IEEE Journal on Selected Topics in Signal Processing
Far-Field Speech Processing in the Era of Deep Learning

November 9th 2018

Dear JSTSP Editor,

We are respectfully submitting a manuscript for consideration of publication in IEEE Journal on Selected Topics in Signal Processing – Far-Field Speech Processing in the Era of Deep Learning, entitled "*Overlapped speech detection and competing speaker counting – humans vs. deep learning*".

The manuscript extends the work published in the Proceedings of INTERSPEECH Conferences in 2014, 2015, 2017:

1. V. Andrei, H. Cucu, A. Buzo, C. Burileanu, "Detecting the number of competing speakers – human selective hearing versus spectrogram distance-based estimator," INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association Proceedings, 2014, pp. 467 – 470
2. V. Andrei, H. Cucu, A. Buzo, C. Burileanu, "Counting competing speakers in a timeframe – human versus computer", INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association Proceedings, 2015, pp. 3999-4003
3. V. Andrei, H. Cucu and C. Burileanu, "Detecting overlapped on short timeframes using deep learning", INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association Proceedings, 2017

The scope of our research is to develop methods for automatic counting of competing speakers on overlapping speech timeframes, with the goal of assisting applications like speech source separation, end-to-end speech recognition, speaker diarization, etc. The current paper is a complete overview of our entire work in this domain, with 50% of the manuscript being dedicated to new, unpublished methodologies and results.

The manuscript summarizes the findings in [1] - [3] and then focuses on the new content being introduced: counting competing speakers on sub-second timeframes using deep-learning. While in [1] and [2] we approached the same problem of competing speaker counting, we did so by using non-deep-learning strategies, mainly based on statistical analysis and by computing similarity degrees to reference templates. Also, in [1] and [2] we were targeting different practical applications like crowd-sensing by analyzing voice signals generally longer than 5 seconds and using a limit of 10 concurrent speakers.

In this manuscript, we are more concerned by the adoption potential of our models and, therefore, we aim for concurrent speaker counting on frames up to 25 milliseconds, on speech mixtures with up to maximum 4 sources, and we are using deep-learning techniques to accomplish this goal. One section of the paper reiterates the results of a perception experiment [1], [2] designed to quantify human level performance, that can be of interest to cross-domain researchers that study human selective auditory attention.

All the source code used to produce the results in we propose in this paper are open-source: *https://github.com/valentinandrei/saa_experiments*

Thank you very much for your consideration.

Yours sincerely,
Valentin Andrei, Ph. D.
University Politehnica of Bucharest, Romania / Intel Corporation, CA, USA
valentin.andrei@intel.com
+1 408 601 8125