Search Problems for Speech and Audio Sequences

Dissertation Defense
Eugene Weinstein
Advisor: Prof. Mehryar Mohri
June 9th, 2009
Department of Computer Science
Courant Institute of Mathematical Sciences
New York University

Audio Collections

- Large audio/video collections exist on the Internet
- Many containing music and/or speech
- How can search engines effectively index this content?

Motivation

- Metadata, e.g., obamaspeech.mov is insufficient
  - We want to enable searching the content
  - To accomplish that, transcribe with text-like units
- Issues: data is highly variable, transcription is difficult
- Primary challenge: uncertainty due to e.g., imperfect
  - Statistical models in speech recognizer
  - Music transcription in terms of notes or sounds

Main Results

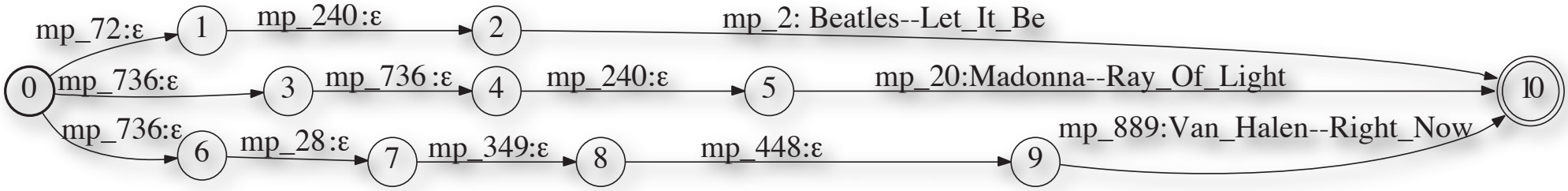
1. Music identification: content-based search for songs
2. Automata: bounds and algorithms for efficient search
3. Topic segmentation: topicality of speech streams

Music ID Overview

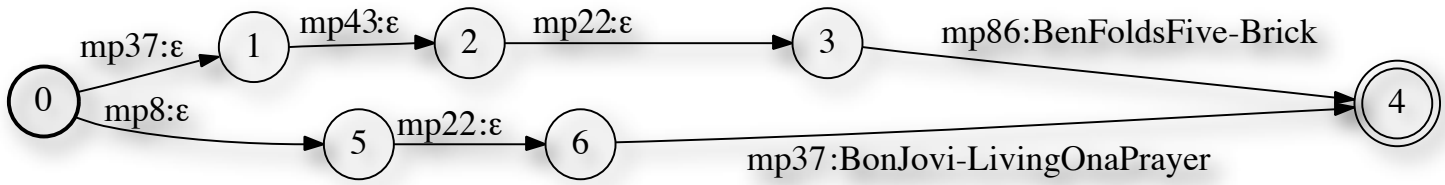
- Music ID scenario: match a few seconds of possibly corrupted or noisy audio to large song database
- Most previous work uses hashing, e.g., [Haitsma et al. '01]
- For a database of 15K+ songs (1,000+ hours of audio), we
  - Automatically learn music phoneme set and a unique phoneme sequence for each song
  - Generate compact mapping from phoneme sequences to songs using weighted finite-state transducers
  - Identify songs using Viterbi decoding

Full Song Recognition

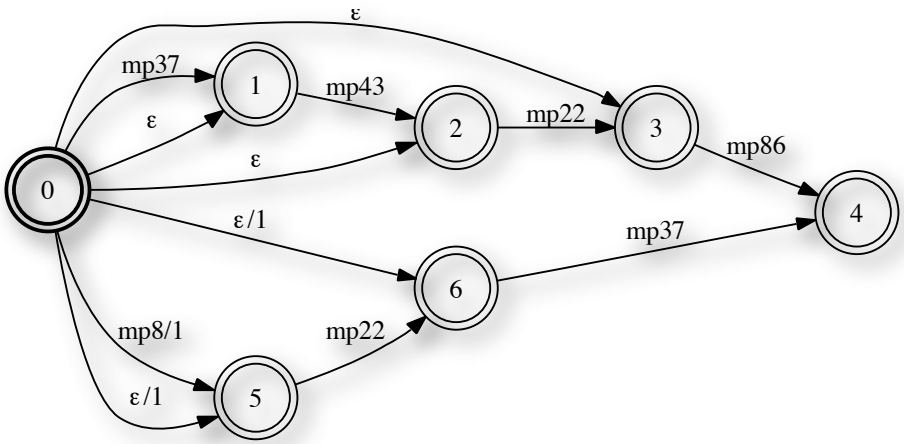
- Want transducer mapping complete music phone sequences to corresponding songs (no snippets for now)
  - Idea: one state chain per song
  - Transition to final state has song identifier as output label (all other output labels are ε's)
  - Using generic automata operations, we construct a deterministic minimal transducer for efficient search

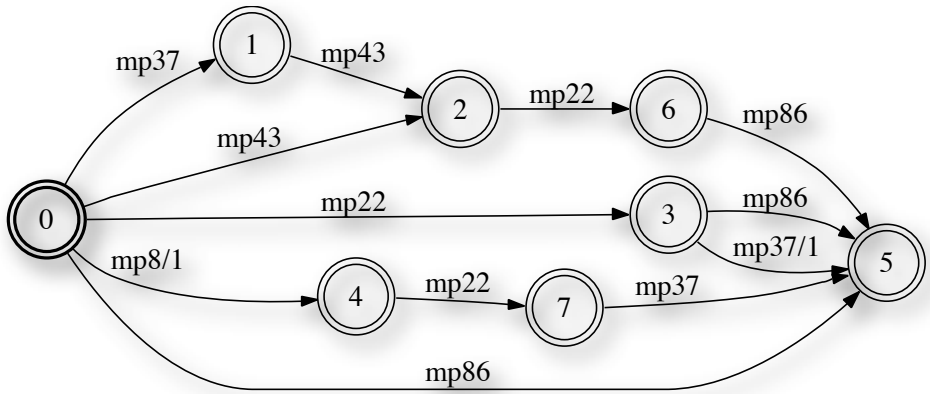


Weighted Factor Acceptor



- Use numerical song id's as weights on transitions
- Add epsilon transitions, make every state final
- Optimize while preserving total path weight [Mohri '97]





Main Results

1. Music identification: content-based search for songs
2. Automata: bounds and algorithms for efficient search
3. Topic segmentation: topicality of speech streams

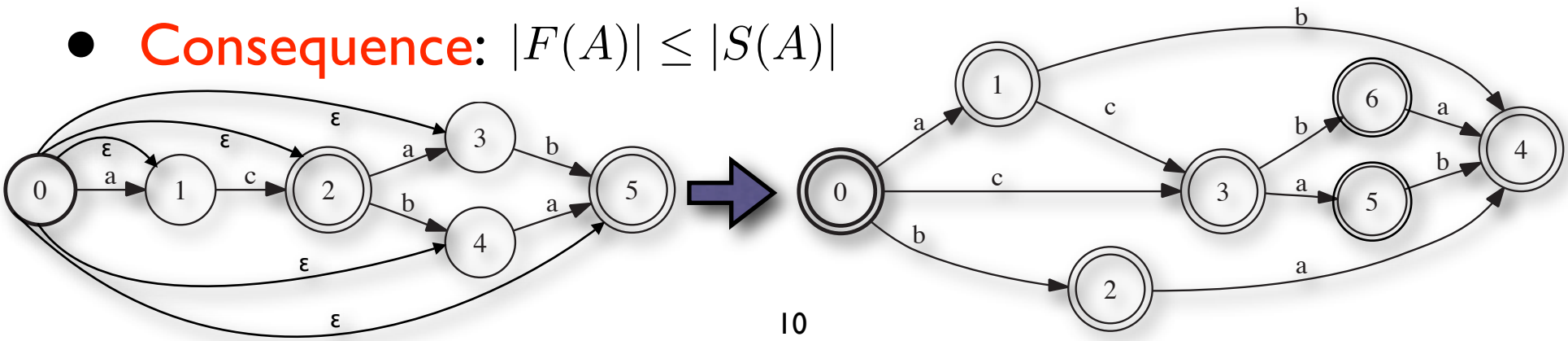
Automata Overview

[CIAA '07, TCS '09, TASLP '09]

- Factor automata enable efficient indexing and search, even when inputs are uncertain
- Music ID: 15,000 songs, 1,700 average phones per song
  - # possible factors = $15{,}000 \times 1{,}700^2 \approx 43 \times 10^9$
- We give new size bounds for the smallest deterministic automaton accepting the factors of a set of strings U
- Or of an automaton A accepting U
- We also give new efficient algorithms for the construction of suffix and factor automata

Suffix & Factor Automata

- We start out with an automaton A recognizing strings in U
- Let $S(A)$ and $F(A)$ be the deterministic minimal automata recognizing the suffixes and factors of A, respectively
- To construct $S(A)$ make each state of A initial (by adding epsilons), determinize, minimize
- To construct $F(A)$ make each state of $S(A)$ final, minimize
- Consequence: $|F(A)| \leq |S(A)|$



Size Results

[CIAA '07, TCS '09]

- Automaton A is k-suffix-unique if no two strings accepted by A share the same k-length suffix. Suffix-unique if $k = 1$
- Theorem: If A is suffix-unique, deterministic and minimal then its suffix and factor automata are bounded in size as

$$|F(A)|_Q \leq |S(A)|_Q \leq 2|A|_Q - 3$$

- Strong improvement vs. [Blumer et al. '87]: $|F(U)|_Q \leq 2||U|| - 3$
- When A is k-suffix-unique, deterministic and minimal, and accepts n strings and A_k is the part of A after removing all suffixes of length k

$$|S(A)|_Q \leq 2|A_k|_Q + 2kn - 3 \qquad |F(A)|_Q \leq 2|A_k|_Q + 2kn - 3$$

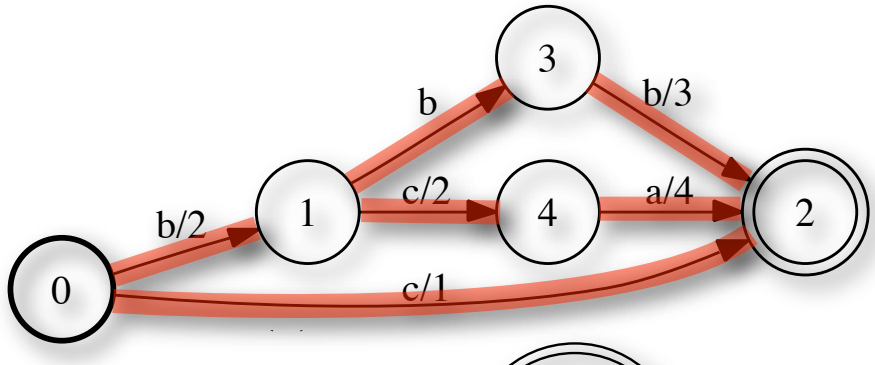
Suffix Algorithm

[TCS '09, TASLP '09]

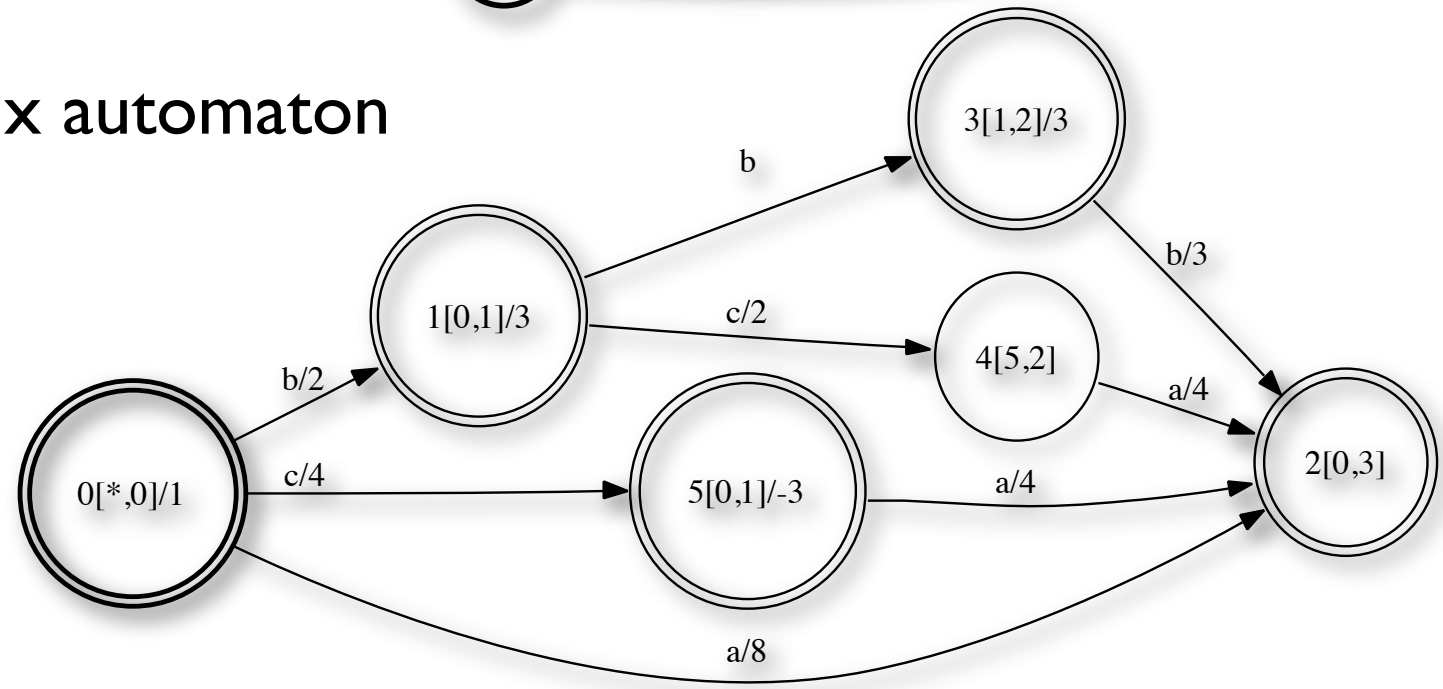
- New bound shows that size of suffix or factor automaton is linear in size of the input automaton
- But constructing this automaton requires the use of generic weighted determinization and minimization
- We have a new linear-time algorithm specifically to construct the suffix automaton directly
  - Can be converted into factor automaton in linear time
  - Builds output automaton on the fly as input is traversed

Suffix Algorithm Example

- Original (input) automaton



- Resulting suffix automaton



- Traversed:

Algorithm Properties

- Complexity of suffix automaton construction algorithm is linear: $O(|S(A)|) = O(|A|)$
  - Substantial improvement over previous suffix and factor automaton algorithms (determinization-based)
  - For music ID task: 17x faster than previous
- Input automaton can be traversed in any order, and can operate in an online fashion
- Operates on suffix-unique input automata; non-suffix unique automata can be encoded by adding final symbol $\$_i$ to each input string x_i of A

Main Results

1. Music identification: content-based search for songs
2. Automata: bounds and algorithms for efficient search
3. Topic segmentation: topicality of speech streams

Topic Segmentation

- ...this protest has brought out thousands of serbs calling for the end of the milosevic regime. opposition leaders are confident milosevic's days in power are numbered. on capitol hill tonight the senate approved 600,000 visas for skilled high technology workers...
- Question: can we segment spoken language by topic even with imperfect transcriptions?
- We give a novel topic segmentation quality measure
- We also develop new discriminative algorithms for topic segmentation of speech and text

Topic Segmentation

- Previous work: many papers threshold the cosine distance
  - Vocabulary $V = \{w_1, \ldots, w_n\}$, counts $C_1(w_i), C_2(w_i)$

$$\frac{\sum_{i=1}^{n} C_1(w_i) C_2(w_i)}{\sqrt{\sum_{i=1}^{n} C_1(w_i)^2 \sum_{i=1}^{n} C_2(w_i)^2}}$$

- Cosine distance is effective [Hearst '94] but compares only counts of a given word
  - e.g., if one text mentions only "football" and another mentions only "sports", according to cosine distance they are not similar

Similarity Measure

- Similarity for words: co-occurrence in a large corpus T

$$\mathrm{sim}(x, y) = \frac{C_T(x,y)}{C_T(x)C_T(y)}$$

- Two segments a, b:

$$K(a, b) = \sum_{w_1 \in a, w_2 \in b} C_a(w_1)C_b(w_2)\mathrm{sim}(w_1, w_2)$$

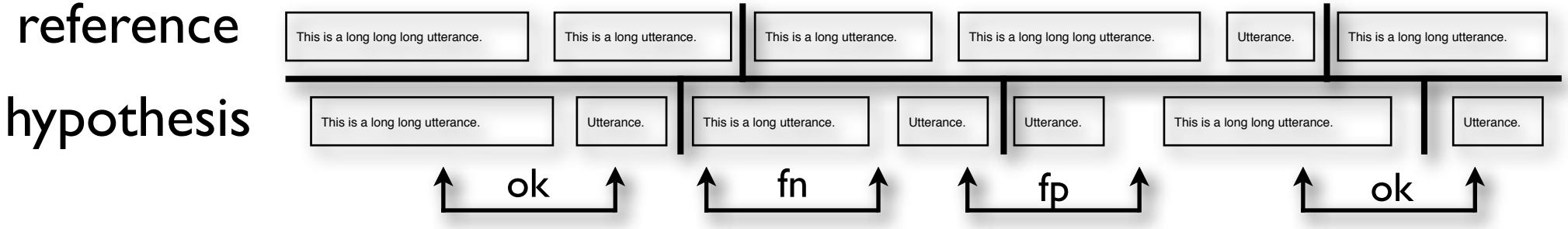
- Normalize:

$$K_{norm}(a, b) = \frac{K(a, b)}{\sqrt{K(a, a)K(b, b)}}$$

- Well behaved: range [0,1] and $K_{norm}(a, a) = 1$

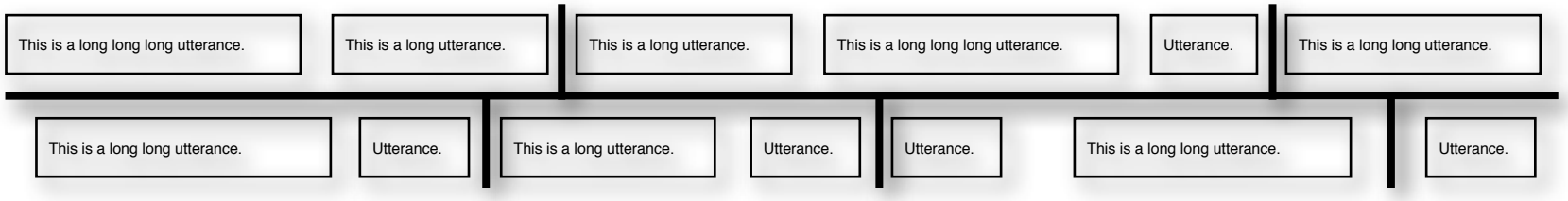
Quality Criterion: CoAP

[Beeferman et al., '99]



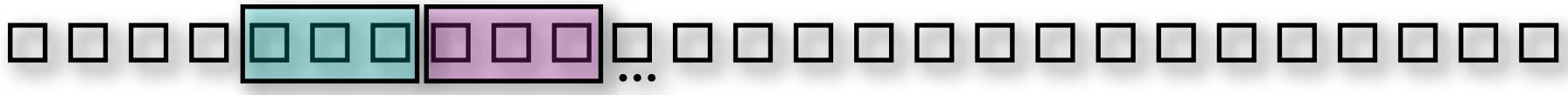
- Move sliding window across text, measure fraction of agreements between reference and hypothesis
- Limitations
  - Does not take word content into account
  - Very dependent on window size
  - Incompletely defined for speech case

New Quality Criterion



- $R = (r_1, \ldots, r_k)$ and $H = (h_1, \ldots, h_l)$ are reference and hypothesis segments, respectively
- New quality measure: $$\mathrm{TCM}(R, H) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} Q(i,j) K_{norm}(r_i, h_j)}{\sum_{i=1}^{k} \sum_{j=1}^{l} Q(i,j)}$$
- $Q(i,j)$: indicator, one if reference segment i overlaps with hypothesis segment j
- Spurious and missing segmentations penalized (a la CoAP)
  - Word content of overlapping segments included in score
  - Can use similarity scores other than K_{norm}

Segmentation Algorithm



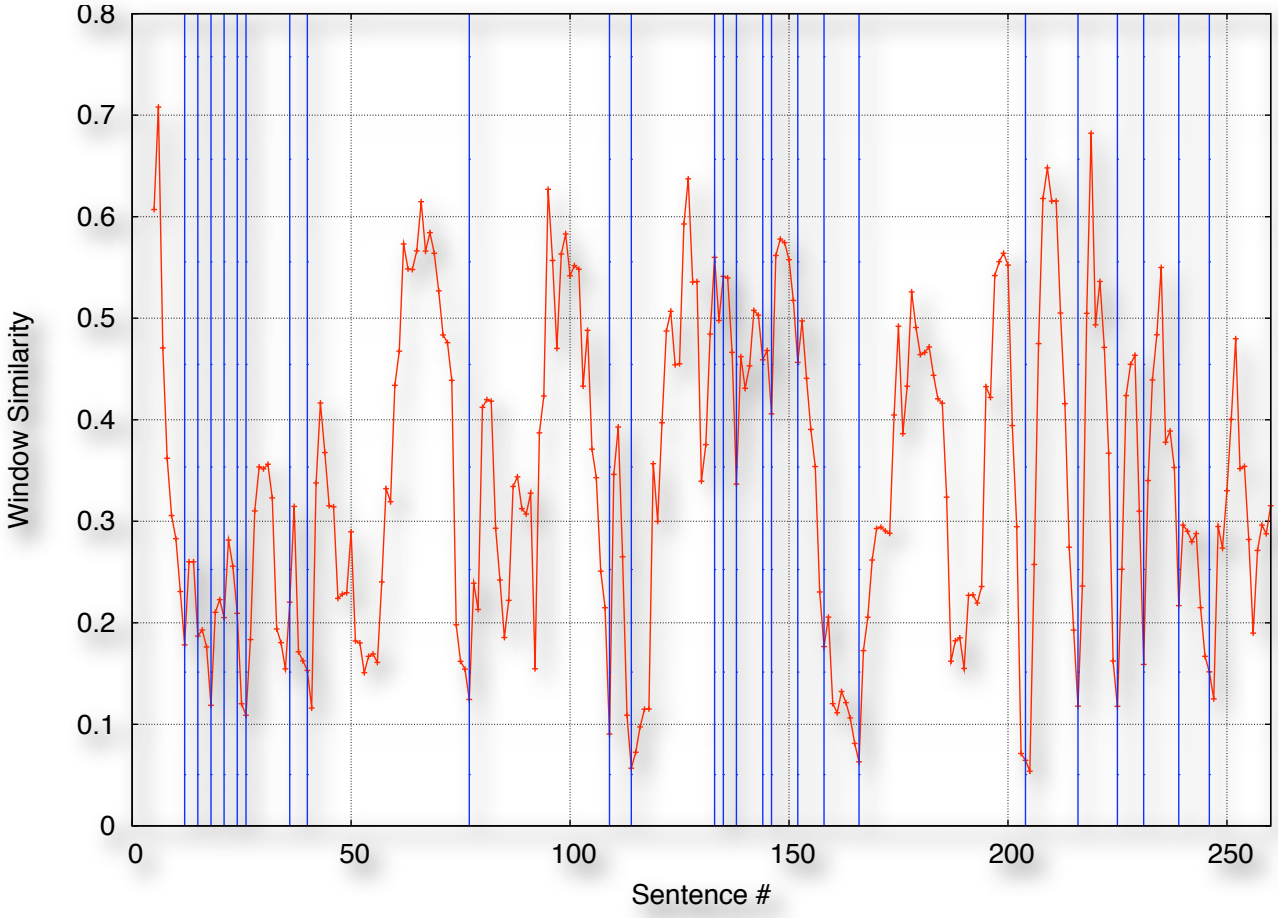
- A generic algorithm: window word counts, evaluate similarity, topic boundaries where similarity is small
  - $w_i = \{x_{i-\delta+1}, \ldots, x_i\}$: window of size δ
  - Our algorithm: use new similarity measure K_{norm}
  - Let $s_i = K_{norm}(w_i, w_{i+\delta})$, boundary set $b = \{i : s_i < \theta\}$
- For robustness, look for local minima,

$$b = \{i : s_i < \theta \wedge s_i = \text{rmin}(s, i - \lfloor \delta/2 \rfloor, i + \lfloor \delta/2 \rfloor)\}$$

$$\text{rmin}(s, i, j) = \min(s_i, \ldots, s_j)$$

Ground Truth Comparison

- Plot of $K_{norm}(w_i, w_{i+6})$, blue lines: reference boundaries



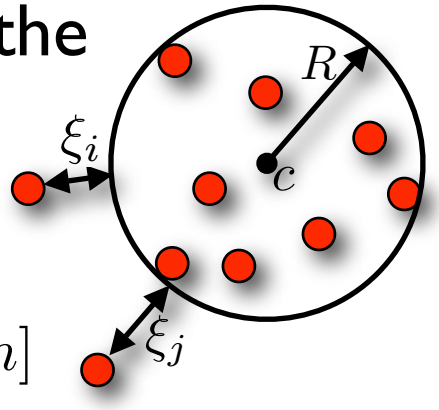
A Problem

- Consider the following text (human transcript):
  - A powerful explosion tore through a cafe frequented by Russian soldiers, south of Chechnya's capital Grozny, Sunday, killing at least eight people, including the owner of the cafe and four Russian soldiers. Russia's Interfax News Agency quotes security sources as saying Chechen rebels have claimed responsibility for the attack. That's our news summary till now. I'm David Coller, VOA News.
- Or this one (speech recognition):
  - the program one five emmys and carson was awarded the presidential medal of freedom and all the signed a law in nineteen ninety two more than fifteen million viewers tuned in to watch and say goodbye issue very hard.
- Substantial off-topic content (i.e., noise)
- This affects our similarity score: need to filter out noise

Removing Outliers/Noise

[Tax and Duin '99, Schölkopf '01]

- Given a set of observations $x_1, \ldots, x_m \in \mathcal{X}$ and mapping into feature space $\Phi: \mathcal{X} \mapsto F$, find compact description
- Sphere in feature space separates the bulk of the observations from the outliers



$$\min_{R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^m, c \in F} \quad R^2 + \frac{1}{\nu m} \sum_i \xi_i$$

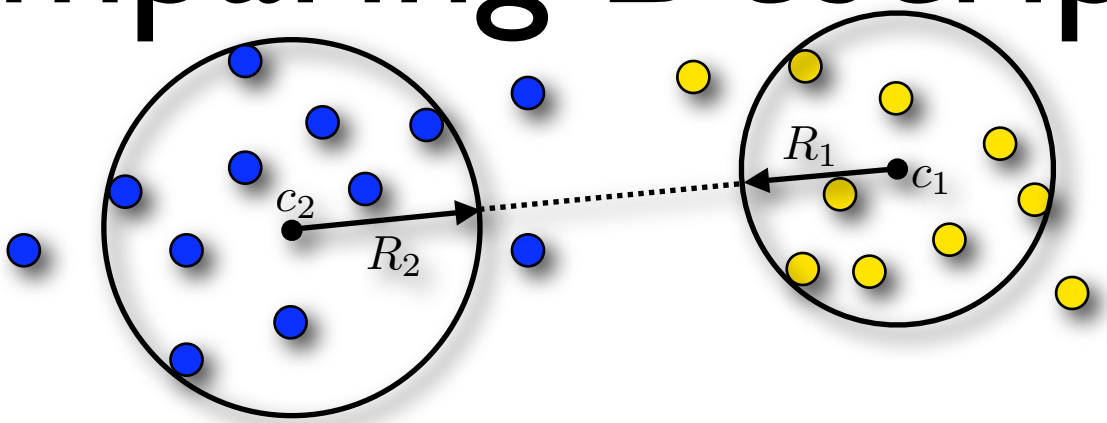
$$\text{subject to} \quad \|\Phi(x_i) - c\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0 \text{ for } i \in [1, m]$$

- Kernelized version: $K(x, y) = \Phi(x) \cdot \Phi(y)$, dual problem

$$\min_{\boldsymbol{\alpha}} \quad \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^m \alpha_i K(x_i, x_i)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu m}, \quad \sum_{i=1}^m \alpha_i = 1.$$

Comparing Descriptors



- Solution to optimization: center $c = \sum_i \alpha_i \Phi(x_i)$
  - Radius is recovered from classifier form

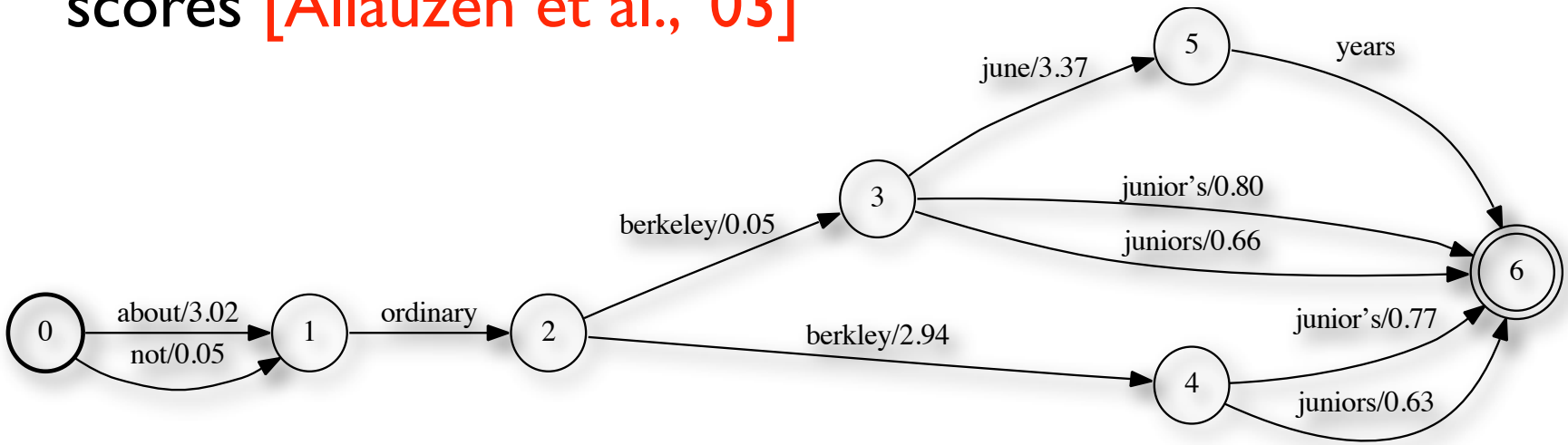
$$f(x) = \text{sgn}\left(R^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) + 2\sum_{i=1}^{m} \alpha_i K(x_i, x) - K(x, x)\right)$$

- To compare the "true" content (i.e., that excluding outliers) of two windows, we compare their descriptors
- Using the kernel, find shortest distance in feature space

$$\text{dist}(w_1, w_2) = \|c_1 - c_2\| - (R_1 + R_2)$$

Lattices

- In a high error rate setting, the one-best recognition hypothesis is not informative enough
- Recognizer can output a lattice of competing hypotheses, each with an associated probability
- Can process to yield expected counts and confidence scores [Allauzen et al., '03]



Algorithms and Evaluation

- Sim algorithm: $s_i = K_{norm}(w_i, w_{i+\delta})$, find local minima
- SV algorithm: $s_i = \mathrm{dist}(w_i, w_{i+\delta})$, find local maxima
- TDT: news speech (VOA, CNN, etc.) and text (NYT, etc.)
- Training set: 1,314 news streams with 21,420 stories
  - Human transcriptions and text streams
- Development: 41 shows, 957 stories
- Testing: 69 shows, 1,674 stories

Experiments

- Compare with hidden topic Markov model (HTMM) [Gruber et al. '07]
  - Generative context-dependent topic model
- Noise is a problem for Sim
- SV is able to overcome noise and beat HTMM
- Lattice-derived information helps

Input Type	Algorithm	Quality Measure	
		CoAP	TCM
Text	HTMM	66.9%	72.6%
	Sim	72.0%	75.0%
	SV	76.6%	77.7%
One-best	HTMM	65.0%	61.5%
	Sim	60.4%	62.8%
	SV	68.6%	66.0%
Counts	HTMM	65.5%	62.4%
	Sim	59.4%	63.4%
	SV	68.5%	66.5%
Confidence	HTMM	68.3%	64.2%
	Sim	59.7%	63.8%
	SV	69.2%	66.8%

References

- Eugene Weinstein and Pedro Moreno. Music Identification with Weighted Finite-State Transducers. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, Hawaii, 2007.
- Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. Factor automata of automata and applications. In 12th International Conference on Implementation and Application of Automata (CIAA), Prague, Czech Republic, July 2007.
- Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. Robust music identification, detection, and analysis. In International Conference on Music Information Retrieval (ISMIR), Vienna, Austria, September 2007.
- Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. General suffix automaton construction algorithm and space bounds. To appear in Theoretical Computer Science, 2009
- Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. Efficient and Robust Music Identification with Weighted Finite-State Transducers. To appear in the IEEE Transactions on Audio, Speech, and Language Processing, 2009
- Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. A New Quality Measure for Topic Segmentation of Text and Speech. In submission, 2009.
- Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. Discriminative Topic Segmentation of Text and Speech. In submission, 2009.

Acknowledgments

Prof. Mehryar Mohri, advisor and reader

Dr. Pedro Moreno, host at Google and reader

Prof. Ralph Grishman, reader

Prof. Juan Pablo Bello

Prof. Dennis Shasha

Thanks to everyone for attending!