# Speech supported interaction with E-learning systems for online presentations

Georg Schneider
University of Applied Sciences Trier
Schneidershof
54293 Trier, Germany
+49 651 8103 580
schneider@informatik.fh-trier.de

Jörg Röpke
University Trier
Universitätsring 15
54286 Trier
+49 651 201-3583
roepke@uni-trier.de

## ABSTRACT
This paper describes a system, which helps to interact with E-learning platforms via speech recognition. The system observes the presentation of a speaker and helps the listeners to orientate in the presented material by highlighting text passages, which correspond to the currently referred topics.

## 1. INTRODUCTION
E-learning systems have helped to cross the barriers of time and space in order to access learning materials. Especially the idea of recording lectures has become very popular over the last years. Many systems support this feature, like Lecturnity [1], TeleTask [2] or Producer 2003 [3]. From the point of view of the teacher, there is relatively easy way to interact with the presented material using e.g. a laser pointer. Using these so called cross-modal references (from modality speech to modality text) she can reference passages on the material she is presenting, which helps the auditor to easily follow the references between speech and material. A pointing device gives her also the possibility to present the material in a natural way since it does not impose restrictions on her behavior. She can move around in the auditorium, she can use gestures, etc., which all contributes to a lively presentation.

However the drawback is that the viewers, which are not present at the lecture room or which look at the recorded presentation will not see this interaction since the slides are usually grabbed directly from the computer and therefore they cannot see the marks on the projection. Therefore these viewers cannot benefit from this explicit referencing i.e. pointing on the projection and they have to establish the link between speech and text on their own. This point is especially crucial, when the learning material is a longer text passage.

One way to overcome this problem would be to manipulate the material only in a way that it would be visible in the computer as well, e.g. using the mouse. Consequently this would tie the teacher to the place where the computer is located. She can no more move around and the presentation style would become relatively static.

A better solution to overcome the problem of visually supporting the referencing between presentation and speech would be the use of speech recognition systems. The system should follow the remarks of a presenter. When she is referencing or citing text passages, the system should highlight the belonging sentence so that the viewers can relate the utterances to the belonging passage. When she is freely explaining or deepening certain facts the system should recognize this as well and continue working in the background until it recognizes another text excerpt.

This example illustrates already some key points, such a system has to deal with. We need a robust speech recognition, which can handle wrongly recognized but similar words and relate them to the text. Furthermore the sentences recited from a speaker do not have to be exhaustively the same words in the same ordering than in the text passages. However the system should recognize that a certain sentence is referenced.

In the following we will illustrate a concept and describe our approach to build a system that supports a presenter in a way that her presentation, i.e. her speech, is automatically associated to the electronically presented content she deals with.

We will start with a related work section. Afterwards we will describe our concept and the system architecture. Then, we will continue with the description of our implementation. The paper concludes with a resume and an outlook of our further developments.

## 2. RELATED WORK
There are different issues that have to be regarded in the context of our approach. On the one hand there are systems that use explicit gestures in order to recognize cross modal references.

One approach is presented in [4] where a user interacts with a projection using a laser pointer whose movements are tracked with a video camera. The SmartKom system [5] for example uses

gesture recognition in order to recognize different input modalities. Both systems need a special instrumented environment, which is camera equipped. Furthermore the cameras have to be adjusted properly in order to recognize gestures. For this reason these systems are only mobile in a limited way.

The advantage of speech recognition in order to make cross modal references visible is, that it does not impose many restrictions on the place. The environment has not to be instrumented highly with cameras that follow the presenter in order to recognize her gestures.

Speech recognition on the other hand is a very powerful technology. However the target of our system is different compared to sophisticated natural language understanding systems (e.g. [6]). In these systems the focus lies on the understanding of spoken natural language, given certain knowledge about the domain. We are targeting somehow an easier task. We already know what the output should be and therefore we have only to decide, if a given utterance relates to the text or not. Like that we can make use of systems with a smaller footprint.

However the task to relate utterances to text passages for E-learning systems is more complicated than the tasks in IVR (Interactive Voice Response) systems (e.g. [7]), where VoiceXML [8] has become a very popular solution. The E-learning scenario needs a more flexible mechanism than EBNF-like grammars since the presenter could possibly start a sentence independently from the text and join the text at a later point. In this case the system should recognize the relation nevertheless.

# 3. CONCEPT AND ARCHITECTURE

In this section we will shortly introduce the E-learning system MOVII [9], which served as a demonstrator for our implementation. Afterwards we will characterize the different technologies that are needed in order to build the system. Finally we will show the system architecture.

## 3.1 The E-learning system MOVII

In this paper, we will only give a short introduction into the ideas of MOVII, further information can be found at the project website (www.movii.de).

The MOVII system is a web based E-learning system. It hierarchically structures the learning content into the following concepts: Module, Act, Scene, Core and Entity.

The so called core is an atomic structure, which refers to a certain learning content. A core can play different roles. It can be the core content to learn, as the word already implies. A core can also be an exercise; it can be more detailed information to this topic or a link to a completely different topic area. Furthermore the same core can play different roles in different learning scenarios. Whereas it can be an exercise in one scenario, it might be the relation to a different topic field in another setting. These different potential roles are called the entities of the core.

Related cores are grouped into scenes. Scenes belong to acts, which again belong to a module. A module covers a coherent topic field. In order to create a certain learning path a sequencing tool is used, where the teacher can arrange cores using a graphical user interface.

## 3.2 Accessing the content

In order to visualize cross-modal references between speech and text it is necessary to access the contents that are on display. Since we are working with a web based system, we can access and manipulate the information via the DOM (Document Object Model) Interface [10]. The DOM interface provides the possibility to fetch the complete content of a web page and to process it in our system. Like that, we know the text the presenter is talking about and we also have the possibility to manipulate its layout, e.g. highlighting text passages or change its color.

Figure 1 shows an example where a sentence has been identified and highlighted in the text.

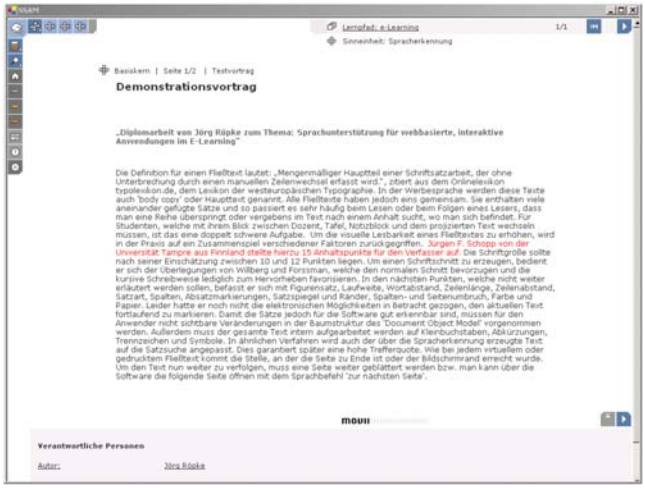


**Figure 1. MOVII with highlighted text passage.**

## 3.3 Robust speech recognition

In order to identify utterances from the presenter, we need a speech recognition engine. The advantage in our scenario is that we already know what we want to recognize. However the speech recognition engine might recognize a similar sounding word instead of the correct word.

In order to deal with this issue we have chosen to use the Levenshtein distance algorithm [11]. This algorithm calculates transitions from one word to another in form of substitutions, deletions and insertions. Using a certain threshold for the Levenshtein distance we can influence the tolerance of the speech recognition. In our experimentations the following equation has showed to produce good results:

$$\mathbf{Length(x)*3/4 <= lev(x,y),} \qquad \mathbf{(1)}$$

where **x** is the result from the speech recognition,

**y** a word from the text,

**Length** is the number of characters in the word and

**lev** is the Levenshtein distance.

If this equation holds, **x** is replaced by **y**.

## 3.4 Identifying utterances in text passages

In the preceding paragraph we have presented a robust approach to identify words in the text. However this is not sufficient in order to identify complete sentences or paragraphs in a text. For

this reason the following factors are also incorporated in the decision if a sentence is selected:

1. Frequency:

   If a word occurs rarely in a text passage, it is likely that this passage is referenced if the word is recognized. Sentences with rare words receive a high certainty value. This approach is also widely used in compression algorithms (e.g. [12]).

2. Co-occurrence:

   Recognized words that also occur together in a sentence make it more likely that a certain sentence is selected. Sentences with many recognized words receive a high certainty value.

3. Sequence:

   If the recognized words appear in the same ordering in the text, it is likely that a certain sentence is selected. However many sentences might have similar subsequences. This search obviously depends on the length of the utterance. If the speech recognition generates a short sentence, it must match more precisely a text passage in order to be selected; longer sentences are allowed to have more inappropriate words.

   The following example explains this idea. A sentence in the text is: "**We will be able to reach our goal next year**". The presenter says: "**And like I said before, the goal will be reached next yea**r".

   First we calculate the relative positions of the words that appear in both sentences in tuples, where the first number is the position in the text and the second number is the position in the utterance.

   **R={(1, 7), (2, 8), (7, 6) (8, 10), (9, 11)}** **(2)**

   Now we calculate the difference between the two values in order to see how the words relate to each other in the different texts.

   **D=(-6, -6, 1, -2, -2)** **(3)**

   This shows that 4 words occur in the same relative position in both sentences: **-6, -6, -2, -2,** which is: **will, be, next, year** and only one word is in a different order: **goal**.

   The algorithm can now be parameterized in two different ways. On the one hand the user can select a value how many percent of the words must appear in the right ordering (in our case we have 4) compared to the original sentence (which has 10 words) and how many recognized words can be ignored (in our case we would allow a threshold of at least 1, if we want to select the sentence from our example) in order to highlight the sentence.

The three different factors mentioned above are evaluated independently and the results are combined. This procedure offers a high flexibility since the values of the single search concepts can be parameterized independently and the combination of the results offers a further possibility to fine tune the system.

In order to combine the results, we use the intersection of the sets of sentences and the amplification of the certainty values of the resulting sentences that have been selected in different search approaches. Finally, if a certain threshold is reached the appropriate sentence is selected.

The search starts again from the beginning if the threshold has not been reached and the speech recognition detects the end of a sentence.

## 3.5 Architecture

Figure 2 shows the system architecture. The user accesses an E-learning content form the MOVII system via her web browser. Our system is hooked up with the browser and accesses the DOM as soon as a new page is retrieved.

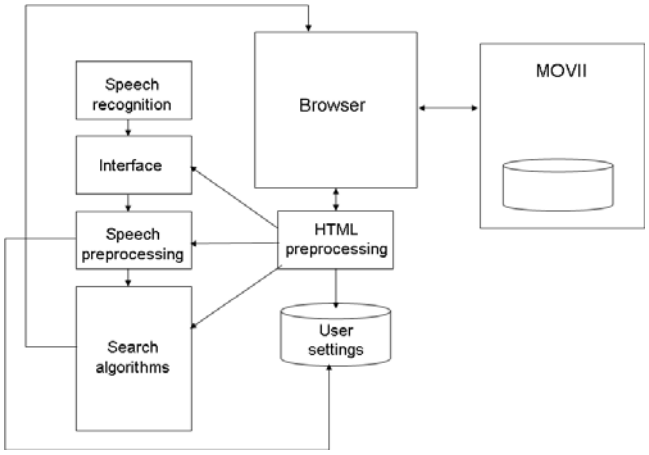


**Figure 2. System architecture.**

The HTML preprocessing fetches the sentences of the web page and builds word lists and calculates the word frequencies. As soon as the preprocessing is successfully completed, the interface on the left hand side is activated.

Now the content from the speech recognition engine can pass to the speech preprocessing. The speech preprocessing uses the user settings in order to identify and substitute abbreviations.

Finally the calculation of the probabilities and the decision, which sentence will be selected is executed in the search algorithm component.

As soon as a sentence is selected, the DOM representation of the browser content is accessed and the selected sentence will be highlighted.

# 4. IMPLEMENTATION

The system has been implemented as a C# application. The web browser control has been wrapped in a way that our system can be seamlessly integrated into the E-learning environment and that we are aware of any events that occur in the browser. Like that we also have full access to the DOM and can therefore easily fetch the content of a web page and manipulate the layout in the same way.

As speech recognition software we have used Dragon Naturally Speaking 7.1[13].

Furthermore we have implemented a graphical user interface in order to configure the system, e.g. supply abbreviation tables and

manipulate the thresholds for the speech recognition engine and the parameters for the voting of the search components.

In order to freely interact with the system we have used a wireless clip-on microphone.

## 5. CONCLUSION AND OUTLOOK

In this paper we have presented a way to support E-learning presentations, especially the visualization of cross-modal references between utterances of a speaker and belonging text material in a seamless way. We have experimented with different search and voting strategies for the results of the searches, which led to architecture that can be easily extended with additional search strategies. Especially the parameterizations of the system helped to adapt the behavior to our needs.

Our experimentations showed that the system can be productively used in a classroom presentation with only relatively few wrongly or not recognized utterances. This however depended on the learning material.

In the future we want to refine the search strategies and also test new algorithms in order to further improve the search results of the system. Additionally we want to test the system in a broader way with different users in varying settings.

## 6. REFERENCES


[1] Lecturnity, Retrieved January 14, 2007, from http://www.lecturnity.de/.

[2] TeleTask Retrieved January 14, 2007, from http://www.tele-task.de.

[3] Producer 2003, Microsoft, Retrieved January 14, 2007, from http://www.microsoft.com/office/powerpoint/producer/prodinfo/default.mspx.

[4] Kirstein, C. Muller, H., Interaction with a projection screen using a camera-tracked laserpointer, in Proc.: Multimedia Modeling, MMM '98, Lausanne, Switzerland, pp. 191-192 1998.

[5] Engel, R., Pfleger, N., Modality Fusion in: Wahlster, W., (ed) SmartKom: Foundations of Multimodal Dialogue systems, Springer, Berlin, Heidelberg, p 223-235, 2006.

[6] Wahlster, W., Verbmobil: Foundations of Speech-to-Speech Translation Springer, Berlin, Heidelberg, New York, 2000.

[7] Pereira, C., DTMS Solutions, Retrieved January 14, 2007, from http://www.dtms-solutions.de.

[8] W3C, Voice Extensible Markup Language (VoiceXML) Version 2.0, March 16. 2004, Retrieved January 14, 2007, from http://www.w3.org/TR/2004/REC-voicexml20-20040316.

[9] Kluge, F., Haberkorn, M., Regueiro-Lopez, H., 2004. movii: Medien bilden – Medienkompetenz oder die Befähigung zum Bild. In: *Bett, K., Wedekind, J., Zentel P., (Eds.): Medienkompetenz für die Hochschullehre. Medien in der Wissenschaft, Bd. 28*, Waxmann-Verlag, Berlin, Germany, 2004.

[10] W3C, Document Object Model (DOM), Retrieved January 14, 2007, from http://www.w3.org/DOM.

[11] Sushmita Mitra, Tinku Acharya, Data Mining: Multimedia, Soft Computing, and Bioinformatics, Hoboken, NJ, USA, John Wiley & Sons, pp 171, 2003.

[12] Henning, P., A., Taschenbuch Multimedia, Fachbuchverlag Leipzig, pp. 34-37, 2001

[13] Nuance Inc., Dragon Naturally Speaking, Retrieved January 14, 2007, from http://www.nuance.com/naturallyspeaking, 2007.