

e-Scribe: Ubiquitous Real-Time Speech Transcription for the Hearing-Impaired.

Zdenek Bumbalek, Jan Zelenka, and Lukas Kencl

R&D Centre for Mobile Applications (RDC)
Department of Telecommunications Engineering
Faculty of Electrical Engineering, Czech Technical University in Prague
Technicka 2, 166 27 Prague 6, Czech Republic
{bumbazde,zelenj2,lukas.kencl}@fel.cvut.cz
<http://www.rdc.cz>

Abstract. Availability of real-time speech transcription anywhere, anytime, represents a potentially life-changing opportunity for the hearing-impaired to improve their communication capability. We present e-Scribe, a prototype web-based online centre for real-time speech transcription for the hearing-impaired, which provides ubiquitous access to speech transcription utilizing contemporary communication technologies.

Key words: Hearing-Impaired, Real-Time Text, Voice over IP

1 Introduction

Speech plays a basic role in communication between people, in education, in sharing ideas and in maintaining social contacts. The hearing-impaired have to challenge communications barriers in a mostly hearing-capable society. This barrier is especially serious when communicating with authorities or courts of law where inability to understand is a major obstacle. According to Czech law [18], hearing-impaired citizens have the right to choose a communication system that matches their needs. Today, this law is impossible to put in practice. One of the goals of the e-Scribe project, a joint activity with the Czech Union of the Deaf, is to remedy this situation. It may be overcome by using on-the-spot real-time speech transcription. Services of fast typists or sign-language interpreters may be employed. The concept of physically present transcribers is used for example in the project of the Czech Union of the Deaf: "Simultaneous transcription of speech" [17]. However, there is great shortage of educated transcribers, costs of these services are high and they are restricted to the particular location, therefore they can be offered to a limited amount of users only. An alternative is using Automated Speech Recognition systems (ASRs). Yet, ASRs are limited in recognition accuracy, especially of colloquial speech and of difficult national languages such as Czech.

The main contribution of the e-Scribe system is in designing a prototype centre for ubiquitous, cost-effective real-time speech transcription, first using human transcribers and later combining ASR with human error-correctors to

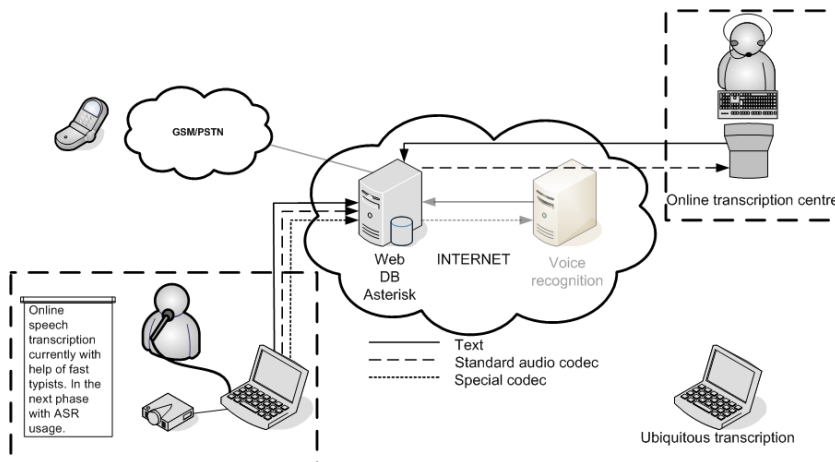


Fig. 1. e-Scribe Overview. Speakers's voice is transmitted via Voice over IP (VoIP) or telephony to the online transcription centre. The voice is transcribed by quick typist using MS Word[®], which is available via web form on the server. The transcribed speech is displayed on a projection screen at the conference using a web browser connected to the server. However, the transcription is available from any location with Internet connectivity.

automate the process of transcription (see Figure 1). The goal of the project is to set up an online transcription centre for the hearing-impaired, technically based on IP telephony and displaying the transcription online on a web page.

2 e-Scribe and Related Work

Using telephony for the deaf or hard-of-hearing has been possible by using text-phones known as Telephone Devices for the Deaf (TDDs) [7]. The disadvantage of TDDs is that both communicating users must have these textphones - yet only few hearing people own TDDs. For communicating with the hearing people, the Typetalk relay [14] could be used. The Short Messaging System is probably the greatest innovation in telephony for deaf people, but it can still not replace the real time flow of text, so to use SMS for longer transcription is impossible.

The topic of automated transcription can be explored from many points of view. An interesting enhancement in using ASR is the so-called captionist (transcriptionist) [3], [5] which is a person who re-speaks the speech that has to be transcribed, resulting in more intelligible and less noisy input for the ASR system. Experiments show [3] that re-speaking may significantly increase ASR efficiency. However, such a system needs at least one trained person for re-speaking the ASR input. ASR systems can be applied for example in instant messaging where the lower accuracy of transcription (about 10 % word error rate) [1], [5] has no fatal effect in understanding of sentence context. This approach is not

convenient when the transcription is supposed to be exploited in more public actions like lectures or technical meetings etc. where exact understanding is required. An interesting way to deal with transcription accuracy is complementing the ASR system with a person capable of error-correcting [4] of the ASR output. This can be done directly during the automated speech recognition. If it is not possible to do the correction immediately it can be done ex-post with the help of lecture audio record if necessary. Similar research in ASR for university lecture rooms is being carried out by the Liberated Learning Project (LLP) [15],[8]. The goal of LLP is enabling students with hearing impairment to access university lectures. In cooperation with LLP, IBM develops the ViaScribe [16] software that is specifically designed for real-time captioning.

The Villanova University Speech Transcriber (VUST) [6] was designed to improve the accessibility of computer science lectures using real-time speech recognition software. The VUST system is based on client-server architecture and consists of three major components: speech recognition, a dictionary enhancement tool, and a transcription distribution application. It is accessible locally via a wireless microphone in the classroom. In contrast, the e-Scribe is a ubiquitous system accessible from wide range of communication equipments such as laptops, PDAs, mobile or ordinary phones and real-time captioning is supported by skilled stenographers from the online transcription centre.

The idea of a ubiquitously available transcription tool was introduced in Scribe4Me: Evaluating a Mobile Sound Transcription Tool for the Deaf [2]. Upon pressing a button on a Scribe4Me PDA, the last 30 seconds of sound is uploaded, transcribed and sent back as a text message. The SMS communication presents a potential problem in interactivity due to the potentially significant delay. Additionally, the 30 seconds transcription is severely limiting. With today's mobile phones or PDAs, e-Scribe aspires to provide a real-time ubiquitous captioning via the phone web browser.

3 Implementation and Architecture

The currently operational e-Scribe solution is based on remote transcription carried out by quick typists. Voice is transmitted by VoIP telephony from the venue of the conference for the hearing-impaired (or another event) to the online centre, and then anywhere to the transcriber. Transcription is carried out by the specially trained typists who use a large list of abbreviations which are expanded into words or sentences. The typists are currently using Microsoft Word[®] software, due to historic, familiarity and performance reasons. Using MS Word[®] for textual input was one of their basic conditions for developing the application for transcription because for transcribers this environment is familiar and they are accustomed to specific layout, use of abbreviations and well-known behavior.

For the e-Scribe system MS Word[®] does not fulfill the requirements well due to the long delay when characters are transmitted. We are currently developing a web-based text editor simulating the behavior of MS Word[®] according

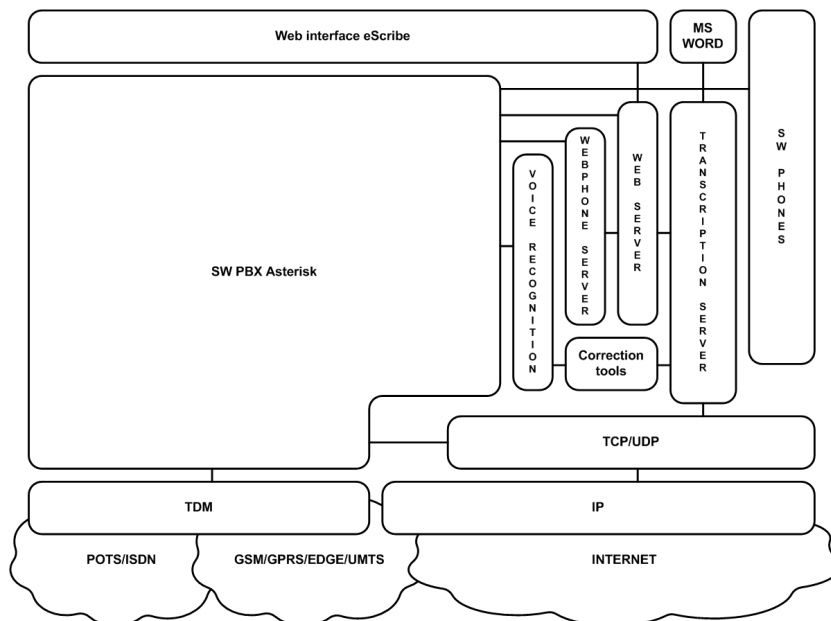


Fig. 2. e-Scribe Architecture. The core of the system, which represents the communication part of the project, is the software PBX Asterisk (Private Branch Exchange) [20]. The other block is the Transcription server which cooperates with the Apache web server [21]. Access to the system is possible from ordinary telephones, mobile phones but also from hardware SIP [11] telephones and software clients. The easiest way to access the system is using a web phone which makes the entire system available through the web browser interface, without necessity of any installation or configuration.

to the requirements of transcribers. At this time we have designed a web application for testing purposes capable of text transmission, backward text editing, changeable text window, adjustable burst of characters to send, etc.. However, user transcriptions are currently still performed in MS Word[®].

The e-Scribe system is now successfully implemented and tested at lectures for the hearing-impaired students at CTU in Prague and it was also used at several public presentation events (e.g. the e-Scribe Press Conference in January 2010). The transcription system is available on the website [19] of the e-Scribe project and we are preparing to start official performance testing. The overall solution architecture is described in Fig. 2.

IN IP telephony, Quality of Service is a frequent topic of discussion. Generally, in transcription systems it is very important to reach high-quality and reliable data flow to avoid information loss between speaker and transcriber. This demand requires not only a good internet connection, but also an errorless software chain and reliable hardware, if used. It is necessary to conduct testing in the of form e.g. conversation tests which examine the whole communication chain and stability of software, hardware phones or of any other devices involved.

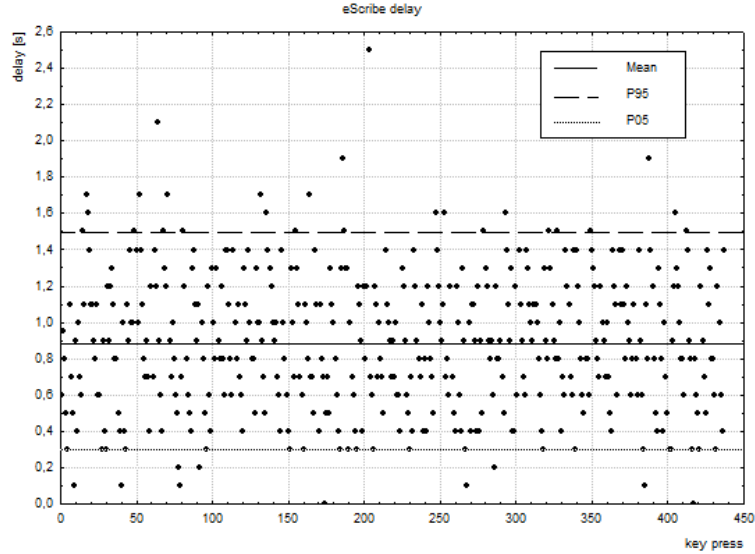


Fig. 3. Total delay from pressing a key to its displaying on a screen (450 values, mean and 5th and 95th percentile).

4 Performance Evaluation

In the process of transcription there are generally two main parameters for evaluation. The first parameter is the accuracy of the transcription. In the human-based transcription, this parameter depends on experience and dexterity of the typist. The second parameter is the delay between pressing a key and its displaying on a monitor or a projecting screen. To demonstrate the differences between situations where the typist is physically present and situations where he or she works remotely using the eScribe system, we prepared an experiment. We observed the total delay from pressing a key to its displaying on a screen. The delay was measured by evaluating the video record (25 frames/second) of keys pressing and displaying the transcription on the web page.

The experiment was carried out on a Windows-based laptop with MS Word[®] and a web browser. For video recording, the SW tool CamStudio [23] was used and the record was analyzed by the VirtualDub 1.9.8 [24]. The test was performed outside of the local network where the eScribe server is placed.

The total delay T_T consists of three partial delays:

$$T_T = T_W + T_N + T_D . \quad (1)$$

The T_W parameter represents the delay caused by MS Word[®] (time between pressing a key and sending it to the network) and due to native features of MS Word[®] the maximum is 1 second. The T_N parameter represents the Round Trip

Table 1. Statistical evaluation.

Total characters	Mean	Median	Modus	σ	σ^2	P_{05}	P_{95}
437	0.89	0.9	1	0.38	0.15	0.3	1.5
Network delay: Respons to ping request							
average delay			min		max		
22			10		329		

Time of the network and it was measured by the network debugging tool PING. The parameter T_D represents delay caused by periodical refreshing of the web page with transcription. The refresh rate was set up to 550 ms in our experiment. The transcription was made by an averagely skilled typist, to whom an article of approximately 450 characters was dictated via VoIP telephony.

The average observed delay was 0.89 seconds with standard deviation 0.38 seconds. The delay distribution is depicted in Fig. 4 and the statistical parameters are summarized in Tab. 1. The delay variation is caused especially by the fact that the transcribed text is displayed in batches (corresponding to the amount of characters which the typist types during 1 second), but each character is compared with an individual key pressing.

In the presented experiment we have evaluated the architecture for online web-based transcription. Although according to the subjective experience of test users the total delay is still acceptable, it does not fulfill the requirements of the ITU-T.140 [25] recommendation for multimedia application text conversation specifying that a maximum buffering time may be 500 ms (300 ms recommended). It is evident that the crucial delay is caused by MS Word[®]. To reduce this, we are developing a web based editor simulating the MS Word[®] environment.

5 Future Work

The emphasis will be henceforth paid on ubiquitous web-based access. The current e-Scribe system is based on the standard client-server architecture using the open source Asterisk software [20] and the proprietary transcription application. The future goal is to propose the architecture and build a prototype of a universal web-based services, portable across various common mobile terminals, providing access to these assistive voice services to the handicapped anytime and anywhere, based on the emerging paradigm of Cloud Computing.

We will also integrate both voice and transcribed text into Asterisk using Real-Time Text (RTT) [13]. RTT is generally streaming text that is sent and received on a character-by-character basis and its benefits are summarized in [9]. In the Internet environment, RTT is represented by RFC 4103 [10] known as ToIP (Text over IP). The Session Initiation Protocol (SIP) [11] and the Real-Time Transport Protocol (RTP) [12] is used in ToIP for real-time text transmission. The ToIP framework is designed to be compatible with VoIP and Video over IP environments and could be implemented in current communication IP systems such as SW PBX Asterisk. With using VoIP and ToIP, a universal and

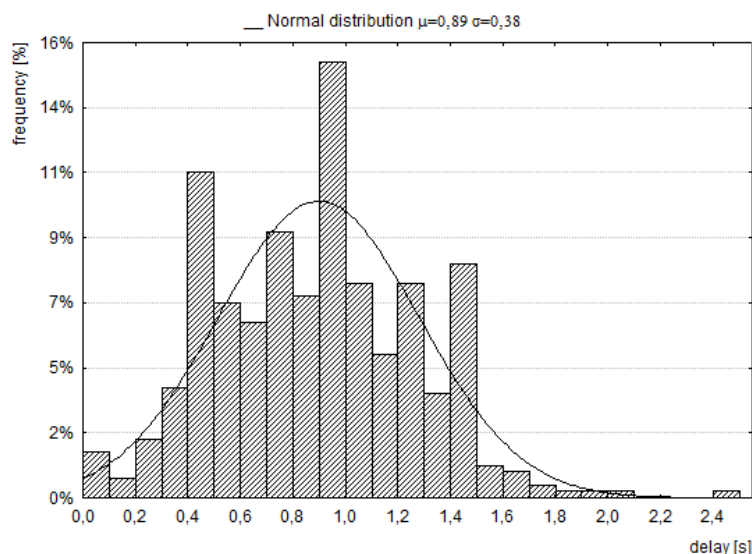


Fig. 4. Delay distribution.

standardized interface for interconnecting with the current and future ASR systems will be created, which is the main goal of the project. But because current speech-recognition software was found to be unsuitable for live transcription of speech (especially for national languages such as Czech), the transcription will be based on ASR software in co-operation with online human error correctors. This will make the transcription more cost-efficient and accessible to more hearing-impaired.

6 Conclusion

The widely accessible online transcription centre enables to provide services to much larger community of the hearing-impaired people. Remote transcription can offer this service with lower cost than if the typist were physically present at the conference or another event for the hearing-impaired people. This implies the amount of "transcribed actions" can be increased. Thanks to the online transcription, communication barriers will be minimized and cultural, educational, social or other events could be accessible for the hearing-impaired. Online transcription may also be used for communication with authorities or courts of law, where inability to understand is one of the most significant problems, or, thanks to its ubiquitous availability, eventually in day-to-day communication. Thanks to the e-Scribe system, communication barriers will be reduced and quality of life of the hearing-impaired people will improve dramatically.

Acknowledgments. The e-Scribe project is generously sponsored by Vodafone Foundation Czech Republic [22].

References

1. Wald, M., Bain, K.: Enhancing the Usability of Real-Time Speech Recognition Captioning Through personalised Display and Real-Time Multiple Speaker Editing and Annotation. In: Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 446-452, 2007.
2. Matthews, T. et al.: Scribe4Me: Evaluating a Mobile Sound Transcription Tool for the Delf. In: Ubicomp 2006, LNCS 4206, pp. 159 - 176, 2006.
3. Miyoshi, S., Kuroki, H., Kawano, S., Shirasawa, M., Ishihara, Y., Kobayashi, M.: Support Technique for Real-Time Captionist to Use Speech Recognition Software. In: ICCHP 2008, LNCS 5105, pp. 647-650, 2008
4. Wald, M.: Captioning for Deaf and Hard of Hearing People by Editing Automatic Speech Recognition in Real Time. In: ICCHP 2006, LNCS 4061, pp. 683-690, 2006.
5. Forman, I., Brunet, T., Luther, P., Wilson, A.: Using ASR for Transcription of Teleconferences in IM Systems. Universal Access in HCI, Part III, HCII 2009, LNCS 5616, pp. 521-529, 2009
6. Kheir, R. and Way, T.: Inclusion of Deaf Students in Computer Science Classes Using Real-Time Speech Transcription. In Proceedings of the 12th Annual SIGCSE Conference on Innovation & Technology in computer Science Education (Dundee, Scotland, June 25-27, 2007). ITiCSE '07. ACM, New York, 192-196.
7. Edwards, A. D. N.: Telephone access for deaf people, (in) Home-Oriented Informatics and Telematics, Proceedings of the IFIP WG 9.3 HOIT 2005 Conference, A. Sloane (Ed.), New York: Springer, pp. 235-244
8. Bain, K., Basson S. and Wald, M.: Speech recognition in university classrooms. Proc. of the Fifth International ACM SIGCAPH Conference on Assistive Technologies, ACM Press, pp. 192-196, 2002.
9. Proposal R1 for Implementation of Real-Time Text Across Platforms, <http://trace.wisc.edu/docs/2008-RTT-Proposal/>
10. Hellstrm, G.: IETF RFC 4103: RTP Payload for Text Conversation, the Internet Society, 2005
11. Rosenberg, J.: IETF RFC 3251: SIP: Session Initiated Protocol, the Internet Society, 2002
12. Schulzrinne, H.: IETF RFC 3550 RTP: A Transport Protocol for Real-Time Applications, the Internet Society, 2003
13. <http://www.realtimetext.org>
14. <http://www.mid-typetalk.org.uk>
15. The Liberated Learning Consortium, <http://www.liberatedlearning.com/>
16. IBM ViaScribe, http://www-03.ibm.com/able/accessibility_services/ViaScribe-accessible.pdf
17. Czech Union of Deaf, <http://www.cun.cz>
18. Czech law no. 155/1998 on comm. systems for the hearing-impaired, as amended by Act no.384/2008
19. <http://www.escribe.cz>
20. Meggelen, J. - Madsen, L. - Smith, J.: Asterisk: The Future of Telephony. 2nd Edition. Sebastopol: O'Reilly Media, 2007
21. The Apache Software Foundation server <http://www.apache.org/>
22. The Vodafone Foundation Czech Rep. <http://www.nadacevodafone.cz/>
23. <http://www.rendersoftware.com/products/camstudio/>
24. <http://www.virtualdub.org/>
25. ITU-T recommendation T.140: Protocol for multimedia application text conversation, ITU-T, 1998