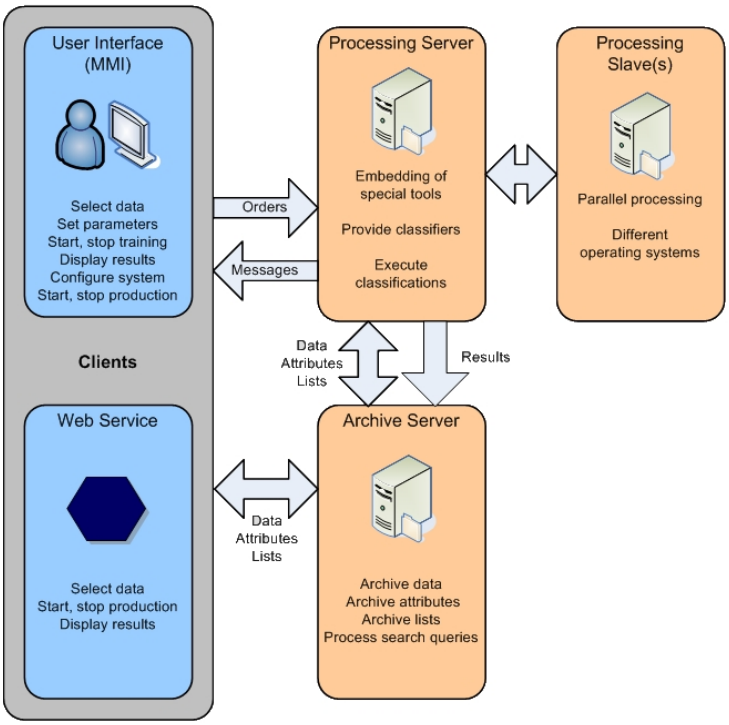
SIPAC

SIGNALS AND DATA IDENTIFICATION, PROCESSING, ANALYSIS, AND CLASSIFICATION



SIPAC KEY FEATURES

SIPAC is a framework for automated and interactive processing of signal and data files. Processing tasks are user definable by taking advantage of splitting the task in sub modules, which are combined in so called signal flows. The system is scalable with respect to processing power and processing software. Data and result files produced by the SIPAC Processing Server can be stored in the SIPAC Archive Server for later use.

SIPAC offers a pre-defined set of processing and analysis modules and organizes the handling of separate software solutions, in particular those of different suppliers, running on machines with different operating systems. An expert for each of the modules defines, configures and administrates all necessary elements of the implementation on the SIPAC platform. After this step, the module can easily be used within a certain work- flow on the basis of web services. The integration of the modules into the environment of the BPMS (business process management system) solutions is possible without all know-how of the dedicated implementation.

Additionally, SIPAC organizes the scalable server farm with regard to load distribution and error handling.

SIPAC follows the SDIA principle (Software Defined Intelligence Architecture) which stands for flexibility, modularity and scalability on standard server hardware.

The Processing Server is the central processing unit of the system. The processing capabilities can be extended using SIPAC slaves for additional hardware. The Archive Server is the data storage of the system. The clients run the graphical user interfaces for the processing and archive services.

SIPAC can be accessed and controlled also via web services from other applications. In particular the integration into BPMS systems is possible. In fact it means that SIPAC offers automatic services for the processing and analysis of signal and data files, which alternatively could be handled by human operators. SIPAC offers sophisticated robot services especially for the content processing and analysis of data, signals and information from different media.

SIPAC FRAMEWORK

- Platform independent
- Scalable in size for required hardware configurations
- User definable flow
- Easy integration of user-owned and external software

SIPAC SERVER FARM

- Parallel processing in a pool of slave-computers
- Automated processing of mass data
- Interactive analysis

SIPAC PROCESSING SERVER

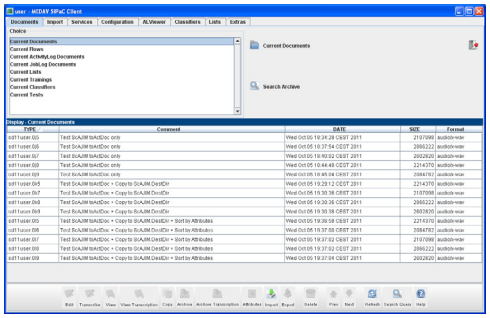
- Different analysis, processing and classification methods available
- Trainable classifiers
- Extendable by a user development kit (UDK)

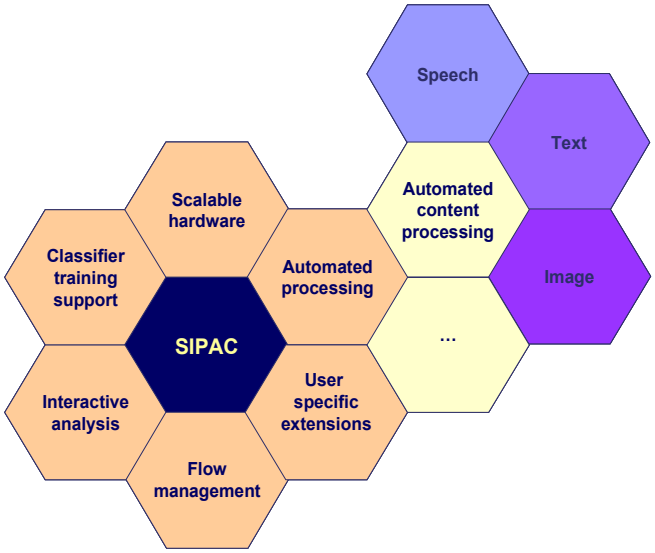
SIPAC ARCHIVE SERVER

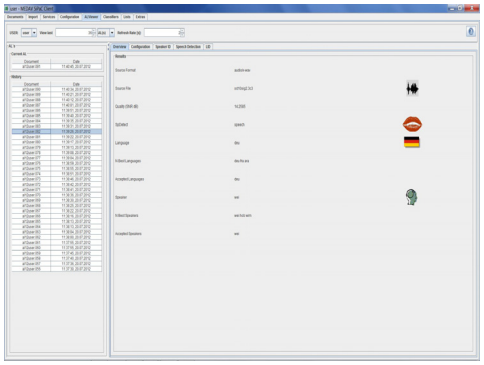
- Storage of data (e.g. inputs and processing results) and metadata
- Intuitive retrieval of results

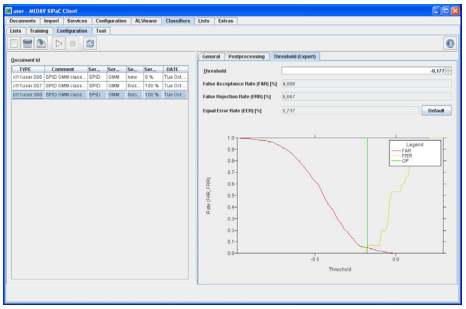
SOFTWARE DEFINED INTELLIGENCE ARCHITECTURE (SDIA)

IMPRESSIONS









BENEFITS

FOR THE USER

- One interface for processing all types of information (speech, text, image etc.)
- Support for automated processing

FOR THE ORGANISATION

- Higher throughput of processed data
- More time for more complex tasks due to savings on routine tasks
- Easy data communication / exchange possible between the departments
- Amortisation of a SIPAC investment after a short time – overall money saving

REFERENCES

SIPAC has been installed for various customers with different requirements spanning diverse system sizes from the stand-alone workplace system based on a single high-end notebook to the scalable system with up to 30 client workplaces on high-performance cluster architecture.

SIPAC was introduced to the market in 1999. New releases are launched every year.

THE SIPAC SYSTEM

SIPAC is a system for automated and interactive processing of data. The processing can be made using different software, either provided by Saab Medav Technologies, by the user or by a third party. The system is scalable depending on the user requirements.

THE SITUATION

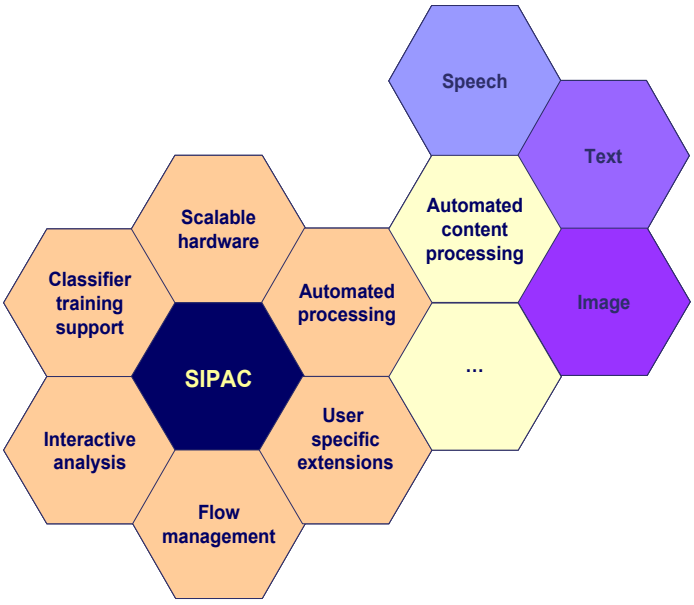
More and more data must be analysed by the user. Experts are busy with routine tasks. In addition, they cannot improve existing solutions. Basic users cannot apply the solutions designed by the experts.

SIPAC is the instrument for automated and interactive processing of data. The expert may define new flows interactively. Basic users may invoke the automated processing of mass data defined by the experts.

SIPAC production and results can also be controlled from other software using web services.

SIPAC PHILOSOPHY

The SIPAC system is very flexible and can be configured according to the needs of the users. The main components are shown in the following picture:

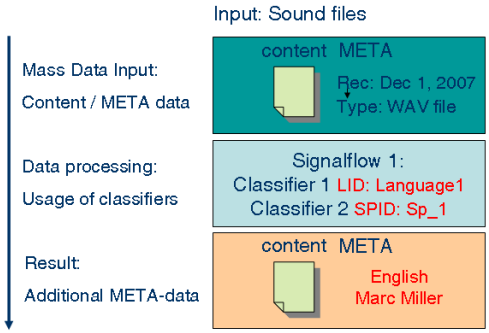


- In the SIPAC Processing Server, automated processing is performed using different classifiers, analyzers and procedures. Saab Medav Technologies provides a variety of processing elements such as classifiers, analyzers and procedures for content processing of speech, text and image.
- Scalable hardware provides flexible computational power involving different computers (even with different operating systems). According to the task and the current situation, computers can be added to or removed from the system. In the SIPAC Archive Server, all relevant data (e.g. input data, results and even the processing elements) and metadata can be stored.
- Also user-owned software can be used inside SIPAC. This software can be embedded by the user himself or by Saab Medav Technologies.
- Flow management typically tries to compose procedures and processing elements. The goal is to automatically analyze, filter, dispatch, channelize and process masses of data. Hereby, the challenge is to not only plainly concatenate the stages of processing but to set the processing into a context and relationships. This works up to complex tasks and achieves high throughputs, efficient information mining and useful limitation of the output to concentrate important information. For example, if the language of a file is determined, the following software may use this knowledge for language specific processing.
- SIPAC can be used for interactive analysis, automated processing and classifier training by calling the provided services. Services are for example file content analysis or language detection of speech samples.
- Trainings are performed to create or improve classifiers. A training environment is provided to support the handling of trainable software and classifiers.

WORKING WITH SIPAC

Each data asset processed by SIPAC is handled as two inseparable files: a body file as binary data and an attribute file containing metadata.

When sending input data to SIPAC, some attributes (e.g. date, time and source) are created automatically and handled in the metadata file. During processing, the results of analyses or classifications are added to the metadata. Other important information that may come from the flow management can also be carried as metadata through the whole processing chain and is available afterwards for further use.



Following pages describe

- the parts of a SIPAC system, especially the archive server, the processing server and the clients
- the software that can be embedded into SIPAC,
- the configuration of the SIPAC system, and
- the technical data.

SYSTEM PARTS

A SIPAC system consists of the following software parts:

- The SIPAC processing server – for data processing, analysis and classification
- The SIPAC archive server – for storing and retrieving data
- SIPAC slave(s) – for distributed processing tasks to different computers and/or operating systems
- SIPAC client(s) – interface for users to control the servers and tasks
- User's development kit – for implementing own extensions into SIPAC

For one SIPAC system one archive server, one processing server and one client is necessary. Bigger systems may consist of for example one archive server, one processing server, 30 slaves and 10 clients.

SIPAC PROCESSING SERVER

The SIPAC processing server is the central processing unit of the system. More hardware can be added for faster processing or extended processing capabilities. For each computer attached to SIPAC, a slave software license is required. If SIPAC slaves are available with the required functionalities, the SIPAC processing server is able to distribute tasks across the attached computers for optimal performance.

The flow management of the SIPAC processing server is based on user definable program flowcharts (simply also known as flows) written in a XML-based command meta-language. The command set covers all functions needed for controlling the automatic processing. The interface for changing or creating flows is very intuitive and the system can support the user when manipulating the flows.

The SIPAC server provides an Application Programming Interface (API) with connection via TCP/IP.

SIPAC ARCHIVE SERVER

The SIPAC archive server is the data storage of the system. All input data and metadata can be stored there. For fast data access, it provides an up-to-date instrument of retrieval, such as categorisation of data and full-text-search.

The core functions of the archive server are:

- Labelling the documents with an internal unique ID for reference
- Saving and compressing files as archive documents
- Storing metadata for documents (also known as attributes)
- Protection of documents against changes
- Configurable permission for the deletion of documents ensure archive integrity
- User and password management
- Access documentation by log files
- Capacity and access performance is limited only by hardware
- Application Programming Interface (API) – connection via TCP/IP

SIPAC CLIENTS

The clients run the graphical user interfaces for both servers, the SIPAC processing server and the SIPAC archive server. The clients are customized and provide specific graphical interfaces for the different tasks (classifier training, flow definition, service application, monitoring, results and data retrieval). Functions for system administration are also available.

For each operator workplace, a client license is necessary. For additional tasks, additional licenses are necessary:

- administrator/expert interface for system setup and control
- training interface for the training of the involved software

All components communicate via TCP/IP protocol. If distributed machines are in use, LAN connection is required accordingly.

TOOL SETS

Different tool sets can be plugged into SIPAC. The tool sets can be bought from Saab Medav Technologies as commercially available packages (available tool sets: Common, Speech, Text, ARGUS, Network – see next pages) or created and compiled on demand either by Saab Medav Technologies or by the user himself. Third party products can be part of a tool set. Available tool sets are described on the next pages.

USER DEVELOPMENT KIT

Optionally, a user development kit (UDK) is available. Using the UDK, the expert can embed own software into the SIPAC environment. After including the customized software, the new features will fit seamlessly into SIPAC and can be used for processing.

The UDK works command line-oriented. Programming knowledge is necessary. An expert training is available.

HARDWARE

The use of hardware is very flexible: notebooks or standard computers can be used for small systems and rack-type servers for large systems. Different operating systems are supported.

PROCESSING ALGORITHMS

A large variety of processing algorithms / software can be embedded into SIPAC.

Often it is recommended or even necessary to train the parameters of the respective software on data of the application in order to obtain optimal performance. For audio signals for example, the channel characteristics or language models may be trained separately.

Software can be provided by

- Saab Medav Technologies: offers tool sets for classification of audio, text and image processing (others on request),
- The customer himself: e.g. task specific software developed by the user,
- Third party: software with a special focus, e.g. for automatic translation

To start the automated processing, files that have been sent to SIPAC are placed in a file directory (this is a common way of production). From the file directory the files are subsequently (and/or parallel) processed according to the user requirements. The result of processing is added to the meta-information of the respective file.

Interactive processing provides more breakpoints for the operator to deal with inputs, procedures and results all the way long from the original data in question to the completely processed result. In most cases, procedures are developed, tested and evaluated in an interactive way before they are finished to an accomplished production flow.

TOOL SETS

Tool sets are used as extension modules of the SIPAC system. Depending on the requirements of the users, different tool sets can be added. Experts can create own tool sets using the User Development Kit (UDK).

Available tool sets are described in the following:

TOOL SET - COMMON

This tool set is included in all SIPAC systems. It provides the following functions:

- File type detection
    - More than 1.100 file types can be classified
    - Depending on the file type, further processing can be made.
- Packing and unpacking support, e.g. for zip files.
- Basic classification: Case Based Reasoning (CBR)

TOOL SET - TEXT

The text tool set is used for the classification of files containing text (it partly includes third-party products)

- Language identification
- Entity recognition
- Topic spotting
- Word spotting
- Term translation and
- Full text translation
- Optical Character Recognition (OCR)

Some of the algorithms are language specific and, therefore, training data must be collected beforehand.

TOOL SET - SPEECH

The speech tool set is used for the classification of audio files. Referring to different tasks, the audio signal is analysed / processed with respect to the following criteria:

- Speech detection
- Gender identification
- Language identification
- Speaker identification
- Topic spotting
- Word spotting
- Speech-to-phoneme
- Speech-to-text

Some of the algorithms are language specific and, therefore, training data must be collected beforehand to improve the classifiers for optimal performance.

TOOL SET – ARGUS

ARGUS is a special analysis tool set for images. It provides the following functions:

- Support Vector Machines (SVM) as a general binary classification method
- Concept of generic content based classification (collecting positive and negative samples, defining feature extraction, automated SVM-training and -test, SVM-classification)
- Image file format specific analyzers (check for file manipulations; for jpeg, png, bmp, others upon request)
- Trained classifiers for bitmap- and jpeg-steganography (StegoBitmap and StegoJPG; check for hidden information in images)
- Trained classifier for high textual content (HiTex) in images (independent of the lettering system)
- Heuristic malware detection (retrieval for indication of potential malicious code without using signatures)
- Optical Character Recognition (OCR)

TOOL SET – STRIX

STRIX is a special analysis tool set for captured network data streams. It provides the following functions:

- Detection of encrypted traffic in binary network streams
- Support of Deep Package Inspection (DPI)

WORKING WITH SIPAC

Following, an example of how to work with SIPAC using the tool set speech is given.

- Employ the needed software
- Define the order of processing

The configuration is done in the following steps – both Saab Medav Technologies and the users are able to configure the system.

TASK ANALYSIS

The task has to be defined exactly with respect to classifiers and data volume.

Example: A system shall be configured for the classification of the prevailing language in audio files. The signals are obtained from the telephone line.

- Languages of interest: English, German, Spanish, French, Italian, Danish
- Expected data volume: 1000 hours of audio signals to be produced per day.

This leads to a need of speech classifiers for the requested languages and a calculation of the necessary amount of processing hardware.

CHOICE OF AVAILABLE CLASSIFIERS

If trained classifiers are available for the given task, nothing has to be done. Otherwise, training data must be collected (or purchased) and the classifiers must be trained.

Example: Trained classifiers are available in telephone quality for all languages but Danish. Thus, Danish telephone data must be collected. A classifier for Danish must be trained.

CONFIGURATION OF THE FLOW

The workflow describing the sequence and conditions of the automated processing must be established.

Example: After speech detection, language identification is performed for the parts of the audio signal containing speech.

TEST AND EVALUATION

The flow is activated including the embedded classifiers. Testing is performed in order to ensure correct function and high quality.

Example: Run SIPAC with a set of test files. The evaluation of the language classification results is performed.

After these steps, the SIPAC system is ready for use and can be started in the user scenario.

For optimal performance, it is recommended to perform the evaluation on a yearly basis in order to maintain the high accuracy with respect to changes of the incoming data (e.g. different telephone codecs).

TECHNICAL DATA

GENERAL

- The framework is scalable from a standalone solution to a server farm
- Internal communication via TCP/IP

SUPPORTED OPERATING SYSTEMS

- Windows: XP, Server 2003, Vista, Server 2008, 7
- Linux
- Solaris
- Others upon request (depending on Java platform)

SIPAC PROCESSING SERVER

- Parallel processing of a not unlimited but sufficient number of processes
- Communication with the SIPAC Archive Server (read, write)
- Local storage of intermediate results
- Integration platform and framework for external tools
- Slave:
    - distributed processing according to slave configuration / load
    - implemented tools can run on different operating systems

SIPAC ARCHIVE SERVER

- Storage of data and assigned metadata
- Storage of flows
- Storage of classification and intermediate results
- Storage of processing elements such as classifiers, analyzers and procedures
- Search functionality

SIPAC CLIENT

- Client application (Fat Client)
- Graphical user interface for system control
- Simple role concept: administrator, user

USER DEVELOPMENT KIT

- Apache Ant based tool for compiling tool sets
- Command-line-oriented

Saab Medav Technologies GmbH

Gräfenberger Str. 32-34, 91080 Uttenreuth, Germany
Homburger Platz 3, 98693 Ilmenau, Germany
Phone +49 9131 583-0 • Fax: +49 9131 583-11
www.saab.com • www.medav.de