YourSpeech: Desktop speech data collection based on crowd sourcing in the Internet

António Calado, João Freitas, Pedro Silva, Bruno Reis, Daniela Braga, Miguel Sales Dias
MSFT, Av. Prof. Dr. Aníbal Cavaco Silva, Ed. Qualidade C1, C2, TagusPark, 2744-010 Porto Salvo, Portugal
{i-antonc, i-joaof, i-pedros, i-dbraga, i-brreis, midias}@microsoft.com

Abstract. YourSpeech is an online platform based on crowd sourcing that aims at collecting desktop speech data at negligible costs for any language, in order to provide larger training data for Automatic Speech Recognition (ASR) systems. The platform provides means that allow users to donate their speech: a quiz game and a personalized TTS system. We have already collected more than 25 hours of pure speech for European Portuguese (EP) and created more than 90 personalized EP voices, since the system went live in late December 2010. After manually analyzing 12% of the collected corpus we found a Word Error Rate of 1.9%.

1 Introduction and System's Description

ASR systems based on statistical methods require vast amounts of transcribed and annotated speech data in order to achieve acceptable accuracy rates. However, these corpora are expensive and recruiting speakers has proven to be also quite costly [1]. The Yourspeech platform (http://pt.yourspeech.net) at MSN, aims at collecting speech at negligible costs for any language. The concept behind this system is to provide the user with an entertainment reward in exchange for his/her speech. This collection is based on crowd sourcing approaches [2] and invites the users to aid in the development of a new ASR technology, while at the same time they are entertained by playing a quiz game (JustSayIt), or by obtaining audio files pronouncing phrases with their own synthetic voices. The platform is online since December 2009 for European Portuguese (EP). The system architecture is based on the client/server paradigm (Fig. 1). The client accesses the platform through a website and is identified by the user´s Windows Live ID. Once there, the user chooses to play the quiz game or create his/her own personalized synthetic voice. In either case, an ActiveX client-side control is locally installed. This control enables the system to access the Windows audio pipeline and record audio using any of the installed devices. Then, the client goes through a setup to select the recording device and to guarantee the quality of the recorded audio. To perform this quality analysis, minimum Signal-to-Noise Ratio (SNR) values and server-side recognition accuracy rates results are considered. When the user chooses the quiz game branch, a Silverlight application is loaded by the client and the game is presented to the user. The difference from this quiz to other quizzes

found across the web [3] is that the questions are read by the default Text-to-speech (TTS) voice installed in the client system and the answers are recognized by a EP speech recognition engine installed in the server. After the answer is spoken, the recordings automatically stop and a wave file is streamed to the server for recognition. In the server, a dynamic grammar with the answers is generated and fed to the engine. If the recognized answer matches the correct answer the user scores points based on the answer difficulty. The quiz contains 18 thematic and generic questions split by easy, medium and hard difficulty. If the user selects the Personalized TTS option, a recording platform is presented. A minimum of 200 recorded sentences, taken from a phonetically rich set is required, in order to guarantee complete phoneme coverage. When enough sentences are recorded, the user may choose to generate his/her personalized voice and download audio files containing his/her synthesized voice. The quality of the voice is increased by the number of recorded sentences. At the time of writing of this paper, YourSpeech is online for 3 months and we have collected more than 25 hours of pure speech in EP. After manually analyzing 12% of the collected corpus (at a stage were 19,2 hours of pure speech were collected), we got 1.9% or Word Error Rate, which shows that our data collection approach is promising and effective.

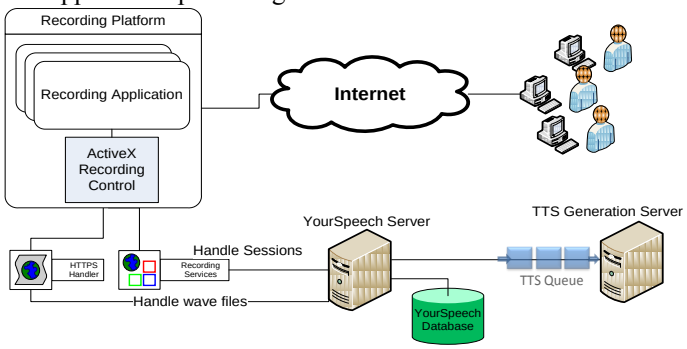

Fig. 1. High level system architecture diagram

We have been receiving extremely positive feedback from the users. The quiz game attracts much more speakers (460 completed sessions) than the Personalized TTS application (90 completed sessions); however a session of the latter produces 24 times more pure speech than the other. The quiz game ended up luring users for the personalized TTS.

References

1. Calado, A., Freitas, J., Braga, D., Dias, M.: Multi-Language Telephony Speech Data Collection and Annotation. In: Braga et al. (eds.) Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies, September 10, 2008, Curia, Portugal (2008)
2. Daren C. Brabham: Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. In: Convergence The International Journal of Research into New Media Technologies, Vol. 14(1), pp. 75-90 (2008)
3. http://www.funtrivia.com/, last seen on 03/08/2010