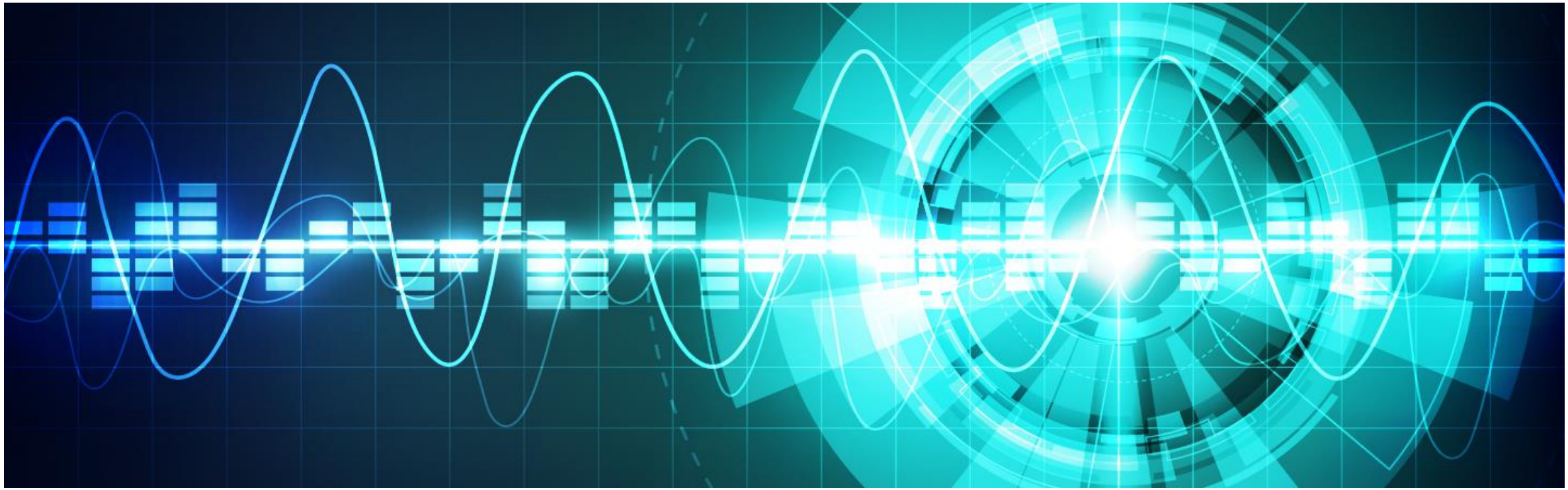
STRUCTURAL ANALYSIS IN AUDIO DATA

Detect structure and information in non machine readable data

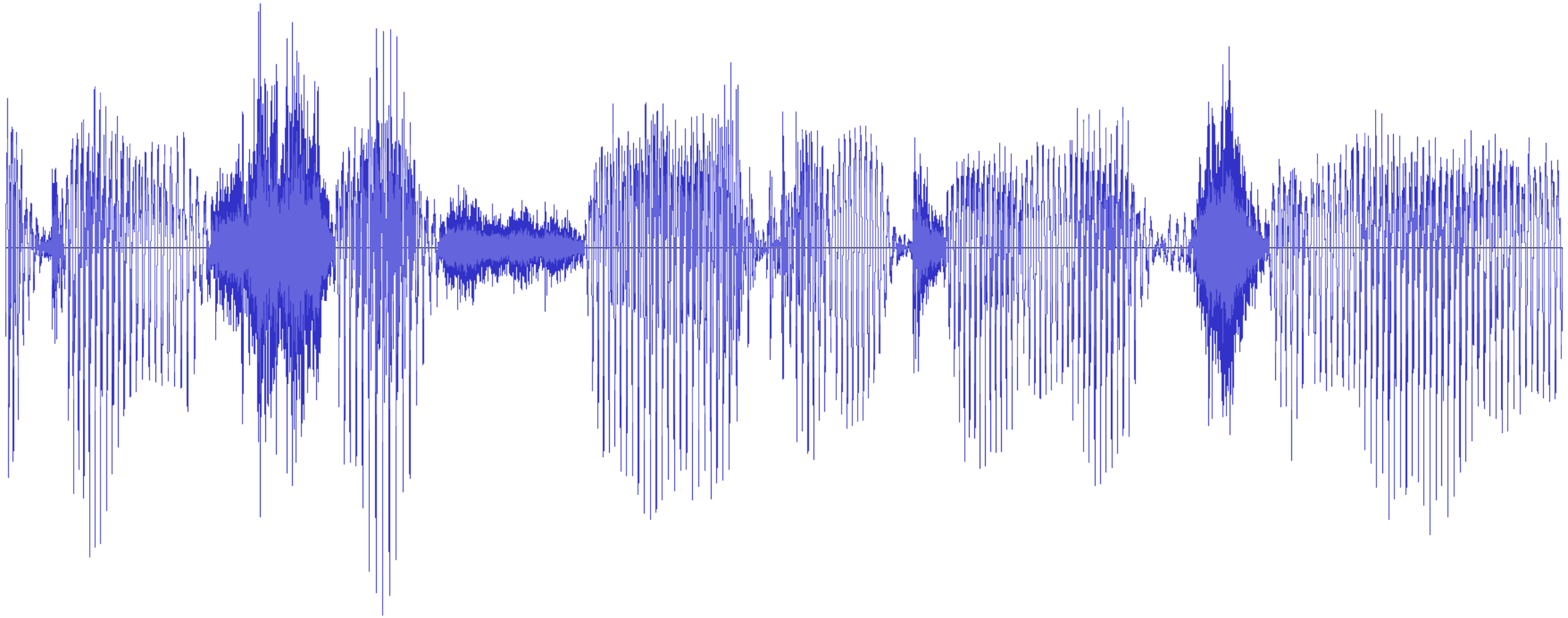


Fraunhofer IAIS

AGENDA

- Introduction into Audio Mining algorithms
- Architecture of an Audio Mining system
- Classification Models
- Challanges for structural analysis and concept detection
- Benchmarking
- Demo
- Open Source

THE ROAD FROM DIGITS TO MEANING



Fraunhofer IAIS

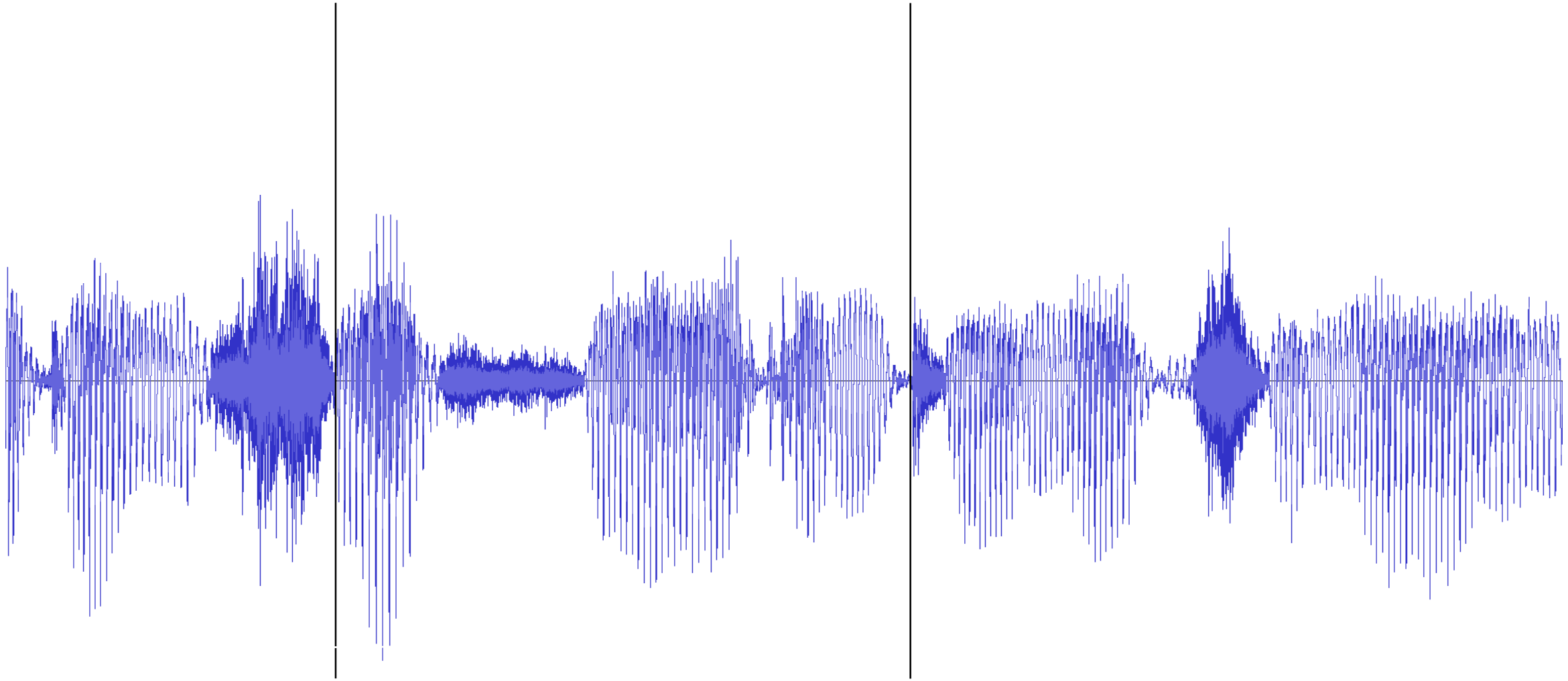
The road from digits to meaning

- The representation of audio signals in a computer are series of number. This representation is very well suited for the playback of audio files, but it is not suitable for search and retrieval tasks. Therefore the information, which is contained in the audio signal, needs to be extracted. This means that the desribing information of a media file gets extended by analysis results. The extraction process is called audiomining.

- The automatic audiomining process extracts information from audio material. The same task can be performed by humans. However, this requires lots of resources. Thats why – even though the quality is usually less than that of a human – audiomining is performed, when not enough resources are available.

Temporal Segmentation

0:00 – 0:10 | 0:10 – 0:35 | 0:35 – 1:00

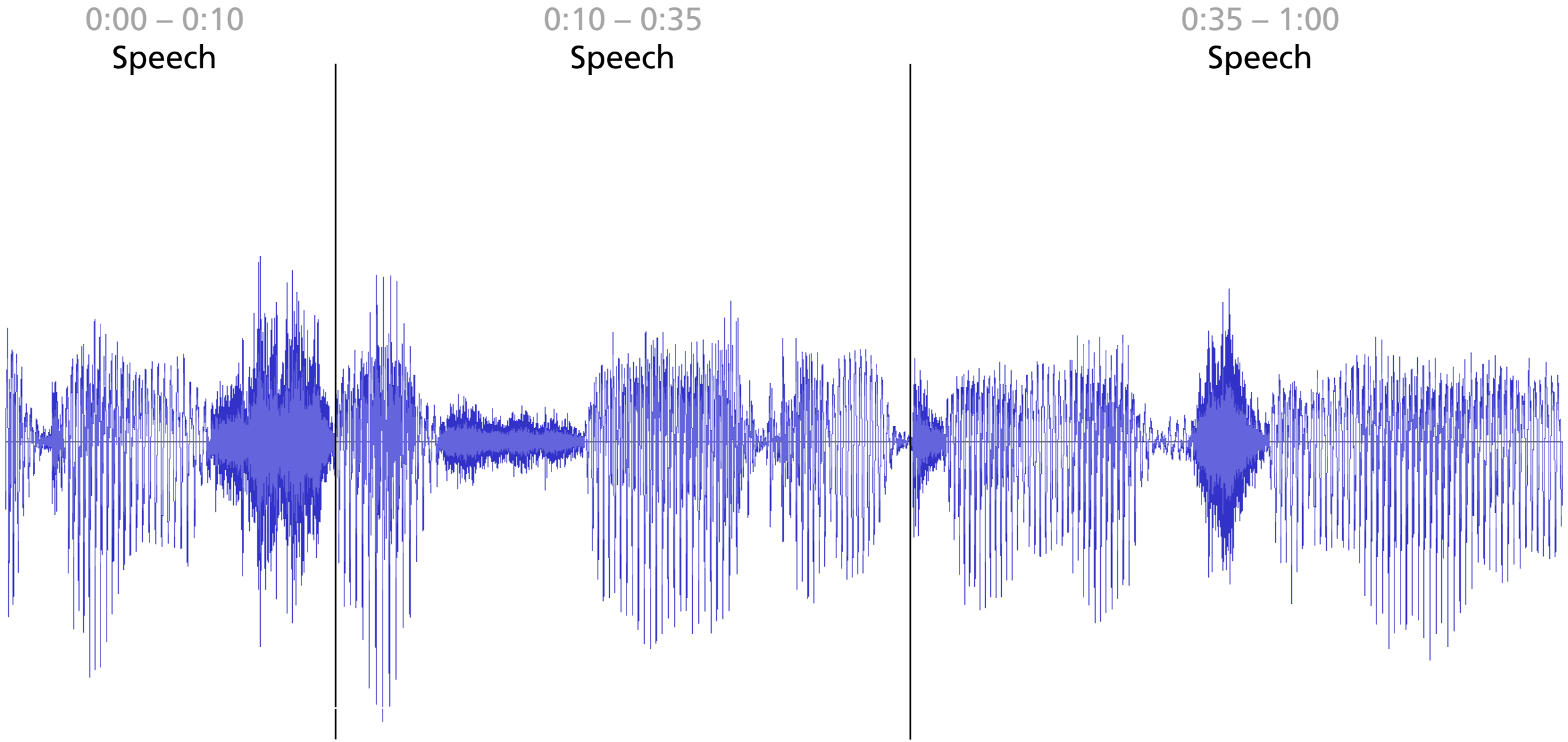


Fraunhofer IAIS

Segmentation

- The first step in audio process is a segmentation. Its goal is to devide the audio signal into sections, which are related to certain features. Such features can be speaker information, speech/non-speech, utterances or topic.
- The found sections, also call segments, are the basis for all subsequent algorithms.
- A combination of multiple different segmentation methods is possible; the subsequent algorithms should then work on a suitable segmentation.

Speech/Non-Speech Detection

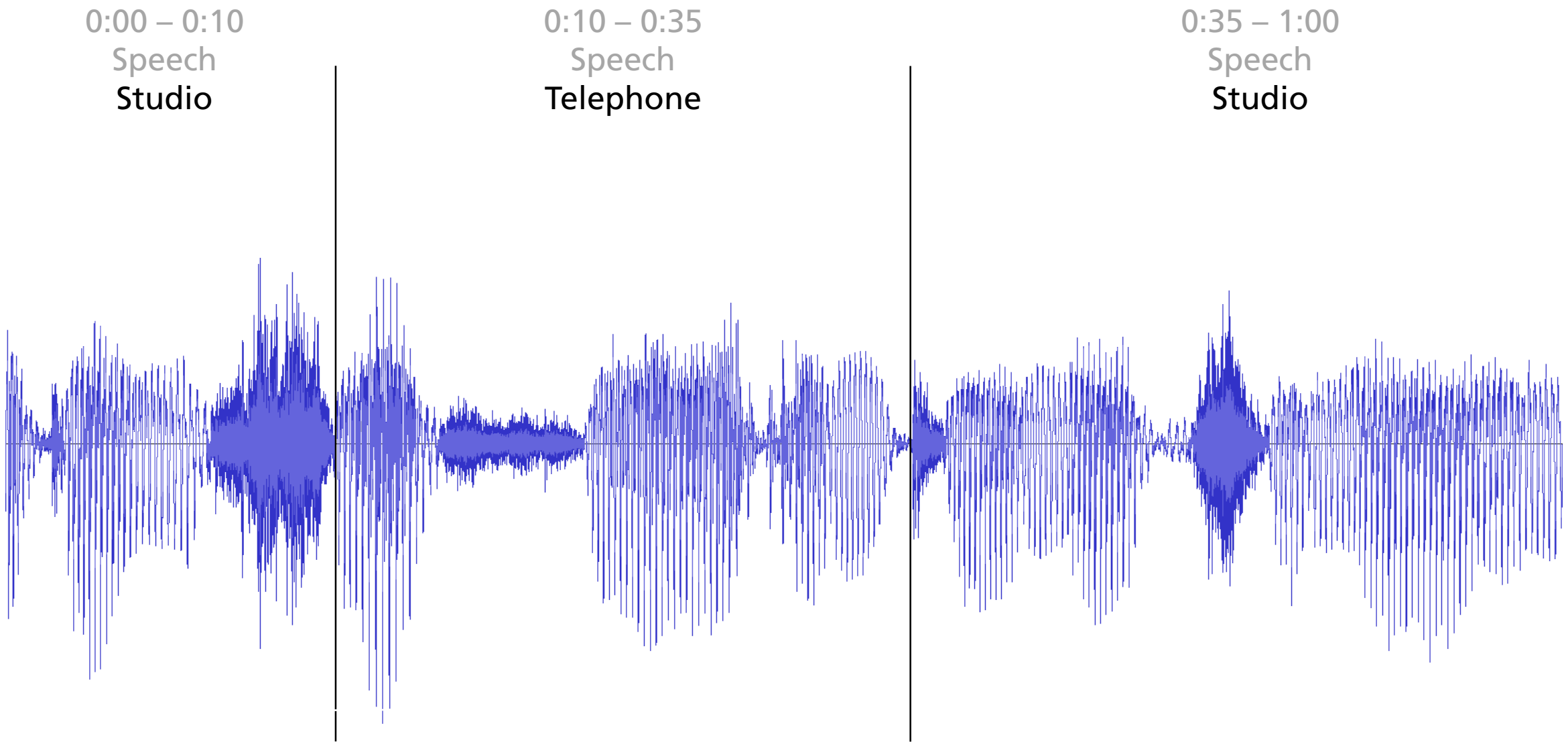


Fraunhofer IAIS

Speech/Non-Speech Detection

- Of importance for the quality of most algorithms is, that those are only applied on suitable data. An algorithm for clustering of segments according to speakers will not work correctly, if it is applied on segments which do not contain speech and thus no speaker. By applying speech/non-speech detection it is possible to filter out such irrelevant segments.
- However if this algorithm does errors, the following algorithms will either not be performed at all or need to work on non speech segments, which they werent trained on

Concept Detection



Fraunhofer IAIS

Channel Detection

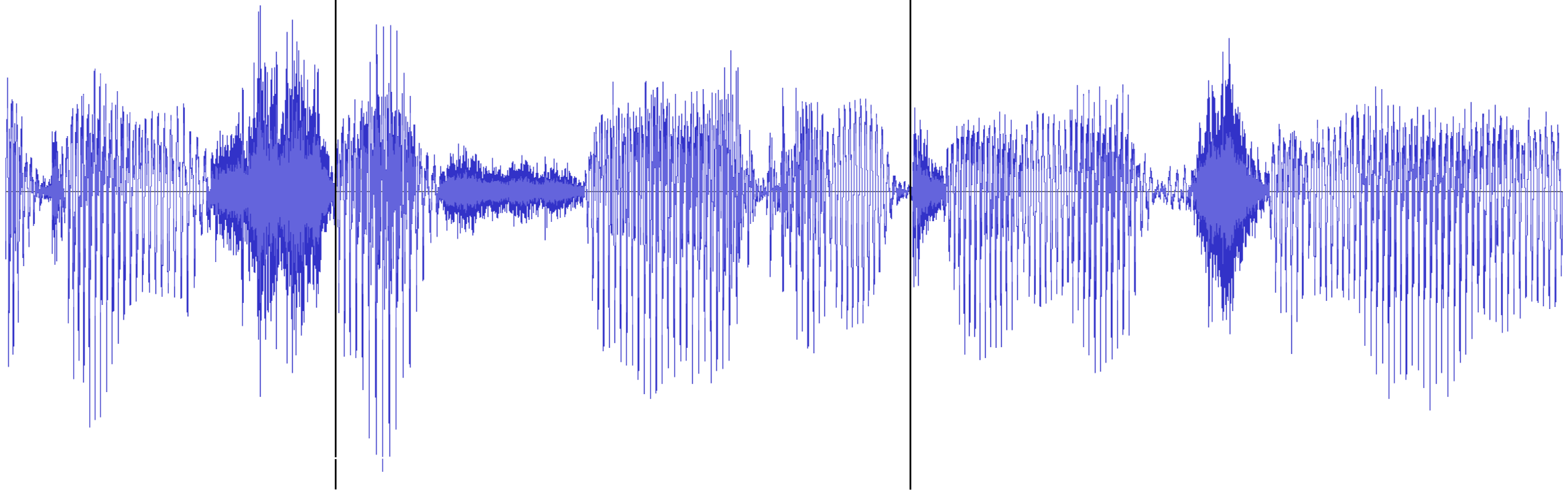
- Goal of the channel detection is to extract the acoustic situation in which a audio signal was recorded. Such situations could be studio environments, telephon situations and many more. The extraction of this information provides the possiblity to adopt speech and speaker recognition models to improve the quality of those models.

Language Recognition

0:00 – 0:10
Speech
Studio
German

0:10 – 0:35
Speech
Telephone
German

0:35 – 1:00
Speech
Studio
German



internal

Fraunhofer IAIS

Language Recognition

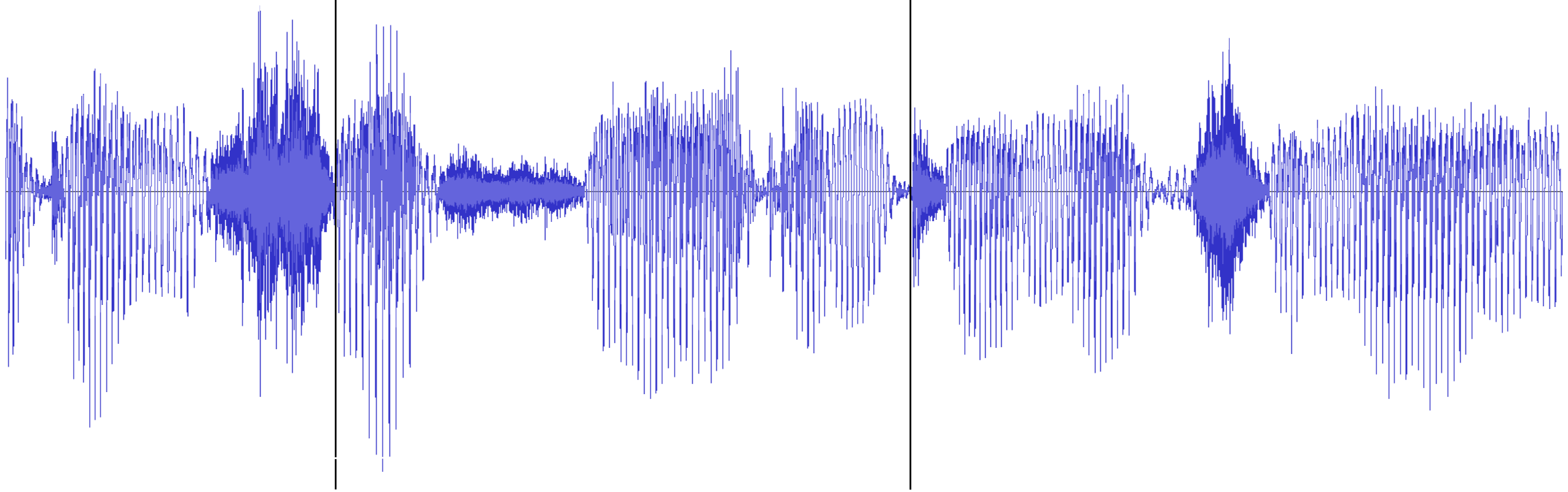
- The language detection determins the language a speaker in a specific segment is speaking. On basis of this information a selection of a specific speech recognition model is possible. If the speech recognition is performed using the incorrect language, it would generate output. However, as its the incorrect language, the output isnt of any use.
- A requirement for a successful language recognition are speech segments, which contain spoken words of only a single language. In cases, where multiple languaes are spoken during a single segment, the algorithm cannot decide correctly and thus the result will be erronous.

Gender Detection

0:00 – 0:10
Speech
Studio
German
Male

0:10 – 0:35
Speech
Telephone
German
Female

0:35 – 1:00
Speech
Studio
German
Male



internal

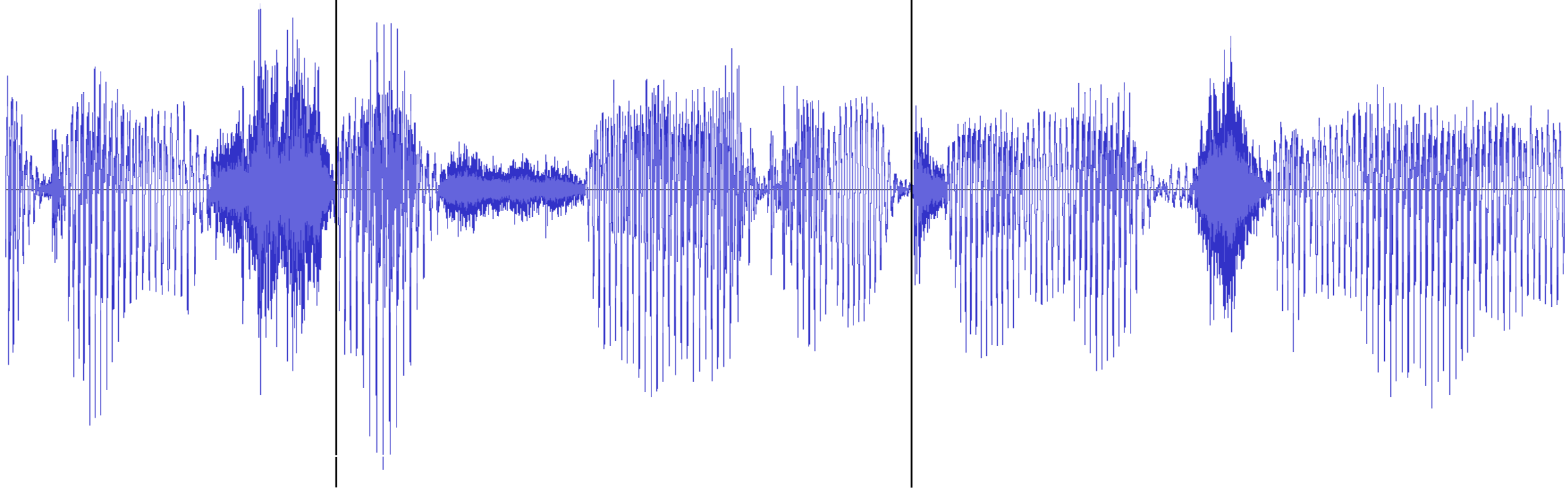
Fraunhofer IAIS

Gender Detection

- The task of gender detection is to determine the gender of the speaking person in a segment. This information can be used for subsequent analysis steps, but can also be used for search requests. For example it would enable the user to search for files which contain only male or only female speakers which talk about a certain topic.
- Based on the results of gender detection the speaker which are used for speaker recognition can be restricted to the correct gender. This would also prevent conflicts between the results of both algorithms.
- Speaker clustering can use the gender information to prevent the clustering of segments, in which different genders were detected, into a single cluster.

Speaker Clustering

0:00 – 0:10	0:10 – 0:35	0:35 – 1:00
Speech	Speech	Speech
Studio	Telephone	Studio
German	German	German
Male	Female	Male
Speaker A	Speaker B	Speaker A



Fraunhofer IAIS

Sprecher Clustering

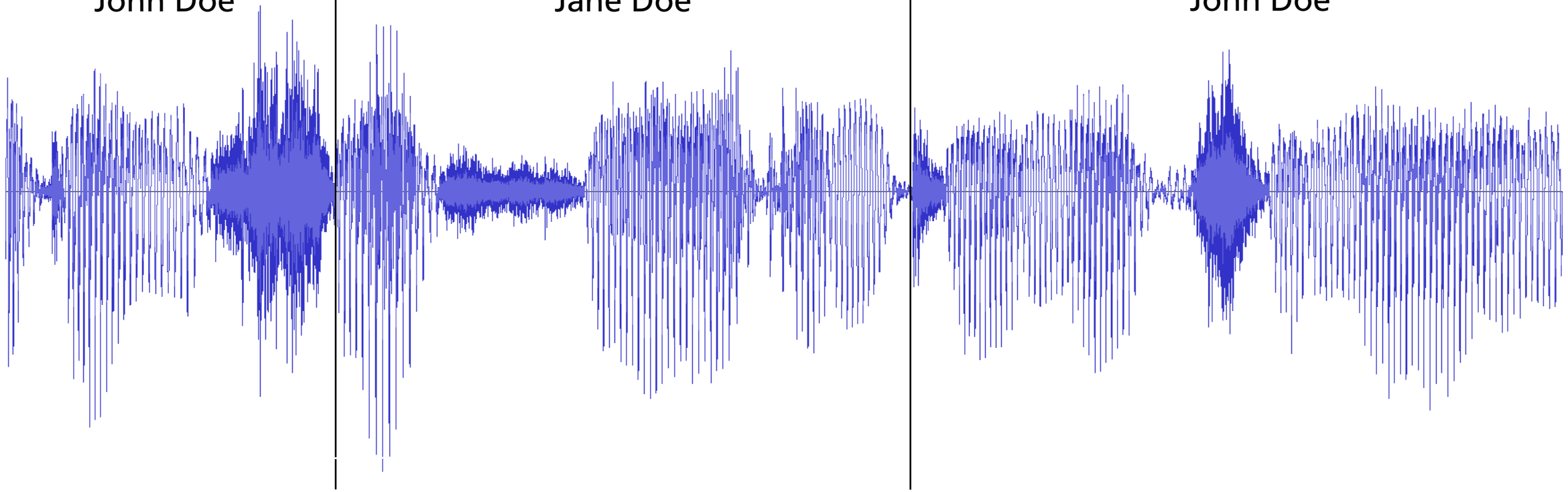
- Speaker clustering assigns each segment of an audio signal to a group, depending on the speaker which speaks in each segment. The algorithm assigns every segment of a single speaker the same group. For each different speaker, different groups are used, thus there should be a correspondence between speaker and a group.
- The information of the speaker clustering results are only relevant within a media file and are not relevant for search requests. But the information, if visuallized, provide the user with additional means for navigation and enabling them to skip unrelevant speaker segments and speed up their work with the media file.
- Results from the gender detection can be used during speaker clustering to prevent the assignment of two segments into a single cluster, where the gender does not match.
- The speakers which are to be grouped do not need to be known beforehand.

Speaker Recognition

0:00 – 0:10
Speech
Studio
German
Male
Speaker A
John Doe

0:10 – 0:35
Speech
Telephone
German
Female
Speaker B
Jane Doe

0:35 – 1:00
Speech
Studio
German
Male
Speaker A
John Doe



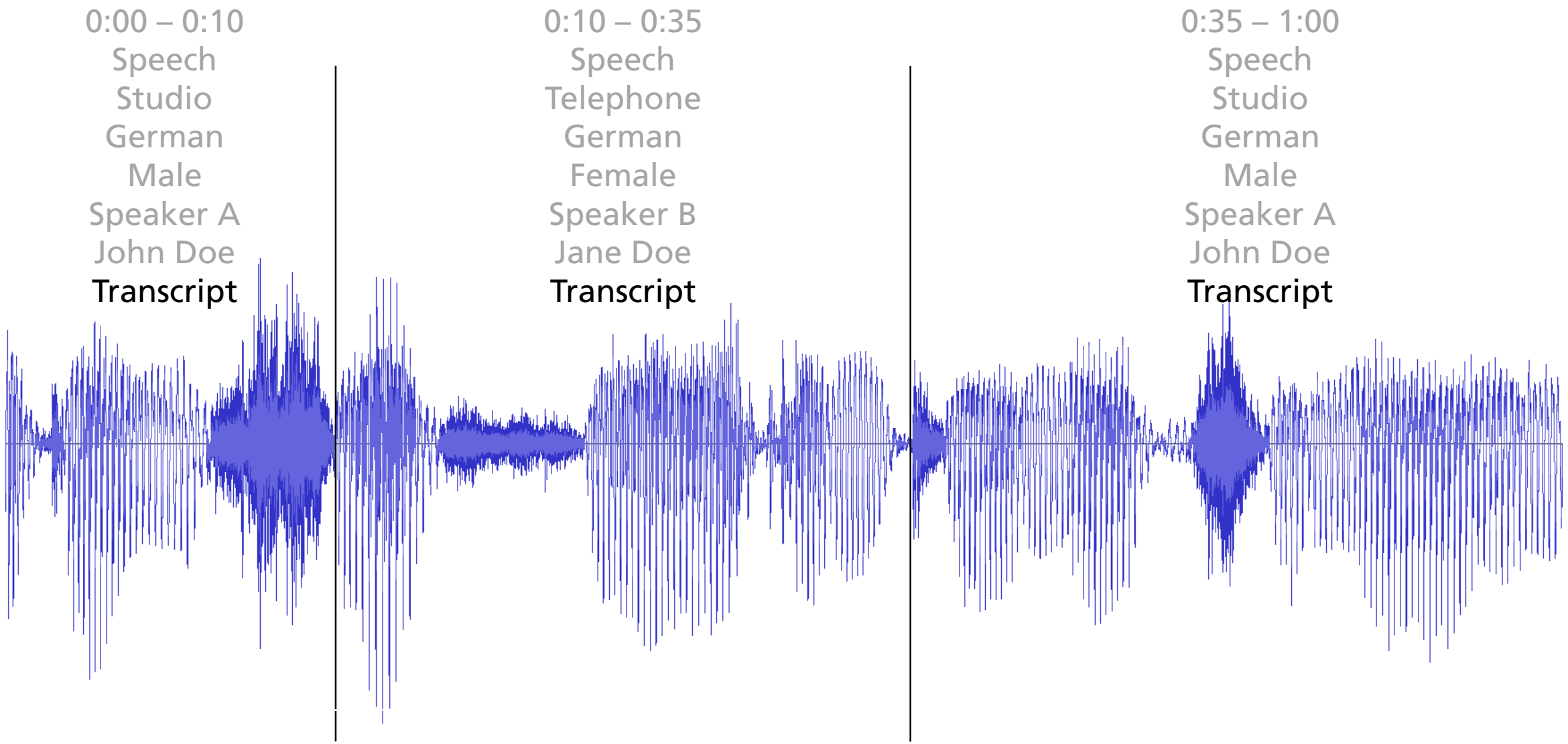
internal

Fraunhofer IAIS

Speaker Recognition

- The speaker recognition compares the speaker of a specific segment with all those speakers present in a speaker database. It enables the selection of either a very specific person or of labeling the segments speaker as „unknown“.
- The speaker information is usable across multiple media files and especially useful for search applications. It also enable the user to perform searches for audio quotes, by searching for speaker and transcript information.
- Instead of the speaker clustering results, the correct speaker can be displayed in a visualization.
- As the algorithm itself does not depend on a mapping between a speaker model and its identifier, its possible to use pure names, ids or links into different databases to identify a certain speaker. This makes it possible to store only relations to different databases to store all speaker related information.

Speech Recognition



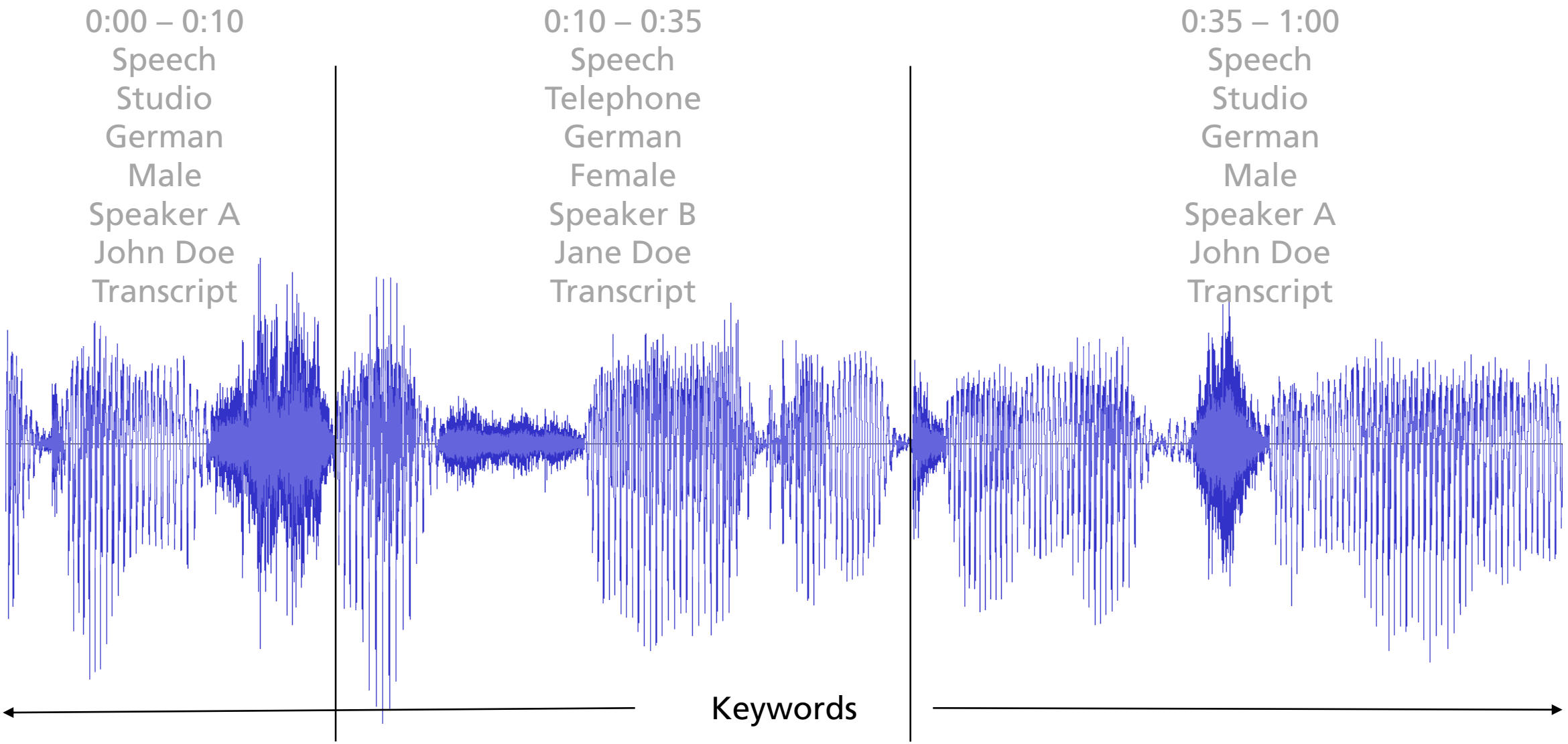
internal

Fraunhofer IAIS

Speech Recognition

- Speech recognition is the most known audiomining application. It transforms the speech signal into text. The resulting text can then be used for full text searches or text mining algorithms which are put on top of the audiomining.
- The speech recognition is usually optimized for a single language and should only be applied on segments, in which this language is spoken.
- More information about speech recognition in the slides of „Automatic Speech Recognition“.

Textmining



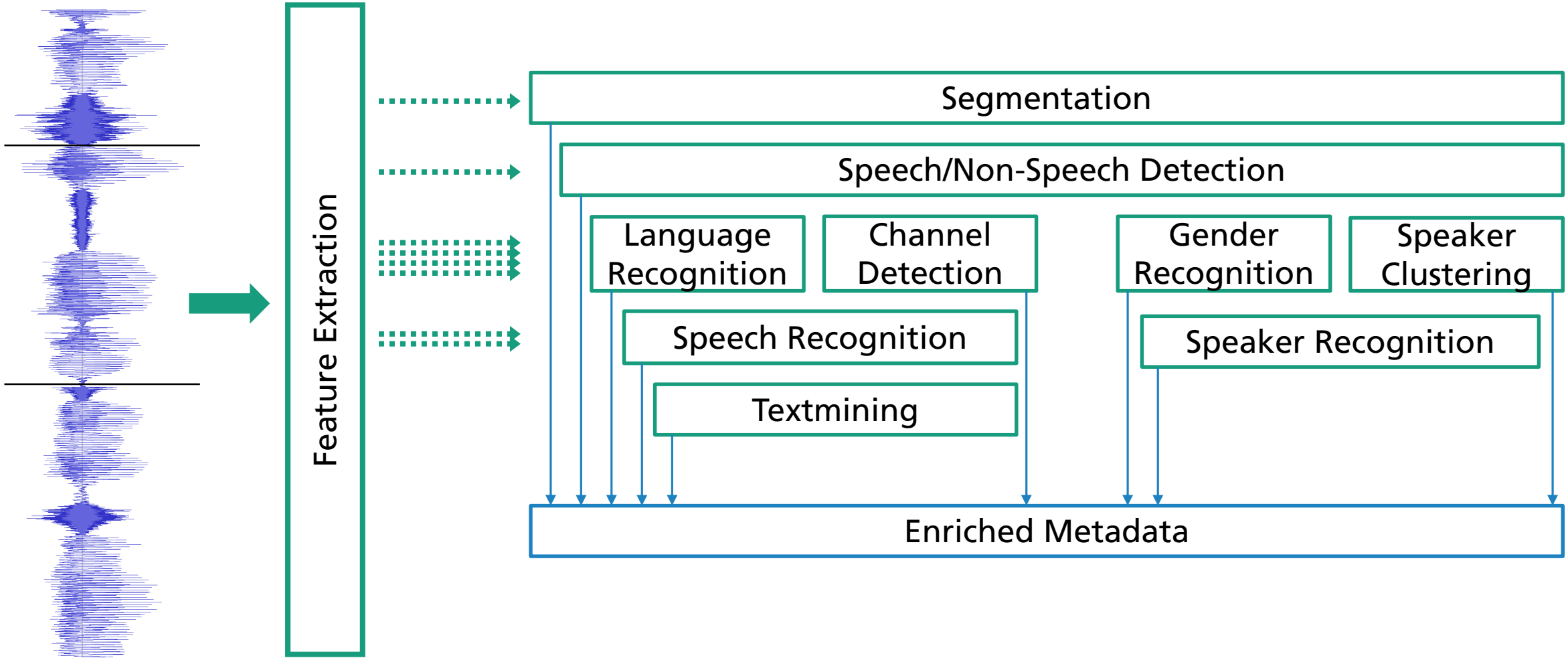
internal

Fraunhofer IAIS

Textmining

- Based on the result of speech recognition, different kind of textmining algorithms can be applied. Some examples include the extraction of keywords, which are especially relevant for a media file and thus describe the content, the classification of entities like persons, places, organizations and so on. Those extracted informations are also available for search and recommendation applications.
- Based on the transcript of a spoken text a topic segmentation can be perform to improve the quality of the search and recommendation applications and search results.

Architecture



Fraunhofer IAIS

Architecture

- Each applied algorithm usually uses individuell features, which are optimized for a specific task, which are extracted from the audio signal.
- Some algorithms are dependent on the output of other algorithms, other are not and can thus be performed in parallel. Algorithms which can be performed in parallel are displayed next to each other, dependent algorithms are placed below those algorithms on which the depend on.
- Each analysis extends the metadata of a file. After applying all different algorithms, the media file is described in a way which is suitable for search and recommendation applications as well as for visualization in user interfaces.

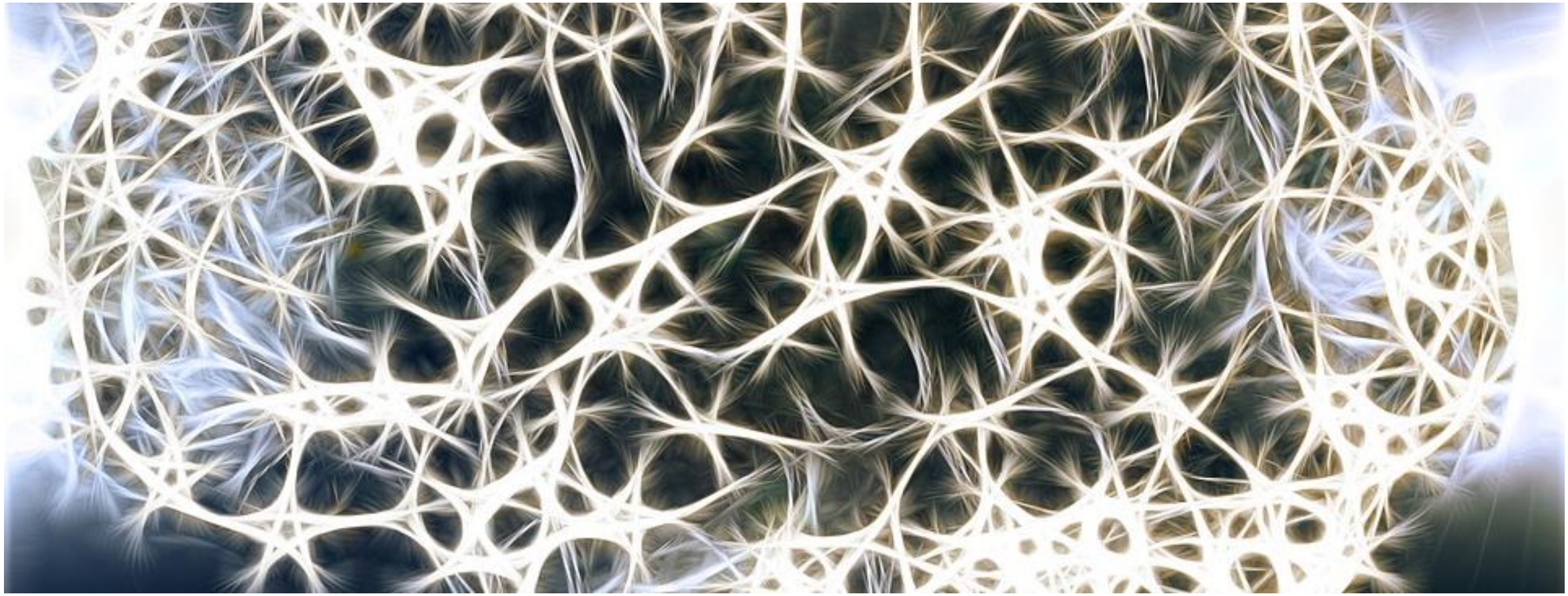
Models – Gaussian Mixture Model

Fraunhofer IAIS

Models – Gaussian Mixture Model

- Based on the training data a linear combination of gaussian distribution is estimated, which described the training data. The gaussian distributions are described by their mean (vectors) and their variance (covariance matrix). During training, for each different training class an own gaussion mixture model is calculated.
- Each gaussian mixture enables us to calculate the probability of a new example being of the tested class.

Models – Neural Networks



Fraunhofer IAIS

Modelle – Hidden Markov Modell

- See foto of the white board explanations.

Challanges

Fraunhofer IAIS

Challanges

- For all algorithms
  - double talk situations
  - Not well adjusted audio signal
  - Background noise
- Segmentierung:
  - Very fast speaker switches
- Sprache/Nicht-Sprache
  - Very quiet speech
- Sprachenerkennung
  - Language switches within a speaker segment

Benchmarking

Fraunhofer IAIS

Benchmarking

- The objective evaluation of the quality requires testdata and reference annotation for that testdata.
- As testdata can be fundamentally different, the evaluation of the algorithm should be done on data which is present in the final application of the algorithms.
- It needs to be decided, what exactly is correct.

Reference Annotation

Fraunhofer IAIS

Reference Annotation

- The automated measurement of the quality can only be done with ground truth references. Those references need to be created by hand. The more test material is used for the evaluation, the more accurate is the estimation of the algorithms quality.
- The measurement should be performed on different examples to provide some variance of the data, to prevent hitting a sweet spot of the algorithm or vice versa.

Whoms quality is measured?

Fraunhofer IAIS

Whoms quality is measured?

- It is essential to measure the correct quality. The best quality measurements are useless if they are performed on data which does not fit to the data which is present in the later system.
- Here two example transcripts are presented, which give you an idea of how different files can differ in theire vocabulary.
-

  foto domain:
  this means I take the animals in the seeker, into the visor and place the AF field with the help of the joystick on the position

  broadcasting domain:
  Rescue services are waiting for daylight to assess the full impact of Hurricane Michael, which made landfall on Wednesday afternoon as a category four storm with 155mph (250km/h) winds

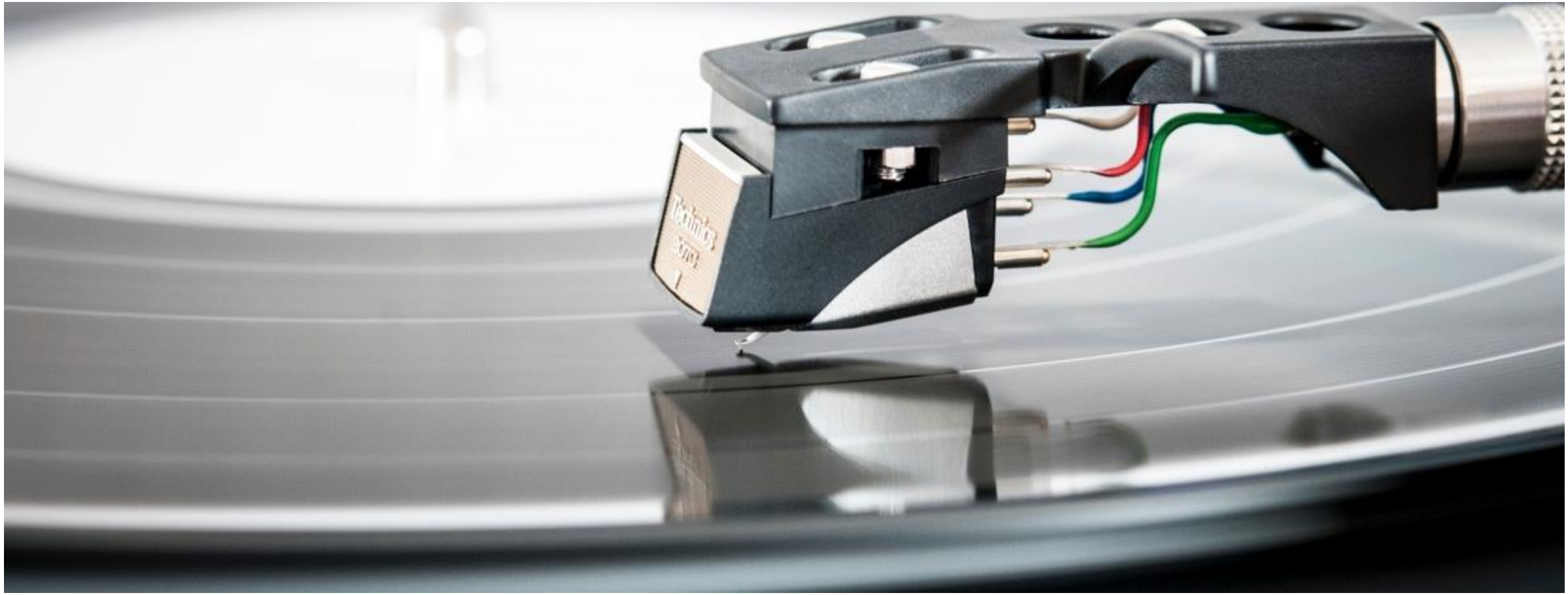
Whats correct?

Fraunhofer IAIS

Whats correct?

- For each application you are required to decide, which result is the actually expected result. In many cases this is not clear.
- Example:
  For some applications a speaker based segmentation is more useful than an utterance based segmentation. If you algorithm does the one while you require the other, the benchmarking should reflect this.
  Especially the segmentation results require the use of tollerances, as even the human cannot decide where exactly the borders of two segments are. The algorithm should provide similar results, but not to the millisecond.

Example Segmentation



Fraunhofer IAIS

Example Segmentation

- „Guten Tag, hier ist die Tagesschau in 100 Sekunden. USA fordern Übergangschef für internationalen Währungsfonds.“

- 2 Segments, a single segment each
  Guten Tag, hier ist die Tagesschau in 100 Sekunden.
  USA fordern Übergangschef für internationalen Währungsfonds.

- 1 segment, as there is only a single speaker present.
  Guten Tag, hier ist die Tagesschau in 100 Sekunden. USA fordern Übergangschef für internationalen Währungsfonds.

- Which do you require in your application? Measure that!

Fraunhofer IAIS

Which is the correct error measure?

Fraunhofer IAIS

Which is the correct error measure?

- The selection of a suitable error measure is crucial for the validity of the benchmarking. A speech recognition can have a low error rate, even though it does not recognize all words of importance.
- Usage of word error rate versus entity error rate for speech recognition: Depending on the application you need to decide which errors are of importance and which can be tolerated. Only those information which are used later on need to be of good quality.
- Another example is the assessment of speaker recognition. Assume the reference data contain the name of each present speaker; the speaker database which is used during benchmarking only those which are relevant for the later application. In such a case you need to decide how to work with such discrepancies between the annotated data and the application data.

Individual vs. wholistic evaluation

Fraunhofer IAIS

Individual versus wholistic evaluation

- Partial evaluation:
  - Each algorithms quality is measured independently based on the assumption, all information it is based on is correct.
  - This evaluation is useful for the selection of different models of an algorithm, as other sources of errors are eliminated.
  - It also provides means to track down the source of recognition errors.
- Wholistic evaluation:
  - Error are accumulated throughout the chain of all algorithms. If a speech/non-speech recognition does an error, the speech recognition is applied according on that result. The result corresponds to the result in a real application.
  - The search of the cause of errors is hardly possible, as it is not always clear which original error lead to which result.

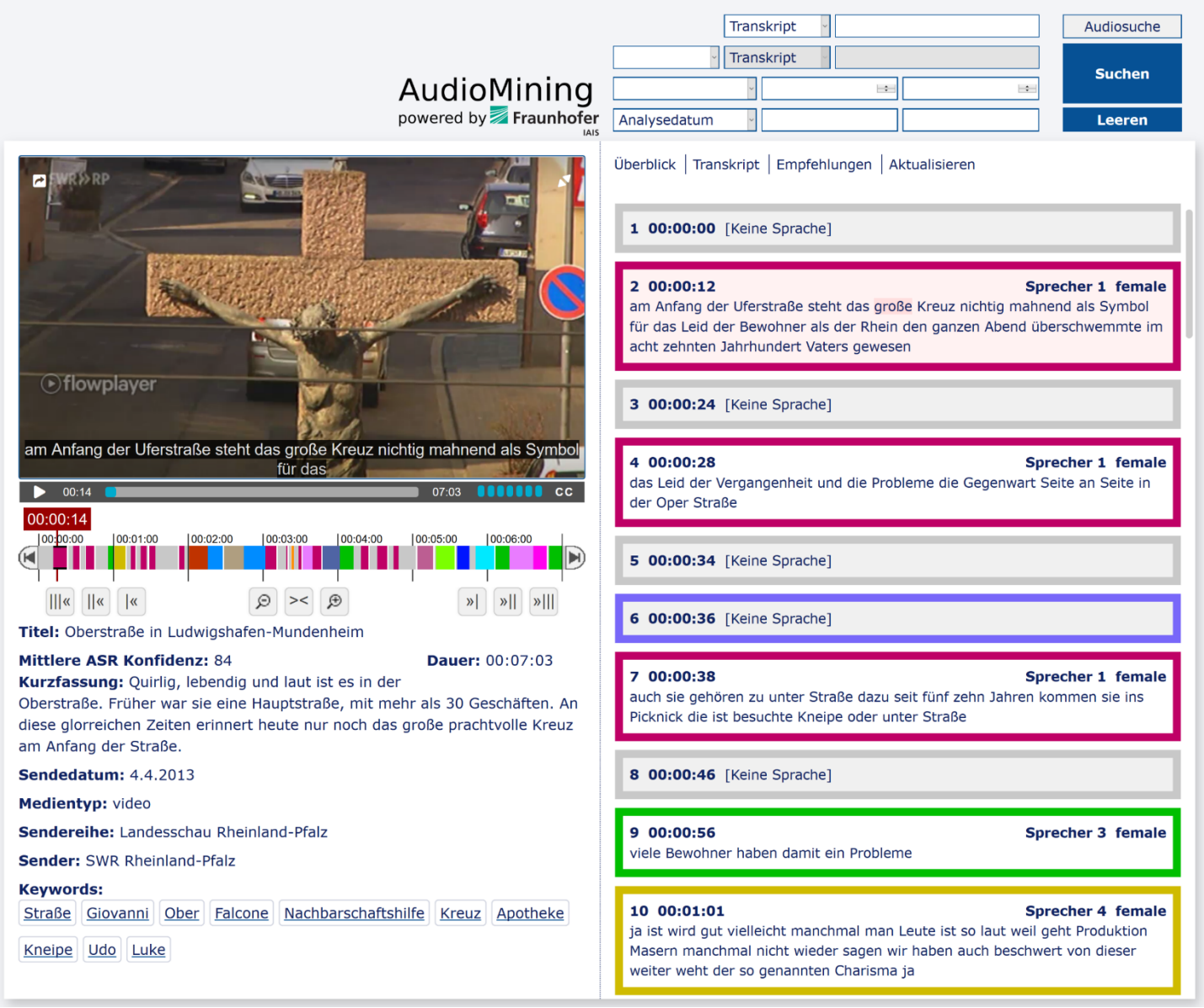
Benchmarking – Does it even makes sense?

Fraunhofer IAIS

Benchmarking – Does it even makes sense?

- Yes it does. Dependent on your application, decide which metadata are really relevant for it. Based on those information, find a way to describe those objectively and measurable. Decide on an error measure to measure the relevant information, in opposite to data which is produced by the algorithm but not required in the end. Then automate the benchmarking, so that you can repeat the benchmarking when adopted any part of the system.
- Benchmarking is used as quallity assurance during software changes and to estimate the performance of the system on specific data.

Demo



Fraunhofer IAIS

Demo

- Depending on the application a good indexing or visualization of the analysis results is required.
  - The information of segmentation is mostly a basis information for algorithms which are build on top, an indexing of segmentation information is only useful in combination with further metadata.
  - The visualization of segment information is important for the user to improve his navigation within a media file and to understand the structure of the file faster.

Open Source Software

Fraunhofer IAIS

Open Source Software

- Segmentation:
  - Keras (MIT licence), Tensorflow (Apache 2.0 licence), theano (3-Klausel BSD licence)
- Concept detection:
  - Keras (MIT licence), Tensorflow (Apache 2.0 licence), theano (3-Klausel BSD licence)
- Speaker-Clustering and speaker recognition
  - Alize, available under LGPL Licence
  - Kaldi, available under Apache Licence 2.0
- Speech Recognition
  - Kaldi, available under Apache Licence 2.0

Who am I?

- **David Laqua**
- Research Engineer
- Telefon: +49 2241 / 14 2725
- Email: david.laqua@iais.fraunhofer.de

Fraunhofer IAIS

Disclaimer

Responsible contact: David Laqua
E-mail: david.laqua@iais.fraunhofer.de