

## A speech classification system

**Dr. Ulla Uebler**  
MEDAV GmbH  
Gräfenberger Str. 32-34  
D-91080 Uttenreuth  
Germany  
[ulla.uebler@medav.de](mailto:ulla.uebler@medav.de)

### ***ABSTRACT TITLE***

*This paper presents the technique of speech classification and the usage of this technology in a set of scenarios. Speech classification reduced the amount of labour necessary for the analysis of incoming audio signals by classifying the signals to the spoken language, the speaker etc.*

*Different scenarios are presented showing how speech classification can be used in order to deliver the maximum efficiency and accuracy in the scope of information processing.*

### **1.0 WHAT IS SPEECH CLASSIFICATION?**

In times of mass data being available from different sources, the fast and efficient analysis of signals of different types is becoming more and more important.

Speech classification is a means for automatic classification of audio signals. Using these techniques, the incoming speech signals can be classified, sorted and prioritized. Accordingly, human operators will be able to analyse the important signals first and will only have to work on those signals that are relevant to their work.

Speech classification is based on statistical techniques; therefore, it is necessary to provide suitable training material for the task. For example, performing language identification for English and French, requires speech signals in both languages to determine the parameters of the system.

Statistical approaches have a recognition accuracy that depends on the difficulty of the task (distinction between similar languages, the amount of different speakers etc).

This paper describes the analysis of speech data, the deployment of the algorithms and results in information fusion. After some theoretical background, the available classifiers for language identification and speaker identification are presented. The main part of the paper describes possible scenarios for the usage of the classifiers.

### **2.0 THE TECHNIQUES FOR SPEECH CLASSIFICATION**

Most speech classification algorithms, especially the ones described here, are based on statistical measurement. Using this method makes it easy to develop classifiers for a new task or for a special

## A speech classification system

situation. Only a little knowledge of a language or of the physiology of speakers is necessary.

Instead, it is necessary to have training material available for each application. For example, if a classifier is trained to distinguish the spoken language according to say English, French and Spanish, it is necessary to have sufficient labelled training material for each of the three languages. For our approach, “labelled” means that for each speech signal, the spoken language has to be assured; a transcript is not necessary.

For each application or scenario, a classifier is trained according to the particular situation. A general recommendation is that the speech signals used for training should be as similar as possible to the task in hand. Thus, if the speaker shall be determined from speech signals coming from a telephone line, the use of telephone speech for classifier training is recommended.

For special applications, different classifiers can be combined with each other. For example, first the parts containing speech are determined. The language identification classifier then processes the parts containing speech. Those parts containing English are then processed by the word spotting classifier in order to find certain important words.

### 3.0 AVAILABLE CLASSIFIERS

The classifiers in the following description need training material according to the classes that are to be distinguished. Each classifier will be described according to its properties and some typical results given.

#### 3.1 Speech detection:

Speech detection: the incoming audio signal is classified according to speech/non-speech. The result is a time labelling for each speech signal.

Non-speech may consist of data transmission and/or of noisy parts of speech. The training material, especially for the non-speech part, should be carefully chosen according to the application.

In case of data transmission, the complete signal is usually classified as “data” and not further used for speech classification.

Telephone signals, however, consist partly of speech and partly non-speech, the latter being either “noise” such as dial tone or “silence” if the other side of the conversation is not transmitted. The result of speech detection is in this case a time-labelling of the signals giving those parts of the speech that are classified as speech.

Typical accuracies for speech detection are shown in Table 1. Thresholds can be set in according to priorities e.g. to the detection of noise or speech. In this table, the parameters are set optimally not to miss speech, leading to a detection of 98 % of speech. At the same time, the number of false alarms rises, i.e. classifying non-speech as speech.

	speech	noise
speech	<b>97.98</b>	2.02
noise	16.35	<b>83.65</b>

**Table 1 Speech detection: accuracies**

### 3.2 Speech enhancement:

A speech signal can be distorted by bad channel quality, background noise, or by overlap with other signals. Speech enhancement algorithms help to improve speech signal quality. These algorithms consist of different types of filters that separate speech from the distorting signals to make the speech easier to understand. The separated-out noise signal may also be listened to for further analysis

### 3.3 Language Identification

Language identification classifiers determine the prevailing language in a speech signal. Training material is needed for each of the languages to be identified.

The training material must be chosen according to the application domain. Example: for robust identification of English, both American and British English should be used for training.

The result of the classification can either be a decision for the complete speech signal, if the language does not change, or a time-labelling that gives the points where the language changes within the signal.

Typical results are shown in Table 2.

	english	german	italian	arabic	spanish
english	<b>83.67</b>	16.33	0.00	0.00	0.00
german	2.86	<b>91.43</b>	0.00	5.71	0.00
italian	0.00	1.41	<b>91.55</b>	1.41	5.63
arabic	0.00	0.00	0.00	<b>100.00</b>	0.00
spanish	0.00	2.35	15.29	4.71	<b>77.65</b>

Table 2 Language identification: accuracies

### 3.4 Speaker identification

Speaker identification classifiers determine the speaker in speech signals. The result is the decision for a certain speaker or a rejection in case of an unknown speaker.

A set of speech signals for each of the speakers to be identified is needed for training. This algorithm is language independent so the speaker will be identified irrespective of the language used. Unknown speakers cannot be identified and will be rejected by the algorithm.

Typical results can be seen in Figure 1, using the set of the '99 speaker identification by NIST. An overall error rate (the so-called equal error rate EER) of 12 % is achieved using a set of about 100 speakers.

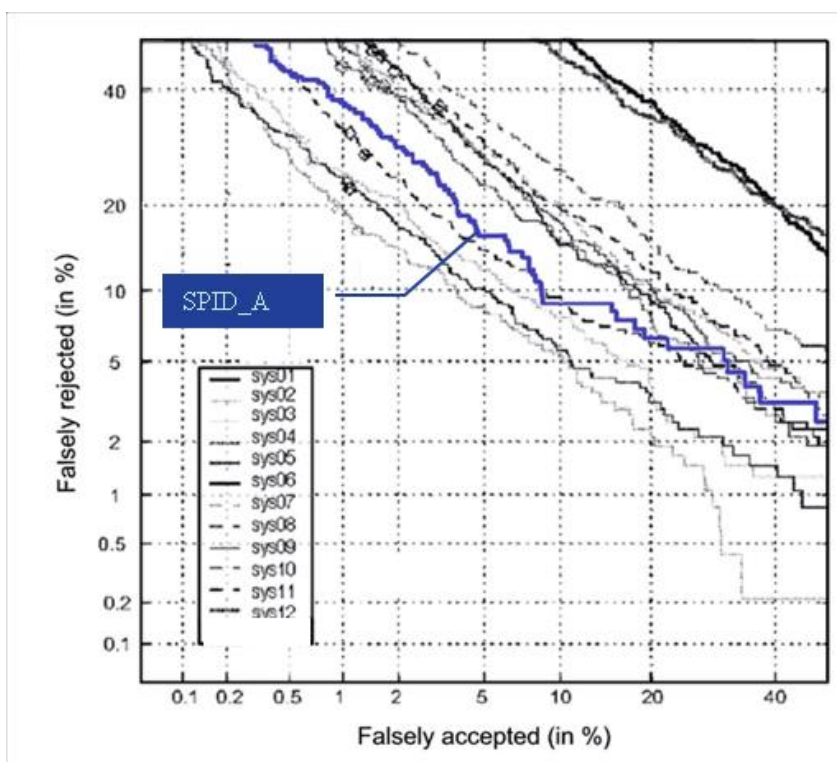


Figure 1 Speaker identification: accuracies

### 3.5 Topic Spotting

The objective of topic spotting is to determine the domain or topic of a speech signal without performing a linguistic analysis of the utterance. The result is the decision for a topic like “sports”, “politics” etc.

Speech signals with the topics of the application are needed for training. The classification result can be given for the complete file or, if the topic changes within a speech signal, for the points in time delimiting a certain topic.

Typical accuracies are around 80 %, where the performance for humans is only slightly better with signals that can be classified to two different classes simultaneously.

### 3.6 Word Spotting

The word spotting classifiers detect special words in speech signals.

In a first step, the user enters the words that are important to a certain application, either by typing or speaking them. The classifier searches for the words inside the given speech signals. During usage, the detected words are presented together with the points in time when they occur within each speech file. There is no need for a complete orthographic transcription of the speech signal when this approach is used.

## 4.0 CLASSIFICATION SYSTEMS

Speech classifiers can be provided in a complete system environment that includes an archive for storing signals and classifiers, accessible through a graphical user interface. This complete system is called SCOOTY (Speech Classification Online and Offline Technology).

In addition, speech classifiers can be used as single modules to be embedded in users' software via standard interfaces. These modules are part of MELANIE (Medav Language Intelligence Environment).

### 4.1 SCOOTY – speech classification online and offline technology

SCOOTY provides a complete system for the analysis of speech signals. It comprises.....

- an archive for storage of incoming audio signals and trained classifiers. Results and additional data can also be stored for the signals classified during use of the system
- a signal processing module containing the algorithms for the classifiers, where the classifiers are processed, and.....
- a graphical user interface, where training, testing and production are started and observed.

The three parts of SCOOTY can run on the same PC or on different machines, depending on the users' needs. The computers may have different operating systems. An example of usage on three machines is given in Figure 2.

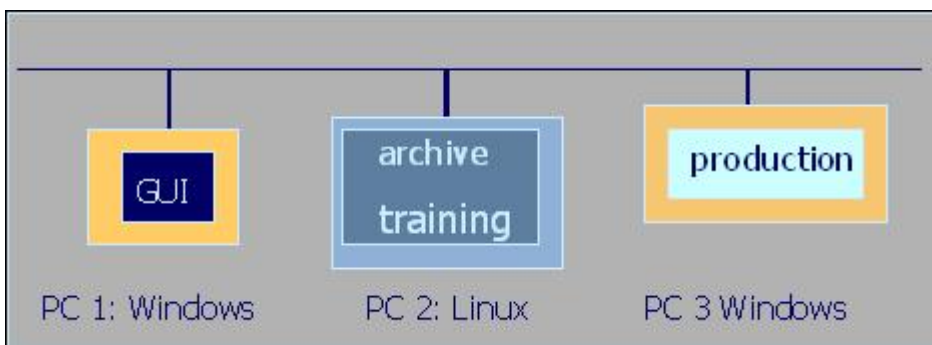


Figure 2 SCOOTY system architecture

The classifiers for the different characteristics may be used sequentially. An example of the graphical user interface is shown in Figure 3. The incoming audio signal was first passed through the speech detection algorithm. The result of this classifier was “speech”, which evoked the following classifier for language identification. In this case, the language was determined as “Spanish” which again evoked the following classifier for speaker identification. The speaker was unknown to the system; therefore the result was described as “rejection”.

The results of the classifiers are visualized on the right side of the GUI, here showing the Spanish flag for the language. The text and images corresponding to certain results can be chosen by the user.

In addition users can listen to the speech signals and perform further interactive analysis like spectrum and formant analysis.

The speech signal can be archived into the SCOOTY database together with the results obtained and

## A speech classification system

additional “meta-information” like time and caller phone-number.

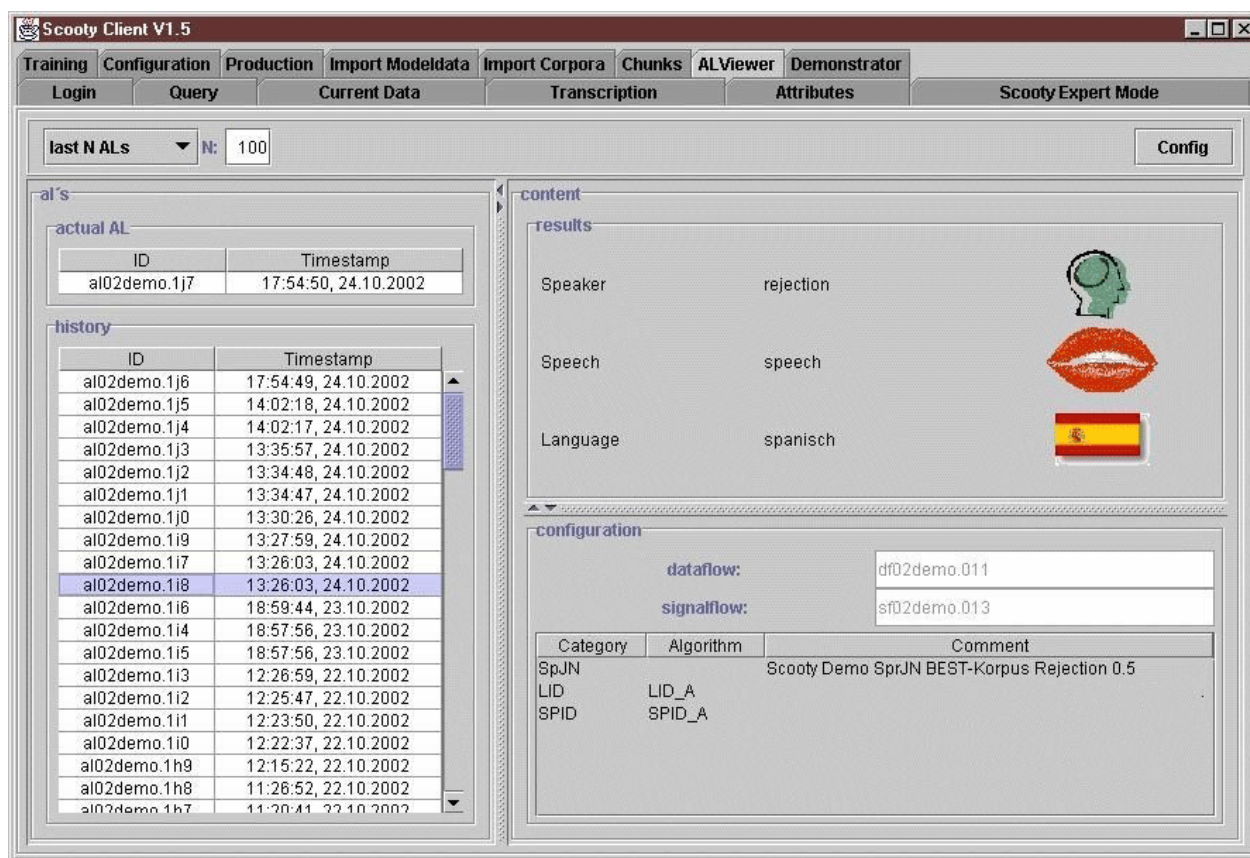


Figure 3 SCOOTY graphical user interface

### 4.2 MELANIE – Medav language intelligence environment

MELANIE can be used as encapsulated software with well-defined interfaces, when speech classifiers are to be embedded into an existing user environment

Training of the MELANIE classifiers is based on lists of speech signals that are provided by the user. The classifiers can be used both on and offline. The audio signals can be provided as files or data streams.

All classification results can be accessed by the user of the system through the well-defined API (Application Programmer Interface).

## 5.0 THE USE OF SPEECH CLASSIFICATION SYSTEMS IN INFORMATION FUSION SYSTEMS

Speech classification systems can analyse incoming signals according to.....

- speech vs. noise, data transmission,

- the language that is spoken in the speech parts of the signal,
- the speaker in a speech signal,
- the topic of speech,
- find certain words in a speech signal.

How can these capabilities be used best in the field of mass information and mass incoming signals?

Depending on the task of the user and the overall system architecture, three scenarios are presented to show different ways of using speech classification systems.

The first scenario shows the general use of speech classification in an information fusion system where different types of information are arriving from different locations (section 5.1). Another scenario shows the use of speech classification system in order to evaluate speech signals and to prioritize them according to their importance (section 5.2). A third scenario shows the usage when searching for specific information, in this example for the occurrence of a certain person (section 5.3).

With the SCOOTY system (Speech Classification Online and Offline Technology), all available information from e.g. the recording (like phone number, time and date of the phone call) is stored as additional attributes. The results of the classification are also saved as meta-information. All this meta-information can be used for further analysis.

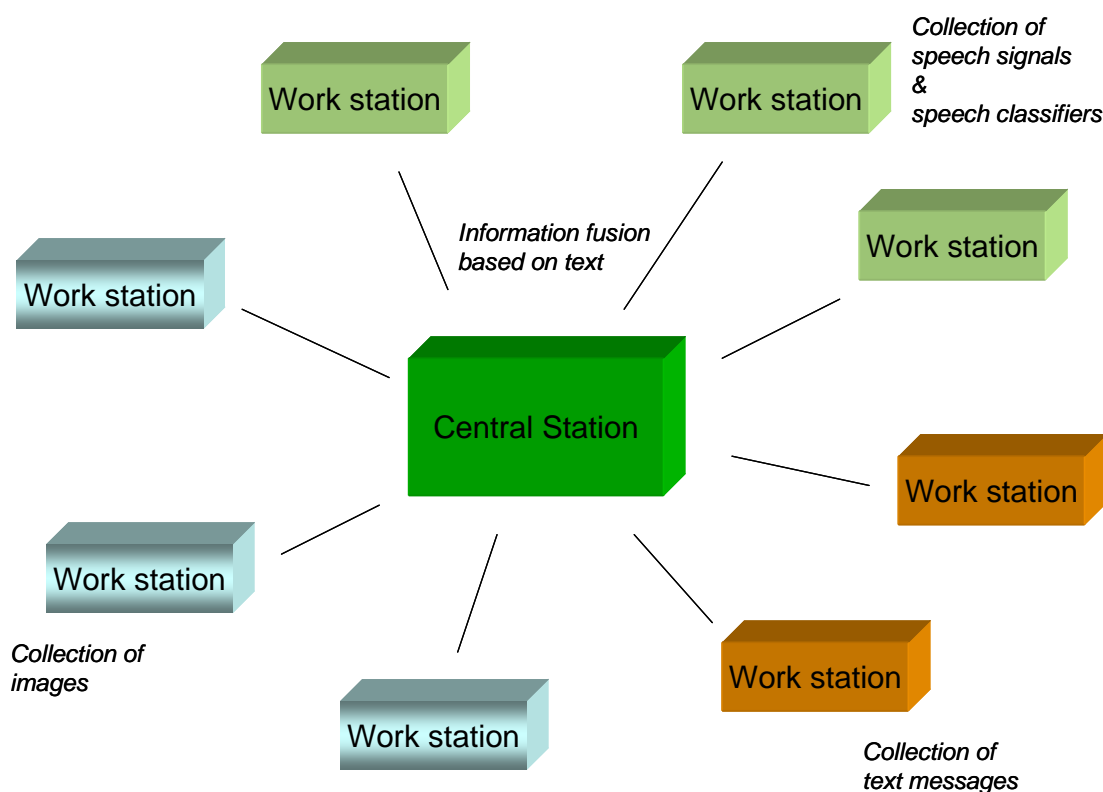
The information from the speech classifiers for each of the three scenarios is used together with information that has been gathered by other available analysis tools.

### **5.1 Scenario 1 – Central Storage and Automatic Analysis**

In this scenario, there are multiple sources of input that are separated geographically or by the current task. The signals can be analysed at their sources (here called work stations), but will normally be sent to a central station. At the central station, all data are stored and processed. This scenario is shown in Figure 4. At the same time, information consisting of texts and images will also arrive at the central station.

All available classifiers automatically process all incoming audio signals. The results for each file will be stored with the signal in the central archive; the language, the speaker, the topic and some important words will be also be stored. This information is called meta-information. Images and text will be analysed in a similar manner.

The meta-information will be evaluated according to a specific task in the central station,. For example, it is possible to make statistics for the frequency of occurrence of a certain language; the daily ratio of Arabic speech under certain conditions may be evaluated. A higher amount of Arabic on a certain day may trigger specified actions for the operators. In the same way, the frequency of a certain speaker may be evaluated, again leading to specific actions.



**Figure 4 Scenario with central station**

The information gained from speech, text, and images can be evaluated at the same time, if this information is stored in a similar manner. Consequently, the information about a person can be analysed from the information gathered from speaker identification in speech signals

Similar analysis and evaluation can also be performed locally at the work stations for the data available there. However, the largest amount of data is found in the central station. This reservoir of different types of data shares the same meta-information format. Statistics, analysis and evaluation can be done across the whole data set, in the same way and with the same tools.

Using data from mobile work stations adds another dimension to the analysis by taking into account the geographical information with the recording. Merging information from mobile and conventional telephony gives additional information about moving entities.

Last but not least, having all available data in the central station makes it possible to retrain classifiers for very specific tasks such as rare signal qualities and codings.

## 5.2 Scenario 2 – Prioritize messages

Another scenario specializes on the application that human operators must evaluate speech signals themselves by listening to the signals in their language.

Usually, the operators analysing messages have a set of languages they understand. The signals coming in will be classified according to the language beforehand with automatic classification and will be passed to

the appropriate operator.

For some languages, however, it may difficult to classify the language correctly. Therefore the operators may encounter speech data that do not correspond to their linguistic capabilities. Knowing that usually a certain percentage of the data volume consists of a certain language, all incoming signals can be rated with a reliability score and sorted accordingly. In this way, the operator will deal first with the best-rated n % of a language before listening to speech that has been classified wrongly. An example is shown in Figure 5.

Language 1		Language 2	
0.77	Language 1 usually has 10 % occurrences	0.62	Language 2 usually has 6 % occurrences
0.71		0.62	
0.67		0.61	
0.67		0.61	
0.66		0.60	
0.65		0.59	
0.61		0.58	
0.55		0.58	
0.54		0.57	
0.54		0.57	

Figure 5 Prioritizing messages

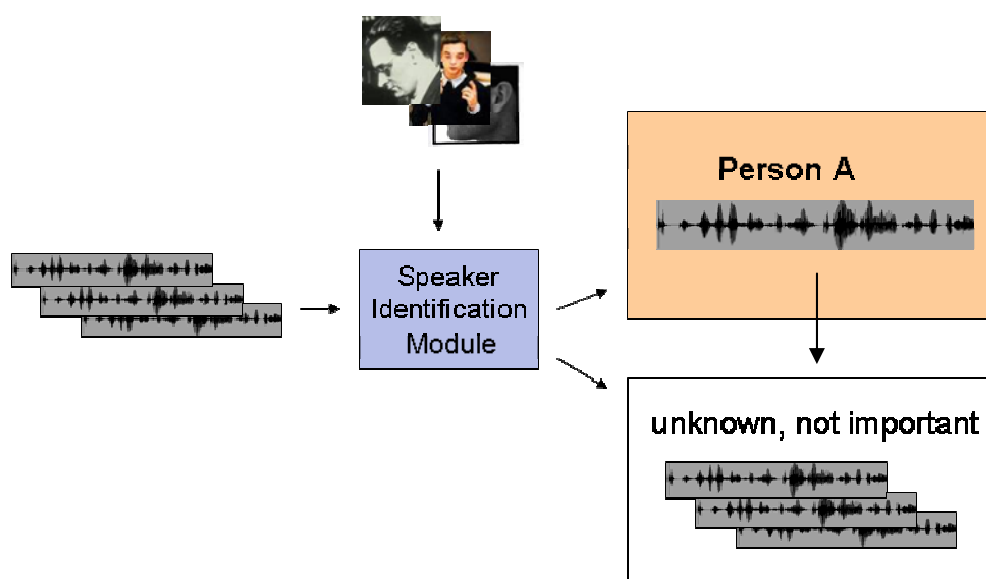
This figure shows the files that are each rated highest for language 1 and 2. Since language 1 has a higher average frequency in the total amount of signals, the 10 best-rated signals have to be analysed. For language 2 with 6 % occurrence only 6 files (out of 100) have to be analysed. Usually the best-rated signals are those that contain that specific language (language 2) with the highest confidence value.

Those signals with the highest score are the ones that have the highest probability for correct recognition. In times of operator scarcity, it seems best to analyse first the signals that are of the right language with a high confidence value.

Since the availability of language specialists is limited, best working performance is achieved when using this scenario because operators will almost always listen to speech in their chosen languages.

### 5.3 Scenario 3 – Search for persons in text and speech messages

This scenario uses speech classification for specific tasks that arise due to special actions. Example: a person A is known from other sources and his activities are to be analysed. Using some speech material of that person, a speaker identification module is generated. This module can now be used to search all incoming signals (and also the signals already in the archive) for that person. For every speech file that is detected, a specific action is evoked. An alert may be given for incoming signals during online processing. Statistics may be generated relating to the behaviour of that person by running a search in the database archive. This scenario is shown in Figure 6.



**Figure 6 Searching for a specific person**

Again, thresholds can be set in order to change the false alarm and the miss rate, respectively. Using this technology, specific individuals can be found effectively. It should be noted that this algorithm is language independent and the person will be identified regardless of the spoken language.

If the person's name is known, it can be searched at the same time (word spotting) in speech signals and available texts from other sources.

## 6.0 CONCLUSION

The three major benefits to be gained from using speech analysis systems are:

- Time and cost reduction for analysis,
- Analysis and ranking according to importance,
- Joint analysis of information from different sources.

Speech classification greatly aids the analysis of audio signals by providing important information about the content of speech signals. There will still however always be the need for humans to perform the final analysis and draw conclusions from the results. Using speech classification leads to less information loss, even during high incoming data levels. Furthermore, the user can concentrate on the important task of analysis and instigating action.