UNIVERSITÄT PADERBORN

Fakultät für Elektrotechnik, Informatik und Mathematik
Institut für Elektrotechnik und Informationstechnik

Prof. Dr.-Ing.
Reinhold Häb-Umbach
Leiter Fachgebiet Nachrichtentechnik

Assessment

of the dissertation with the title

Far-Field Speaker Verification Incorporating Multi-channel Processing

Submitted by Ladislav Mosner

Topic and Context

The processing of speech spoken at a distance from the microphones, known as far-field speech processing, has witnessed a significant increase of attention in science and industry. It became an important research topic, not the least because of the emergence of digital home assistants, such as Amazon's Alexa, and the advent of different community challenges, such as the CHiME series. Speech recorded at a distance is affected by various distortions, such as room reverberation and the capture of unwanted acoustic sources. An important tool to combat such signal degradations is multi-channel processing, i.e., sampling the acoustic wave at different positions in space and exploiting the spatial diversity of desired and undesired signal components to clean up the signal for further processing.

While multi-channel signal enhancement for human-to-human communication and for automatic speech recognition (ASR) has been widely investigated, the application studied in this dissertation, far-field speaker verification has received comparatively little attention sofar. Nevertheless, it is of significant relevance, as one can imagine many application scenarios where a speaker must be verified by signals recorded by multiple microphones at a distance, while the enrollment data may have been recorded with a single microphone. For these reasons I consider the topic to be very relevant and appropriate for a dissertation.

Thesis Contents

After a short introduction illustrating the evolution of the author's research on this topic and a summary of his main contributions, Mr. Mosner gives a comprehensive introduction into the topic of speaker verification, formally defining the task, giving an overview of the core techniques and defining the evaluation metrics.

Since multi-channel processing is the main tool to approach the signal degradations caused by recording speech from a distance, Chapter 3 is devoted to the essentials of microphone array signal processing and spatial filtering. It starts with a discussion of the physics of spatial filtering, which, in my view, is not so important for the remainder of the thesis and therefore could have been left out or at least be significantly shortened. The next section on data-dependent and adaptive beamforming is much more relevant for what follows. It presents a thorough introduction of the objective functions and the signal

processing for beamforming, followed by a description of frequency-domain broadband beamforming, which is particularly relevant for speech signals.

Machine learning relies on the availability of appropriate data. The author of this thesis contributed significantly to the definition, collection and dissemination of appropriate multi-channel data to conduct research on speaker verification. His efforts are detailed in Chapter 4. It describes the evolution from the BUT ReverbDB corpus to the MultiSV dataset and, finally, the MultiSV2 corpus. While the former included non-freely available data, MultiSV and MultiSV2 are based solely on publicly available data. MultiSV defines development and evaluation data, while MultiSV2 is much larger and additionally provides multi-channel training data and augmentations with a diverse set of noise signals. As a compromise between pure simulations and original recordings, the datasets consist of retransmissions of existing corpora.

The following three chapters describe the main algorithmic contributions of the candidate.

Chapter 5 discusses multi-channel speech separation techniques. Mosner distinguishes three ways to employ neural networks for spatial covariance matrix (SCM) estimation: direct mask prediction for mask-based beamforming, indirect mask prediction for mask-based beamforming, and mask-less beamforming. The first employs a neural spectral mask prediction network to estimate the SCMs of speech and noise from the noisy speech signal. From those SCMs the beamforming coefficients e.g., of a MVDR beamformer, can be readily computed. The second technique first determines preliminary estimates of the target speech signal and the distortions by a time-domain separation module. The speech and noise masks for beamforming are then estimated from those preliminarily separated signals, and those masks then serve to estimate the SCMs and then to compute the beamformer coefficients. Finally, the mask-less beamforming takes the speech and noise estimates of the separation module directly to compute the SCMs required for the computation of the beamforming coefficients. Note, that maskless beamforming is kind of a misnomer, because the intermediate speech and noise estimates are obtained by mask multiplication. While the first technique is well-known, the other two are slightly different to techniques earlier proposed in the literature.

An original contribution of the author is the Reference Channel Attention (RCA) Combiner, which is a technique to exploit multi-channel information already for the preliminary speech separation module of the above second and third covariance matrix estimation method. The RCA Combiner contains a module, the Cross-Frame Reference Channel (CFRC) Attention that can deal with time-misalignment between the channels, which is an important aspect in an ad hoc scenario, where the microphones are distributed in an environment and do not form a compact array. The experimental evaluation shows that the RCA leads to a slight improvement in SDRi (Signal-to-Distortion Ratio improvement over the input) and the STOI measure of speech intelligibility. It would have been interesting to see how it compares to an explicit time synchronization of the microphone channels. The experiments further reveal that direct mask prediction leads to smaller SDR improvements than indirect mask estimation. This comes for me to no surprise, since it is well-known that beamforming (without postfiltering) leads to poorer noise suppression than mask multiplication for source extraction. Interestingly, the better noise suppression of this second technique does not come at the cost of introducing more artefacts, as the intermediate speech and noise estimates are only used for mask estimation, while the final signal estimates are still obtained by beamforming, which is known to hardly introduce any artefacts.

Although this chapter is entitled “Speech Separation for Speaker Verification’’, the presented techniques are by no means bound to be used only for subsequent speaker verification, they rather present a thorough investigation into speech versus noise separation with multi-channel techniques.

The next chapter, Chapter 6, investigates the use case of speaker verification, as a downstream task to the preceding beamforming. Furthermore, beamforming is complemented with a preceding dereverberation stage. Experiments are carried out for “classical” speaker verification based on i-vectors and PLDA, as well for more recent neural speaker verification based on speaker embedding vector extraction. Cascading signal enhancement (beamforming) and speaker verification can lead to suboptimal performance because the two modules are designed to optimize different objectives. To relax this potential source of degradation, Mosner investigates fine-tuning the enhancement module by backpropagating gradients computed from a speaker verification objective, be it cross entropy or mean-squared error. To be able to propagate the gradient through the eigenvector decomposition of the Generalized Eigenvalue (GEV) beamformer, he reformulates the eigenvalue problem via Cholesky decomposition of the noise SCM. This leads to simplified formulas compared to the original approach published by Boeddeker et al. (2017). Experimental results verified the performance benefits obtained by this fine-tuning.

Chapter 7 is devoted to the use of originally single-channel self-supervised learned feature extraction methods for multi-channel speaker verification, in particular the use of WavLM. Ladislav Mosner presents an interesting multi-channel extension of such pretrained single-channel models, which he termed METRO. It fuses information from other channels at the frame level. So-called M-M blocks allow each channel to benefit from information from other channels, while the M-S blocks compress the multi-channel input to a single-channel output. Information at different layers of the WavLM encoder are gathered by multi-head factorized attentive (MHFA) pooling. This novel multi-channel extension of originally single-channel SSL models led to consistent improvements in speaker verification performance.

The next chapter draws some conclusions and presents an outlook on future work. It is followed by the bibliography and two appendices.

Concluding Assessment

While multi-channel speech enhancement for automatic speech recognition has been extensively investigated, the topic of this thesis, multi-channel processing for far-field speaker verification is a relatively unexplored field. However, one might wonder what impact the different downstream task has on the preceding multi-channel processing. Initially not so much, as the different SCM estimation methods discussed in Chapter 5 could be used for other downstream tasks without modification. But even here, despite the large research body on multi-channel speech processing, Mr. Mosner could add original contributions by developing the reference channel attention combiner.

While joint fine-tuning of the cascade of beamforming and ASR has been done before, the joint fine-tuning of beamforming and speaker verification developed by Mr. Mosner is new. To ease gradient backpropagation he applied a Cholesky factorization to the noise SCM, and he could show that this end-to-end training leads to consistent performance improvements.

Of particular originality is the multi-channel extension of large pretrained models described in Chapter 7. While this was applied to speaker verification its usefulness goes well beyond that: it can be applied to many more downstream tasks and SSL models.

Overall, this clearly demonstrates the ability of the author to conduct independent research at an international level. Also, the design, development and provision of corpora for multi-channel speaker verification research is an important scientific effort and a service to the scientific community.

The thesis is exceptionally well written. It is well structured, precise and at the same time concise. It is self-contained.

I wish to emphasize that I consider the references to prior work to be exceptionally good. In the cases where I believe I can judge the quality, the author always gave the best references. This shows to me that Mr. Mosner is a person with an outstanding research erudition.

He has published his research at the top conferences in our field (ICASSP, Interspeech, ASRU).

To conclude, in my opinion the doctoral thesis meets all requirements of the proceedings leading to a PhD conferment.

Finally, at the defense I would like the candidate to address the following questions

- The MultiSV data sets contain retransmissions as the development and evaluation data. While this is a good compromise between pure simulation and real recordings of speakers, I wonder what the candidate's opinion is about the validity of drawn conclusions for an application with real speakers. Do you expect head movements and thus time-varying transfer characteristics to be critical, or why do you consider them to be not critical?
- Considering a beamforming frontend for either speaker verification or ASR, do you expect the downstream task to influence what will come out as the best performing front end (not considering fine-tuning of front-end with gradients from downstream task)?
- Fine-tuning the beamforming front-end with a speaker verification objective led to improved performance. Can you see any difference in the obtained beampatterns with and without fine-tuning?
- What is your overall conclusion/recommendation: Should we use a beamforming front-end or employ a multi-channel extension of pretrained models, such as WavLM, instead? Compare the two both in terms of verification performance and computational complexity.

Paderborn, November 13, 2025

(Reinhold Häb-Umbach)