# Open Vocabulary Speech Analysis in Vitalas


Daniel Schneider

Speech Group, Fraunhofer IAIS


Fraunhofer Institut
Intelligente Analyse- und
Informationssysteme

# Outline

- Vitalas Scenario: Broadcast News Audio Indexing
- Structural Audio Analysis
- Open Vocabulary Speech Recognition
- Demo: AudioMining

# Challenge in Vitalas: Large Scale Broadcast News Indexing

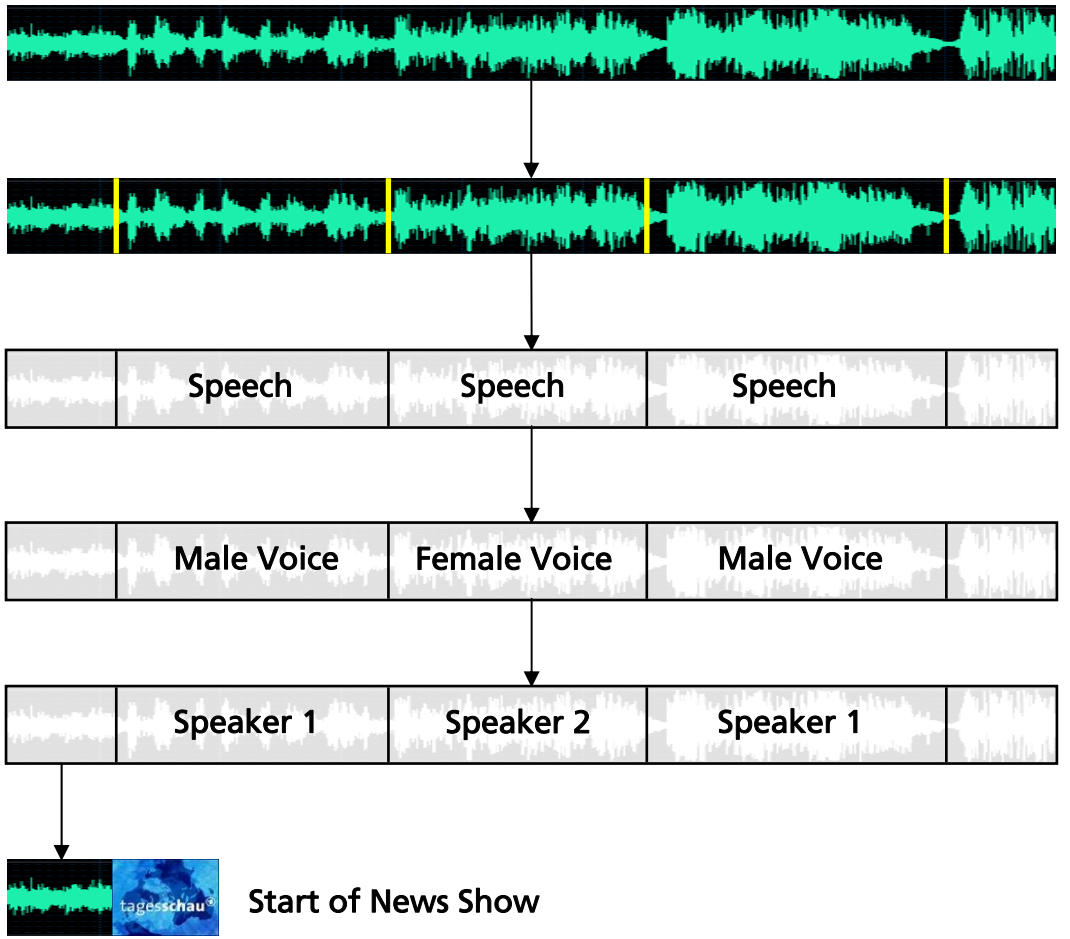
- Huge amount of data (> 10.000 hours)


- Heterogeneous material
  - From various sources of unknown type
  - High topic variability
  - Huge vocabulary
  - Multilingual data


- Requires efficient and robust algorithms for…
  - Information extraction
  - Information retrieval

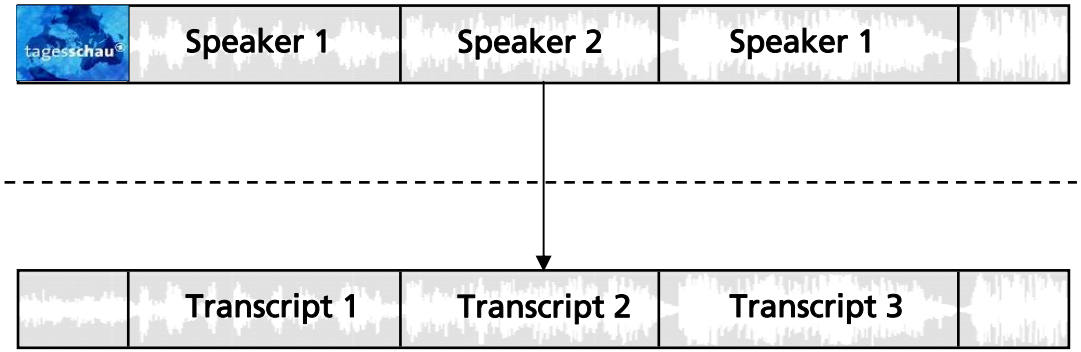
# Structural Audio Analysis in Vitalas

- Unstructured Audio Data
- Homogeneous Segmentation
- Speech Detection
- Gender Detection
- Speaker Clustering
- Programme Identification via Jingle



# Speech Recognition

- Structual Analysis



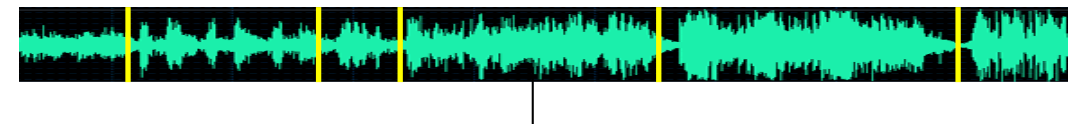

- Speech Recognition

- Transcripts can be used for…
  - Search in entire archive (“Audio-Google”)
  - Media observation (Alert if keyword occurs)
  - Input for text mining (e.g. Topic Detection)

# Speech Recognition Challenges

- Out-of-Vocabulary problems with classic word based ASR of broadcast data
  - New and popular words (e.g. Gammelfleischskandal - „rotten meat scandal“)
  - Proper names (companies, cities, people)

- Compound words in German (climate conference – Klimakonferenz)
- Huge Lexica required – large effort



**“... der Klimakonferenz der Vereinten Nationen ...”**

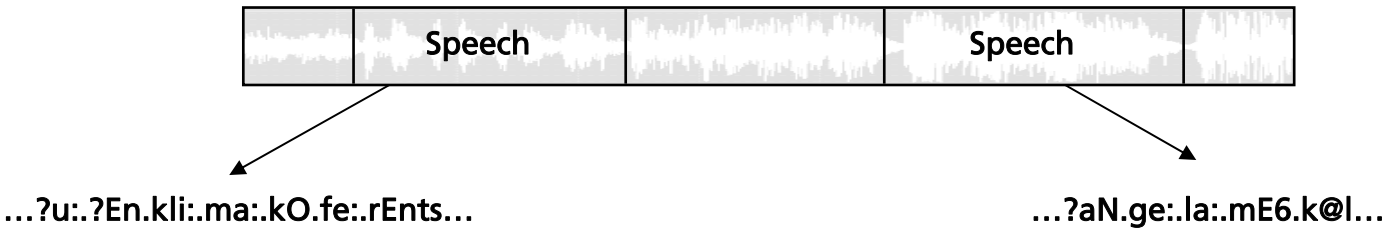
der Kiel Markt Konferenz der Vereinten Nationen

# Phonetic Approach to Open Vocabulary Indexing

- Idea:
  - Search on phonetic subword level instead of word level
  - Search for a sequence of sounds instead of words

# Phonetic Approach to Open Vocabulary Indexing

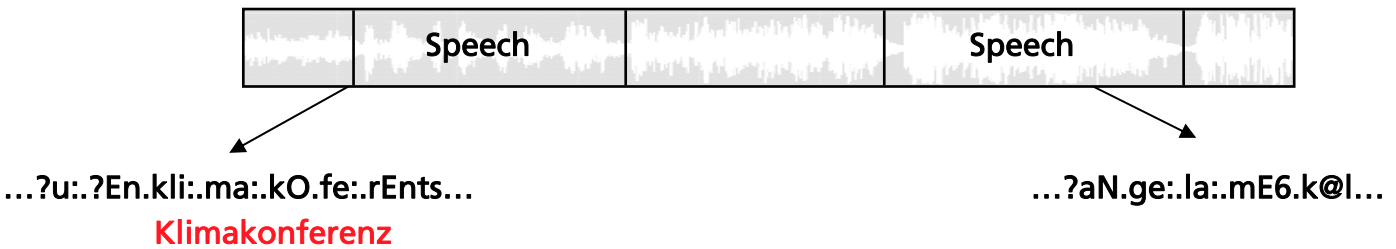
1. Generate transcription on subword level (phone or syllable)



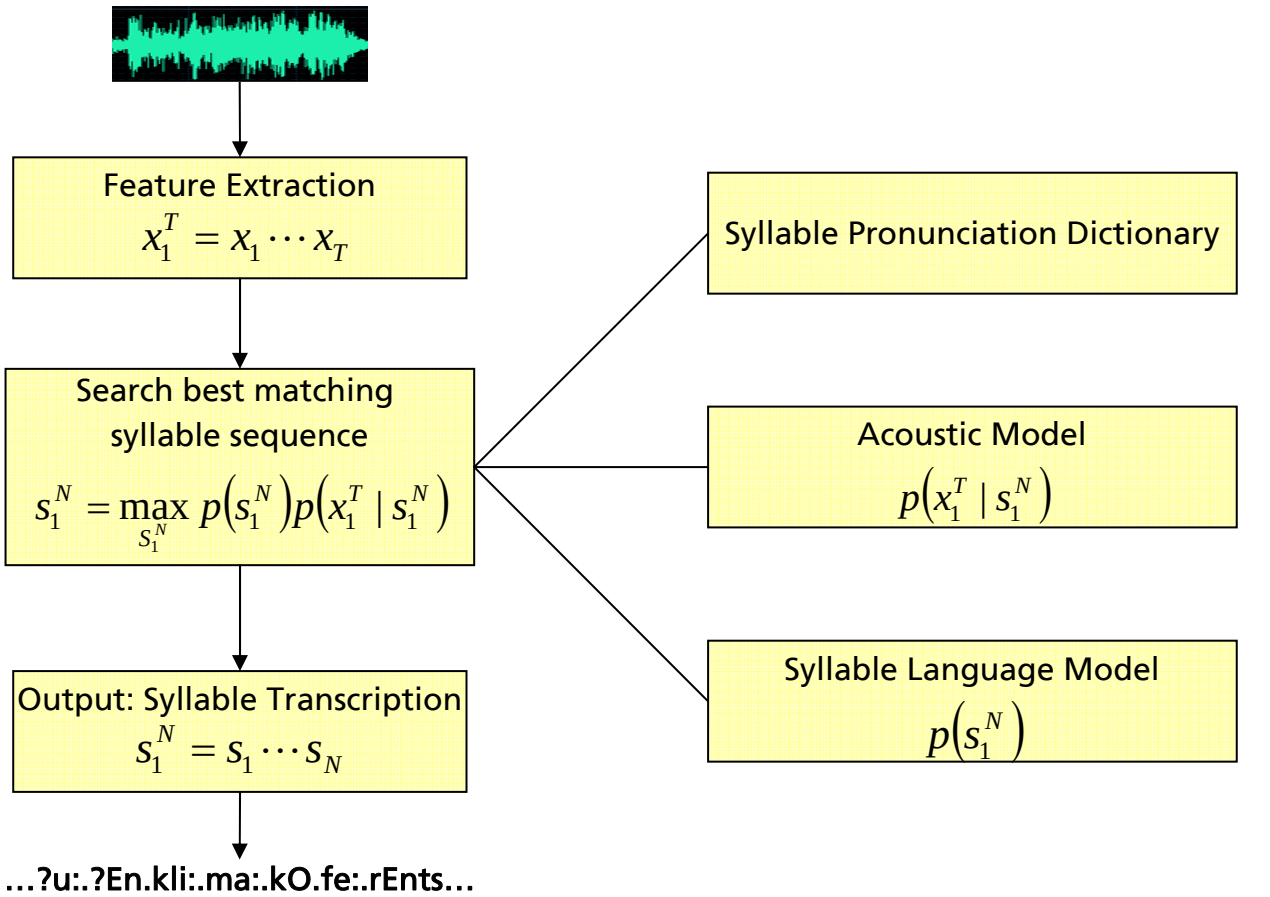

2. Break down search term into subword units

   Klimakonferenz → kli:.ma:.kOn.fe:.rEnts

3. Fuzzy Phonetic Search



# Phonetic Approach (1): Generate Subword Transcription




- Cross-Word Triphone Hidden Markov Models
- Bi- / Trigram Models

# Phonetic Approach (2): Fuzzy Syllable Search

- Break down search term: Klimakonferenz → kli:.ma:.kOn.fe:.rEnts
- Goal: Retrieval of documents containing similar syllable sequences
- Fuzzy search based on Levenshtein Distance between
  - Single syllables
  - Syllable sequences

Examples distances between single syllables:

| | | |
|---|---|---|
| d_e:_s_ | d_e:_s_ | zero |
| d_e:_s_ | k_O_n_ | high |
| d_e:_s_ | d_i:_s_ | low |
| d_e:_s_ | d_I_s_ | medium |

Examples distances between syllable sequence:

| | | |
|---|---|---|
| k_l_i:_ m_a:_ | k_l_i:_ m_a:_ | zero |
| k_l_i:_ m_a:_ | k_l_i:_ n_a:_ | low |
| k_l_i:_ m_a:_ | k_l_i:_ n_6_ | high |

- Solution based on Dynamic Programming (c.f. Speech Decoding)

# Properties of Phonetic Subword Approach

- The set of subword units is finite and (rather) small
  - Complete vocabulary coverage (no OOV)
  - 10.000 syllables compared to 300.000+ words
  - Compact ASR search space

- Implicit decomposition of compounded words
  - *kli:.ma:.kOn.fe:.rEnts* gives 100% hit for the search terms
    *Klima, Konferenz, Klima Konferenz, Klimakonfernez*

- Implicit stemming capabilities of fuzzy search
  - Skandal – skan.da:l
    Skandals – skan.da:ls (less important to learn genitive explicitly)

# Experiments: Fraunhofer AudioMining Corpus

- High Quality Studio Data
  - Accurate sentence level transcriptions
  - (Almost) no background noise
  - Only one speaker per segment
- 14 hours of carefully annotated training data
- 3 hours of evaluation data (disjoint from training set)
- Main Challenges
  - Speaking rate (interview vs. read speech)
  - Spontaneous Speech in interview situation

Data: German News Shows

Comparable to VITALAS

data sets from IRT and INA


Broadcast News


Broadcast Conversation

# Experiments: Model Setup

- Acoustic Models
  - Maximum Likelihood Reestimation
  - Phonetic Clustering of triphones
  - 7300 triphone HMMs with up to 16 Gaussian mixture components

- Language Models trained on 2000-2006 newswire data with CMU SLM toolkit
  - 80 million running words
  - Text transformed to syllables
  - Corpus Topics: Politics, Economy, Culture, Sports

- Pronunciation Lexicon: 10000 most frequent syllables from LM training

# Current Results - Speech Recognition

- Task: Syllable Transcription of 3 hours of Broadcast Data (Radio Shows)

| Syllable Error Rate | ASR Realtime Factor |
|---|---|
| 34.3 | 1.5 |

- High error rates (test set includes several BC shows)
- Example for frequent substitution error:
  - Reference: U_n_t_ (and)
  - Recognized: U_n_ (an')
- Errors partly covered by fuzzy retrieval

# Current Results – Fuzzy Phonetic Retrieval

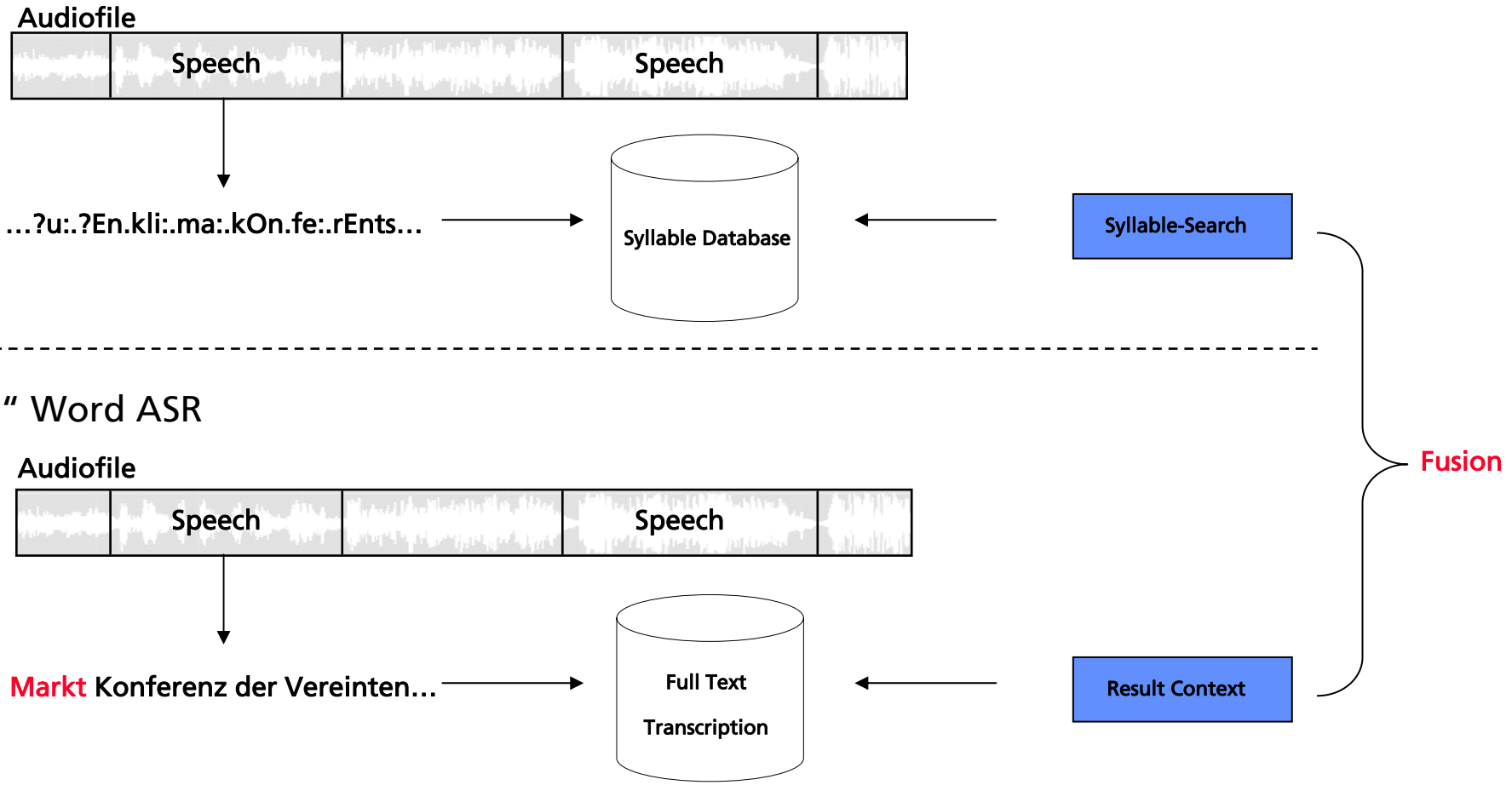
- Task: Detect 213 keywords and keyphrases in recognition results
- Confidence thresholds of the fuzzy search can be chosen depending on the application

| Confidence Threshold | Precision | Recall | Remark |
|---|---|---|---|
| 0.70 | 0.66 | 0.65 | Equal Error |
| 0.85 | 0.91 | 0.53 | Tuned for Precision |

- Some errors due to…
  - Search term is substring of actual spoken compound word (Klima – Klimakonferenz)
  - Short search terms consisting of highly frequent syllables (Mutter – mU.t6)

# Additional Word Context for Enhanced Display of Results

1. Vocabulary Independent Syllable Recognition



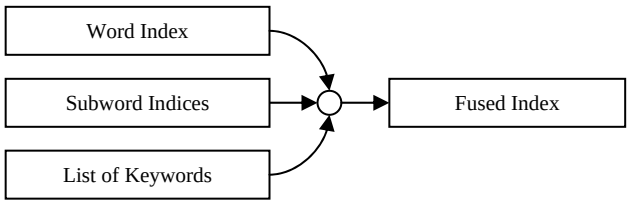
# Demo: AudioMining

# Next Steps

- **Evaluate the syllable approach on other languages**
  - Vitalas End-Users: IRT (German) and INA (French)

- **Improve Recognition Accuracy**
  - Use information extracted by structural analysis
  - Speaker / domain / progamme adaptivity

- **Improve Information Retrieval Accuracy**
  - Fusion of word, syllable and phoneme recognition results
  - Exploit ASR output graph instead of 1-Best




- **Consider Scalability**
  - Current search approach not applicable for 10k hours archive
  - Evaluate efficient implementations and alternatives

# Thank you!