NTT, INC.
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0237 Japan.
Telephone: +81 (0)50 3482 1753

Kyoto, November 7th, 2025

Review of a Doctoral Thesis titled: Far-field speaker verification incorporating multi-channel processing

Name of the doctoral student: Ladislav Mošner

Name and institution of the reviewer:
Marc DELCROIX,
NTT Communication Science Laboratories,
NTT, Inc., Japan

Please find below my review of the thesis.

I. Thesis review

This thesis investigates the development of robust speaker verification (SV) systems for far-field speech acquisition, where microphones are positioned at a distance from the speaker. In such conditions, the microphone signal is corrupted by acoustic interference such as background noise, reverberation and eventually interference speakers which significantly degrade signal quality and, consequently, SV performance. The research specifically explores multi-channel (multi-microphones) recording scenarios, leveraging microphone array processing techniques such as beamforming to enhance the captured speech signal and improve verification accuracy.

Appropriateness and relevance

The robustness of speaker verification (SV) systems against acoustic interference is a practical and significant topic in speech processing research. It has many potential applications with the rapid development and adoption of speech interfaces. Microphone array technologies have demonstrated beneficial to improve noise robustness of automatic speech recognition and other speech-related tasks, making their application to SV both logical and promising. Despite its importance, this area has received limited attention, rendering the candidate's contribution both timely and highly relevant.

NTT, INC.
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0237 Japan.
Telephone: +81 (0)50 3482 1753

A summary of the contributions of the thesis

The thesis covers extensively the topic of multi-channel SV. The first three chapters cover the fundamental theory and technologies, while the later chapters introduce novel contributions, including a new dataset, novel multi-channel neural speech separation approaches, the joint training of multi-channel front-end and SV back-end, and an integrated multi-channel SV based on SSL representation. Let me summarize the contributions of each chapter below.

Chapter 1 provides a brief introduction to the problem.

Chapter 2 introduces the details of the SV problem. It defines the key evaluation metrics used throughout the thesis, i.e., equal error rate (EER) and minimum detection cost function (minDCF). It also reviews the important techniques used for SV, with a focus on recent neural approaches. The explanations are clear and detailed, demonstrating that the candidate has a solid grasp of the field.

Chapter 3 introduces the basic concepts of microphone array signal processing, focusing on beamforming, using both rigorous mathematical derivations and simple figures to illustrate the key concepts. This chapter provides the fundamentals of the theory behind beamforming and extends to practical approaches, such as MVDR and GEV beamformers. The chapter is well-written and shows a deep understanding of the topic. At this point, the formulation mostly ignores reverberation, which will be covered briefly in Chapter 6.

Chapter 4 describes one of the key contributions of the thesis, which consists of corpora designed for multi-channel SV. In the age of data-driven processing, the design of high-quality corpora is crucial for advancing technological development. Designing good speech corpora is a tedious task, and the candidate took significant consideration in the design of the MultiSV and MultiSV 2 corpora. In particular, MultiSV 2 consists of a large training dataset, making it appropriate for training neural-based multi-channel SV systems.

Chapter 5 introduces a good overview of neural speech separation. The chapter reviews single-channel neural approaches and their application to multi-channel recordings, through mask-based beamforming or mask-less beamforming. It also introduces two key network

NTT, INC.
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0237 Japan.
Telephone: +81 (0)50 3482 1753

architectures that will be used in all investigations, TasNet and DPT net. In particular, the parameters of the beamformer, such as the frame length, the type of masks, the strategies for combining the masks, and the methods for estimating them, have been thoroughly investigated, using both oracle and estimated experiments.
In general, many multi-channel speech separation approaches rely on single-channel speech separation followed by classical beamforming for multi-channel processing. One of the key novel contributions in this chapter is the introduction of the reference channel attention (RCA) module, which allows multi-channel separation.

Chapter 6 shows the core investigation of the thesis and evaluates the multi-channel pre-processing for SV. The first part of the chapter investigates the cascade of speech dereverberation (with WPE) and beamforming with classical (ivector-based) SV systems. It concludes that WPE and beamforming are beneficial for improving multi-channel SV performance, approaching the performance of clean speech.
The next investigation focused on the use of multi-channel pre-processing with more modern neural SV systems. In particular, it investigates separate and joint-training approaches, introducing a novel formulation for joint-training with the GEV beamformer, and also exploring joint-training with MVDR beamformer. Finally, combining the multi-channel model proposed in Chapter 5 with RCA and joint training yields superior SV performance even in challenging conditions, significantly outperforming a single-channel system. This is one of the major findings of the thesis.

In Chapter 7, the candidate introduced an alternative approach for handling multi-channel recordings by extending a single-channel embedding extractor to handle multi-channel inputs. This approach was based on the SSL models that have been shown to provide robust speech representations, but are mostly limited to single-channel processing. By introducing channel communication blocks, the multi-channel SV system can largely outperform the single-channel-based system. Although conceptually simpler than the approach presented in Chapter 6, this approach relies on larger models. The thesis offers thus different options with different properties. Therefore this chapter complements the investigation of multi-channel SV by presenting a fundamentally different way of handling spatial information than in the previous chapters of the thesis.

NTT, INC.
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0237 Japan.
Telephone: +81 (0)50 3482 1753

Finally Chapter 8 summarizes and concludes the thesis by providing potential future works showing the impact of the presented investigations extends beyond the SV task.

Novelty and significance:

The thesis presents several notable contributions. Most significantly, it explores the application of microphone array technologies to speaker verification (SV), a topic that had previously received limited attention. Recognizing the scarcity of suitable data, the candidate developed dedicated datasets for training and evaluating multi-channel SV systems. Furthermore, the candidate provided a comprehensive coverage of the multi-channel SV subject, introducing multiple novel frameworks and demonstrating their effectiveness through rigorous experimentation.

Evaluation of the formal aspects of the thesis:

As described in the overview, the thesis covers both theory and practice of multi-channel SV. It is clearly written and demonstrates that the candidate has a solid comprehension of the topic and could develop novel technologies.

Quality of publications

The candidate has published his proposed methods at several peer-reviewed international conferences, such as Interspeech and ICASSP, which are the top venues for researchers in speech processing and SV.

Comments and questions:

1. The approach taken by the candidate aligns with the research in multi-channel robust ASR, which has yielded very promising results. However, I wonder if this is the optimal approach for SV. Indeed, ASR needs to recover all parts of the speech to transcribe each uttered word accurately. In contrast, SV may not need to recover the whole speech content to capture the speaker's identity. Therefore, it may be better to completely ignore unreliable parts of the captured signal, caused by loud noise, etc. Could the candidate comment on this point?
2. In Chapter 5, although RCA seems to bring improvements in the case of the conv-TasNet configuration, the improvement appears small, and its significance is not measured. Note

NTT, INC.
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0237 Japan.
Telephone: +81 (0)50 3482 1753

that in this chapter, all techniques are evaluated in terms of speech enhancement metrics, which, as Chapter 6 reveals, are not well correlated with SV performance. Could the candidate comment whether speech enhancement metrics are useful for selecting the front-end for multi-channel SV?

3. The tendency of the results is not always the same for the dev and eval sets. Could the candidate comment on these differences and how to practically choose the multi-channel SV system configuration?

Minor comments:

1. In Chapter 2, the candidate could explain the rationale behind the choices of some techniques, such as MHFA or the losses used. Is it because the chosen techniques achieved state-of-the-art at the time, or for some other reason?
2. In chapter 4, I would have appreciated a table summarizing the main characteristics of the prior and proposed datasets to clarify what problems are addressed by the proposed datasets. Especially, it would make the chapter easier to read if the candidate would provide a brief section summarizing the key differences between MultiSV v1 and v2 and VOiCES datasets, as it seems that the dev and eval sets of MultiSV and VOiCES datasets are pretty similar (as far as I understood, subsets of the Voices dev and eval datasets).
3. About data generation in Chapter 4, I would like to know the impact of data filtering on performance. As the VoxCeleb data may not be completely clean, the candidate removed samples with SNR below 20 dB or 15 dB for MultiSV2, using an SNR predictor. The details of the SNR prediction seem to be missing. Besides, what is the impact of using noisier speech sources for multi-channel SV? On one hand, I agree that the multi-channel simulation scheme relies on clean speech sources. However, by filtering the number of samples, I assume that you reduced the number of possible training speakers, which may also have an impact on performance.
4. Another aspect of the data simulation, presented in Chapter 4, is the use of point sources for the noise, whereas in practice, noise may be diffusive. I wonder if the candidate has tested whether this assumption is reasonable?
5. In Chapter 6, I could not understand why the candidate did not perform dereverberation when conducting the investigations with the neural SV systems, despite it appearing to provide promising results in the preliminary experiments.

NTT, INC.
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0237 Japan.
Telephone: +81 (0)50 3482 1753

II. Student's overall achievements

Overall R&D activities evaluation:

The thesis demonstrates that the candidate possesses a profound understanding of the field, encompassing both its theoretical foundations and recent developments. This is evident in the candidate's ability to summarize the key concepts of microphone-array and SV technologies. In addition to the theoretical background, the candidate demonstrated his capabilities in innovative thinking by proposing novel approaches for SV that achieved promising results. He also demonstrated his scientific abilities in designing a novel dataset and a high level of logical thinking by the design of his investigations, where key parameters were tested rigorously and compared with appropriate baselines.

He has regularly published his results at international conferences, showing his productive research activity. Besides, in addition to his first-author publications, he also has several co-author papers, showing his capability in teamworking, which is important for research.

Therefore, the candidate has proven his high scientific erudition.

Assessment of other characteristics (optional):

The candidate has performed internships at several institutions, leading to publications. He took part in several NIST evaluations and was also received several awards. All these achievements reveal the candidate has acquired and diverse research experience within the course of his studies.

III. Conclusion

The conclusion should contain an explicit statement saying whether, in your opinion, the thesis and the student´s achievements until now meet the generally accepted requirements for the award of an academic degree (in accordance with Section 47 of Act No. 111/1998 Coll., on higher education institution).*

* Short overview of both the Act and corresponding internal BUT regulations is enclosed.

I have carefully reviewed the doctoral thesis of Mr. Ladislav Mošner.

Despite a few recommendations and some points I raised for discussion, the thesis represents a significant contribution to the field of SV and will provide new opportunities for research

NTT, INC.
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0237 Japan.
Telephone: +81 (0)50 3482 1753

and technological development. The work achieved is original and considerable. The investigations were carried out with great diligence to details. The material of the thesis is based on numerous peer-reviewed international conference papers, submitted at top conferences in the field. For all these reasons, I believe that the doctoral thesis meets the requirements of the proceedings leading to the PhD title conferment.

Kyoto, November 7th, 2025

Marc Delcroix
Distinguished Researcher
Signal Processing Research Group
NTT Communication Science Laboratories
Kyoto, Japan
Email: marc.delcroix@ntt.com