Crowd-sourcing platform for large-scale speech data collection

João Freitas[1], *António Calado*[1], *Daniela Braga*[1], *Pedro Silva*[1], *Miguel Sales Dias*[1]

[1] Microsoft Language Development Center, Portugal
{i-joaof, i-antonc, i-dbraga, i-pedros, Miguel.Dias}@microsoft.com

Abstract

This paper presents an online platform based on crowd sourcing for speech data collection, named YourSpeech. This platform 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. YourSpeech provides means for users to donate their speech through a quiz game and a through a platform that allows the deployment of a personalized TTS (Text-to-Speech) system. We have already collected more than 25 hours of pure speech for European Portuguese (EP) and achieved a Word Error Rate (WER) of 1% over 10% of the collected corpus.

Index Terms: Speech data, crowd sourcing, speech donation, Text-to-Speech, Automatic Speech Recognition

1. Introduction

It is well known that Automatic Speech Recognition systems based on statistical methods require vast amounts of transcribed and annotated speech data in order to achieve acceptable accuracy rates. Acquiring a lot of speech data is particularly difficult when addressing less-resourced languages or even any other language that is not amongst the big five in terms of market economic relevance (English, Spanish, French, German and Italian). The main reason for this is that these corpora are expensive and recruiting speakers has proven to be quite costly [1] and hard to manage. Besides, some speech databases lack quality because of the bad recording conditions, sample rates inconsistency, erroneous, inconsistent or inexistent transcription, etc. [2].

This paper describes a solution to tackle this issue by using a crowd-sourcing approach. Crowd-sourcing is a term used to describe the leveraging of vast amounts of people to achieve a specific goal in a collaborative manner over the Internet. Many crowd-sourcing initiatives have been made possible due to the availability of Web 2.0 technologies, which enable massive collaboration projects to take place. Crowd-sourcing can be considered as a distributed process for the resolution of problems. Typically the process is as follows: an entity has a problem and needs to solve it in a cost effective way. The entity publishes the problem in the web and usually provides the tools to solve it. Users (the crowd) respond to the call and propose solutions to the problem. The publishing entity chooses the winning solution and rewards the user/users accordingly. Rewards vary from money incentives to just public recognition. The publishing entity will own the final winning solution. Multiple solutions can be found across the web in order to digitize old books [2], transcribe speech [4], classify tunes [5], classify galaxies from the Sloan sky survey [6], find ideas for proposed problems [7], image [8][9] and video [10] tagging and even build a summary of the entire Human knowledge [11].

The Yourspeech platform (http://pt.yourspeech.net) at MSN [12] 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 [13] and invites the users to aid in the development of new ASR technology, while at the same time they are entertained by playing a quiz game (JustSayIt), or by obtaining audio files containing phrases pronounced by their own synthetic voices. European Portuguese is the language used in this prototype, but our goal is to scale it to other languages.

This paper is organized as follows: section 2 describes the system architecture, how the quiz game is built and how the personalized TTS voices are produced. In section 3, the media repercussion and users' experience and feedback is discussed. In section 4, the current results are depicted and in sections 5 and 6, future work and conclusions are presented respectively.

2. System description

The system architecture (Figure 1) is based on the client/server paradigm. The client application accesses the platform through a website and it is uniquely identified by the user's Windows Live ID. Once there, the user chooses to play the quiz game or create the user's own personalized synthetic voice. In order to access the client's operating system resources such as recording and playback devices, an ActiveX control is installed in the local machine. This control provides access to the Windows audio pipeline, thus enabling audio recording using any of the installed audio input devices.

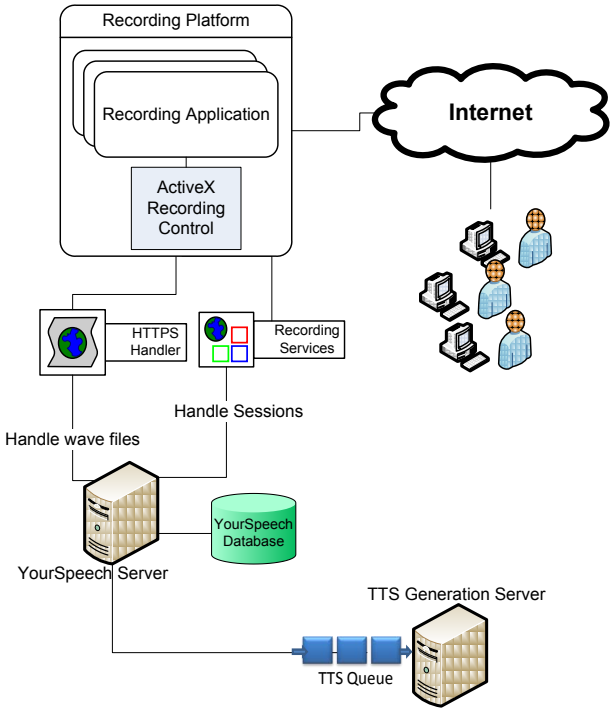

Figure 1: *High level system architecture diagram.*

After installing the ActiveX control, the client undergoes a setup phase in order to select the recording device and to guarantee the quality of the recorded audio. To perform this quality analysis, a Signal-to-Noise Ratio (SNR) value and server-side recognition accuracy rates' results are calculated before allowing the user to donate speech. After successfully passing the setup phase, the user is ready to donate speech.

2.1. Quiz game

When the user chooses the quiz game branch, a Rich Internet Application (RIA) based on Silverlight is loaded (Figure 2) and the game is presented to the user. The difference from this quiz to other quizzes found across the web [14] is that the questions are read by the default Text-to-speech voice installed in the client system and the answers are recognized by the correspondent speech recognition engine installed in the server. After the answer is spoken, the audio recording automatically stops and a wave file is sent to the server for recognition.

At the server side, 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 accordingly to the answer's difficulty. Each quiz contains 18 thematic and generic questions split by easy, medium and hard difficulty, randomly extracted from a total pool of 160 questions. This pool of questions was entirely generated taking into account the phonetic richness of the answers as well as content variability, i.e. we tried to create questions from various common themes such as, sports, history, geography, literature, mathematics, physics, etc.

Figure 2: *Quiz game.*

2.2. Personalized TTS

When the user selects the Personalized TTS option, a different Silverlight based interface is presented (Figure 3). A minimum of 200 recorded sentences, taken from a phonetically rich set of prompts, are required in order to guarantee complete phone coverage and the success of the TTS generation process. When the required number of sentences is recorded, the user may choose to generate his/her personalized voice. Upon the generation of his/her personalized voice, the user is able to create and download audio files containing his/her synthesized voice uttering any input text of his/her choice. The quality of the voice rises as the number of recorded sentences increases.

The personalized TTS system relies on a secondary server to generate the personalized voices. This server receives requests to generate a user personalized voice over Microsoft Message Queuing (MSMQ). Tests indicate that the server is able to generate over 20 voices simultaneously, however the performance impact increases the generation time for more than 24 hours.

Figure 3: *Recording platform for the Personalized TTS option.*

Due to this fact, the server only generates 12 voices simultaneously, leaving other voice requests on hold. The process of generating a voice with 200 recorded prompts takes approximately 20 minutes. A synthetic voice generated with 200 prompts is intelligible, but may lack naturalness, because more prompts are required and its quality may be degraded, since the audio capture may be done using a common laptop and simple recording hardware and software.

The TTS system used to generate the user's voices is based on HMM's [15]. The front-end is dictionary-based, being composed by a lexicon with around 135000 words, phonetically annotated by a professional linguist with phonetic transcriptions, stress marks and syllable boundaries, and with Part-of-Speech (POS) information. The stress and syllable marking was automatically assigned using linguistic rule-based algorithms, specially developed for the European Portuguese language. The front-end is also composed by the text analysis module, which involves the sentence separator and word breaker components, including several other files, such as phone set and features and the POS tags set. It also includes a rule-based TN (Text Normalization) module [16], as well as a homograph ambiguity (also polyphony) resolution algorithm [16][18], a stochastic-based LTS (Letter-to-Sound) converter, used to predict phonetic transcriptions for out-of-vocabulary words and the prosody models, which are data driven using a prosody tagged corpus of 2000 sentences and a POS tagger who provides morpho-syntactic contextual information. The front-end outputs phonetic transcriptions that are subsequently input of the TTS runtime engine or back-end, which then outputs synthetic voice.

A voice font is built with the users' voice samples recorded through Yourspeech online platform. The recorded waves must be coincident with the scripts used in the EP HTS-based synthetic voice. The voice font creation involves wave processing, automatic phonetic annotation using ASR acoustic models, alignment with orthographic input and compiling. In order to be able to dynamically create voices, the whole

pipeline process was automated. Figure 4 illustrates the system workflow.

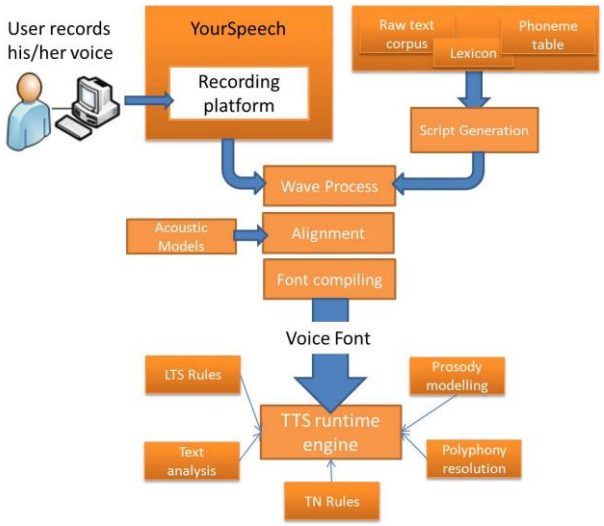

Figure 4: *Synthetic voice creation platform using YourSpeech.*

3. Media and User feedback

The success of this project is also dependent on the number of users that visit the site and actually record at least one prompt. For that reason, the user interface should attract the user's curiosity, but special emphasis should also be given to the dissemination and advertisement of the website. Figure 5 illustrates a clear correlation between media publication and number of visitors in the website. This graphic shows that YourSpeech started to be disseminated in some important European Portuguese blogs.

Later, it was integrated into MSN website and the highest peak of visits occurred when MSN announced and featured the Quiz Game in their entertainment area.

After that, YourSpeech appeared in several magazines publications [19][20][21], online TV stations [22] and it was referred in the television cable channel news TVI24 [23]. Site traffic statistics indicate 8682 visits and 1308 registered users.

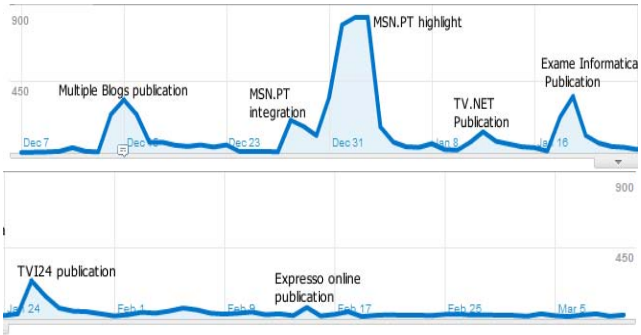

Figure 5: *Website visits from December to March.*

The website also allows users to comment on their experience. User feedback was extremely positive being reported that: "The game didn't have any issues in recognizing the answers, even when I misspelled some word"; "Good initiative! Keep up the good job and I'm looking forward for the final result"; "Very interesting game and a good cause for the Portuguese people"; "Excellent application. Good job".

4. Results

The YourSpeech platform went live at 14th of December 2009. Table 1 shows the speech/session results obtained until April 14th 2010.

Table 1. *Speech/Session results*

	Quiz Game	Personalized TTS	Total
Pure Speech (hours)	3.87	21.4	25.27
Total audio (hours)	11.9	48	59.9
Completed Sessions	473	94	567
Incomplete Sessions	205	223	428
Utterances	18300	9463	27763

Based on the results shown on Table 1, we can see that the Quiz Game is only responsible for 15% of the accumulated amount of pure speech, although it has 83% of the completed sessions. The reason for this to happen is due to the amount of time used when playing the Quiz Game (typically around 5 minutes per session) when compared with the time required to record the minimum 200 sentences for the personalized TTS voice (typically around 20 minutes per session).

From the 93 completed sessions, the total number of personalized TTS voices generated with success was 63.

A quality analysis was performed on the audio that has been collected during the first month, in order to check how well the speakers read what was proposed. The analysis was done with the help of a native Portuguese transcriber that listened and transcribed a sample of the collected corpus. The transcription process discards wrong utterances and corrects misspelled ones. The achieved results are described on Table 2.

Table 2. *Word Error Rate*

	Quiz Game	Personalized TTS	Total
WER	10.3%	0.05%	1%
Insertions	79	46	125
Deletions	92	103	195
Substitutions	36	47	83
Words	2010	40119	42129

Based the results from Table 2, we can see that the WER for the Quiz Game is much higher that for the Personalized TTS mostly due the number of insertions. Users tend to add articles or other words to the expected answer in order to complete their answer. In the personalized TTS system, the opposite happens. The user is told to read a sentence and we have seen that it is more common for users to omit or misspell words.

5. Future work

This platform has proven that crowd-sourcing can be a reliable and powerful tool to collect speech with a very small cost if the right motivation is provided and if a good marketing and

advertisement structure is in place. Future steps include: transcribe and annotate all the collected corpora, retrain existent acoustic models by adding the collected data and verify any changes in the ASR accuracy rate.

We would also like to create content-specific games that are focused on certain groups of words (e.g. city names, numbers, etc.) in order to have acoustic models specialized in specific grammar types.

Improvements to the platform include: increase the number of questions available in the quiz, provide a better user experience by using Silverlight 4 and its built-in microphone recording features; have the platform available beyond the browser, this is, online and offline or migrate totally or partially the platform to Microsoft's Azure platform [24] in order to use cloud computing. It is also planned to use this concept in order to collect other languages.

6. Conclusions

YourSpeech demonstrates how crowd-sourcing can be used to expand speech resources. This platform is essentially divided into two applications: a quiz game with a speech interface and a recording platform that allows generating the user's own personalized synthetic voice.

The quiz game attracts 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 proved to be a lure for the personalized TTS system. At the time of writing of this paper, YourSpeech is online for 4 months and we have collected more than 25 hours of pure speech in European Portuguese.

After manually analyzing 10% of the collected corpus, we got 1% total WER, which shows that our data collection approach is promising and effective. According to the experience obtained from the European Portuguese campaign, YourSpeech is a viable platform for obtaining speech data at marginal costs given the fact that appropriate marketing and advertisement actions are taken.

This concept can also be expanded to a multi-lingual platform and applied using more sophisticated games.

7. Acknowledgements

The authors would like to thank to the Public Relations department from Microsoft Portugal for all the marketing support and to all YourSpeech users for donating their speech and feedback.

8. 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] Neto, N., Patrick, C., Adami, A.G., Klautau, A. G., "Spoltech and ogi-22 baseline systems for speech recognition in Brazilian portuguese", in Teixeira, A., Lima, V., Oliveira, L., Quaresma, P. (eds) Propor 2008, LNCS (LNAI), vol. 5190, pp. 256-259, Springer, Heidelberg, 2008.

[3] Re-captcha website, Online: http://recaptcha.net/ accessed on 20th April 2010.

[4] Google new account creation, Online: http://mail.google.com/mail/signup accessed on 20th April 2010.

[5] Tag a Tune, Online: http://www.gwap.com/gwap/gamesPreview/tagatune/ accessed on 20 April 2010.

[6] Galaxy zoo, Online: http://www.galaxyzoo.org/ accessed on 20th April 2010.

[7] Idea Bounty, Online: http://www.ideabounty.com/ accessed on 20th April 2010.

[8] Squigl, Online: http://www.gwap.com/gwap/gamesPreview/squigl/ accessed on 20th April 2010.

[9] Google Image labeler, Online: http://images.google.com/imagelabeler/ accessed on 20th April 2010.

[10] PopVideo, Online: http://www.gwap.com/gwap/gamesPreview/popvideo/ accessed on 20th April 2010.

[11] Wikipedia.org, Online: http://en.wikipedia.org/wiki/Wikipedia accessed on 20th April 2010.

[12] MSN Portugal, Online: http://pt.msn.com/ accessed on 20th April 2010.

[13] Brabham, D. C., "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.

[14] Funtrivia website, Online: http://www.funtrivia.com/, accessed on 20th April 2010

[15] Braga, D., Silva, P., Ribeiro, M., Henriques, M. and Dias, M. "HMM-based Brazilian Portuguese TTS", Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies, September 10, 2008, Curia, Portugal,2008.

[16] Ribeiro, M., Braga, D., Henriques, M., Dias, M., Rahmel, H., "Resolução de Ambiguidades na Normalização de Texto em Português Europeu", in Actas do XXIV Encontro Nacional da Associação Portuguesa de Linguística. Textos Seleccionados, Braga, Portugal, 2009.

[17] Braga, D., "Algoritmos de Processamento da Linguagem Natural para Sistemas de Conversão Texto-Fala em Português" PhD Thesis. A Coruña University, Spain, 2008.

[18] Braga, D.; Coelho, L.; Resende Jr., F. G. V., "Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems", in Proceedings of Interspeech 2007. Antwerpen, Belgium. pp.1761-1764, 2007.

[19] Interview for Exame Informática (top business magazine in Portugal) magazine. Online: http://aeiou.exameinformatica.pt/a-microsoft-quer-por-os-portugueses-a-falar-para-o-computador-video=f1004730 accessed 20th April 2010.

[20] Exame Informática article. Online:http://www.pt.cision.com/O4KPTWeb/ClientUser/GetClippingDetails.aspx?id=8b7e0270-395c-4a9c-a8fc-4429c32d9b72&analises=1 accessed 20th April 2010.

[21] Expresso Online article. Online: http://aeiou.expresso.pt/gen.pl?p=stories&op=view&fokey=ex.stories/563758 accessed 20th April 2010.

[22] Interview for TV.net (on-line TV station in Portuguese). Online: http://tvnet.sapo.pt/noticias/video_detalhes.php?id=53031 accessed 20th April 2010.

[23] TVI24 interview. Online: http://www.agenciafinanceira.iol.pt/consola.html?id=1132691 accessed 20th April 2010.

[24] Microsoft Azure: http://www.microsoft.com/windowsazure/ accessed 20th April 2010