The SWARA Speech Corpus:
A Large Parallel Romanian Read Speech Dataset

Adriana Stan[a], Florina Dinescu[b], Cristina Țiple[b], Șerban Meza[a], Bogdan Orza[a], Magdalena Chirilă [b]and Mircea Giurgiu[a]

a) Communications Department, Technical University of Cluj-Napoca, Romania
b) Department of Otorhinolaryngology, Iuliu Hatieganu University of Medicine and Pharmacy, Romania

UNIVERSITATEA TEHNICĂ DIN CLUJ-NAPOCA

UMF UNIVERSITATEA DE MEDICINĂ ŞI FARMACIE IULIU HAŢIEGANU CLUJ-NAPOCA

Overview

Introduction

Recording process

Data segmentation

Results

Conclusion

Introduction

The SWARA Project

Mobile System for Rehabilitative Vocal Assistance of Surgical Aphonia

"SWARA will provide a portable, fast and easy to use assistive speech synthesis system for laryngectomized patients, enabling them to interact in an almost natural manner with other social participants by using a customised voice."

MINISTERUL EDUCAȚIEI NAȚIONALE

PN-II-PT-PCCA-2013-4 No. 6/2014

http://speech.utcluj.ro/swara

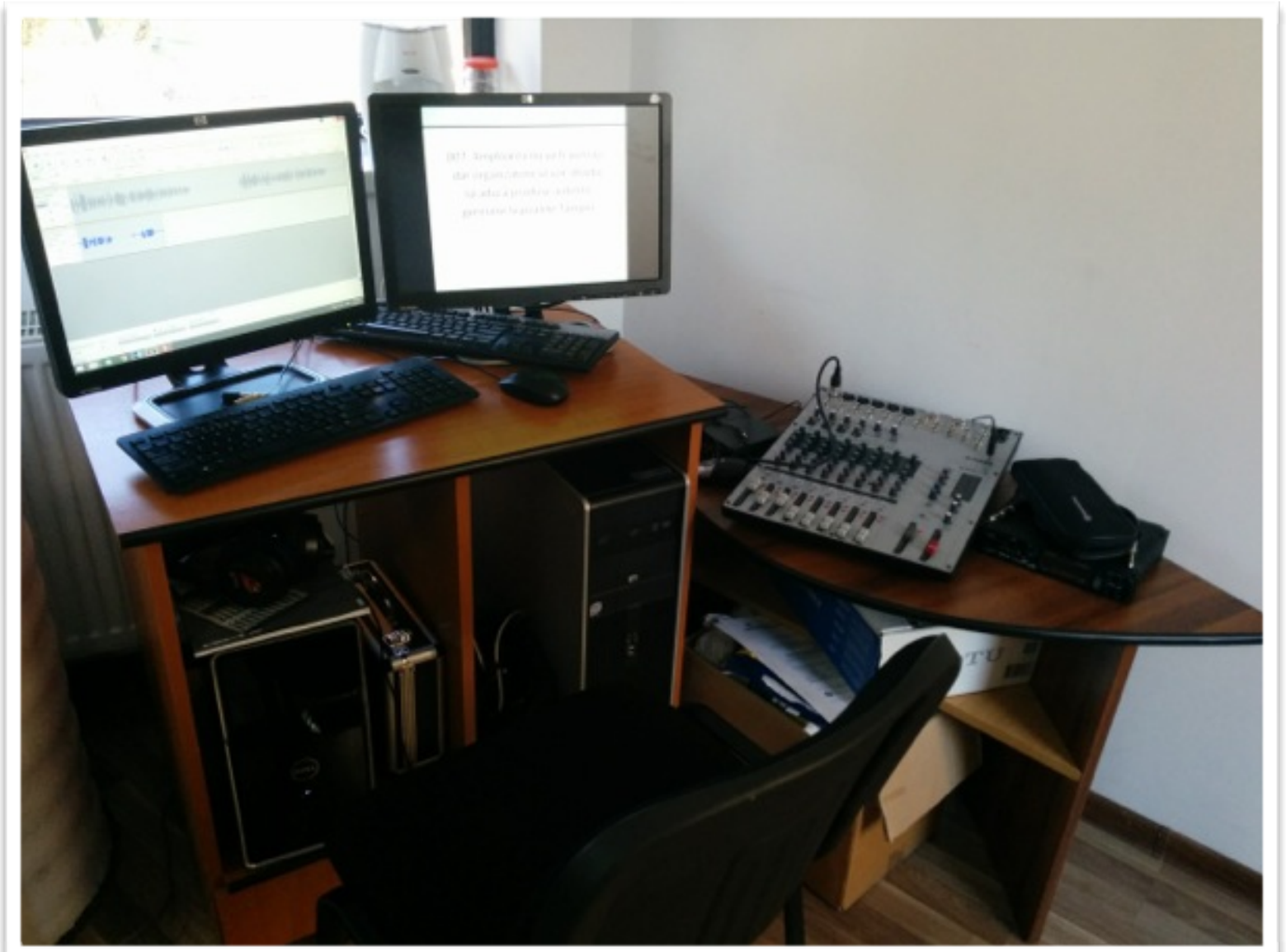
Romanian speech datasets

- **RSC** corpus for LVCSR - 100 hrs of speech, 157 speakers *(Cucu et al., 2014)*
- **Romanian Speech Synthesis (RSS)** corpus - 3.5 hours of speech, 1 speaker *(Stan et al, 2011)*
- **Romanian Anonymous Speech Corpus (RASC)** - 3000 utterances, multiple speakers *(Dumitrescu et al., 2014)*
- Romanian version of the **GRID** corpus - 400 utterances, 12 speakers *(Kabir and Giurgiu, 2011)*
- Romanian version of the **EUROM 1** database - 10 hours of data, 100 speakers *(Boldea et al., 1998)*
- **IIT** corpus - 45 minutes of speech, 3 speakers *(Bibiri et al., 2013)*

Recording process

Recording booth

Monitoring panel



Technical specifications

Technical specifications

- Sound proof booth

Technical specifications

- Sound proof booth
- AKG C214 large diaphragm microphone

Technical specifications

- Sound proof booth
- AKG C214 large diaphragm microphone
- communication via headphones with the outside supervisor

Technical specifications

- Sound proof booth
- AKG C214 large diaphragm microphone
- communication via headphones with the outside supervisor
- MOTU UltraLite MK3 sound card

Technical specifications

- Sound proof booth
- AKG C214 large diaphragm microphone
- communication via headphones with the outside supervisor
- MOTU UltraLite MK3 sound card
- Yamaha MW12c digital mixer

Technical specifications

- Sound proof booth
- AKG C214 large diaphragm microphone
- communication via headphones with the outside supervisor
- MOTU UltraLite MK3 sound card
- Yamaha MW12c digital mixer
- Audacity

Technical specifications

- Sound proof booth
- AKG C214 large diaphragm microphone
- communication via headphones with the outside supervisor
- MOTU UltraLite MK3 sound card
- Yamaha MW12c digital mixer
- Audacity
- 48kHz sampling rate at 16 bit depth

Technical specifications

- Sound proof booth
- AKG C214 large diaphragm microphone
- communication via headphones with the outside supervisor
- MOTU UltraLite MK3 sound card
- Yamaha MW12c digital mixer
- Audacity
- 48kHz sampling rate at 16 bit depth
- no pauses between utterances

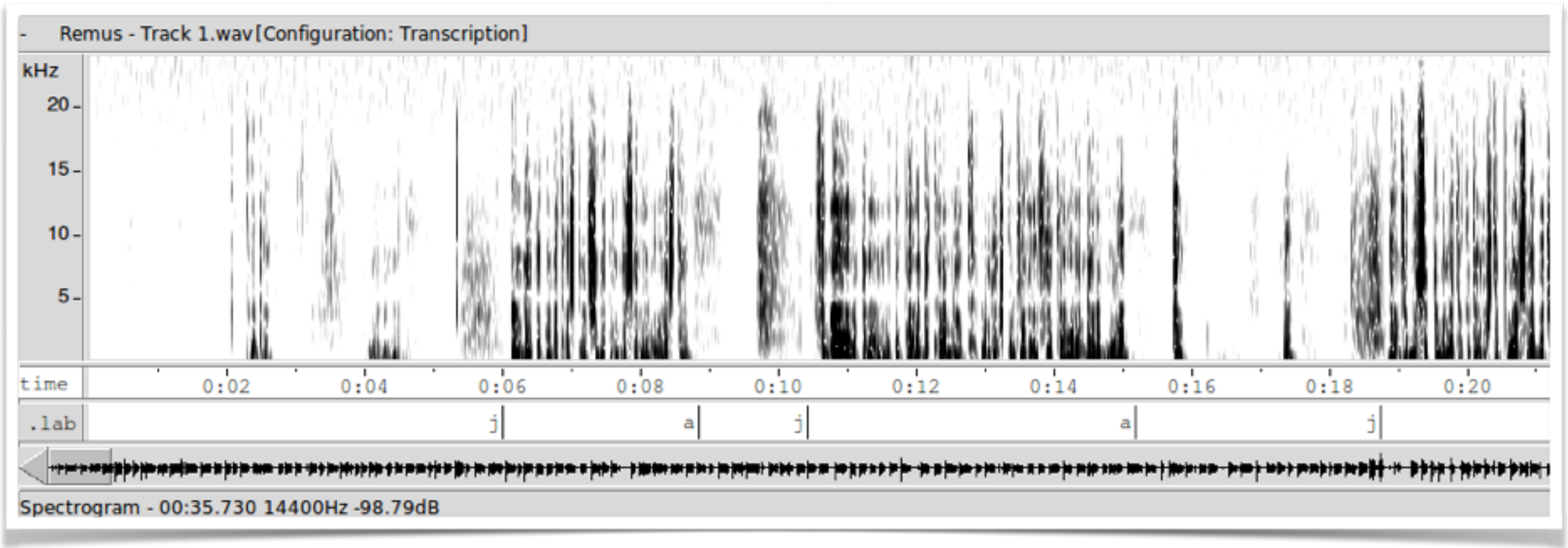
Recording prompts

RO	*Suntem una dintre cele mai vechi familii din Eforie.*
EN	We are one of the oldest families in Eforie.
RO	*Se poate face asta, dar depinde cum o faci, fiecare are modul lui de a vedea lucrurile.*
EN	This can be done, but it depends on how you do it, because everybody has his own way of looking at things.
RO	*Guvernul Britanic a comandat şaizeci de milioane de doze.*
EN	The British Government ordered sixty million doses.
RO	*De pildă, aş face un brand din brânza şi gemurile locale.*
EN	For example, I would brand the local cheese and jams.

Data segmentation

Utterance-level
Phone-level

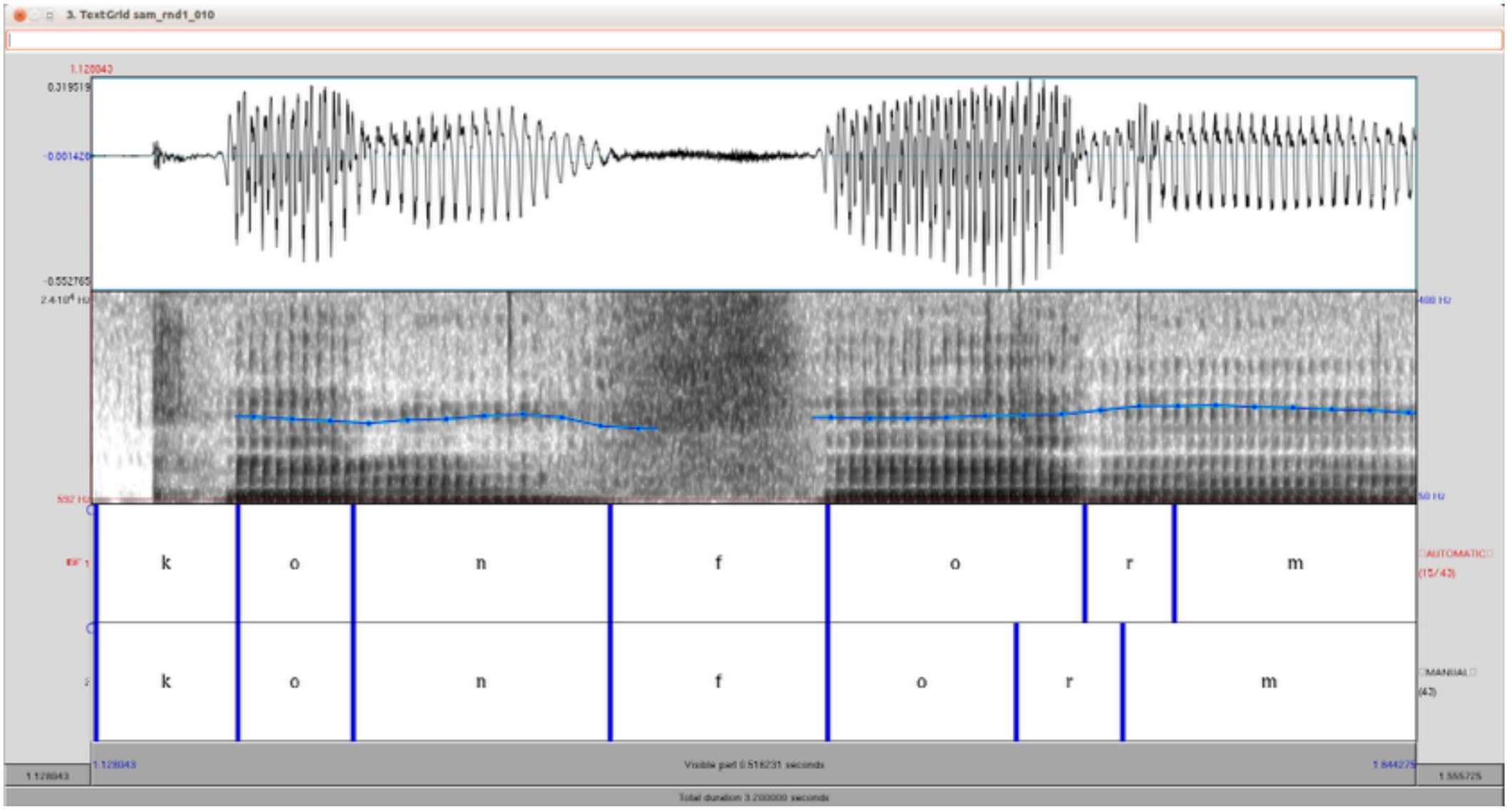
Utterance segmentation



Phonetic alignment

- SWARA front-end phonetic transcriber ~96% accuracy
- HMM-based acoustic models ~93% accuracy
- HTK, 5 state left-right configuration, 8 re-estimations, no state tying and a flat start, no speaker adaptation strategies

Phonetic alignment



Results

Copus contents
Synthetic voice building

SWARA Corpus Contents

- **17 speakers**: 7 male and 10 female
- aged between 20-35 years old
- with no self-declared hearing or speaking impairment
- mild regional accents
- **21 hours** and 19 minutes
- **19,292** utterances
- **880 common** utterances

SWARA Corpus Contents

No.	Speaker ID	Sex	Duration	No. of utts
1	BAS	F	1h34’	1493
2	CAU	F	1h11’	996
3	DCS	F	1h50’	1493
4	DDM	F	1h09’	996
5	EME	F	1h53’	1493
6	FDS	M	0h57’	996
7	HTM	F	1h06’	981
8	IPS	M	0h58’	996
9	PCS	F	1h08’	996
10	PMM	F	1h01’	921
11	PSS	M	1h27’	1486
12	RMS	M	1h08’	996
13	SAM	F	1h43’	1493
14	SDS	M	1h01’	996
15	SGS	M	0h55’	996
16	TIM	F	1h09’	973
17	TSS	M	1h01’	996

SWARA Corpus Contents

No.	Speaker ID	Sex	Duration	No. of utts
1	BAS	F	1h34’	1493
2	CAU	F	1h11’	996
3	DCS	F	1h50’	1493
4	DDM	F	1h09’	996
5	EME	F	1h53’	1493
6	FDS	M	0h57’	996
7	HTM	F	1h06’	981
8	IPS	M	0h58’	996
9	PCS	F	1h08’	996
10	PMM	F	1h01’	921
11	PSS	M	1h27’	1486
12	RMS	M	1h08’	996
13	SAM	F	1h43’	1493
14	SDS	M	1h01’	996
15	SGS	M	0h55’	996
16	TIM	F	1h09’	973
17	TSS	M	1h01’	996

SWARA Corpus Contents

No.	Speaker ID	Sex	Duration	No. of utts
1	BAS	F	1h34’	1493
2	CAU	F	1h11’	996
3	DCS	F	1h50’	1493
4	DDM	F	1h09’	996
5	EME	F	1h53’	1493
6	FDS	M	0h57’	996
7	HTM	F	1h06’	981
8	IPS	M	0h58’	996
9	PCS	F	1h08’	996
10	PMM	F	1h01’	921
11	PSS	M	1h27’	1486
12	RMS	M	1h08’	996
13	SAM	F	1h43’	1493
14	SDS	M	1h01’	996
15	SGS	M	0h55’	996
16	TIM	F	1h09’	973
17	TSS	M	1h01’	996

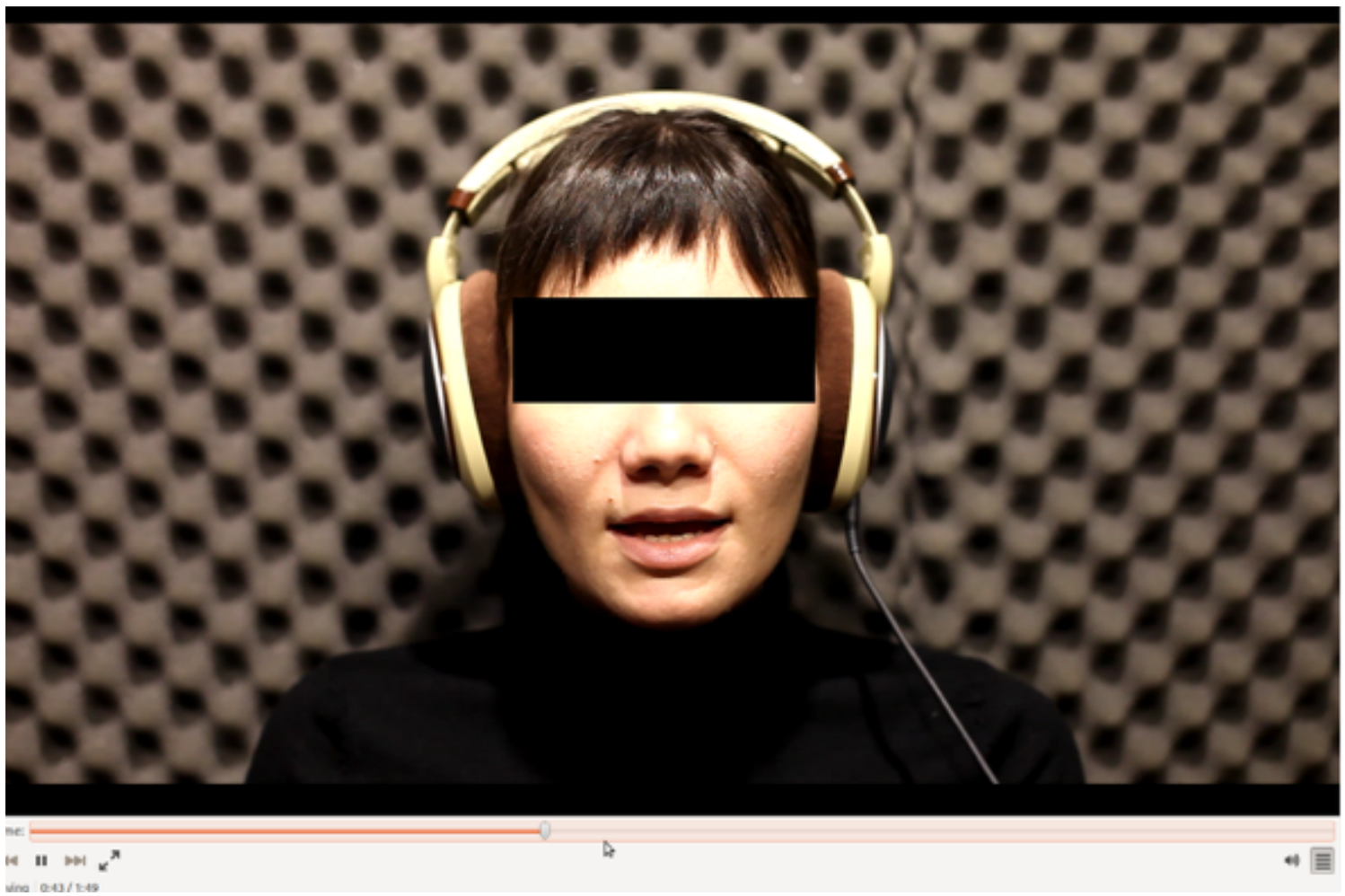
SWARA Corpus Contents

No.	Speaker ID	Sex	Duration	No. of utts
1	BAS	F	1h34’	1493
2	CAU	F	1h11’	996
3	DCS	F	1h50’	1493
4	DDM	F	1h09’	996
5	EME	F	1h53’	1493
6	FDS	M	0h57’	996
7	HTM	F	1h06’	981
8	IPS	M	0h58’	996
9	PCS	F	1h08’	996
10	PMM	F	1h01’	921
11	PSS	M	1h27’	1486
12	RMS	M	1h08’	996
13	SAM	F	1h43’	1493
14	SDS	M	1h01’	996
15	SGS	M	0h55’	996
16	TIM	F	1h09’	973
17	TSS	M	1h01’	996

SWARA Corpus Contents

No.	Speaker ID	Sex	Duration	No. of utts
1	BAS	F	1h34’	1493
2	CAU	F	1h11’	996
3	DCS	F	1h50’	1493
4	DDM	F	1h09’	996
5	EME	F	1h53’	1493
6	FDS	M	0h57’	996
7	HTM	F	1h06’	981
8	IPS	M	0h58’	996
9	PCS	F	1h08’	996
10	PMM	F	1h01’	921
11	PSS	M	1h27’	1486
12	RMS	M	1h08’	996
13	SAM	F	1h43’	1493
14	SDS	M	1h01’	996
15	SGS	M	0h55’	996
16	TIM	F	1h09’	973
17	TSS	M	1h01’	996

Video recordings

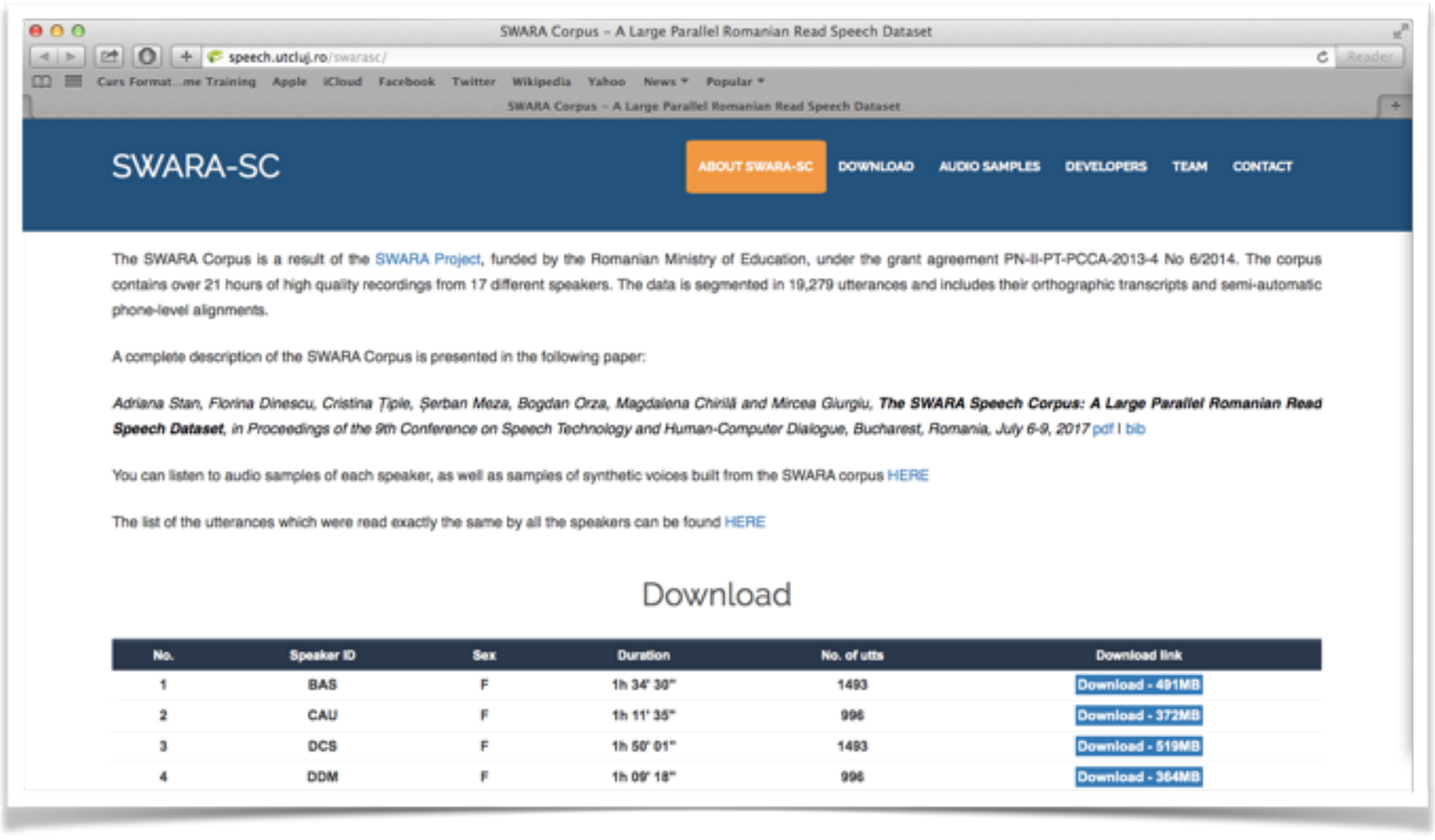


Synthetic voice samples

- HMM and DNN-based text-to-speech systems
- STRAIGHT and WORLD vocoders
- http://speech.utcluj.ro/swarasc/samples/

Download

http://speech.utcluj.ro/swarasc



Conclusions

Conclusions

- one of the largest Romanian spoken datasets freely available for both academic and commercial use
- 21 hours of data at 48kHz sampling rate
- recorded in a professional environment
- includes semi-supervised phonetic alignment
- future work: automatic speech recognition and speaker adaptation for TTS systems

Thank you for your attention!

Questions?

adriana.stan@com.utcluj.ro
www.speech.utcluj.ro/swarasc

References

(Cucu et al., 2014) H. Cucu, A. Buzo, L. Petric, D. Burileanu, and C. Burileanu, “Recent improvements of the SpeeD Romanian LVCSR system,” in Proc. of The 10th International Conference on Communications (COMM), May 2014, pp. 1–4.

(Stan et al., 2011) A. Stan, J. Yamagishi, S. King, and M. Aylett, “The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate,” Speech Communication, vol. 53, no. 3, pp. 442–450, 2011.

(Dumitrescu et al., 2014) S. D. Dumitrescu, T. Boros, and R. Ion, “Crowd-sourced, automatic speech-corpora collection - building the Romanian Anonymous Speech Corpus,” in Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL2014),Reykjavik, Iceland, May 2014, pp. 90–94.

(Kabir and Giurgiu, 2011) A. Kabir and M. Giurgiu, “A Romanian Corpus for Speech Perception and Automatic Speech Recognition,” in Proceeding of 10th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications, 2011, pp. 323–326.

(Boldea et al., 1998) M. Boldea, C. Munteanu, and A. Doroga, “Design, Collection, and Annotation of a Romanian Speech Database,” in Proceedings of 1st Conference on Language, Resources and Evaluation, 1998.

(Bibiri et al., 2013) A.-D. Bibiri, D. Cristea, L. Pistol, L. A. Scutelnicu, and A. Turculet, “Romanian Corpus For Speech-To-Text Alignment,” in Proc. of the 9th International Conference on Linguistic Resources And Tools For Processing The Romanian Language, 2013, pp. 151–162.