Speech Indexing

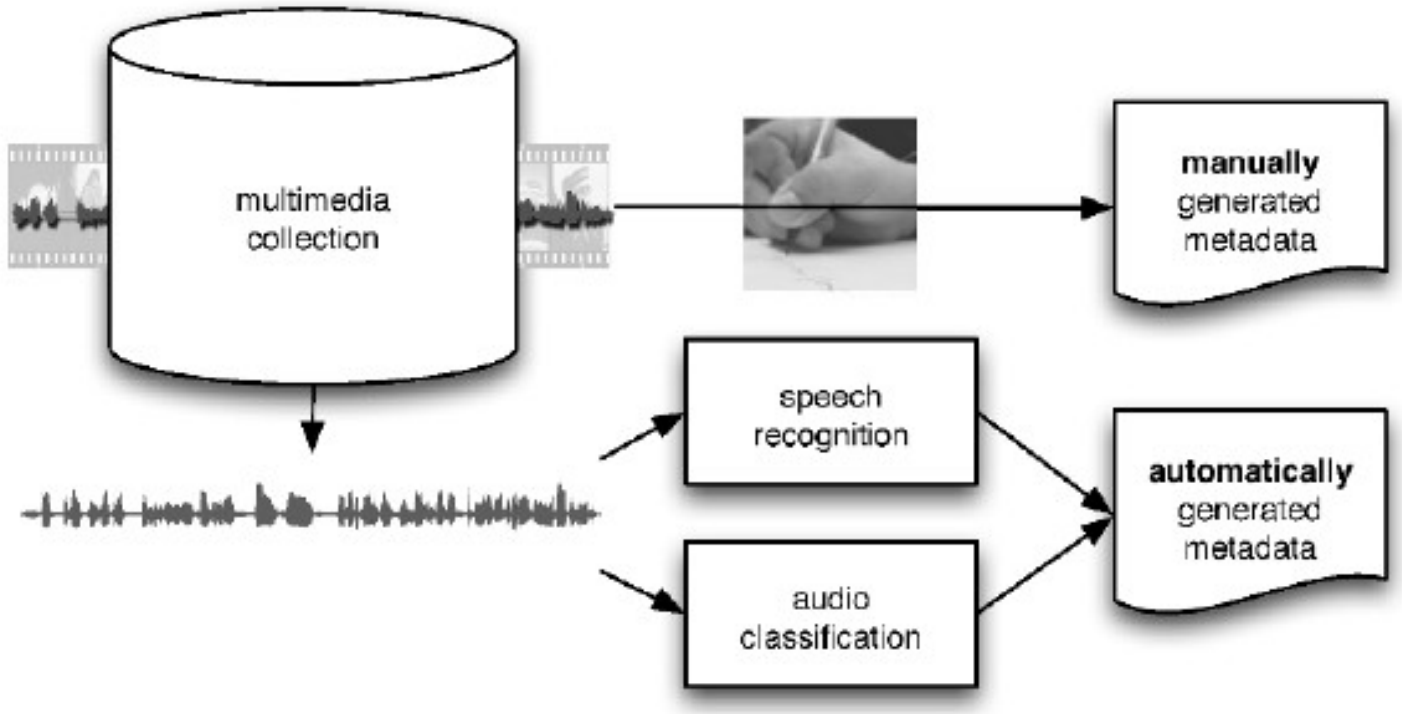
Luis De Alba
ldealbar@cc.hut.fi
29.2.2008

Outline

- Introduction
- Speech Recognition
  - Features extraction
  - Acoustic & language modeling
  - Dictionary
- Spoken Document Retrieval
  - Manual VS Automatic
  - Techniques
- Robust Speech Recognition & Retrieval
  - Query and Document expansion
  - Vocabulary optimization
  - Topic-based language models
  - Acoustic adaptation
- Cross-media Mining

Introduction

- Every Organization chooses how much metadata attaches to its multimedia collections.
- “Information is in the audio, video is for entertainment” Richard Schwartz.
  - Using Automatic Speech Recognition Technologies:
    - Speech -> Text + Linguistic annotations.



- Henk Blanken, et al.

- Other Clues;
  - Information about the speaker:
    - Gender, Background, Emotional State
- Then, if the objective of the recognition of the words spoken is to support retrieval it is called *Spoken Document Retrieval* (SDR)

Speech Recognition

- Feature Extraction
  - Digital acoustic signals transformed into a compact representation that captures the characteristics of the speech signal.
  - Vector of features for every 10ms in 16ms overlapping windows.

Speech Recognition

- The coefficients can be the MFCC (Mel Frequency Cepstral Coefficients)
  - Take the Fourier Transform of selected window.
  - Map the amplitudes into the Mel scale.
  - Take the Discrete Cosine Transform.

Speech Recognition

- Acoustic Modeling
  - Observations represented as O
  - Task:
    - Find the sequence of words $W = \{w_1, w_2, \ldots, w_N\}$ most likely to match O
  - Choosing the highest probability:

$$\hat{W} = \arg\max_{w} P(W|O)$$

  - Then:

$$P(W|O) = \frac{P(O|W) \cdot P(W)}{P(O)}$$

Speech Recognition

- Popular approach in speech recognition is the use of hidden Markov Models (HMMs)
- HMMs represent connected states each one having a transition probability.
- The problem goes $P(O|W)$ to $P(O|M)$
  - M represents the sequence of a word associated to W. Each model M can be a *phone* (smallest unit in speech)

Speech Recognition

- Language Modeling
  - P(w) can be expressed as:

$$P(W)=\prod_{i=1}^{n} P(w_i | w_{1,}\ldots, w_{i-1})$$

    - The probability of w_i being spoken preceding the sequence (w_1 ... w_{i-1})
  - For long history it is not feasible.
  - Use Markov assumption:
    - “The probability of a future event can be predicted by looking at its immediate past”

Speech Recognition

- *N*-grams, number of previous words. Usually 1 or 2.
  - Two-word history, trigram models can be generated as:

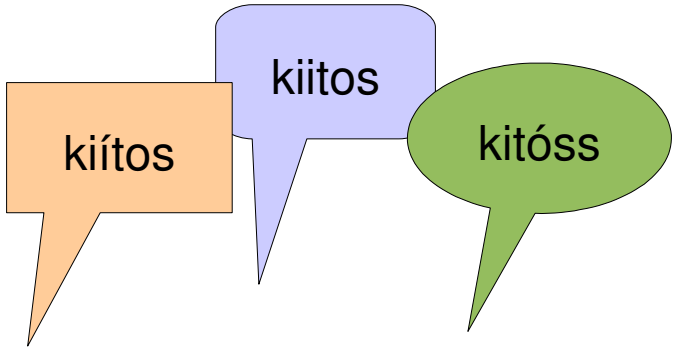
$$P(w) \approx P(w_0) \cdot P(w_1 | w_0) \cdot \prod_{i=2}^{n} P(w_i | w_{i-1}, w_{i-2})$$

Speech Recognition

- Dictionary
  - Vocabulary: List of all the words in the language model.
  - Pronunciation Dictionary: Link between the acoustic model and the language model.
  - Usually represented as lexical trees.

Speech Recognition

- How to obtain word pronunciations:
  - Manually.
    - Time Consuming
  - G2P tools (Grapheme to Phoneme)
    - Produces the *average* pronunciation.
    - Pronunciation varies according to age, gender and dialect. 40% of the words are not correctly pronounced.
    - To override this issue include pronunciation variations in the lexicon.

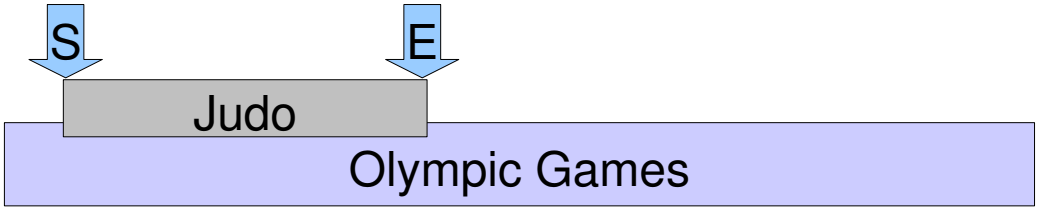


Spoken Document Retrieval

- To support document retrieval is necessary to create a textual representation.
- Depending on companies and/or libraries management, it may exist:
  - Human generated metadata.
  - Bibliographic (tombstone) information.
- However, it is still difficult to find specific passages.

Silver Medal ...

55 kilos ...



Spoken Document Retrieval

- Manual VS Automatic
  - Instead of speech recognition technology do it manually.
  - *Minutes* for meetings are done this way and, if there is video, the annotations can be very useful.
  - Manually means people, people means money. Therefore, manually is expensive.
  - ASR (Automatic Speech Recognition) has largely improved lately.

Spoken Document Retrieval

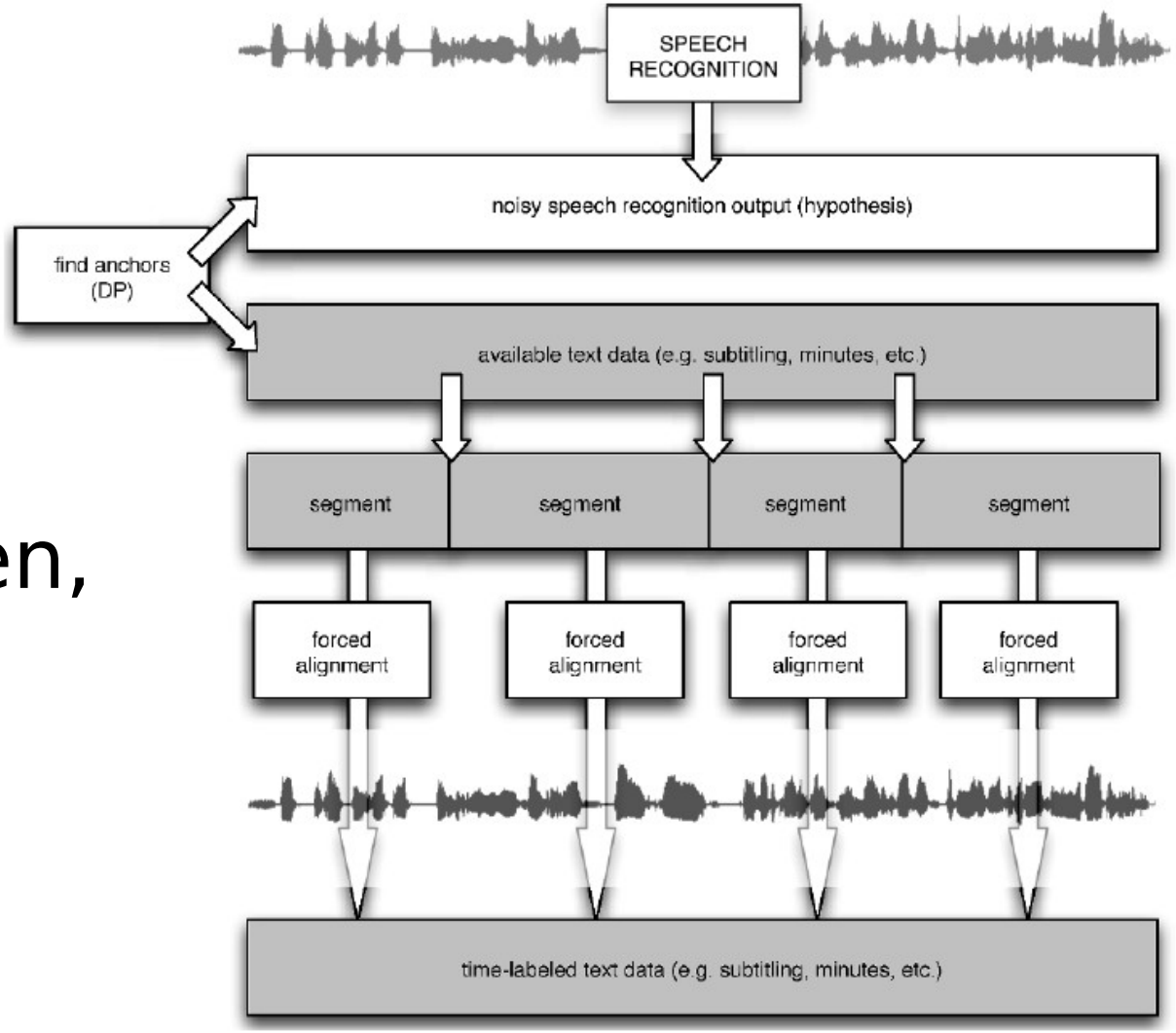
- Techniques
  - Synchronization of available textual resources.
    - Collateral textual resources that are closely related can be exploited.
      - e.g., Subtitle information for the hearing impaired. (some even provide boundaries)
    - The time labels from the sources are crucial for the indexing.
    - If there are no time labels some synchronizations shall be done.

Spoken Document Retrieval

- Synchronization technique:
  - A SRS is used to generate a “inaccurate” transcript of the speech -> Hypothesis
  - The Hypothesis is aligned to the Minute using dynamic programming.
  - Where the Hypothesis and Minute match generate an “anchor”
  - Using the timing from the SRS and the “anchors” generate Segments.
  - Finally, individual segments of audio and text are accurately synchronized.

Spoken Document Retrieval

- Henk Blanken, et al.



Spoken Document Retrieval

- Large vocabulary speech recognition.
  - When a system produces errors successful retrieval will be doubtful.
  - A speech recognition performance of 50% is the minimum for a useful performance.
  - Todays systems require:
    - Speaker-independent, trained using large amount of example audio from the domain.
    - Large Vocabularies: 65,000 words.
  - Out of the Vocabulary
    - When a word is not in the vocabulary it can not be recognized and will not appear in the minute.

Spoken Document Retrieval

- e.g., BBN technologies has 1.8M of words for training and an America English GigaWord News corpus of 1 Billion words of text.

Spoken Document Retrieval

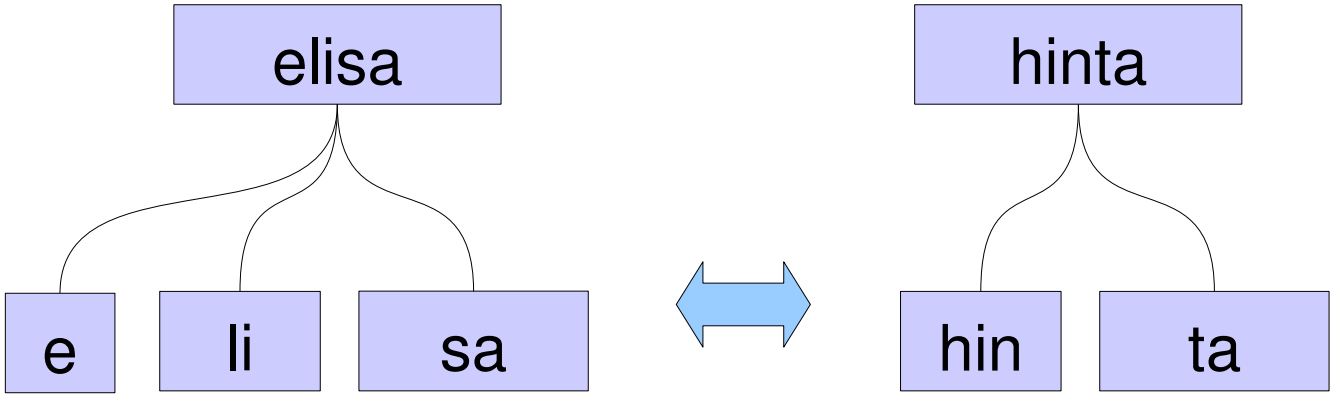
- Keyword spotting.
  - Feasible approach when computer power is limited.
  - Keywords are usually fixed in advance.
  - Method works for a restricted domain.
    - e.g., weather reports
  - Useful when heavy-weigh speech recognition is not feasible.

Spoken Document Retrieval

- Sub-word unit representative.
  - Previous presented approaches focus on words as units of the decoded speech.
  - An alternative is to use sub-word units, such as *phones*.
  - The document is represented in terms of these sub-words units.
  - Phone recognizer requires only an acoustic model and a small phone grammar. It is much faster than large vocabularies approach.

Spoken Document Retrieval

- Is less sensitive to out-of-vocabulary words.
- However, tends to produce higher error rates. Because it is based solely on acoustic information.



Robust Speech R & R

- For retrieval purpose it is important to have the *content words* right.
  - e.g., nouns, names, adjectives
- The indexing process will discard the rest of the words.
  - e.g., articles
- Analyzing the global word error may not be adequate.

Robust Speech R & R

- *Reference* is the original transcript and the *Hypothesis* is the generated transcript.
- The word error rate WER is calculated as follow:

$$WER = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Total words in } \textit{Reference}}$$

Robust Speech R & R

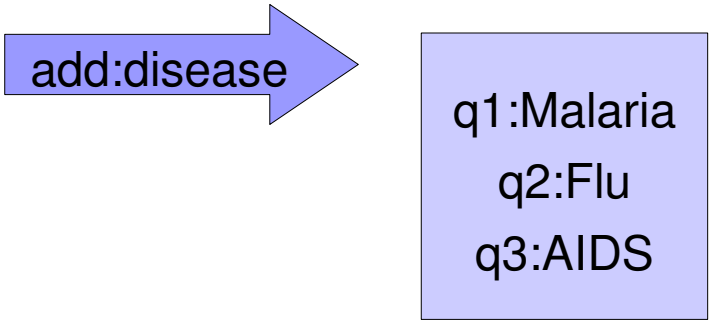
- The term error rate TER is calculated as follow:

$$TER = \frac{\sum_{t \in T} |R(t) - H(t)|}{\sum_{t \in T} R(t)}$$

  - Where *t* represents a given query.
- The TER gives a more accurate measure of speech recognition performance.

Robust Speech R & R

- Query and Document expansion
  - This technique simply adds words to the query in order to improve retrieval.
    - Process:
      - After initial run from the top *N* most relevant documents,
      - Select the top *T* terms and add them to the query to enrich it.



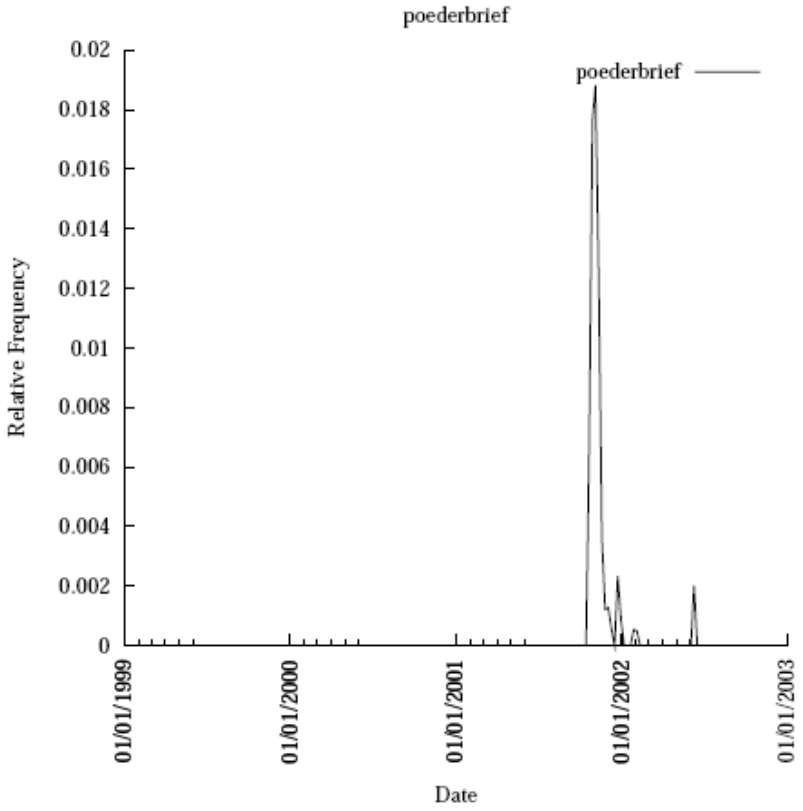
Robust Speech R & R

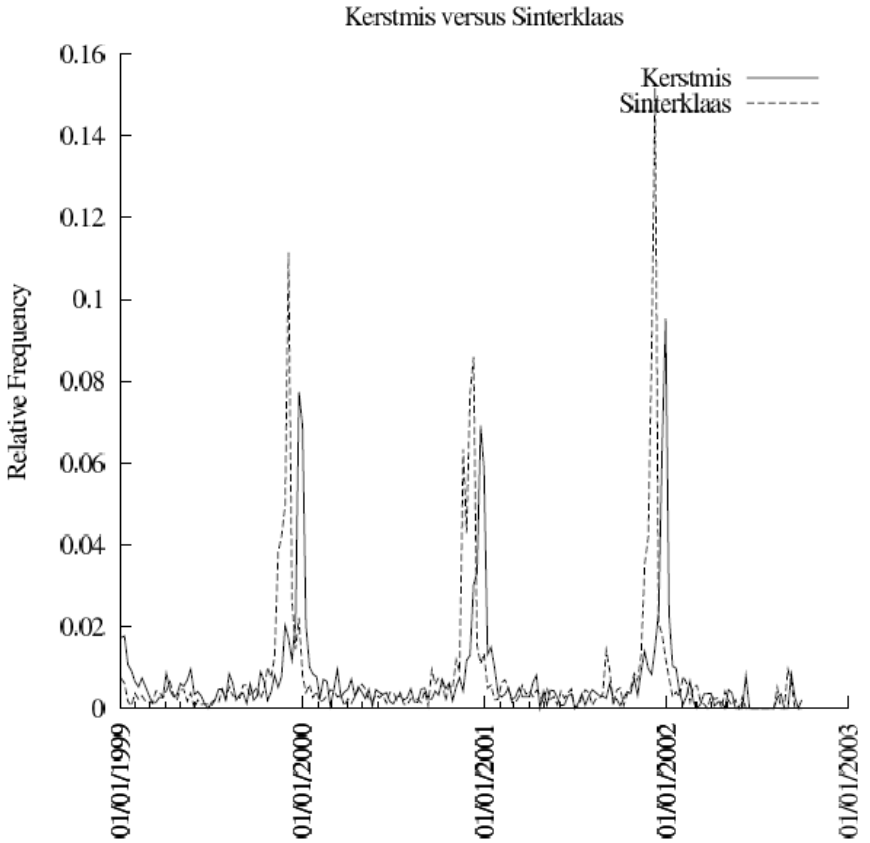
- Vocabulary Optimization
  - For success is necessary to minimize the out-of-vocabulary words.
  - Selection of appropriate set of vocabulary words reassembling the domain.
  - The maximum number of words that can be included in the dictionary is restricted.
    - Typically to 65,000 words.

Robust Speech R & R

- With topics changing is necessary to revise the vocabulary and update accordingly.
  - Include new words (new topics)
  - Remove obsolete words (old topics)
- There are words that occur in certain times of the year.
- Some words appear suddenly and can not be “foreseen”

Robust Speech R & R





- Henk Blanken, et al.

Robust Speech R & R

- Topic-based language models
  - Words are selected with a focus on a specific segment of an audio.
  - Include the *n*-grams specific to the topic.
    - *bank*
      - *the interest rate in the bank* more
      - *the bank of the river* less

Robust Speech R & R

- Topic-based requires 5 steps:
  - 1) Segmentation of the audio file.
    - Segments can be interpreted to be on the same topic.
    - In practice, segmentation is not known. Then change of speaker, silence intervals, etc. are used.
  - 2) Speech recognition on the segments.
  - 3) Definition of the *topic*
    - Using the transcripts and collateral text.
    - Collateral text can come from different sources:
      - e.g., newspaper, topics database

Robust Speech R & R

- 4) Generation of the topic specific language model.
  - Using a ranked list of similar documents.
- 5) Speech recognition using the topic-based language model.

Robust Speech R & R

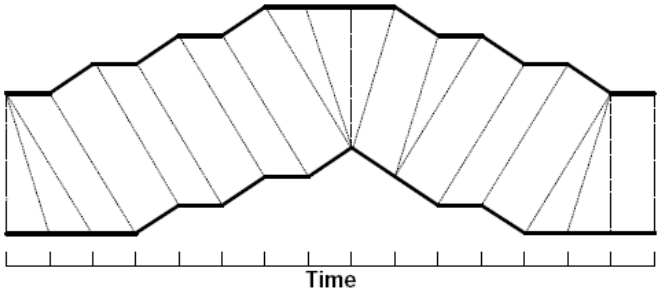
- Acoustic Adaptation
  - To be robust a speech recognition system shall have good performance even when the quality of the input is low.
    - Background Noise, Cross-Talk, Low Audio Quality
  - Speaker to Speaker characteristics vary due to:
    - Vocal tract, Age, Speaking style, etc.

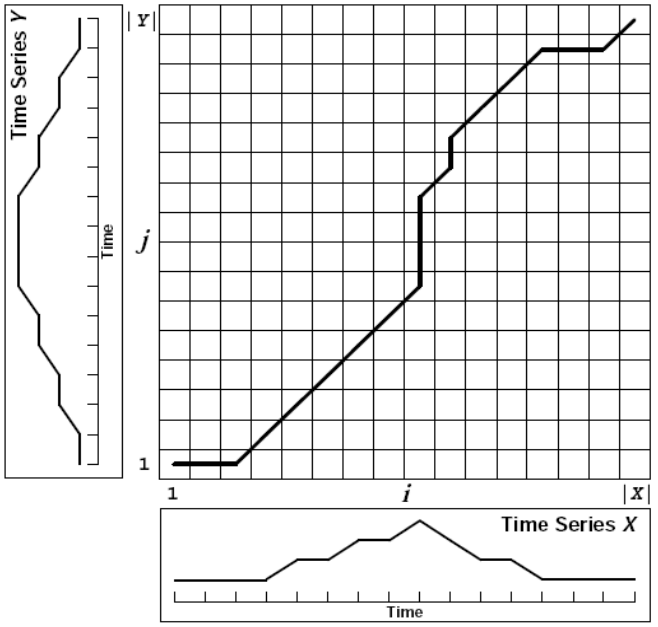
Robust Speech R & R

- To override the problem more Training can be done, or
- apply *normalization* and *dynamic adaptation* procedures.
  - Start with a stable baseline and tune it to the specific conditions in the task domain.
- A) Vocal Tract Length Normalization
  - The average length of the tract is 17cm, but it varies with dimension of the person.

Robust Speech R & R

- Vocal tract length normalization technique aim is to compensate acoustic differences.
- Normalization from the cluster of speakers to the “generic” speaker.
- The normalization is done by *warping* the frequency axes.





- Stan Salvador, Philip Chan. FastDTW. 2007.

Robust Speech R & R

- B) MAP and MLLR Adaptation
  - Maximum A Posteriori
  - Maximum Likelihood Linear Regression
  - These methods aim at adjusting the *model parameter* not the *spectral information*.
  - Model adaptation can be done off-line or at recognition time (online)
  - MLLR aims to capture the general relationship between the speaker independent modal set and fit it to the adaptation data.

Robust Speech R & R

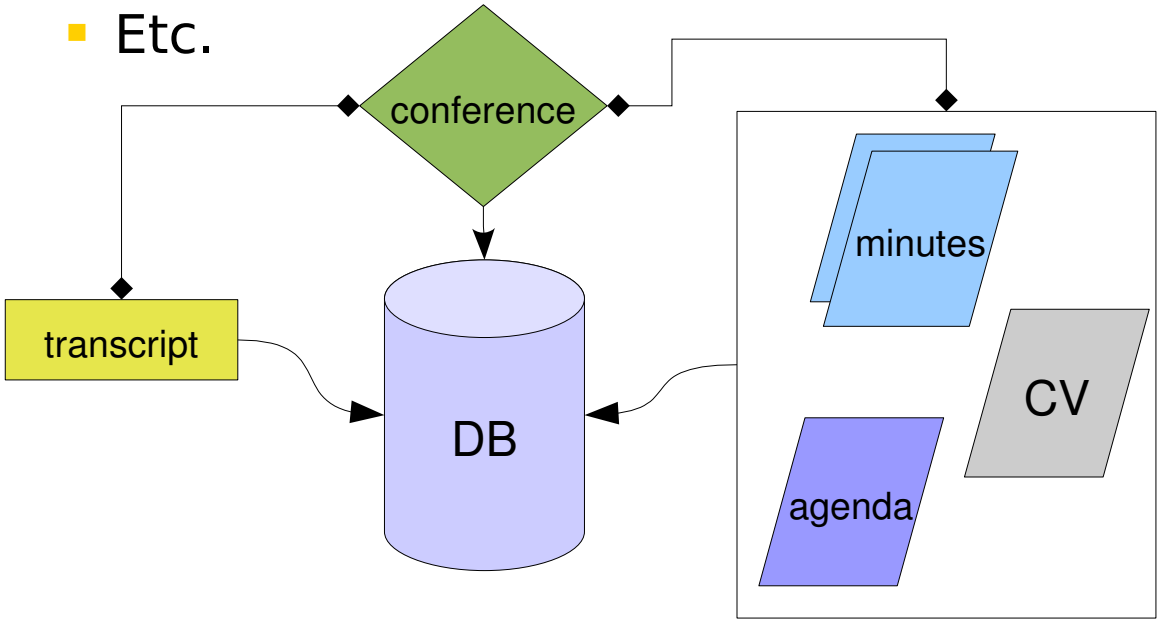
- MAP combines the information from the adaptation data and some prior knowledge about the model.
- A disadvantage is that large number of adapted models may be generated.

Cross-media Mining

- Go one step further:
  - Exploit collateral or text resources that are accessible.
  - Mining these resources can support access to different perspectives on the available data.
    - Linking newspaper articles with broadcast items.
  - Available semantic annotations for documents with similar profiles can be exploited to improve the searching.

Cross-media Mining

- Example:
  - Video conference in a company:
    - Speech Recognition of the Video (transcript).
    - Minutes.
    - Reports.
    - Information of people attending the meeting.
    - Etc.



Conclusions

- For Speech Indexing and Retrieval it is necessary to have a **transcript** of the contents of the multimedia data.
- How:
  - Speech Recognition.
  - Synchronization of Data Sources.
  - Cross-media Mining.
- Do not forget:
  - That every approach has certain advantages and disadvantages. Which to use depends on the application.

Bibliography

- Henk Blanken, et al. Multimedia Retrieval. Springer. 2007. [Speech Indexing Ch. 7]