Crowd-sourced, automatic speech-corpora collection – building the *Romanian Anonymous Speech Corpus*

Stefan Daniel Dumitrescu
Tiberiu Boros
Radu Ion

RACAI

RESEARCH INSTITUTE FOR ARTIFICIAL INTELLIGENCE

Summary

- Introduction
- A Proof of Concept Platform - Step 1
- Future development - Step 2
- Conclusions

Motivation

- Major bottleneck for ASR/TTS research is the lack of free speech resources
- Even more so for Romanian
  - According to the MetaNet White Paper Series (outcome of the FP7 MetaNet umbrella projects), Romanian language is classified into the *fragmentary* support class (2nd lowest out of five), together with 14 other languages for speech and text resources.

Motivation

- Major bottleneck for ASR/TTS research is the lack of free speech resources

Speech Processing

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	▪ English	▪ Czech ▪ Dutch ▪ Finnish ▪ French ▪ German ▪ Italian ▪ Portuguese ▪ Spanish	▪ Basque ▪ Bulgarian ▪ Catalan ▪ Danish ▪ Estonian ▪ Galician ▪ Greek ▪ Hungarian ▪ Irish ▪ Norwegian (Bokmål, Nynorsk) ▪ Polish ▪ Serbian ▪ Slovak ▪ Slovene ▪ Swedish	▪ Croatian ▪ Icelandic ▪ Latvian ▪ Lithuanian ▪ Maltese ▪ Romanian ▪ Welsh

Motivation

- Major bottleneck for ASR/TTS research is the lack of free speech resources

Speech and Text Resources

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	▪ English	▪ Czech ▪ Dutch ▪ French ▪ German ▪ Hungarian ▪ Italian ▪ Polish ▪ Spanish ▪ Swedish	▪ Basque ▪ Bulgarian ▪ Catalan ▪ Croatian ▪ Danish ▪ Estonian ▪ Finnish ▪ Galician ▪ Greek ▪ Norwegian (Bokmål, Nynorsk) ▪ Portuguese ▪ Romanian ▪ Serbian ▪ Slovak ▪ Slovene	▪ Icelandic ▪ Irish ▪ Latvian ▪ Lithuanian ▪ Maltese ▪ Welsh

- http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

Our response

- Create the missing resources!
- Use the example of VoxForge, but adapt to our requirements, and then extend
- Use crowd-sourcing to build a free-speech, time-aligned, multi-user corpus.
  - Such a corpus is difficult to find for English, and virtually non-existent for Romanian.
- The platform needs to be autonomous and self-improving (~zero maintenance effort)

Expected Goals and Outcomes

- Time-aligned speech corpus
  - Used to train better ASR/TTS systems; used to automatically improve the platform itself.
- Free-speech unannotated corpus
  - Used to create test-sets (ex: multi-user ASR gold standard)
- Improved ASR and TTS algorithms
  - Allows us to experiment with the algorithms themselves as we have better corpora on which to train them

First step – is it feasible?

- Development of the website
  - Minimal user pre-requirements
  - Tech used: Javascript, HTML5 and Flash
  - Backend: PHP & SQLite
- Sentence database creation
  - 10K+ balanced set of sentences
    - Most were extracted from Wikipedia
    - Short sentences only
    - Properly terminated
    - No names, numbers, etc.

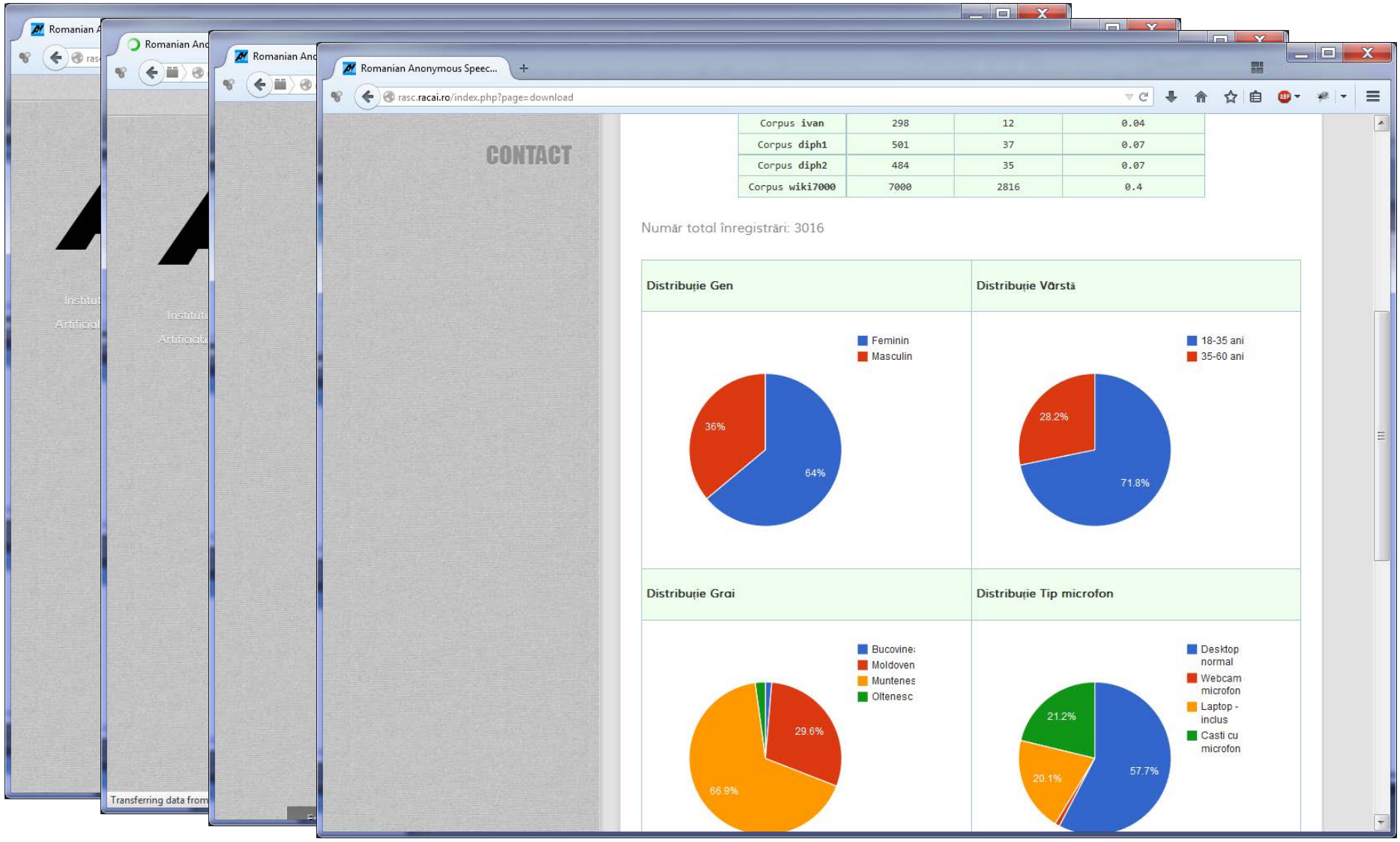
First step – is it feasible?

o Results:

- Around **4.3K sentences** in two distinct experiments so far.

- Relatively equal distribution male-female
- Three quarters of users are under 35 y/o
- Skewed distribution of mostly local users (Muntenia region), a third Moldovan users, and negligible number from Transylvania.
- Normal desktop microphone used most often, followed by headsets.

rasc.racai.ro – as it is now



Concept

- How can we make the platform grow faster? – Make it interactive!
  - Leisure is an important factor to take into account for users willing to spend time on our site
  - Attempt to create the “viral” factor. Make it fun and users will share it on.
- Data is gathered similarly : users speak predefined sentences, but in different settings, helped by built-in ASR and TTS modules.

Proposed games

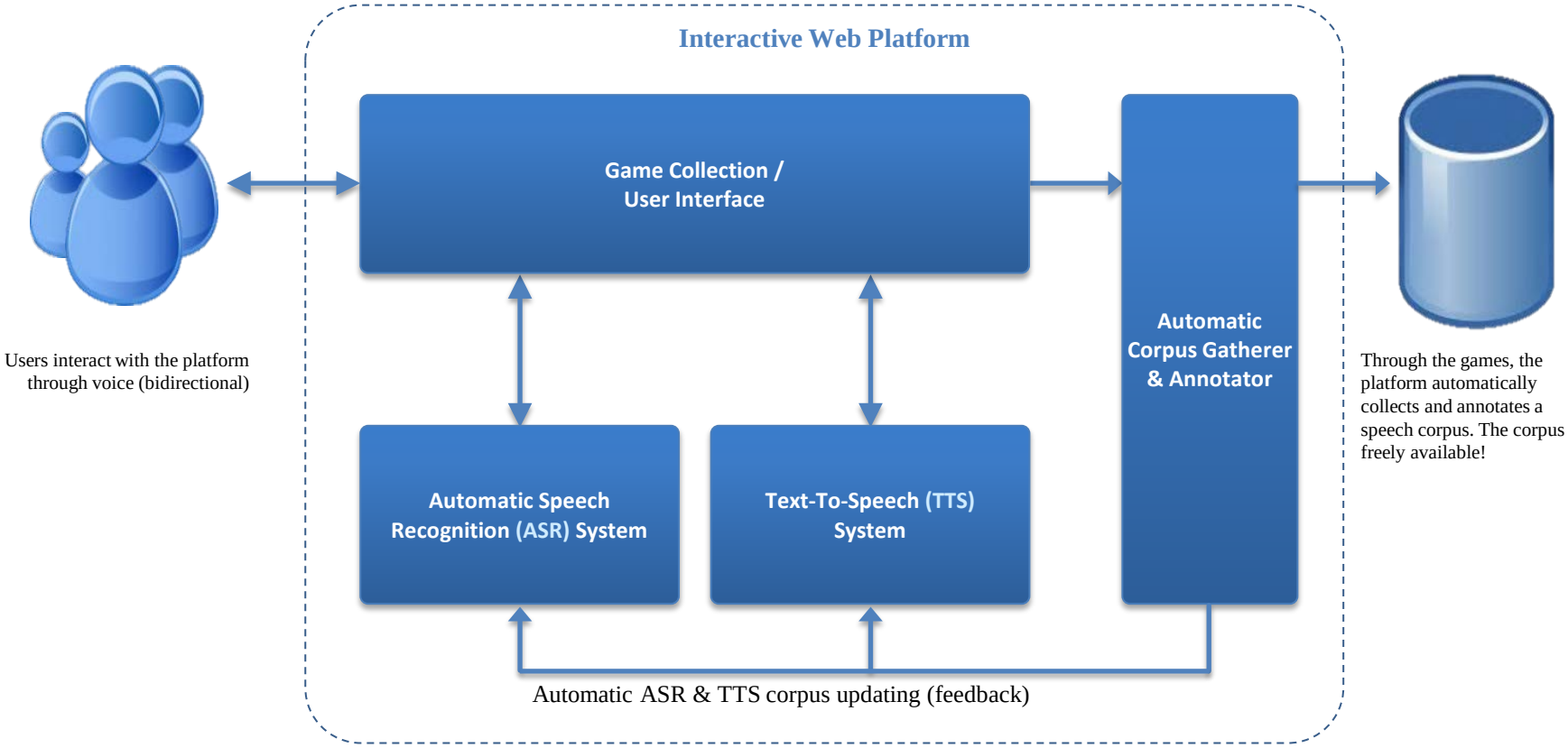
- Game 1 - Voice mimicking
  - after voice adaptation, the system will allow the user to input text and play it back using the user's own voice (including effects like pitch shift)
  - the user can save or share the results.
- Game 2 - Voice-morphing karaoke: the user will read lyrics (without singing them) from various karaoke songs.
  - We will modify his/her voice parameters to match that from the song, generating his voice on the recordings (just like a normal karaoke system).

Proposed games

- Game 3 - Voice-chat with a computer robot (bot)
  - This game will be a prototype bidirectional speech-to-speech system between the user and a computer bot.
  - Based on available online bots, we could sustain a mildly reasonable “conversation” with a user.

- We must note that all the game ideas are not new!
- The methods and technologies used to power them have been tried and successfully tested before in different scenarios.

Architecture



Current state

1. Voice mimicking. Done. The module is undergoing tests to see how it scales to multiple users.

2. Voice morphing karaoke. Almost done. Most of the basic components of the system are operational. We are currently working on integration of the modules and are checking karaoke licensing for some Romanian songs.

3. Voice chat. Under development. The NLP intermediary module has raised a number of issues, the major challenge being that we need it to work for the Romanian language, while almost all current implementations are in English. Other reasons: good rule-based chat-bots have to be trained automatically.

Current state

Sound modules:

o ASR

- Training: RASC sentence-level alignment + domain LM → Sphinx → acoustic model
- Running: Wav input → Sphinx → text

o TTS

- Better speech segmenting, better prosody pattern prediction, regional accent analysis, other indirect TTS improvements

Conclusions

- Currently, Romanian suffers from a lack of speech resources. As such, we are creating an online, self-sustainable, self-improving platform.
  - Initial proof-of-concept shows promise.
  - Currently working on second stage game-enabled platform.

- Goal: deliver the Romanian speech community a free, time-aligned speech corpus.

Thank you !