Review for Katerina Zmolikova's PhD Thesis titled "Neural Target Speech Extraction"

by Hakan Erdogan

May 31st, 2022

This is a review for the PhD thesis titled "Neural Target Speech Extraction" written by Katerina Zmolikova, submitted to Brno University of Technology Faculty of Information Technology. The thesis was done under the supervision of Prof. Dr. Ing. Jan (Honza) Cernocky.

The thesis is addressing the problem of "target speech extraction" using neural networks. In a mixture of multiple speech sounds, the task is to extract the one corresponding to a target speaker's speech from the mixture. To give information about the target speaker, an auxiliary recording ("enrollment utterance") from that speaker is provided to the model. This thesis reviews this problem and many approaches that have appeared in the literature for solving it. It also compares this approach with the alternative approach of speech separation which tries to separate all speech sounds in a mixture. Finally, the thesis also reviews how these models can be used to improve speech recognition in overlapping speech conditions and how these models can be used together with diarization models to improve speech recognition.

The author, Katerina (Katka) Zmolikova was the one who first proposed this task and posed a possible solution to it. Katka had many follow-up papers and studies, most of which are summarized in this thesis. Having followed some of these publications, it was nice to see a well-written concise version in the form of a PhD thesis that comprehensively covers the methods explored in these studies. The thesis also attempts to compare earlier methods with recent approaches that appeared in the literature and that makes it a nice reference for future studies in the field.

The topic of the thesis is very appropriate and timely since speech separation and extraction have been an area of interest both for academia and industry recently. These models potentially have many applications in today's advanced communication devices and for intelligent processing of real-life audio recordings. The techniques described in the thesis are state-of-the-art techniques which are currently an important and ongoing topic of study.

This thesis includes original work that has been performed by the author and goes beyond what has been presented in earlier papers by the author. I definitely appreciate the effort made by the author to not take the easy route by collecting earlier published papers as different chapters

which could also have worked as a thesis. However, the author took the time to prepare an original thesis that sorts and combines the information from earlier papers and does new experiments to compare the alternative methods introduced by other researchers in the literature. This makes the thesis a document that strongly stands on its own.

The ideas in the thesis have been presented as 7 different papers as mentioned in Section 1.2 of the thesis. These publications have been very well cited and appreciated a lot by the research community and these studies have received a good number of citations (255 times total according to Google Scholar).

The content of the thesis and the list of the publications imply that Katerina Zmolikova is a person with an outstanding research erudition. I appreciate the mathematical clarity in the thesis and even the description of other side topics, such as UBMs, x-vectors, mixture models are very well done in a short but accurate way.

Here are some suggestions and detailed comments about the content of the thesis which may help improve the thesis. I would like the ones in bold to be answered / discussed during the defense.

1. In Equation 2.2 (and 3.2), the STFT domain representation of the time-domain convolution requires some additional assumptions. Maybe a footnote can be added to clarify those assumptions. It is assumed that the STFT frame length is longer than the length of the RIR signal, since otherwise an STFT domain multiplication (followed by an inverse STFT) will not be enough to get the same output as the time-domain convolution. Actually, even that assumption is not enough since multiplication in the STFT domain cannot fully represent a time-domain convolution due to circular convolution and the way inverse STFT is typically done. So, maybe we can say “It is assumed that the STFT domain multiplication followed by an inverse STFT yields a sufficiently close result to the time-domain convolution which may be achieved by choosing the right frame length and FFT size for the STFT.”
2. In 3.4.1, when describing the Lombard effect, instead of “the speech level increases”, we can say “human speakers tend to produce louder and prosodically different speech” to make it clearer what is meant.
3. In 3.4.4, maybe provide some pointers to the future chapters and sections that talk about addressing the domain mismatch problem, namely weakly supervised and unsupervised training approaches.
4. In Equation (4.2), (4.4), it may help to explicitly indicate the time dimension of the layer inputs (l_k) since the \lambda does not have a time dimension and it is concatenated with all time frames, it may help the readers if we use a t subscript for the layer outputs. This is made clearer when talking about the attention based method which uses a time-varying embedding, but I think it may be helpful for the discussion and understanding. If the math is not simplified by doing this or it is difficult to do it in some other sense, then mentioning the fact that the embeddings (\lambda) are concatenated with all frames of the features would also work.

5. In 4.8.3, it is mentioned that the multiplication-based method of informing the network is used. It is not clear at this point why this is done. Giving a pointer to the comparison in Table 4.8 may help.
6. In 4.8.3, mentioning that test time enrollment utterance lengths can be different from 0.5 seconds may be good.
7. **Are all the results in the paper done with 8 kHz data? Would it help to have at least one experiment with 16 kHz data? Do you anticipate any difference in results with respect to the sampling rate?**
8. In Section 4.8.4, “the lowest Si-SDR” -> “the highest SI-SDR”.
9. **In Section 4.8.4, related to the results in Table 4.5, 4.6, for a more clear description of the experiment, please say that all speakers are extracted (one by one) from the mixture to compare with the separation outputs. You can add a simple figure of how the extraction for all speakers is done and how separation is done. For example some information about how the “enrollment utterances” were chosen for a pair in a mixture may be a good addition. Are the enrollment utterances fixed for a given mixture or do they change during training? How about for evaluation?**
10. Some discussion of the assumption that the target speaker exists in the mixture signal and maybe some results showing when the target does not exist in the mixture (for example reporting the energy of the output wrt the input energy) would be nice. I am guessing to handle this case, some new training may need to be done, so to avoid that, I would just add some text about this important assumption when discussing Table 4.6.
11. **In Section 4.8.7, in the third paragraph, the issue about “incorrect identification” seems in conflict with the text in the forth paragraph which says separation is not working in those “failure” cases. It would be good to clarify whether it is an incorrect identification issue (that is the other speaker’s speech is chosen instead of the target) or that separation is totally failing and producing incorrect results in general. A clearer analysis of these failure cases may be a good addition to the thesis.**
12. In Table 4.7 headers, “Target-in**tefe**rence” -> “Target-in**terfe**rence” in two places.
13. In Section 4.8.11, one other way to obtain 512 STFT features with 16-sample frame length would be to perform zero padding to perform a larger size FFT for each frame. The way described in the thesis also looks fine but I wanted to mention this alternative.
14. In Figure 5.6 and/or the corresponding text, please mention the number of CACGMM components. Is it 3?
15. In Section 6.3.2, in Equation (6.3) there is SR, but in the text there is SER. There seems to be a need to correct one of them.
16. In Section 6.3.3, please mention in the beginning that the VBx method only does non-overlapping diarization instead of mentioning it later in the text.
17. In the discussion of Table 6.4, I think “the degradation of DER in the *fair* condition” -> “the degradation of DER in the *forgiving* condition” since there is degradation only in the forgiving condition in Table 6.3.
18. Section 6.5 title should be “Using diarization labels to fine-tune TSE for speech recognition” since the TSE model is fine-tuned, not the ASR model.

19. Section 6.5, if I recall correctly, “mixture consistency” was first introduced in this paper: “Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., & Saurous, R. A. (2019, May). Differentiable consistency constraints for improved deep speech enhancement. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 900-904). IEEE.”
20. In Section 6.5.4, I think instead of “re-training”, a better term would be “fine-tuning” since it is the more common usage I believe. Also, please mention how many iterations are used for the fine tuning since maybe we do not want to deviate too much from the starting model weights.

In conclusion, in my opinion, the doctoral thesis strongly meets the requirements of the proceedings leading to a PhD title. I have been very pleased to review this thesis. Let me know if there are further questions about this review.

Hakan Erdogan
Research Scientist at Google
hakanerdogan@google.com, hakan.erdogan.phd@ieee.org