Is speech technology ready for use now?

Ulla Uebler, Dirk Kolb
MEDAV GmbH
Gräfenberger Str. 32-34
91080 Uttenreuth
Germany
E-mail: {Ulla.Uebler | Dirk.Kolb}@medav.de

ABSTRACT

Research and development in speech technology has been performed for almost 30 years now. Coming from experimental systems, a set of products has been developed in this time. From the view of the potential users, the main question remains: has the technology reached a state where it can be used meaningfully?

This paper will discuss this question - and will give an overview of the tasks that speech recognition can solve and the state for usability for each of the task.

Keywords: Speech processing, automatic processing, language identification, speaker identificaion.

INTRODUCTION

Almost thirty years have passed since computers have been used for the first time to help with the processing of audio signals. Starting from very simple tasks like presenting the amplitude or spectrum of an audio signal, the task of speech processing has increased dramatically.

In the 1980's, when introducing artificial intelligence into the research fields, enormous expectations were made to speech analysis: there have been research projects that should cover not only the automatic understanding of spoken utterances, but also the automatic translation and generation of speech in a different language. These research projects covered high requirements to all parts of the process, e.g. to the grammatical and semantic understanding of utterances.
Nowadays, the used algorithms have been optimized, and the focus of speech recognition lies also on the usability for "real" applications. This means at the same time, that the requirements on the capabilities and the complexity must be reduced. On the other hand, the goal is to obtain accuracy rates and a processing speed that enables the practical use of speech technology in real environments.

This paper is structured in the following way:

- Requirements of potential users of speech technology – often these requirements are orthogonal to technical problems and solutions. Above all, users do not want to need deep insight into the complex technology.
- Tasks for speech classification: the different applications of speech technology are presented; typical applications are listed together with a short insight into the technical side together with accuracies that could be expected.
- Conclusion: here, the needs of the users and the technical possibilities are put together; as a consequence, the different applications are evaluated according to both types of criteria.

REQUIREMENTS BY THE USERS

Speech technology is a very complex matter concerning the algorithms for the processing of audio signals. In order to become an expert in this field, it often takes years for a newcomer to understand the complex procedures.
For users that want to employ speech technology, there is also some effort to understand why some applications are less complex from the point of view of algorithms and why some accuracy rates are unexpectedly high or low.

Typical quality measures for the users are:

- Accuracy of the processing result
- Processing time: the goal is at least to be faster than real-time
- Easy use of the algorithm – use the speech product without being an expert in speech technology

These requirements – high accuracy combined with high speed – are obviously meaningful for the users. These characteristics, however, depend on a set of technical parameters. These technical parameters should be fulfilled at the most in order to provide the best results for the users. Some parameters are for example:

- Signal quality: the better the quality is, the better the accuracy would be.
- Training data: the algorithms for speech technology normally must be trained: the more data and the best suiting training data lead to the best results.
- The complexity of the task: e.g. for language identification, better results are obtained when distinguishing between two than ten different languages.
- The difficulty of the task itself: it is more difficult to distinguish between English and German than between English and Mandarin due to the characteristics of the languages themselves.
- The application itself: if the speaker does not explicitly speak clearly in order to be well recognized (e.g. dictation systems), we speak of ***non-cooperative*** users. These users make the classification task more difficult.

It is important to know that these parameters have a direct impact on the results. Occurring recognition errors can be understood and also the need for providing training data.

For the user it is mainly important that speech processing tools work accurately and fast. The standard user does not want to be involved into the details of the algorithm, although some details might help to improve the accuracy of the system.

TASKS FOR SPEECH PROCESSING

Speech Detection

Task	Detection of those parts of the speech that contain speech.
Challenges	Non-speech parts containing music, murmur -> difficult to classify as non-speech
Technical background	Two different algorithms are possible and used: • Thresholding the energy of a signal: above a threshold, it is regarded as speech – this algorithm is not trained

	• Providing a trained class of each noise, speech, silence etc. The current signal part is compared to the classes; the decision is made for the most similar class. Nowadays, often a mixture of these approaches is used.
Typical recognition rate	Around 90 to 95 % depending on set thresholds and false alarm rates.

Since the beginning of speech classification, these algorithms have been improved and yield now a good accuracy to be used in non-cooperative applications.

Example application:
A room / underground station may be observed with respect to activity using a microphone. Using speech detection, only those possibly few minutes of a day need to be further stored and analysed, leading to a huge reduction in storage and human effort.

Language Identification

Task	Detection of the spoken language in an utterance
Technical background	Examples of the languages of interest are stored and their characteristics are estimated. A new speech sample is analysed in the same manner; an estimation is performed if the new speech sample belongs to one of the trained language classes. Typical algorithms are using Fast-Fourier-Transform (FFT), cepstral analysis for the storage of relevant characteristics of speech. For modelling the temporal characteristics, often Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM) and language models are used. This approach must train the respective language characteristics. Therefore, a set of speech samples (at least 3 hours) is necessary for each language.
Challenges	• Accents and dialects of the users (non-native English speakers, Arabic dialects) • Similar languages (Czech, Slovak) to be distinguished • Speech quality (noise etc.) • Find suitable training material (same dialects, similar signal quality)
Typical recognition rate	• Around 95 % with 5 languages and good quality • Around 80 % with 10 languages and bad signal quality.

For scenarios with a good signal quality, this approach is good for use. In scenarios with bad quality, the algorithm has a higher failure rate.

Example application:

In telephone conversation, speech samples occur in different languages. The operator analysing the speech samples only speaks a subset of the occurring languages. The automatic classification of the languages help the operators to deal mostly with the languages they are interested in and which they understand.

Speaker Identification

Task	Detection of the speaker in an utterance
Technical background	The approach for speaker identification uses the same principles as listed in language identification. One difference is that the timely order of the speech sample bits is not counted as important. In addition this algorithm is independent of the spoken language and could, thus, recognize the same speaker when speaking different languages. For each speaker, at least 2 minutes of speech are necessary.
Challenges	• Speaker speaking in different channels (telephone, HF) • Speaking under different conditions (tired, sick) • Non-cooperative speaker – changes his voice. • Similarity among the speakers (male, 20-25 years old)
Typical recognition rate	Around 85 % with 100 speakers and good quality

This approach is very useful for differing between 100 to 1000 speakers. The results are very good then. Having much more speakers (e.g. 10.000), this algorithm does not provide a high accuracy, but can give a hint of a set of possible speakers and thus lead to a pre-selection of the true speaker. This algorithm works for non-cooperative speaker speaking in different language, it is not necessary that a pre-defined utterance is spoken (no key words necessary).

Example application:

In a discussion, meeting or telephone talk, the current speakers switch from time to time. Using speaker identification, the current speaker can be identified, e.g. out of all employees of a company.

This approach can also be used for training speaker groups. For example, if there is a search profile for young male persons, there is a large number of speech samples that can be ignored because the respective speaker does not show the characteristics of the current search.

Topic Spotting

Task	Detection of the spoken topic in an utterance (sports, politics)
Technical background	There are two different approaches prevailing on the market: • Perform a (partly) transcription of the speech signal first, do a text analysis on the topic in the following • Perform a topic analysis directly on the speech signal The first approach requires a well performing speech transcription (see "Word Spotting" or "Transcription of speech" below, and the availability of a text analysis tool. The second approach works directly on the speech signal, using similar techniques as for language and speaker identification. MEDAV uses the second approach. For this approach, at least 1 hour of speech per topic is required.

Challenges	• Similar or overlapping topics (general vs. specific topic) • Classification to several topics possible • Provide sufficient training material for each topic
Typical recognition rate	Around 80 % with 3 topics and good quality

This special approach does not need especially transcribed training material like other approaches do, therefore it can be used quite well for new languages and topics. The recognition rate helps the operator for a selection of more relevant utterances.

Example application:

Such a technique is widely used for indexing TV programs, e.g. in order to distinguish between the news block and commercials.

Another application is to filter phone calls of interest from small talk phone calls.

Word Spotting

Task	Detection of the certain words
Technical background	For the acoustic part of this task, the techniques used for word spotting are similar to those used for language identification. For the extraction of the words themselves, an in-word grammar and pronunciation rules or a pronunciation lexicon is needed. Usually, emphasis is set on the acoustic representation of the words. Therefore, a huge amount of speech data is needed for the modelling of acoustic units, e.g. phonemes – usually about 50 hours of transliterated speech is needed.
Challenges	• Long words are better recognized than short words • Words with suffixes and prefixes are difficult to be recognized as the basic word. • Find sufficient training material for a language (50 hours of transliterated speech)
Typical recognition rate	Around 80 % with 100 words – eventually a high false alarm rate.

The main challenge is to find sufficient training material for languages other than English – the training material should be similar with respect to quality of speech and dialects: the effort of recording new material is the main difficulty. The words to be recognized must be carefully chosen – only these words can be recognized. Short words are difficult to find.

This algorithm can be well used for good signal quality. With bad signal quality, a high false alarm rate will be set – still for a pre-selection of the utterances, the algorithm is usable.

Example application:

For public radio transmissions, word spotting is used to find certain names in the news block, e.g. “Berlin” or “president”. Using the time of the occurrence of such a word, the relevant utterance is easily found.

Gender Identification

Task	Detection of the gender of the speaker
Technical background	There are two possible approaches: • Trainable: a set of male and female utterances are trained in order to obtain corresponding gender classes – algorithms are similar as those of language and speaker identification • Threshold based: the average frequency of a speaker is measured, together with some more sophisticated values. The decision is made with respect to these values The threshold bases approach is easier to use, since there is no need for a training phase. The performance between the two algorithms is similar.
Challenges	-
Typical recognition rate	Almost 100 percent for most applications

This task is simple – there are only two classes possible. Using the high accuracy of this approach, this algorithm is used to possibly sort out half of the material – data reduction.

Example application:

Similar application as “speaker groups”: if the operators are only interested in male speakers, the currently irrelevant speech signals can be filtered out this way.

Emotion Detection

Task	Detection of the emotion
Technical background	The algorithms base on the algorithms used for language identification. Additionally, information on the melody and prosody of the utterance is used.
Challenges	• So far used in research and first applications like call-centres • No experience with bad signal quality and larger set of emotions
Typical recognition rate	-

The idea is to find people speaking with special emotions – assuming that important news are accompanied by emotions. First results with call centres (anger vs. normal) show some good results. So far no results on the correlation between “relevant event” and “high emotion

score" are known. In future, this could be a good way to rate the importance of an utterance.

Example application:

This algorithm is so far used mainly in call centre applications. Here, the algorithm detects, if the customer is satisfied or is going to complain about products – he might thus be connected to different departments inside the company.

This algorithm might also be used in order to determine if a person is nervous.

Transcription of Speech

Task	Transcription of utterances in plain text
Technical background	The algorithms used are similar to those used for word spotting. In addition, more sophisticated algorithms are employed dealing with the higher number of words (word graphs, word lattices, language modelling). For the acoustic training, 50 hours of transliterated speech is necessary. For the modelling of the word sequences (language modelling), texts of some hundred pages are needed. Sometimes, additional grammars are included in order to reduce the complexity of the task.
Challenges	• Used with cooperative speakers (dictation systems, telephone information system). • No experience with almost unlimited vocabulary, different possible speakers combined and non-cooperative speakers. • Find sufficient training material
Typical recognition rate	• Around 95 % for dictation systems (10.000 words, optimized for only one speaker) • Around 90 % for information systems (1.000 words, different cooperative speakers

This algorithm works very well for cooperative users in well-defined environments. There are only little experiences with non-cooperative users in adverse environments and open vocabulary. Each of these factors reduce the accuracy. First experiments are done in this field – in some time, there may be useful products available treating with all the factors reducing the accuracy at the moment.

Example application:

A typical application are dictation systems, that are used in office environments. These systems work only for one user at a time (they must be adapted to each user with a special training session).

For any type of application, this algorithm would be useful, since it converts the format from "*speech*" to "*text*", and thus makes the content available for further operations on text basis. Algorithms that work on texts can perform a couple of additional tasks, e.g. automatic summarization of texts, indexing of text, visualisation of texts.

CONCLUSION

In the previous section, different applications have been shown for speech technology. These applications have a different degree in complexity and experience in real-life applications. The question of this paper can be answered only when looking at the different parts but not for speech technology in total. In our applications, we mainly have situations with non-cooperative users, i.e. the user does not explicitly help the system to understand him with e.g. clear speech.
MEDAV has more than twenty years of experience in the field of speech signal processing and comes to the following characterisation:

Before classifying the different applications, it is important to remember:

- The algorithms are based on statistics, and therefore may show some classification errors, even for "easy" tasks
- Also humans make error when classifying – the different types and amount of error may be compared.
- We must take into account that the use of automated application will lead to a reduced amount of hours that operators would have to spend on these tasks.

The judgement thus tries to evaluate the possibly higher error rate together with reduced working hours on this task. The break-even point where possibly lower accuracy equals with the reduced number of operators working on this task combined with the fact that the operators can concentrate on the "important" tasks may differ from task to task and also depending on the current mission.
The classification presented here comes from various experiences from our customers and our own experience from more than 20 years.

Algorithms that lead to satisfying results with non-cooperative users

- Speech detection
- Language identification
- Speaker identification
- Gender identification

These algorithms are well developed for many years. In the meantime, they are robust enough in their functionality that they may deal with sudden changes in signal quality and other adverse conditions.
Our customers use these algorithms and they obtain a benefit with this system.

Algorithms that lead to satisfying results with cooperative users also help in non-cooperative environment

- Topic spotting
- Word spotting
- Transcription of speech

These algorithms are more complex than the above mentioned ones and thus more sensible to adverse conditions. In good conditions, they work well and reliably, under adverse conditions the recognition accuracy may drop.
Depending on the user scenario, the use of these algorithms may be useful

- Because they deliver a new measure for the estimation of a message into "important" and "not important".

- Under some conditions, they deliver very good results and can be used as they are (see above).

Algorithms that are in the state of research at the moment

- Emotion detection

This algorithm is already used in a scenario (complaint management), where it is important to distinguish between “angry” and “not angry”. This task is somewhat easier and we do not know about misclassifications in this field.
Still, this task could be interesting in order to classify nervous people etc. However, no reliable data are available at the moment. This task will be open for the future.

Recommendations

The algorithms in the first section have been developed for years and are ready to use without any restrictions.
The algorithms in the second section have been subject to research for many years, but have only be used for a short time with non-cooperative users. Therefore some aspects (like training data for rare languages) have not been completely solved. Still, these algorithms support the operators by indicating where important information could be found.
Finally there are other algorithms that are quite new in research that may be used one day in speech applications containing all the difficulties like many speakers, bad signal quality etc.

When users start with speech technology, they may do their first steps with the well established algorithms, and then expand the system to more complex algorithms, also depending on their way of working.

REFERENCES

[1] Carr, O. and Estival, D. (2003). “Document Classification in Structured Military Messages", Australasian Language Technology Workshop. Melbourne, Australia.

[2] Deloule, F. (2007) “Data Fusion for the Management of Multimedia Documents”. Proceedings of the 10th International Conference on Information Fusion, Quebec, Canada, 2007.

[3] FM 34-8-2 Intelligence Officer´s Handbook, Field Manual, Headquarters of the Army, Washington DC [pf09ko.0c6].

[4] LRE 09, 2009), “The NIST Language Recognition Evaluation Plan”.

[5] SRE 09, (2010), “The NIST Year 2010 Speaker Recognition Evaluation Plan”.

[6] STD, (2006), “The Spoken Term Detection (STD) 2006 Evaluation Plan”.

[7] Uebler, U. (2006) “A Speech Classification System”. In Information Fusion for Command Support (pp. 12-1 – 12-10). Proceedings RTO-MP-IST-055, Paper 12. Neuilly-sur-Seine, France: RTO.

[8] Uebler, U. (2001), “Multilingual speech recognition in seven languages,” in Speech Communication, Aug. 2001, pp. 53–69.