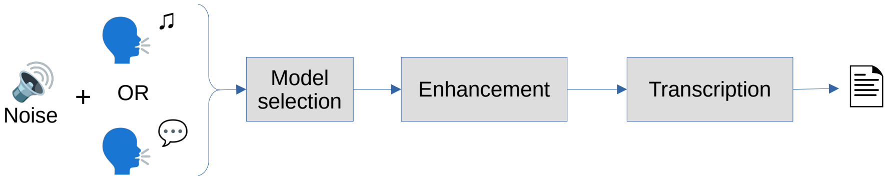
Automatic speaking and singing voice transcription in noisy and reverberant conditions on the edge



Supervisors: Mathieu Fontaine, Louis Bahrman, Sicheng Mao (IDS) - *prenom.nom@telecom-paris.fr*
Minimum number of students per group: 3
maximum number of students per group: 5
How many groups for this project ? 1
Tags: Transcription automatique de la parole, transcription de paroles, débruitage, déréverbératon, edge computing / Automatic Speech Recognition, lyrics transcription, denoising, dereverberation, edge computing

1 Context

The growing popularity of voice assistants has created the need for robust Automatic Speech Recognition (ASR). In real-life conditions, the performance of ASR systems is hindered by degradations, such as reverberation or noise, of audio signals captured by distant microphones. Recent approaches consider end-to-end denoising, dereverberation and transcription, for better performance [1]. However, such approaches represent a huge memory and time complexity at inference, such that they have a significant energy consumption, and a high carbon footprint. In this case, speech signals, which are a private data, have to be processed online. Smaller models [2, 3, 4] can get closer to the performance of large models by preprocessing the audio signals. Preprocessing techniques often involve dereverberation [5] and/or speech enhancement [6]. Furthermore, the generalization performance of such smaller systems on singing voice, for the task of automated lyrics transcription (ALT) for instance, has been only little explored [7].

This project aims at building a privacy-centric, energy-efficient model for the task of Automatic Speech Recognition and, if feasible, Automated Lyrics Transcription.

2 Expectations

Detailed description This project is divided in several parts. First we need to obtain a first impression and fundamental knowledge by a pedagogical task. Then students are free to explore and implement any relevant topics. We recommend the following roadmap:

- Automatic selection of the speech or singing voice transcription model. **An SVM classifier of MFCC features MUST be implemented.**
- Choice and evaluation of an ASR model for clean speech.
- Evaluation of the chosen ASR model in noisy-reverberant conditions
- Choice and evaluation of preprocessing methods on ASR performance
- Evaluation of the chosen ASR method on singing voice
- Choice and evaluation of a ALT method
- (Bonus) Implementation of an interface on smartphone

It is *not* necessary to cover the whole path and students are free to modify them after reasonable discussion with the supervisors. It is also not expected to reimplement from scratch but to make use of existing implementations.

Deliverables

- Demonstrator in form of a laptop application
- Open-sourced code of the demonstrator on github
- (Recommended) Jupyter notebook explaining design choices and some performance metrics
- (Optional) Smartphone interface to record audio and send it to the demonstrator

Expectations The following knowledge is expected after the accomplishment of this project

- Scientific
  - audio features classification: MFCC (TSIA) and SVM (SD-TSIA211)
  - Signal processing (TSIA, SI-101)
  - Speech transcription (TSIA206)
- Tools/soft-skills
  - Python programming language and commonly used libraries e.g. numpy, matplotlib, scikit-learn, pytorch, pykaldi et etc, with package management tools like pip, conda or mamba.
  - Read documentations of a toolkit and make use of it.
  - version control and group organization: git, github, code editors: using one of jupyter lab, vscode, spyder et etc.
  - Bibliography
  - Time management

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, PMLR, 23–29 Jul 2023.

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF, IEEE Signal Processing Society, 2011.

[3] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014.

[4] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[5] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1717–1731, Sept. 2010. Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[6] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, “Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7407–7411, 2022.

[7] K. Vijayan, X. Gao, and H. Li, “Analysis of speech and singing signals for temporal alignment,” in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1893–1898, 2018.