

# Multimodal Interface for Data Retrieval During Conversation

Roman Ženka ([zenkar1@fel.cvut.cz](mailto:zenkar1@fel.cvut.cz)) and Pavel Slavík ([slavik@fel.cvut.cz](mailto:slavik@fel.cvut.cz))  
Czech Technical University in Prague, Czech Republic

## **Abstract**

We propose a tool for aiding the scientific conversation by providing the participants an easy access to discussed data. Our tool utilizes sketch and speech recognition capabilities of a TabletPC, which allows it to identify the data being discussed and to visualize them in proper context. The entire progress of the conversation is being recorded for future offline viewing and archiving.

Combination of two natural means of communication – speech and sketching – gives our tool several advantages compared to conventional means of human-computer interaction.

The richness of speech allows fast and comfortable specification of required data, without the burden of navigating through conventional GUI, such as menus or directory trees. The command for retrieving the information can be even given within the conversation itself (i.e. as a part of a longer sentence) so handling the computer does not interrupt the speaker.

The sketching significantly raises the robustness of speech recognition. Compared to speech it also allows fast specification of 2D positions and relationships. The permanently visible sketches also complement the transient character of speech, providing a visual feedback about the progress of the conversation.

Finally, by simultaneous sketching and talking, the scientists automatically create a graphical representation of the recorded conversation. By clicking on the sketches it is possible to quickly navigate to the related part of conversation. This way the results can be distributed and watched offline by other members of the research team.

An important advantage of our approach is also its social factor – our tool requires nearly no direct users' attention, which can be spared for other participants. Users can speak fluently – without delays caused by complicated interaction with the computer – and even maintain direct eye contact most of the time, while presenting their data in high quality.

Our approach has been tested on a sample application for illustrating the ongoing conversation in real time. The user tests have shown that this tool can be valuable especially for managing larger meetings, where it allows faster and more fluent communication, saving valuable time of the participants.

## **Introduction**

Our goal is supporting communication between people, especially scientists. During conversation, people often have to provide data to support their standpoints or decisions. The data is usually visualized to be easily understandable, mainly in form of images, charts and tables.

If the communication is to be effective, the data should be easily and quickly accessible. It is often not possible to prepare a presentation in advance. On the other hand, seeking for the data during the conversation introduces unpleasant delays. We need a method for fast and efficient retrieval of data that could be used directly during the conversation. The conventional directory structures full of files require a lot of

time to navigate through, as well as a lot of user attention, which should be used elsewhere. Associating hot-keys with particular data does not work with large databases and introduces considerable strain on user's memory. Our goal is an interface that lets the user request the data in a fast and natural way.

## **Multimodal interfaces**

For achieving this goal, we decided to use a multimodal user interface that combines speech and sketch recognition.

So-called multimodal interfaces let the user control the computer by several different means at once. This can lead to higher comfort for the user as well as higher work productivity. Since we are interested in supporting conversation, a natural means of controlling our tool would be speech recognition. The computer could listen to the conversation and provide relevant data.

Since the users are engaged in a conversation with each other, the computer must not be the primary focus of their attention. User attention should be devoted solely to the discussed topic, not to controlling a computer. In best case, the computer would be controlled seemingly without any effort. This requires a specific type of user interface – we call such user interface “ambient”, since the user even does not have to be aware of its function most of the time.

During our tests we found out that sole speech recognition cannot be used in practice, due to numerous recognition errors and also due to complexity of such task. Instead, we couple the speech with sketch recognition. The users can scribble a very rough symbol for the data they demand on the computer screen while they speak. The combination of these two modalities allows much more precise and robust recognition. Our tool thus acts as a “magical paper”; able to turn very rough drawings into the objects they stand for.

Our tool also records the entire conversation (speech and sketches), which can be replayed afterwards. Since the conversation can be lengthy and difficult to skim, each drawn sketch allows fast skipping to the time when it was created. This way the user can immediately access the conversation relevant to any particular data.

## ***Previous work***

Our work is related to research of multimodal user interfaces. The idea of combining of speech and sketches has been already studied in the context of combined modalities interfaces [4]. Our work differs to this approach by presenting an interface, which is used indirectly, while the user performs a different task.

Simultaneous recording of speech and sketches is described in [7]. However, our approach goes beyond that, since we recognize the drawn shapes as well as the spoken words to give the user extra advantages.

The sketches to be recognized can be taught to the system using a method described in [6]. The fast skimming of recorded conversation using sketches is similar to approach described in [1] and [8]. However, we rely mainly on the sketches instead of recognizing words that denote switches in conversation topic.

Since our users are in fact creating a presentation on the fly, some of our inspiration also came from tools for rapid design, such as [3]. The entire user interface of our tool is very informal and easy to learn. An evaluation of benefits of informal user interfaces can be found for instance in [2].

## ***Our approach***

For our work we used TabletPC platform which provides capabilities of both speech and sketch recognition. The user can sketch directly on the computer screen with a stylus. The TabletPC that can be set to have the display facing up resembles a notebook, roughly the size and shape of a notebook and thus feels very natural during the conversation. Another possibility is using a SmartBoard [5] for sketching, to give the users more space.

## **Collecting data**

The system is used for very fast data retrieval and presenting. Since the data itself is being displayed during the conversation, it has to be converted into a graphical representation. At the moment our tool supports only images, so only data exportable to images can be presented. An easy and straightforward extension would be direct support of other data formats, such as PDF, spreadsheet tables or text documents.

Each image is tagged with one or several keywords and a simple sketch. The sketch consists of a single line, so it can be drawn very quickly and easily. The system supports a basic set of simple shapes and uppercase letters, other symbols can be learned. Images are stored in a simple database.

The database creation is unfortunately necessary to allow rapid data retrieval when the speed is needed. This overhead is compensated by the ability to pick any item directly while talking within a second, instead of browsing large directory structures.

## **Sketching**

The system provides a set of simple default sketches. The user can define new shapes by technique described in [6].

The recognition of the sketches runs parallel to speech recognition. The tool is capable to evaluate several possibilities at once, because sometimes it may not be clear which sketch the user meant to draw. The information about probability of given shape is passed to the rest of the system, where it can be combined with inputs from the speech recognizer.

Although our system is designed to support only single-line sketches, it often happens to the user that the line they draw gets disconnected, for instance if they rise the stylus above the screen shortly while sketching. The system automatically connects broken lines like this in case the line fragments end in very close spatial and temporal proximity. This approach raises the robustness of recognition and turns out to be very useful especially for novice users who tend to sketch on the screen very lightly.

## **Speech recognition**

For speech recognition we used Microsoft Speech API. The speech recognition can be very problematic especially when our tool is used in noisy places. Use of a microphone is recommended. After experiments with several types of headsets, including wireless ones, we realized that even when a quality microphone is used, the recognition tends to produce many “false positives”.

A false positive occurs when the user utters a phrase that is similar to those stored in the database. Our experiments showed that the tool could not be used solely with speech recognition due to these numerous errors. This makes the sketch input necessary.

The speech API is able to calculate several recognition candidates for uttered words together with their probabilities. The recognition is very hardware-intensive process,

so there are observable delays present (about one second). Although the users do notice the delay, it is small enough to make the tool appear interactive. Faster processors might eliminate such delays entirely.

It is important that the user can use the keywords directly within longer sentences. The conversation then appears more natural, since it is not interrupted by explicit commands directed to the computer. However, using words within larger sentences can lower the recognition performance, since the slurred speech can be difficult to fraction. In case the users observe poor performance, they should try to pronounce the keywords more clearly.

### Combination of speech and sketches

By combining speech and sketches, we can achieve higher recognition robustness, especially by eliminating irrelevant speech.

The combining starts with finding recognized sketches and words that seem to belong together. Observing the temporal proximity and picking only words spoken close to the time when the sketch was drawn works reasonably well. In our work we used a 5 second limit on time difference between speaking and sketching.

The various possible meanings of recognized speech and sketches are combined by multiplying their respective probabilities. The most probable result is then used as the proposed recognition.

Except rising the robustness, the combination of speech and sketches helps resolve ambiguities that can appear when several objects are assigned the same sketch. This occurs frequently when the database of objects becomes large. Figure 1 shows an example of ambiguity resolution.

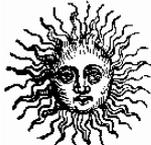
Sketch	Speech	Result
		
		
		

Figure 1: Ambiguity resolving by combining speech and sketches.

### Replay and skimming

Once our tool is started, it automatically records all the conversation. The sketches are also recorded and synchronized with the spoken words. Since the speech does not require high quality recording to be comprehensible, the stored conversations are small enough to be practical.

A conversation can become very long. Replaying it afterwards becomes very time consuming, especially when the user searches for only one particular piece of

information. For this reason, we implemented a method for fast skimming. This method is based on the observation of the conversations.

At the moment some data are brought to the screen, we can tell that the data is relevant to the topic of conversation at that particular moment. Thus, for locating such topic, it is sufficient to bring the conversation to the point when the data were displayed. In practice, the user can click on a particular data and hear the conversation occurring at the time when the data were requested.

Since often the conversation topic sometimes changes although no data is brought to the screen, we further extended our technique to allow the user create a mark which is not recognized and translated. These marks have to be made on the side of the screen, and are treated as personal notes for further easier navigation.

## **Implementation**

Our tool was implemented on a TabletPC using the built-in gesture recognition engine and Microsoft Speech API for speech recognition. This reduced the time needed for implementation drastically and resulted in a tool whose behavior is compatible with other pen-enabled applications.

The TabletPC platform has several disadvantages that limit the usability of our tool. The main problem is the need to draw relatively large sketches compared to the piece of paper due to low resolution of the display and pen positioning. This way the screen soon becomes cluttered. For the future work it would be very interesting to evaluate our tool in connection with another device, for instance with specialized “electronic notebook” [7] or a SmartBoard [5].

## **Results**

Our measurements showed rise of recognition accuracy when sketches were accompanied by spoken words. Table 1 summarizes our results. Due to the nature of our measurement, the values can vary significantly in case of a person with difficult accent or a person that has problems with sketching. Provided values apply to a natural usage of our tool – that means the users did not attempt neither to be well understood by the computer, nor to fool the recognition process.

<b>Method</b>	<b>False positives</b>	<b>Wrong recognition</b>
Speech recognition	80%	30%
Sketch recognition	10%	20%
Combined	<5%	<5%

**Table 1.** Recognition errors

Speech recognition by itself is obviously not sufficient, especially because many false positives – recognition of a familiar phrase happening at time when such phrase has not been uttered. Speech recognition is also not perfect and often leads to errors. On the other hand, the sketch recognition can lead to a false positive only in case the “personal note” of the user is accidentally drawn the way it is recognized, which happens rarely. The errors in sketch recognition depend mainly on the quality of drawing. The more hastily the user draws, the higher is the possibility of error.

Combination of the two techniques basically eliminates the problems with false positives in speech (speech is recognized only in a short time frame before and after sketching on the screen). The combination also raises the robustness of recognition to the point when the tool actually starts being useful.

Despite our effort, the recognition is not 100% correct all the time. In case an error occurs, the user can correct it by repeating the keyword while holding the stylus down at the wrongly recognized object.

## **Conclusions**

We have presented a tool supporting data retrieval during conversation. The combination of sketches and speech leads to extremely fast and effective communication that does not slow down the user in conversation. Fast skimming the recorded conversation is possible by using the sketches as a graphical representation of the conversation structure.

In the future we would like to address especially the database creation, which can appear as extra work to many users. An interesting option would be controlling search engines using our interface, so the data could be retrieved from a wider variety of sources.

## **Acknowledgements**

This project has been supported by Microsoft Research, as well as by internal CTU grant CTU0409513 and by the Ministry of Education, Youth and Sports of the Czech Republic under research program No.Y04/98:212300014 (Research in the area of information technologies and communications).

## **References**

1. Arons, B., SpeechSkimmer: a system for interactively skimming recorded speech, *ACM TOCHI*, v.4 n.1, 1997, 32–38
2. Bailey, B. P., Konstan, J. A., Are informal tools better?: comparing DEMAIS, pencil and paper, and authorware for early multimedia design, *Proc. of the conference on Human factors in computing systems*, 2003, 313–320
3. Müller, R., Ottmann, T., The "Authoring on the Fly" system for automated recording and replay of (tele)presentations. *Multimedia Systems*, v. 8 , Issue 3, 2000, 158–176
4. Oviatt, S., De Angeli, A., and Kuhn, K., Integration and synchronization of input modes during multimodal human-computer interaction. *Proc. of the workshop "Referring Phenomena in a Multimedia Context and their Computational Treatment"*, ACL/EACL'97, 1997, 1–13.
5. Plimmer, B., Apperley, M., INTERACTING with sketched interface designs: an evaluation study, *Extended abstracts of the 2004 conference on Human factors and computing systems*, 2004, 1337–1340
6. D. Rubine, Specifying Gestures by Example, *Proc. SIGGRAPH'91*, 329–337
7. Stifelman, L. J., Schmandt, C. M., The audio notebook: paper and pen interaction with structured speech, Dissertation, Massachusetts Institute of Technology, 1997
8. Whittaker, S., Davis, R. Hirschberg, J., Muller, U., Jotmail: a voicemail interface that enables you to see what was said, *Proc. SIGCHI*, 2000, 89–96