Towards adaptive, multi-domain speech transcription systems

VOCAPIA research

LREC 2022 Industry Day - 22. June 2022

Who Are We ?

- R&D company and software publisher founded in 2000
- Specialized in state-of-the-art speech processing technologies
- Privileged partnership with LISN Univ. Paris-Saclay/CNRS lab
- Participation in (inter)national research projects

Our products

- VoxSigma® Software Suite (SaaS or on-premise)
  - Audio & Speaker Segmentation
  - Language Identification
  - Speech-to-text Transcription
  - Speech-text Synchronization
  - Keyword Spotting
- Applications
  - Telephone speech analytics
  - Media monitoring
  - Transcription (parliament hearings, conference calls. . .)

Current challenges

- Is speech-to-text solved ?
  - Vocal assistants everywhere
  - Open source toolkits for machine learning
  - Lots of linguistic corpora and pre-trained models
- But are current performances well assessed ?
  - Publications often rely on easy benchmarks
  - Performance of systems are over-estimated (Szymański et al, “WER we are and WER we think we are”, Findings of ACL, 2020)
- Challenges of real-life application remain
  - (Highly) noisy acoustic conditions
  - Foreign accents
  - Very spontaneous speech with overlaps
  - Code-switching
  - Under-resourced languages

New languages and dialects

- Low-resource languages
  - Sharing data between (similar) languages
  - Multilingual acoustic/phonetic models
  - Similar trend for linguistic models
- Code-switching
  - Too short for a purely acoustic segmentation
  - Lexical-level has to be taken into account
  - Ideally, a bilingual transcription system
- Data sparcity is always an issue
  - Creative combinations of low-supervised learning and data augmentation

Towards multi-domain systems

- Solutions often specific to applicative domain
  - Broadcast Speech (Radio/TV/internet)
  - Conversational Telephone Speech (CTS)
  - Teleconferences
  - Air trafic control
  - . . .
- Needs to
  - Be more robust to domain changes
  - Evolve towards more generic solutions
- Recent multi-domain developments
  - Performed for several languages (English, French, Arabic. . .)

Use cases

- Multi-domain, multi-dialect Arabic
  - High linguistic variability, code-switching
  - Few written or audio data (tweets/blogs)
  - Transfer learning and adaptive networks
  - Résulting system more efficient than separate specific models
- Noisy speech
  - Challenging VHF/UHF communication conditions
  - 4-fold WER increase on noisy corpus
  - Adaptation of a CTS English transcription system
  - More robust multi-channel models

Conclusions and Perspectives

- Conclusions
  - Continuous improvement of ASR systems
  - . . . but still far from an universal off-the-shelf solution
- On-going scientific progress
  - Optimisation of neural architectures
  - Development of lighly supervised approaches (self-sup. learning)
  - Better data sharing between languages, domains. . .
- Relevant linguistic corpora remain the key to success!