schafer epitomized the importance of having a strong passion. Gregory Abowd, Aaron Bobick, and Christine Mitchell spent a lot of time guiding my research and teaching lessons for life that I will always carry.

Grad school is impossible to navigate without fellow graduate students. First among them is Antonio Haro. He literally introduced me to academic research, and helped me put together my first research paper. Gabriel brostow was another veteran eye-team member who showed me how to cultivate a good team spirit. David Minnen, Yifan Shi, Raffay Hamid, Vivek Kwatra, Nick Diakapoulos, Yushi Jing, Ping Wang, and all others on the 2nd floor of TSRB have been a source of learning and fun. I will forever remember them for keeping me sane through the grad school, and pursuing computational perception along with me.

There are also a few unique people, who followed me as we together pursued wonderful projects. Synedria project team - Robert Rodgers, Jeff Bidzos, Scott McRae, Mike Helman-Darley - was probably the most important one. These guys were the best group of undergraduates I ever worked with. There are so many more undergraduates who worked with me in both Synedria and CALO projects. They taught me something unique early in

my career - art of leading teams. I will remember these early experiences and lessons learnt.

I certainly have to thank my committee - Jim Rehg, Fred Juang, and Irfan Essa. Jim spent a lot of time guiding my early research, and pushed me to explore mathematical basis for my research. Though I did not spend as much time with Fred, he certainly kept me focussed as I worked through this thesis. There is something unique I learnt from Fred - filtering out the research field and focussing on specific problem.

Of course none of this was possible without one person - Irfan. He is my first mentor, my first boss, and my committee chair. I can never thank him enough. Irfan, you gave me so many opportunities and taught me invaluable things early in my career - how to drive home my ideas and how to lead. Most importantly, you taught me the importance of creativity and taking risks in pursuit of my passion. That is something I will always hold on to.

TABLE OF CONTENTS

DEDICATION . . . . **iii**

ACKNOWLEDGEMENTS . . . . **iv**

LIST OF TABLES . . . . **vii**

LIST OF FIGURES . . . . **viii**

SUMMARY . . . . **ix**

I INTRODUCTION **1**

II PREVIOUS WORK **3**

III FRAMEWORK **5**

3.1 Processing slide vocabulary 5

3.1.1 Extracting text from presentations 5

3.1.2 Vocabulary from slides 6

3.2 Processing presenter's audio 8

3.2.1 Audio segmentation 8

3.2.2 Audio transcription 9

IV DISTANCE MEASURES **11**

4.0.3 Euclidean distance 12

4.0.4 KL divergence 12

4.0.5 Exponential distance metrics 14

4.0.6 Summary of metrics 16

V RESULTS **18**

VI CONCLUSION **22**

APPENDIX A — ALGORITHM TO COMPUTE DISTANCE METRIC WITH AN EXAMPLE **23**

REFERENCES **27**

INDEX **29**

VITA **30**

LIST OF TABLES

1 Slide transition detection w.r.t length of audio segment 20

LIST OF FIGURES

1 Interface for browsing indexed presentations. Slides most relevant to current slide are shown on the right. 1

2 Example illustrating process of generating a relevance curve using two streams - words from a slide, and from different audio segments. $WER \sim 20\%$ . . . 6

3 Framework for segmenting presentations into word frequencies 7

4 Overview of vocabulary processing 7

5 Power spectral density of a silent audio segment. Percentage of content above threshold: 3.8% 8

6 Power spectral density of a non-silent audio segment. Percentage of content above threshold: 40.3% 9

7 Relevance score using euclidean distance as a function of word frequencies from different transcripts. Fixed word frequency in the slide = 3 13

8 Relevance score using KL divergence as a function of word frequencies from different transcripts. Fixed word frequency in the slide = 3, total number of words = 30 14

9 Relevance score using normalized exponential distance as a function of word frequencies from different transcripts. Fixed word frequency in the slide = 3 16

10 Relevance curves (colored) for all slides from a presentation. Dashed lines represent slide transition points. Data $WER \sim 20\%$ 19

11 Performance of slide transition detection as a function of word error rates for the exponential distance measure. Curves (colored) represent results from all presentations of the data set. 19

12 Comparing of slide transition detection using different distance measures. Blue: exponential, Pink: KL divergence, Orange: Euclidean. Curves are plotted as a function of word error rates 21

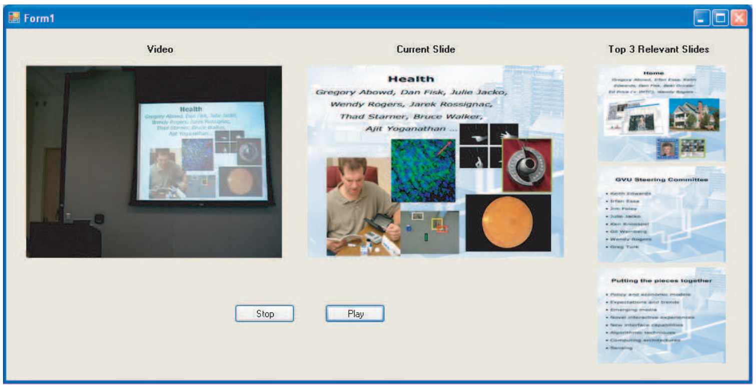
SUMMARY

This thesis presents novel techniques to index multiple media streams in a digitally captured presentation. These media streams are related by the common content in a presentation. We use relevance curves to represent these relationships. These relevance curves are generated by using a mix of text processing techniques and distance measures for sparse vocabularies. These techniques are used to automatically detect slide boundaries in a presentation. Accuracy of detecting these boundaries is evaluated as a function of word error rates.

CHAPTER I

INTRODUCTION

Presentations are rich in different kinds of media content - slides, spoken words, videos and other elements shown by the presenter, interactions with the audience, and the background material. This thesis presents a framework and algorithms for creating an indexing system that takes into account inherent relationships between different modalities - text, audio, and video. These relationships occur naturally in presentations, when the presenter makes references to relevant background material or when he or she introduces upcoming slides or summarizes previous slides. An indexing system that detects and captures these inter-relationships between different media can provide a broader context for querying the information in a presentation. Figure 1 shows an interface developed for this purpose. Additionally this framework is plausible for indexing many forms of digitally captured temporal sequences like meetings, presentations, lectures, and news shows - all of which contain information from multiple modalities like audio, video, and text documents.



Health

Gregory Abowd, Dan Fisk, Julie Jacko, Wendy Rogers, Jarek Rossignac, Thad Starner, Bruce Walker, Ajit Yoganathan ...

Home

Gregory Abowd, Irfan Essa, Keith Edwards, Dan Fisk, Beki Grinter, Ed Price (+ IMTC), Wendy Rogers ...

GVU Steering Committee

- Keith Edwards
- Irfan Essa
- Jim Foley
- Julie Jacko
- Ken Knoespel
- Gil Weinberg
- Wendy Rogers
- Greg Turk

Putting the pieces together

- Policy and economic models
- Expectations and trends
- Emerging media
- Novel interactive experiences
- New interface capabilities
- Algorithmic techniques
- Computing architectures
- Sensing

Figure 1: Interface for browsing indexed presentations. Slides most relevant to current slide are shown on the right.

Indexing presentations is challenging because vocabulary in a presentation slide is considerably smaller than the vocabulary in a regular document. In our framework, we first improve this vocabulary by using text processing techniques like word-stemming, stop-word removal, and adding synonyms and transitive words. We then find common words between this improved vocabulary, and the audio transcript. Since the audio transcript also has a word error rate, the total number of common words between the two vocabularies is further reduced. These small vocabularies, and low word frequencies reduce the effectiveness of traditional text indexing methods like [7] and [9], which need a high-dimensional space for their success. We therefore propose distance measures which can still use these low word frequencies, and assign a score to describe the relevance between a slide and a segment of audio.

Although we assign quantitative relevance between the two data streams, evaluating relevance is very qualitative in nature. Instead, we set up an experiment that uses the relevance computed by our framework to automatically detect slide boundaries on a set of professionally recorded presentations. Performance of these techniques is evaluated over varying word error rates (WER). The ground truth for slide boundaries comes from the presentation video. To measure the accuracy of detected boundaries under different WERs, artificial WERs are introduced by manually transcribing presentation audio. We presents detailed results from normalized exponential distance measures, as it was found to be robust to variation in word frequencies within each slide, and performed better than others. Additionally, we also provide a comparative analysis of two other distance measures.

The rest of this document is organized as follows: chapter 2 gives an overview of previous work, chapter 3 describes the core system framework including vocabulary processing and speech processing, chapter 4 is devoted to a thorough discussion of distance measures, detailed results are presented in chapter 5. Appendix walks through the framework and distance measures using a detailed example.

CHAPTER II

PREVIOUS WORK

Several techniques have been proposed to index content using a single data stream, either audio or text. The SpeechBot project at Cambridge research laboratory [18], and Cambridge university spoken document retrieval system [10] use speech recognition technology to index the audio stream. Different techniques are used to improve the speech recognition - fusion of semantic and acoustic information as described in [12] and [14], and using sub-words for relieving the effect of out of vocabulary words in spoken audio as described in [11]. In our approach for indexing content in multiple streams, we focus on algorithms to find relationships between these streams and not just on improving the accuracy of querying for words.

Latent semantic analysis (LSA) is a well established text indexing method where a large matrix of term-to-document associations is decomposed into a smaller space that reflects key associative patterns in the data [7]. Probabilistic LSA is another such method that builds upon LSA by deriving mixture decomposition from a latent class model [9]. In general, methods based on latent-semantic structure need rich high-dimensional representation for their success [9]. In presentations, the number of keywords per slide and number of slides per presentation is very small. This data sparsity results in a low dimensional space compared to the standard corpora size used in LSA based techniques. Instead of reducing this further by using LSA, we first employ text processing techniques like stop word removal, stemming, and WordNets to improve the vocabulary. Then we generate relevance curves to represent relationships between the audio stream and sparse vocabularies in slides.

Language models for information retrieval have gained popularity due to their simple probabilistic meaning, and efficient performance with large corpora of data [6]. These language models have evolved towards a unified framework proposed by Robertson *et. al* [17] to estimate probability distributions for the document corpora and the search query. Language model based frameworks can also incorporate user feedback to modify probability distributions of documents and queries [5]. While probabilistic techniques are useful in querying large corpora of documents, presentation slides have fewer text to create meaningful probability models. We found that presentations and transcripts have too few word matches to create any meaningful probability distributions. Instead, relevance curves generated using distance measures are able to provide a more intuitive representation for relationships between the audio stream and slides.

Researchers have also presented interfaces to allow user indexing while capturing the presentation. As part of the Classroom 2000 project [4], Pimentel *et. al* [15] presented new hypertext authoring methods that index content by capturing user-interaction with the application's interface during a live session. Hyper-links generated by the interaction can be used to navigate through the captured content. While active user interaction is plausible in a classroom setting, our approach can index multimedia content generated in all general presentations where interaction from the audience is not always forthcoming.

CHAPTER III

FRAMEWORK

The process flow of our framework starts with content from two streams and results in a relevance curve that illustrates relationships between these streams. Figure 2 shows this process using an example. First, vocabulary extracted from $slide9$ is improved using text processing techniques. Second, words common to this improved vocabulary and transcribed audio are put together. As shown in the figure, words from $slide9$ match with three different audio segments. Finally, using word frequencies of matched words, distances between words in $slide9$ and each of the three audio segments is computed. The final relevance curve created using these distances shows different time instances when the presenter referred to the content in $slide9$. The same relevance curve is used to build the interface shown in Figure 1, where the top three relevant slides correspond to three peaks in the relevance curve.

3.1 Processing slide vocabulary

3.1.1 Extracting text from presentations

Figure 3 describes the process of processing a PowerPoint presentation into a set of word frequencies for each slide. Each slide in a presentation can be considered as a document, and the set of slides that form the presentation can be considered as a corpus. Presentation data from PowerPoint slides is extracted into an $NxMx2$ array of word frequencies, where N are the total number of slides in the presentation, M are the number of words in each slide. For more detailed explanation of data formats, please refer to the example described in appendix. Microsoft's Office developer API is first used to split the presentation into slides, then words are extracted from these slides are represented as word frequencies. Word frequency matrices for each slide are combined with similar word frequency matrices obtained from transcript to compute relevance measures.

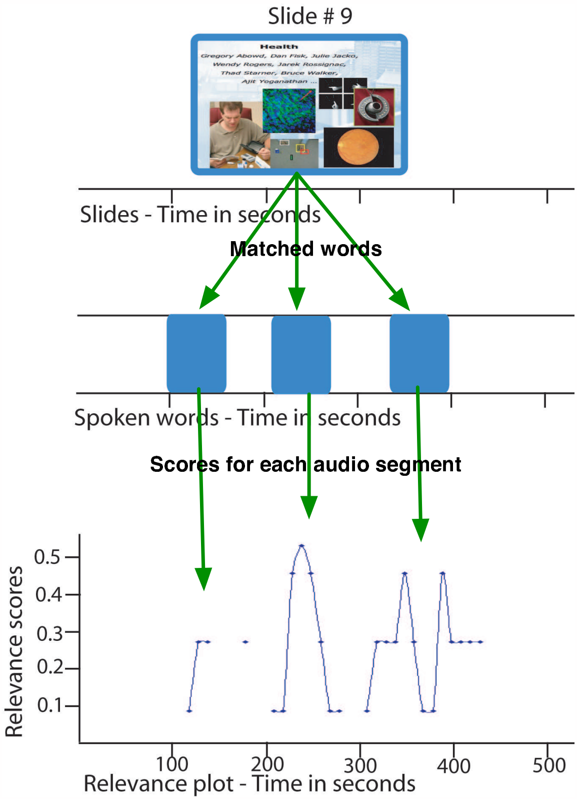

Slide # 9
Health
Gregory Abowd, Dan Fisk, Julie Jacko,
Wendy Rogers, Jarek Rossignac,
Thad Starner, Bruce Walker,
Ajit Yoganathan ...
Slides - Time in seconds
Matched words
Spoken words - Time in seconds
Scores for each audio segment
Relevance scores
0.5
0.4
0.3
0.2
0.1
100
200
300
400
500
Relevance plot - Time in seconds

Figure 2: Example illustrating process of generating a relevance curve using two streams - words from a slide, and from different audio segments. $WER \sim 20\%$

3.1.2 Vocabulary from slides

Word frequencies extracted from each slide are sorted into a set of unique words to create the vocabulary for each slide. A slide vocabulary can contain many words, not all of them relevant to the indexing process. If every term in the vocabulary is considered for indexing purposes, the index will be complete but the percentage of useful entries in the index will be low [3]. New words are added, and some words are filtered out of the vocabulary by balancing its coverage and relevance. For example, words like the, and, or it rarely relate to the meaning and value of a document. Such words can be filtered out from the vocabulary. At the same time, presenters are likely to speak different grammatical forms of words present in slides. Therefore, the quality of vocabulary can be improved by adding such words. Following text processing methods are used to improve this vocabulary: (a) Use of stopwords (e.g. the, in, it) - terms to ignore during text extraction; (b) Minimum term length constraint; (c) Word stemming. Figure 4 shows different steps in this process.

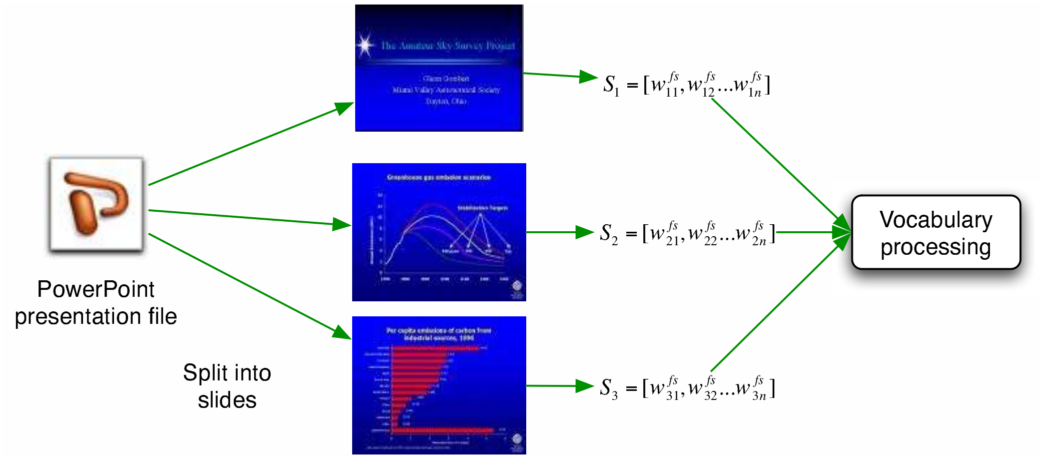

PowerPoint presentation file
Split into slides
The Amateur Sky Survey Project
Greenhouse gas emission scenarios
Per capita emissions of carbon from industrial sources, 1996
$S_1 = [w^{fs}_{11}, w^{fs}_{12} ... w^{fs}_{1n}]$
$S_2 = [w^{fs}_{21}, w^{fs}_{22} ... w^{fs}_{2n}]$
$S_3 = [w^{fs}_{31}, w^{fs}_{32} ... w^{fs}_{3n}]$
Vocabulary processing

Figure 3: Framework for segmenting presentations into word frequencies

First, terms from the slide vocabulary included in the stopword list, and ones that are shorter than the minimum word length are skipped during the text extraction process. Then through stemming, suffixes like "ing", "es" or "er" are stripped to improve search quality [16]. While increasing the likelihood of a successful search, stemming also reduces the size of index by eliminating words that have same prefix but different suffixes. For example, words like "swim", "swimming", "swimmer" will be considered as variations of the same word: "swim".

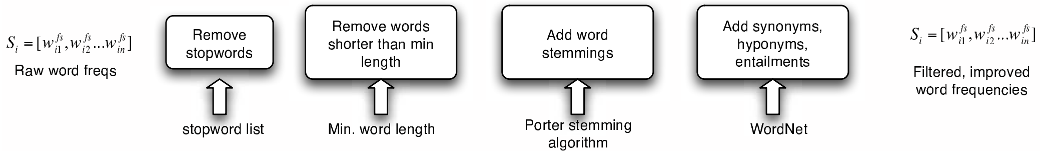

$S_i = [w^{fs}_{i1}, w^{fs}_{i2} ... w^{fs}_{in}]$
Raw word freqs
Remove stopwords
stopword list
Remove words shorter than min length
Min. word length
Add word stemmings
Porter stemming algorithm
Add synonyms, hyponyms, entailments
WordNet
$S_i = [w^{fs}_{i1}, w^{fs}_{i2} ... w^{fs}_{in}]$
Filtered, improved word frequencies

Figure 4: Overview of vocabulary processing

The vocabulary created from the first step is enhanced by adding more words. For example, presenters are likely to use synonyms of words in a slide while describing it's content. They may also use different lexical forms of the same words. Thus, adding these *related* words can potentially improve the quality of index. We use a lexical network called WordNet [13] to find these related words. WordNet can represent several different linguistic relationships like synonymy (similarity), antonymy (opposites), hyponymy (transitives), Meronymy (stemming for nouns), Troponymy (stemming for verbs), and Entailment (relationship between verbs). WordNet also associates words with different grammatical categories - nouns, verbs, adverbs, and adjectives.

3.2 Processing presenter's audio

3.2.1 Audio segmentation

Audio is recorded from a meeting into a single file stream with start-time and duration. To compare slides with different instances of presenter's audio, the audio stream is split into individual sentences. This is accomplished by detecting silent segments in the audio stream. An audio segment can be labeled as silent if it has a very low power spectral density (PSD). Audio recorded in our experiments were sampled at 16kHz. We take 1600 samples at a time, and compute its PSD. Figures 5 and 6 show PSDs of a silent and non-silent segment. In figure 5, only 3.8% of the signal has power above -35dB and thus was marked silent. In case of figure 6, 40.3% of the signal is over -35dB power. In our experiments all segments that have less that 15% of signal over -35dB line as silent. This threshold was conservatively chosen, and reliably detected silent segments in audio from all the presentations.

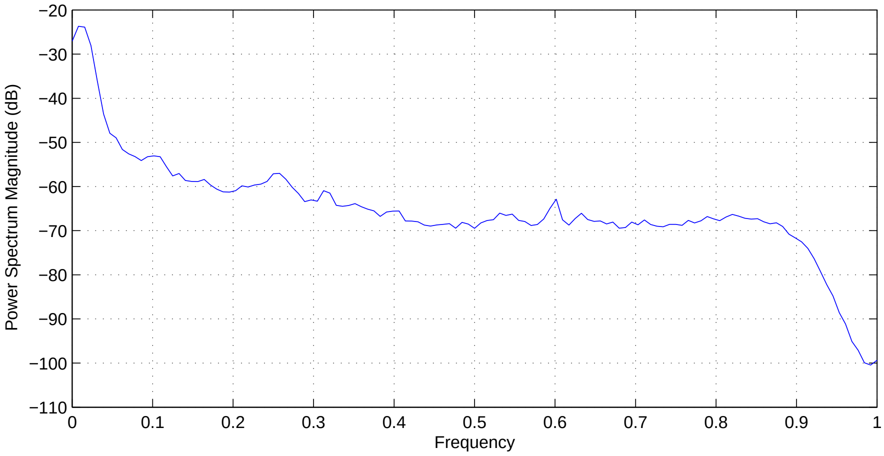


Figure 5: Power spectral density of a silent audio segment. Percentage of content above threshold: 3.8%

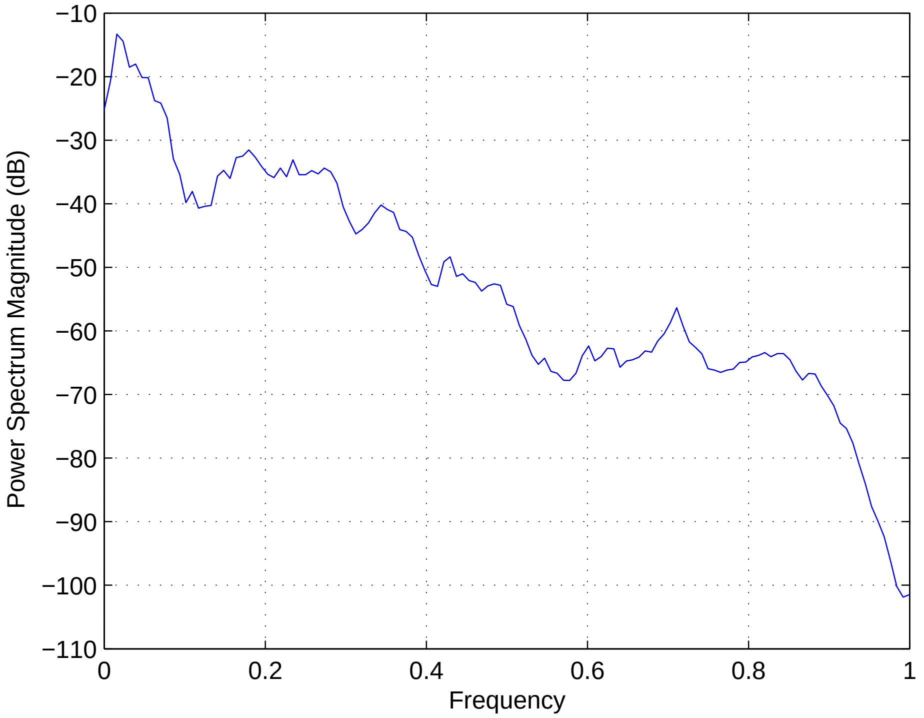


Figure 6: Power spectral density of a non-silent audio segment. Percentage of content above threshold: 40.3%

3.2.2 Audio transcription

Microsoft's speech application programming interface (MSAPI) was used to automatically transcribe presentation audio. MSAPI was chosen over several publicly available open source projects like Sphinx [2] for it's simpler programming interface. When tested under default settings, MSAPI also gave the lowest word error rate (WER) compared to other softwares. Like other softwares, MSAPI can be customized for better performance by changing different acoustic and linguistic parameters. These customizations are specific to application or environment, and cannot be generalized. To maintain a uniform benchmark for the speech recognition, all recorded meetings were transcribed by running the recognition software in it's default mode.

To understand the effectiveness of our approach over different word error rates (WER), we also obtain manual transcripts of presentations. While these transcripts were created by professional transcribers, these transcripts were also not perfect. We inspected these transcripts and found that all of them had WER less than 10%. Each of these manual transcripts are modified to create new erroneous transcripts with a different WER. These erroneous transcripts are used to understand the effectiveness of distance measures under varying word error rates.

CHAPTER IV

DISTANCE MEASURES

After improving the slide vocabulary, and transcribing the audio, distance measures are computed. A distance measure gives a numeric value to represent the relevance between a slide and a segment of transcribed audio. These measures are computed using a 30 second segment of audio stream taken at 5sec intervals. The relevance curve is then obtained by interpolating connected audio segments with non-zero value for their distance measure. The distance measure itself is obtained by first putting together words common to improved slide vocabulary, and audio transcript. Word frequencies of these matched words form the primary data in this problem. Distance measures like squared distance, ratio, dot product, and KL-Divergence were first employed over these word frequencies.

These measures were not able to differentiate between relative occurrence of words in different audio chunks, or they did not comply with mathematical properties of metric spaces [8]. Exponential distance measure was then developed to better represent relationship between transcripts and slide vocabulary. Next three sections give detailed explanation of three metrics - euclidean, KL divergence, and exponential - that were found to have distinct properties for solving our problem.

Once the distance d^M is evaluated for each matched word, the final distance d^M_f for the word is scaled down based on the relative occurrence of the word $f(word)$ with respect to the total number of words T_w in the slide. After performing scaling shown in equation 1, the additive distance is calculated for all matched words in a slide and a segment of transcript.

$$d^M_f = d^M * \frac{f(word)}{T_w} \tag{1}$$

4.0.3 Euclidean distance

The one dimensional euclidean distance between two data points x and y can be defined as follows:

$$d^M = |x - y| \tag{2}$$

In our problem, x and y are word frequency of a matched word in the transcript and the slide respectively. Figure 7 shows the plot of euclidean distance measure using an example from our data set. This plot is obtained for a particular word in a slide, using it's frequency from different transcripts. Since euclidean distance is also the absolute value function, its global minima occurs when $x - y = 0$. This occurs when word frequency from the transcript is same as the word frequency from the slide. This can be a problem because value of zero for distance measure does not have a meaningful explanation. Also, there were several instances when this occurred in our data set.

Additionally, symmetry of this curve causes inconsistencies. Since this function is symmetric about the global minima, several points on the x-axis can lead to the same y-axis value. This means several combinations of word frequencies from slides and transcript can have the same value. This ambiguity is undesirable as it can lead to same relevance score for multiple segments.

4.0.4 KL divergence

Kullback-Leibler divergence (KL divergence) is one of the standard ways to measure similarity between two density functions. Word frequencies in slides and in audio chunk can be interpreted as density functions, and the KL divergence can be used to measure the distance between these. To put it formally, the KL divergence between two discrete probability distributions p and q is defined as:

$$d^M = KL(p, q) = \sum_x p(x) log \frac{p(x)}{q(x)} \tag{3}$$

The ratio $\frac{p(x)}{q(x)}$ in KL-divergence equation is very relevant when comparing two distributions, but can lead to unexpected results in our problem. Consider an example: if user's

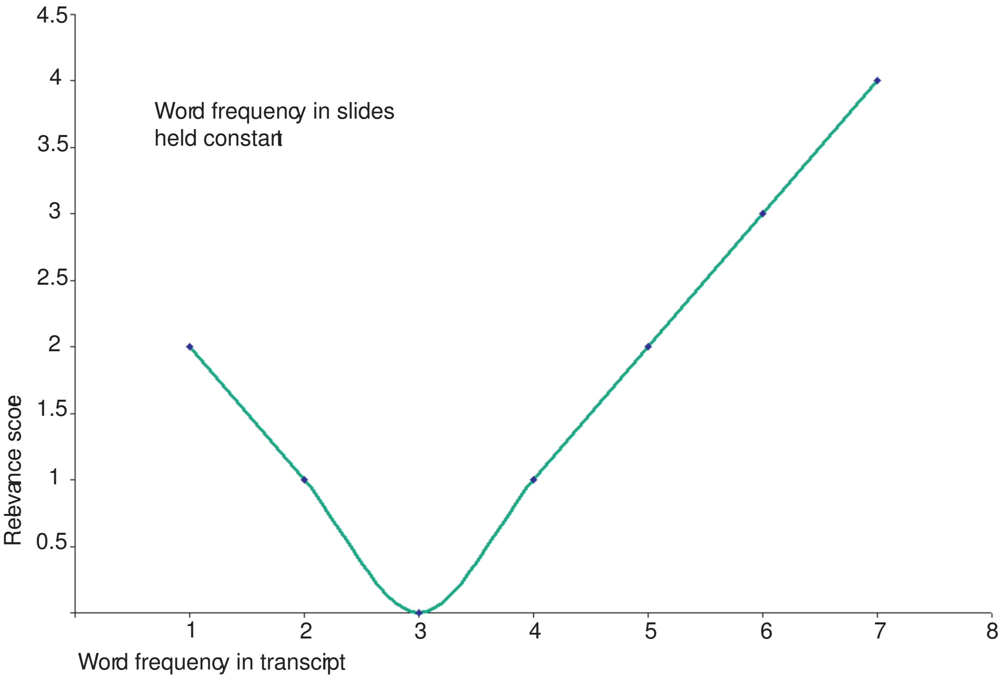


Figure 7: Relevance score using euclidean distance as a function of word frequencies from different transcripts. Fixed word frequency in the slide = 3

audio chunk has very high occurrences of all words in the slide, intuitively we expect to give it a very high score between the chunk and the slide. But, KL divergence only relies upon the relative occurrence of each word in the audio chunk. KL divergence does not distinguish a case with similar relative word frequencies vis-a-vis a case where audio chunk has high word occurrences.

Figure 8 shows an example from our data set. Similar to the plot obtained from euclidean distance, we chose a particular word a slide and plot the variation in KL divergence using it's frequency from different transcripts. Since p and q in KL are relative occurrences of a data point within each distribution, total number of words in the slide and each transcript is also necessary for computing the distance measure. These values are shown in brackets for the transcript.

There are two things to note from this plot. Last three points on the x-axis have higher word frequency than the slide. An ideal distance measure would produce an increasing relevance for the transcript. KL divergence does indeed generate an increasing curve. KL divergence further takes into account relative occurrence of words within the transcript and results in a reducing slope with decreasing relative frequency. This might seem quite desirable. The relative occurrence is important to consider when dealing with large number of matched words. But, in presentations the number of words that match are very low. Scaling illustrated in equation 1 partially takes this into account, and is sufficient for our problem.

Like the euclidean metric, KL divergence also causes problems due to it's symmetry. Like the euclidean distance, KL divergence can give same distance values for two different combinations of slide-transcript word frequencies. In case of KL divergence this point occurs when p and q are equal, i.e. when relative occurrence of a word in both the transcript and the slide are equal.

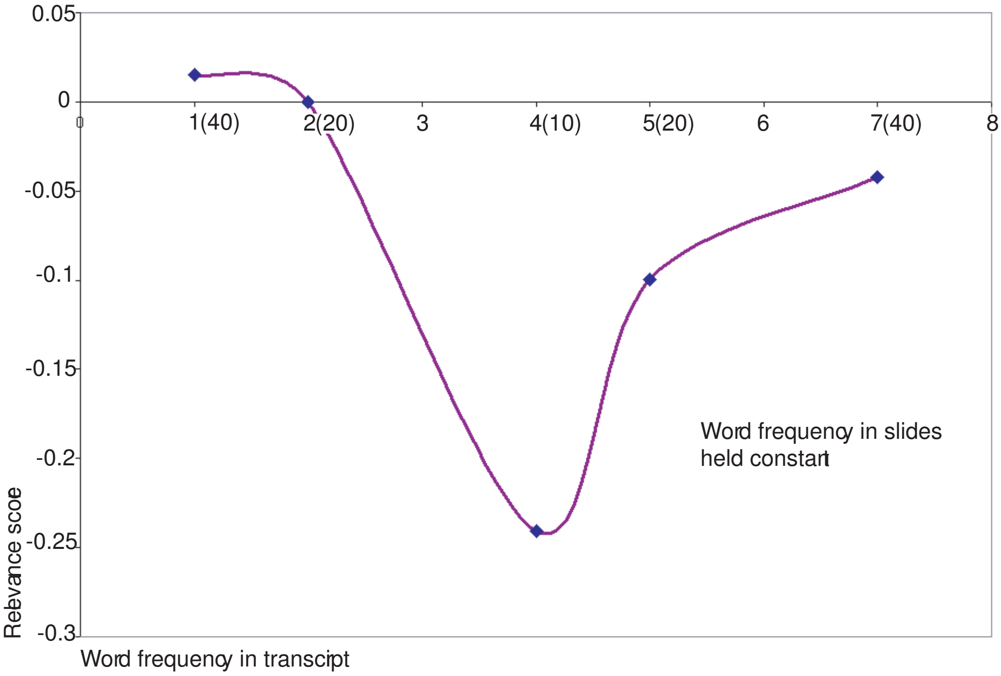


Figure 8: Relevance score using KL divergence as a function of word frequencies from different transcripts. Fixed word frequency in the slide = 3, total number of words = 30

4.0.5 Exponential distance metrics

After analyzing several rudimentary distance measures, we reason that an ideal measure should have the following properties: increase the distance measure of a particular word if that word occurs more frequently in the transcribed audio than it does in the slide, and lower the distance measure if the word occurs less frequently in the transcribed audio than it does in the slide. The optimum match would be when the user spoke the word as many times as it was in the slide. As illustrated with euclidean and KL divergence, an ideal metric must not produce ambiguous values for distance measures. A distance measure with these properties should also possess the standard mathematical properties of a metric space. All these properties can be described by an exponential function shown in equation 4.

$$d^M = \int_0^x \frac{1}{\sqrt{2\pi}\sigma} \exp^{\frac{-(s-\mu)^2}{2\sigma^2}} ds \tag{4}$$

In equation 4, μ is equal to the frequency of a word from a slide's vocabulary, and x is the frequency of the same word from the transcribed speech segment. Clearly, the motivation for this function comes from the cumulative Gaussian probability curve with a given variance.

Figure 9 shows an example from our data set. Similar to the plot obtained from euclidean distance, we chose a particular word a slide and plot the variation in exponential function using it's frequency from different transcripts. The optimum case is when a word occurs equal number of times in both transcribed speech and in a slide. The distance in this case would be 0.5 - peak of Gaussian normal curve. It is the mid point in the figure, and curve is symmetric along y-axis about that point. When a word occurs less number of times in the transcribed speech, it gets a distance less than 0.5 and the point falls to the left of the peak for Gaussian curve. Similarly when a word occurred more number of times in the transcribed speech then the point occurs to the right of the peak and the distance is greater than 0.5.

One thing to take into account with this distance measure are it's asymptotic properties. Equations 5 and 6 illustrate this. At asymptotic points, this measure will gives values that are too close and cannot be distinguished using reasonable bit precision. In the example shown in the figure, any transcript with word frequency > 5 will have too close of a distance measure.

$$\lim_{x \to 0} \int_0^x \frac{1}{\sqrt{2\pi}\sigma} \exp^{\frac{-(s-\mu)^2}{2\sigma^2}} ds = 0 \tag{5}$$

$$\lim_{x \to +\infty} \int_0^x \frac{1}{\sqrt{2\pi}\sigma} \exp^{\frac{-(s-\mu)^2}{2\sigma^2}} ds = 1 \tag{6}$$

Thus, the usefulness or the range for this distance measure will be dictated by it's slope. As shown in equation 7, the slope is inversely proportional to the cube of variance. Thus a large enough variance should suffice. This value can also be chosen easily as total number of words that match in presentation is usually quite low.

$$\frac{d}{dx} \lim_{x \to +\infty} \int_0^x \frac{1}{\sqrt{2\pi}\sigma} \exp^{\frac{-(s-\mu)^2}{2\sigma^2}} ds \propto \frac{1}{\sigma^3} \tag{7}$$

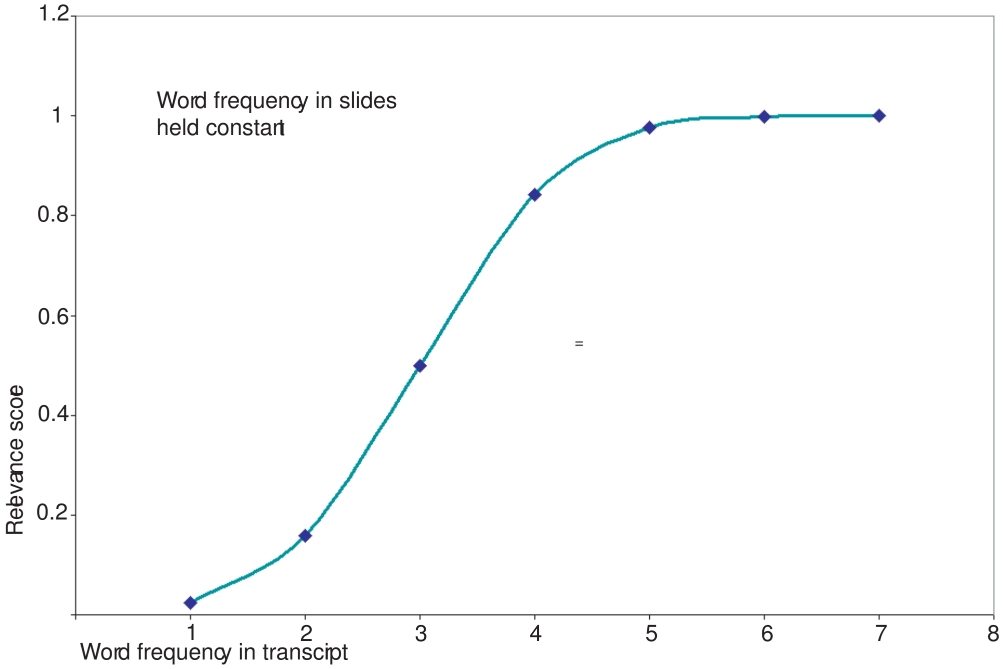


Figure 9: Relevance score using normalized exponential distance as a function of word frequencies from different transcripts. Fixed word frequency in the slide = 3

4.0.6 Summary of metrics

To summarize, an ideal metric should have following properties: (a) It must have a monotonic curve across the range of word frequency values, (b). It should not produce same distance value for two different pairs of word frequency values, (c). The measure must also take into account the optimum match, the case when the user spoke the word as many times as it was in the slide, (d). The metric should not produce a zero value over any range of word frequencies because distance measure of zero does not have a meaningful explanation. As described in preceding sections, well known metrics like euclidean and KL divergence do not have all of these properties. While KL divergence can take into account relative importance of a word within a vocabulary, scaling can also be performed by using much naive methods. Relative importance of a word is not very important in our problem space because re-occurrence of a matched word is very low. Overall the KL divergence loses out because of it's inability to provide a unique value for every slide-transcript word frequency pair. The non-linear exponential function has been fundamental in gaussian probability theory, and it very well has all the properties necessary for a distance measure for solving our problem.

CHAPTER V

RESULTS

The goal of distance measures proposed in this thesis is to quantify relevance between different media streams. But, relevance is a very qualitative term. Instead of evaluating relevance directly, we use relevance curves obtained from interpolating distance measures to automatically detect slide transitions in a presentation. Then, the intersection of relevance curves of any two adjacent slides gives the slide transition point. Knowledge of slide order prevents the need to consider overlaps from other relevant slides the presenter may have referenced. If there are overlaps between adjacent relevance curves, we choose the mid point of the overlap.

We use professionally captured presentations [1] that include the video, slides, and the audio. These are weekly invited talks given by a varied set of computer science professionals. We take 5 of these presentations, each having a unique topic and duration, a varying word density per slide, and different number of slides. To keep the experiment consistent, we only use content from first 10 slides of each presentation for detecting slide transitions.

Audio is transcribed both manually and using automatic transcription. Manual transcripts are modified to simulate different word error rates (WER). These transcripts are used to obtain points on the relevance curves. Each of these points is obtained by computing distance measure for vocabulary in each slide with a 30sec long audio segment taken at 5sec intervals in the audio stream. The relevance curve is then obtained by interpolating connected segments of the transcript with non-zero values for the distance.

Figures 10 and 11 show detailed results based on the exponential distance measure. Later in this section, we compare different distance measures. Figure 10 shows interpolated relevance curves for all slides from a sample presentation. In the ideal case, a monotonously decreasing curve is produced as the presenter is getting to the end of a slide, and a monotonously increasing curve is produced as the presenter begins a new slide. Slide transitions shown as dashed lines are points at which these relevance curves intersect. The ground truth for slide transitions is obtained from the video of the presentation. The absolute error between the detected and observed slide transitions is then computed. Figure 11 shows these errors from all presentations as a function of WER. In table 1, these errors are presented as a function of length of audio segment used for computing distance measure at each point on the relevance curve.

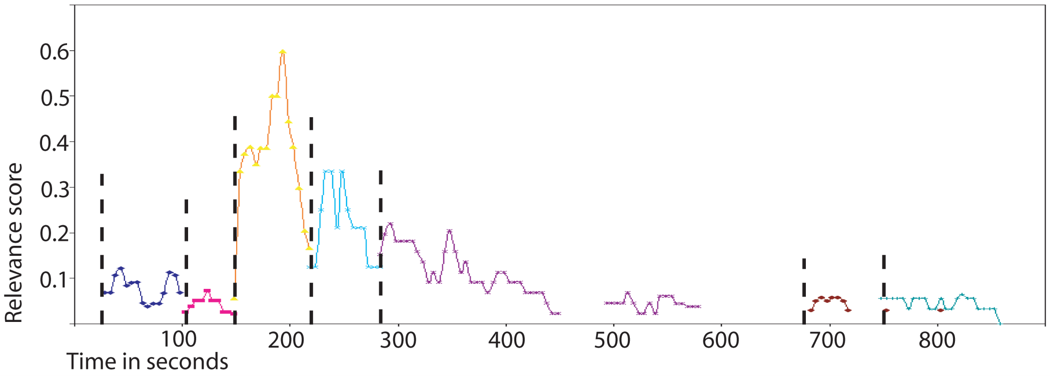


Figure 10: Relevance curves (colored) for all slides from a presentation. Dashed lines represent slide transition points. Data $WER \sim 20\%$

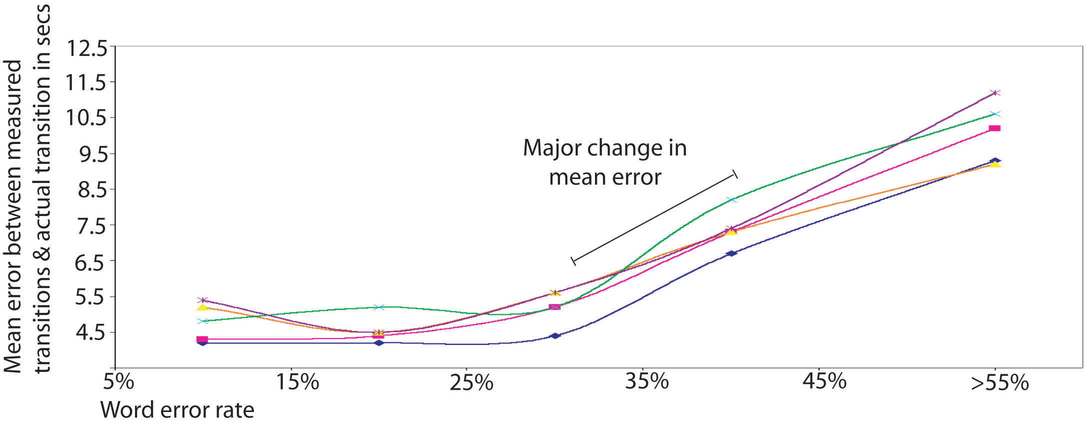


Figure 11: Performance of slide transition detection as a function of word error rates for the exponential distance measure. Curves (colored) represent results from all presentations of the data set.

Several observations can be made from this experiment. In figure 5, intermittent gaps occurred when the presenter was describing an image or a video in the slide. Dominance of images reduces the text in slides needed for computing distance measures. These gaps also occurred when a presenter did not speak any words related to the content in the slides.

Table 1: Slide transition detection w.r.t length of audio segment

Length of audio segment	Error mean	Error std. dev.
20sec	4.25sec	0.58sec
30sec	4.32sec	0.73sec
40sec	7.12sec	2.34sec
60sec	10.25sec	3.53sec

While small gaps in the relevance curve doesn't imply that the presenter is off-topic, a larger gap could imply a digression or that content in slides is not adequately reflecting the talk. As shown in figure 11, presentations with WER under 30% have low detection errors ($< 5sec$). WER over 30% reduces the number of matched words between the slide and the transcript. This in turn adversely effects relevance curves as additive distance measure computed for each audio segment is reduced due to lack of words. Though these errors increase, $10 \sim 12secs$ lag is still usable when detecting slide transitions. Also, WER below 40% can be achieved easily by using a domain specific recognizer. Lastly, from table 1 we can infer that longer audio segments yield higher errors in detected slide transition points. While this is the case, longer segments are in fact useful for presentations with less number of words per slide. In such presentations, longer audio segments can incorporate more words from the audio stream and can compensate for less number of words from slides.

Finally, figure 12 shows error variation between all three distance measures - euclidean, KL divergence, and exponential measure. All three distance measures are used to compute the error between observed and detected slide boundaries from all presentations in the data set. Curves in figure 12 represents the average error for each distance measure over the data set. Several observations can be made from this curve. First, KL divergence mostly follows the exponential measures in it's error. It is separated by a uniform offset through out the range of WERs. This offset arose from ambiguous word frequency pairs when using the KL divergence. Clearly there weren't too many ambiguities to cause higher errors.

Euclidean distance measure on the other hand resulted in significantly larger errors. This happened because euclidean measure evaluates to zero every time there are same number of words in both slide and the transcript for every matched word. Whenever users spoke the word same number of times as it is in the slides, it led to these errors. KL divergence evaluates to zero only when the relative word frequencies in a slide and transcript match up. This is less likely to happen because number of words in a transcript is always more than number of words in the slide.

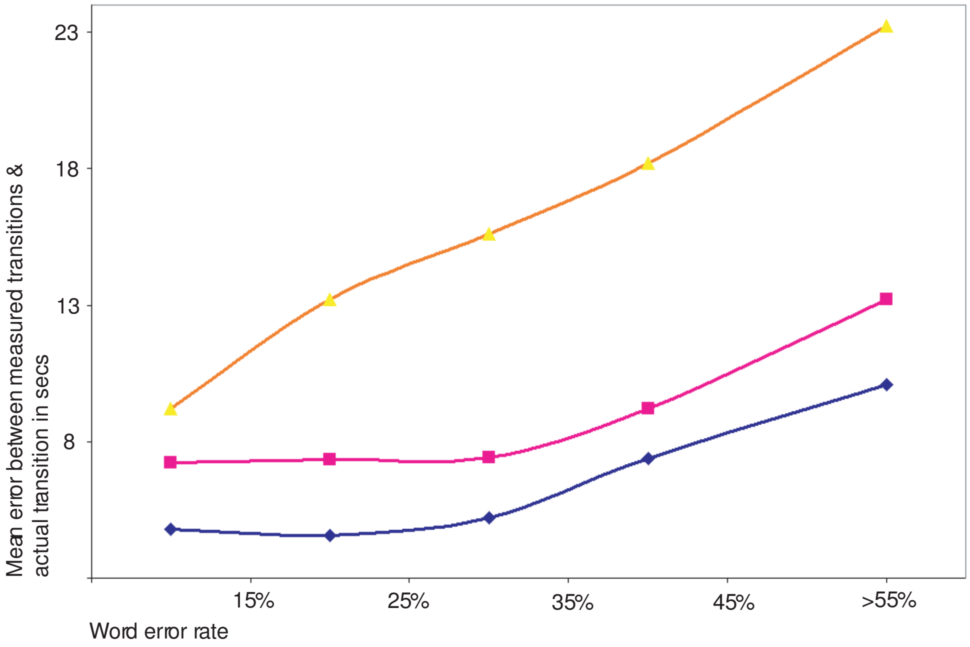


Figure 12: Comparing of slide transition detection using different distance measures. Blue: exponential, Pink: KL divergence, Orange: Euclidean. Curves are plotted as a function of word error rates

CHAPTER VI

CONCLUSION

This thesis presents a framework for creating an indexing system for digitally captured presentations. Using novel distance measures, this framework can take into account natural relationships between different streams of information in presentations. We use relevance curves to represent these relationships. These relevance curves are generated by a mix of text processing techniques and distance measures for sparse vocabularies. Three different distance measures are discussed. We reason that the exponential distance measure had all the necessary properties to give accurate relevance curves.

We demonstrate the ability text processing and distance measures to index multiple media streams in presentations with sparse vocabularies. We apply this approach to automatically detect slide transitions in a presentation. Our experiments with natural presentations show the reliability of our approach to detect slide boundaries under achievable WERs. In future, we plan on incorporating relevant background material - documents, content from the web, or meta data associated with images and videos - to create a more holistic index for the content. We plan on accomplishing this by extending the concept of relevance curves to represent relationships between the presentation and indexable content in the background material.

APPENDIX A

ALGORITHM TO COMPUTE DISTANCE METRIC WITH AN EXAMPLE

A.0.7 Notation

PowerPoint slides with their vocabulary are defined as $S_i = [w_{i1}^{fs}, w_{i2}^{fs}, ..., w_{iN}^{fs}]$, where w_{in}^{fs} is the number of occurrences of a word w_{in} in a slide S_i. Any particular chunk of transcribed speech is defined as $C_i = [w_{i1}^{fc}, w_{i2}^{fc}, ..., w_{iM}^{fc}]$, where w_{im}^{fc} is the number of occurrences of a word w_{1m} in the transcribed speech chunk. d^M is a metric for words that are matched between the transcribed speech chunk and any particular slide. In addition to comparing a single chunk, a group of chunks must also be compared with a PowerPoint slide. The intuition being, several chunks together may still be part of one single slide.

A.0.8 Simple difference and euclidean distance

Lets begin with an example, and define 2 transcribed chunks C_1, C_2 and one slide S_1 to compare them with. Let the vocabulary of a slide be $[parameter, fundamental, color, process]$. Using the frequency of occurrence of these words, we can define $S_1 = [1, 1, 3, 4]$. Let the vocabulary of the input chunk C_1 be $[parameter, fundamental, color, process]$. Using the frequency of occurrence of these words in the input chunk, we can define $C_1 = [1, 1, 3, 1]$. Let the vocabulary of the input chunk C_1 be same as C_2. It has different word frequencies $C_2 = [1, 1, 4, 7]$. Since all words match for both the chunks C_1, C_2, and input slide S_1, the only metric we care about is d^M. d^M, the absolute difference metric between a slide and an input chunk will be

$$d^M = |C_1 - S_1| \tag{8}$$

For slide S_1, and the chunk C_1, the metric would be

$$d^M = |C_1 - S_1| = |[1-1, 1-1, 3-3, 1-4]| = [0, 0, 0, 3] \tag{9}$$

For slide S_1, and the chunk C_2, the metric would be

$$d^M = |C_2 - S_1| = |[1-1, 1-1, 4-3, 5-4]| = [0,0,1,1] \tag{10}$$

If the talk matched what was written on the slides exactly, d^M would have been all zeros. In case of chunk C_1 presenter spoke the word *process* fewer times in his speech. Then the distance for *process* was 3. In case of the chunk C_2, user spoke the word *process* 7 times but the distance is still 3. This discrepancy makes difference metrics unusable for comparing occurrences. Other difference metrics like squared differences, and simple subtraction (allowing negatives numbers), are not any better at describing the distance. Even a ratio of word occurrences between the speech chunk, and PowerPoint slide cannot solve this problem.

A.0.9 Exponential distance

What we need in the metric is the following: increase the distance score of a particular word if that word is spoken more than it is in the PowerPoint slide, and lower the distance score if the word is spoken less times than they it is in the PowerPoint slide. The optimum match would be when user spoke the word as many times as it was in the slide. This behavior can be described by a simple cumulative Gaussian probability curve, with the mean equal to the frequency of word in the slide. Using the example data given in previous section, the word *process* will have following scores:

$$Gaussianparameters : mean = 4, var = 1 \tag{11}$$

$$d^M_{1,1} = cdfofgaussian(x = 1, mean = 4, var = 1) = 0.0013 \tag{12}$$

$$d^M_{1,2} = cdfofgaussian(x = 7, mean = 4, var = 1) = 0.9987 \tag{13}$$

If a word occurs less number of times in the speech than it is in the slide, as it is the case in first example shown above, it gets a lower score. In case of the second example, word occurred more number of times in the audio chunk than in the slide, therefore the score is higher. The additive distance using all words for each slide can be calculated as below: $S_1 = [1, 1, 3, 4]$, $C_1 = [1, 1, 3, 1]$, $C_2 = [1, 1, 4, 7]$ Chunk C_1, word: *parameter*

$$d_{1,1}^M = cdfofgaussian(x = 1, mean = 1, var = 1) = 0.5 \tag{14}$$

Chunk C_1, word: $fundamental$

$$d_{1,1}^M = cdfofgaussian(x = 1, mean = 1, var = 1) = 0.5 \tag{15}$$

Chunk C_1, word: *color*

$$d_{1,1}^M = cdfofgaussian(x = 3, mean = 3, var = 1) = 0.5 \tag{16}$$

Chunk C_1, word: *process*

$$d_{1,1}^M = cdfofgaussian(x = 1, mean = 4, var = 1) = 0.0013 \tag{17}$$

Overall distance score $d_{1,1}$ for chunk $C_1 is[0.5 + 0.5 + 0.5 + 0.0013] = 1.5013$ Chunk C_2, word: *parameter*

$$d_{1,2}^M = cdfofgaussian(x = 1, mean = 1, var = 1) = 0.5 \tag{18}$$

Chunk C_2, word: $fundamental$

$$d_{1,2}^M = cdfofgaussian(x = 1, mean = 1, var = 1) = 0.5 \tag{19}$$

Chunk C_2, word: *color*

$$d_{1,2}^M = cdfofgaussian(x = 4, mean = 3, var = 1) = 0.8413 \tag{20}$$

Chunk C_1, word: *process*

$$d_{1,2}^M = cdfofgaussian(x = 7, mean = 4, var = 1) = 0.9987 \tag{21}$$

Overall distance score $d_{1,2}$ for chunk $C_1 = [0.5 + 0.5 + 0.8413 + 0.9987] = 2.84$ The scores make sense given there were more occurrences of words in the 2nd chunk than in the first chunk. The one issue that comes up here is the choice for variance. The mean sort of makes sense, but what is the intuition behind the variance? Please refer to section 4.0.5 for more detailed mathematical explanations for mean and variance in formulating this metric.

REFERENCES

[1] "Gvu brown bags," *www.cc.gatech.edu/gvu/events/brownbags.*

[2] "Sphinx-4," *http://cmusphinx.sourceforge.net/sphinx4.*

[3] "Establishing a suitable source to search," *Apple developer connection*, 2004.

[4] Abowd, G., "Classroom 2000: An experiment with the instrumentation of a living educational environment," *IBM Systems Journal, Special issue on Pervasive Computing*, vol. 38, no. 4, 1999.

[5] Bodoff, D., "Relevance models to help estimate document and query parameters," *ACM transactions on information systems*, vol. 23, 2004.

[6] Crestani, F., Lalmas, M., Rijbergen, C. V., and Campbell, I., "Is this document relevant?...probably: A survey of probabilistic models in information retrieval," *ACM Computing Surveys*, vol. 30, 1998.

[7] Deerwester, S., Dumais, S., and Harshman, R., "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, 1990.

[8] Etgen, G., Salas, S., and E.Hille, *Calculus: One and several variables.* John Wiley and Sons, 2003.

[9] Hofmann, T., "Probabilistic latent semantic analysis," in *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, (Berkeley, CA), pp. 50–57, Aug. 9990.

[10] Johnson, S., Jourlin, P., and Woodward, P., "The cambridge university spoken document retrieval system," *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, 1999.

[11] Logan, E., Moreno, P., and Deshmukh, O., "Word and sub-word indexing approaches for reducing the effects of oov queries on spoken audio," *Proceedings of Human Language Technology Conference*, Mar. 2002.

[12] Logan, E., Moreno, P., and Prasangsit, P., "Fusion of semantic and acoustic approaches for spoken document retrieval," Tech. Rep. HP200355, Cambridge Research Laboratory, Hewlett Packard Inc., Cambridge, Massachusetts, 2003.

[13] Miller, G. A., "Wordnet: a dictionary browser," *Proceedings of the First International Conference on Information in Data*, 1985.

[14] Ng, K., "Sub-word based approaches for spoken document retrieval," *Ph.D Thesis*, 2000.

[15] Pimentel, M., Abowd, G., and Yoshihide, I., "Linking by interacting: a paradigm for authoring hypertext," *Proceedings of ACM Hypertext*, 2000.

[16] Porter, M., “An algorithm for suffix stripping,” *Program; automated library and information systems*, pp. 130–137, 1980.

[17] Robertson, S., Maron, M., and Cooper, W., “Probability of relevance: a unification of two competing models for document retrieval,” *Information technology - research and development*, 1982.

[18] Thong, J., Moreno, P., and Moores, M., “Speechbot: An experimental speech-based search engine for multimedia content on the web,” *IEEE Transactions on Multimedia*, 2002.

INDEX

VITA

Perry H. Disdainful was born in an insignificant town whose only claim to fame is that it produced such a fine specimen of a researcher.