Cloud-Based Assistive Speech-Transcription Services

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. Real-time speech transcription is a service of potentially tremendous positive impact on quality of life of the hearing-impaired. Recent advances in technologies of mobile networks, cloud services, speech transcription and mobile clients allowed us to build eScribe, a ubiquitiously available, cloud-based, speech-transcription service. We present the deployed system, evaluate the applicability of automated speech recognition using real measurements and outline a vision of the future enhanced platform, crowdsourcing human transcribers in social networks.

Keywords: Hearing-Impaired, Cloud Computing, Voice Recognition.

1 Introduction

Speech is one of the most natural means for sharing ideas, information or feelings. However, speech may also become a communication barrier for those not knowing the language or unable to use it, like the hearing-impaired or the foreign-language speakers. For those hearing-impaired who lost hearing during their life, transcription is a natural way of receiving information. Motivated by the growing number of the hearing-impaired in the Czech Republic (ca 500.000) and the fact that only 1.2% of them use sign language, the Czech Technical University in Prague (CTU) and the Czech Union of the Deaf (CUD)[4] have embarked on a joint project called eScribe [2], with the goal of reducing communication barriers and improving quality of life of the hearing-impaired. As part of the project we have designed and built a prototype cloud-based assistive speech-to-text services platform, providing ubiquitously available real-time speech transcription. The actual transcription may either be provided by real transcribers or by Automated Speech Recognition (ASR) engines.

2 Related Work

Today's speech-to-text services (STTS) are generally provided by one of the following methods: 1. physically present transcribers; 2. remote-transcription

K. Miesenberger et al. (Eds.): ICCHP 2012, Part II, LNCS 7383, pp. 113–116, 2012.

carried out by human transcribers [1][2]; 3. concepts based on ASR [4]; 4. ASR combined with human error-correction [3]. Physically present transcribers are highly limiting because there is shortage of educated transcribers, costs of these services are high and they are restricted to particular locations. An alternative is using ASR. Yet, ASR is limited in recognition accuracy, especially of colloquial speech and of difficult national languages such as Czech. Today, ASR systems are largely trained on literary texts. Dictionaries and hypotheses for national colloquial languages are missing. ASR techniques are sensitive to many miscellaneous characteristics of the input signal. From fundamental signal attributes (signal input level, sample frequency) across disturbing influences (background noise, other voices, music) to the culture of idea formulation (fluent speech with minimal unfinished sentence fragments). Remote transcription and ASR combined with human error-correctors is limited by costs and human resources too.

3 eScribe Implementation and Architecture

The currently operational eScribe platform utilizes the widespread Voice-over-Internet-Protocol (VoIP) telephony. Asterisk, a simple but powerful system for VoIP communication, acts as the core server. Audio and text transmission are controlled by the Session Initiation Protocol (SIP). In the eScribe architecture, Asterisk operates as a transparent gateway between the ASR server and Google Cloud. To interconnect eScribe with the ASR server, a modified SIP MESSAGE method is used. The Cloud part of eScribe is based on the Google App Engine, and communication with Asterisk is carried out by the Jabber protocol. The Cloud environment is used as a text editor, represented by Google Docs, as well as storage space for the transcribed texts. Some of eScribe internal logic is deployed in the Google Cloud too. There are several options how to deliver the audio signal to the system: using a cell phone, a fixed phone, a SIP phone or a webphone. From the user perspective, eScribe is used in two modes: *1. Lecture mode* (one speaker - many listeners), *2. Face-to-face communication mode.* Practical tests have shown that a DECT (Digital Enhanced Cordless Telecommunications) phone used as a wireless microphone by the speaker and a laptop (usually connected to a beamer) as a display is the best arrangement for the Lecture mode. For the Face-to-face communication, tablets or smartphones were the user equipment alternatives.

4 Performance Evaluation and User Experience

One of the key criteria impacting the quality of ASR is the choice of codec and its related parameters. eScribe uses ASR provided by Newton Technologies [9]. Fig. 1 shows recognition success rates comparing various audio codecs at different sample frequencies. The success rate was evaluated according to the methodology described in [6]. Final results were obtained as an arithmetic mean of simple results from a set of 13 utterances. The spectrum of utterances was not optimized in any way, so the utterance set consisted of audio records including many disturbances

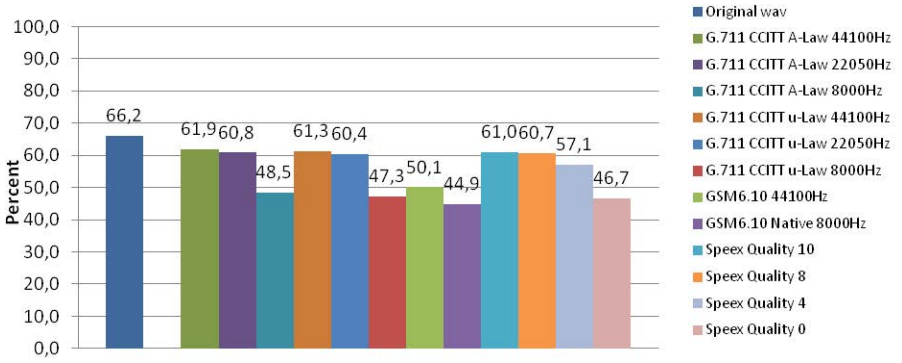

Fig. 1. Recognition Success Rates for Several Codecs and Different Settings. Recognition success rate of the original wav file, using a 44 kHz sample frequency at 8bit coding, is used as reference. The results can be grouped around two success-rate levels. The first group includes utterances with sample frequency higher than 8 kHz (quality parameter higher than 4 in SPEEX codec). Success rates of this group vary at about 60 %. The second group is characterized by success rates of around 50 %. Unfortunately, the common sample frequency for telecommunication audio signals is 8 kHz, which provides perceptibly worse performance than in the case of the other group.

such as background noise, voices, unintelligible parts etc., just as in ordinary everyday speech. In our scenario, the performance of ASR was quite independent of the general type of audio codec, but considerable improvement can be reached by using a wideband codec such as SPEEX or G.722 in VoIP telephony or AMR-WB in mobile telephony. Other methods to enhance results of the recognition process exist, e.g. text correctors or trained shadow speakers [4]. Nevertheless, the role of human transcribers is essential for a number of real-life situations where the accuracy of sentence understanding does play a crucial role in the life of an individual, such as at courts of law. Although eScribe was developed as a prototype and a proof of concept, we have arranged several events, where eScribe provided speech transcription to hearing impaired people. In the Lecture mode, several lectures at CTU were transcribed. Also CUD uses eScribe at their conferences or meetings. Using eScribe in the Face-to-face mode was firstly demonstrated in February 2012 in a cafe at Vodafone headquarter in Prague and later at a Vodafone shop in communication between a shop assistant and a hearing impaired client. [7]

5 Future Work

The limiting factors of today's systems for speech transcription are lack of well-educated transcribers or shadow speakers, and associated financial costs. Until the ASR systems are able to reliably recognize colloquial language or speech in noisy environment, the role of humans will remain irreplaceable. The boom of social networks brings a potential tool to attract transcribers. As demonstrator, we aim to interconnect our cloud-based transcription solution with a crowdsourcing

platform using gadgets built upon G-Talk, which can easily be integrated into social networks or webpages. The platform will support continuous improvement of transcription capabilities by developing a learning database of transcribed texts. Among the challenges to be addressed remain algorithms matching users to groups of appropriate transcribers, techniques of collecting the transcribed data and of creating colloquial dictionaries for ASR, anonymization algorithms for the transcribed speech, and methods for real-time auctioning of transcriber services.

6 Conclusion

In this paper, we have presented a prototype of a ubiquitous real-time speech-transcription service deployed in a cloud environment using both human transcribers and ASR technology. The wide availability of access to this real-time speech-transcription service will enable providing it to a much larger community of hearing-impaired people. High usage is expected in face-to-face communication, where the service of a physically present transcriber (assistant) is difficult to arrange from both financial and logistical point of view. The eScribe solution has a huge potential to minimize communication barriers and to make cultural, educational, social or other events more accessible to the hearing-impaired people. We have also proposed to reduce the limitations of the current solution - the shortage of well educated transcibers or shadow speakers - by crowdsourcing these services within social networks. This approach would help popularize the eScribe project and last but not least provide valuable scientific data for future research and development in the area of assistive voice services.

References

1. Miyoshi, S., Kuroki, H., Kawano, S., Shirasawa, M., Ishihara, Y., Kobayashi, M.: Support Technique for Real-Time Captionist to Use Speech Recognition Software. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 647–650. Springer, Heidelberg (2008)
2. Bumbalek, Z., Zelenka, J., Kencl, L.: E-Scribe: Ubiquitous Real-Time Speech Transcription for the Hearing-Impaired. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010. LNCS, vol. 6180, pp. 160–168. Springer, Heidelberg (2010)
3. Wald, M.: Captioning for Deaf and Hard of Hearing People by Editing Automatic Speech Recognition in Real Time. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 683–690. Springer, Heidelberg (2006)
4. Forman, I., Brunet, T., Luther, P., Wilson, A.: Using ASR for Transcription of Teleconferences in IM Systems. In: Stephanidis, C. (ed.) Universal Access in HCI, Part III, HCII 2009. LNCS, vol. 5616, pp. 521–529. Springer, Heidelberg (2009)
5. Czech Union of Deaf, `http://www.cun.cz`
6. IDIAP Research Institute - On the Use of Information Retrieval Measures for Speech Recognition Evaluation, `http://publications.idiap.ch/downloads/reports/2004/rr04-73.pdf`
7. `http://www.kochlear.cz`
8. `http://htk.eng.cam.ac.uk`
9. `http://www.newtontech.cz/`