Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Utilizing VOiCES dataset for multichannel speaker verification with beamforming

Ladislav Mošner, Oldřich Plchot, Johan Rohdin, Jan Černocký

Brno University of Technology, Faculty of IT, IT4I Centre of Excellence, Czechia
imosner@fit.vutbr.cz, iplchot@fit.vutbr.cz, rohdin@fit.vutbr.cz, cernocky@fit.vutbr.cz

Abstract

VOiCES from a Distance Challenge 2019 aimed at the evaluation of speaker verification (SV) systems using single-channel trials based on the Voices Obscured in Complex Environmental Settings (VOiCES) corpus. Since it comprises recordings of the same utterances captured simultaneously by multiple microphones in the same environments, it is also suitable for multichannel experiments. In this work, we design a multichannel dataset as well as development and evaluation trials for SV inspired by the VOiCES challenge. Alternatives discarding harmful microphones are presented as well. We asses the utilization of the created dataset for x-vector based SV with beamforming as a front end. Standard fixed beamforming and NN-supported beamforming using simulated data and ideal binary masks (IBM) are compared with another variant of NN-supported beamforming that is trained solely on the VOiCES data. Lack of data revealed by experiments with VOiCES-data trained beamformer was tackled by means of a variant of SpecAugment applied to magnitude spectra. This approach led to as much as 10% relative improvement in EER pushing results closer to those obtained by a good beamformer based on IBMs.

1. Introduction

Over the last three years, the speech community has observed significant improvements in speaker recognition (SR). This is mainly due to the advent of embedding extracting neural networks, most notably x-vector extractor [1]. Becoming a new state of the art, they have almost completely replaced i-vectors [2]. While an embedding extractor is typically trained discriminatively, a generative backend in form of Probabilistic Linear Discriminant Analysis (PLDA) [3] is still often used to produce verification scores. With newly proposed loss functions for the embedding network training [4, 5, 6, 7, 8], cosine similarity scoring has, however, been shown to perform better in some scenarios [9].

Despite recent advances, difficulties arise when it comes to far-field recognition. In such a scenario, the microphone or microphone array records all noises and distractors. The recording usually takes place in reverberant enclosures, such as rooms. This adds another level of difficulty since reverberation is a different type of corruption. As opposed to additive noise, the effects of room acoustics can be modeled with a certain level of abstraction by linear filtering. The interest of industry in far-field and also multichannel speaker verification (SV) has been increasing, which lead to more research and consequently to releasing suitable benchmarks and datasets. The HI-MIA dataset [10] has recently been released for the task of text-dependent SV. It contains utterances of "Hi, Mia" in English and "ni hao, mi ya" in Chinese. It was recorded in a real smart home environment by six 16-microphone arrays and one high fidelity close-talking microphone. A few corpora have also been released for the text-independent SV, which we are dealing with in this paper. In [11], authors have introduced a benchmark that is based on a public CHiME-5 corpus which was originally released for an automatic speech recognition (ASR) challenge. The SV benchmark includes multi-speaker and single-speaker enrollment and test recordings. It is designed for both single-channel and multichannel approaches. The dataset of our particular interest is the Voices Obscured in Complex Environmental Settings (VOiCES) corpus [12]. We will analyze it w.r.t. its capacity of being used for designing a multichannel SV benchmark. The dataset itself does not aim only on fostering far-field SV research but its goal is also to support research in ASR, source separation, sound localization, and other areas. By the time of writing this paper, there were two releases, each comprising recordings from two rooms. Totally 15 hours of English read speech from 300 speakers (with the same amount of male and female voices) taken from LibriSpeech [13] were retransmitted in indoor conditions. A loudspeaker replaying the clean utterance was placed on a motorized stand to simulate head movements. The corpus includes four noise conditions according to background distractors playing concurrently with the speech loudspeaker – television, music, babble, and ambient (no replayed noise). Authors of the VOiCES corpus organized a challenge. It had two tasks – SV and ASR – each of which had two conditions according to training data: fixed or open. In this paper, we are dealing with SV, therefore, we will make use of the trials defined for the speaker identity related part of the challenge.

Along with the increased interest in far-field and multichannel benchmarks, research on these topics has also received increased attention. In [14], authors proposed an end-to-end framework utilizing multiple microphones for extraction of speaker embeddings. The method is based on the simultaneous processing of channels by 2- and 3-dimensional convolutional layers. Other studies approach multichannel processing with beamforming [15, 16]. Usually, some type of beamformer supported by a neural mask estimator is used. Mask estimation performed by a neural network, where the masks are subsequently used for cross power spectral density matrix (PSD) estimation, was introduced in the ASR community [17]. In [15], authors examine minimum variance distortionless response (MVDR) and generalized eigenvalue (GEV) [18] beamformers. MVDR is also utilized in [16] where it is combined with dereverberation and finally also with x-vector extractor and trained jointly. Many works use non-standard training and/or evaluation datasets usually created by means of simulation. It points out a long-lasting lack of appropriate datasets for benchmarking. In this work, we examine the potentials and draw-

10.21437/Odyssey.2020-27

backs of using the VOiCES dataset for the evaluation of multichannel SV systems. In our case, an NN-supported GEV beamformer will be used. The problem of mask-based beamformers is that simulated data are needed for training as knowledge of clean speech and additive noise component is required. Recently, we have proposed a solution to mask estimator training for GEV beamformer where only clean reference is known – this requirement is satisfied by the VOiCES dataset. Therefore, we will also asses usability of the VOiCES corpora for training.

2. The VOiCES dataset for multichannel experiments

The VOiCES from a Distance Challenge 2019 [19] offered a track for SV. We focus on a verification that employs information from multiple channels. We will, therefore, make use of the fact that the data in VOiCES are captured by multiple microphones. To this end, we will analyze and redefine both the development and evaluation trials of the challenge and asses the usefulness of the dataset for training and evaluation of multichannel SV systems.

2.1. Breakdown of the VOiCES challenge trial definition

The overall statistics of the original sets of data are shown in Table 1. Development and evaluation sets share some properties: sets of enrollment and test utterances are disjoint and they are recorded in different rooms. Evaluation dataset is more difficult as enrollment recordings contain not only utterances retransmitted in room conditions but also original clean LibriSpeech files. Clean enrollment recordings and those recorded in room 3 contain voices of two disjoint sets of speakers. Sets of 196 speakers in the development portion and 100 speakers in the evaluation set are disjoint as well.

Both development and evaluation trials are created in such a way that, for every enrollment recording, there are always multiple test recordings containing the same content (utterance, speaker, background noise, and room) recorded with multiple microphones. Based on this fact, we will consider the case where speakers are enrolled using a single microphone, which can vary for every speaker. No pre-processing will be applied to enrollment segments. Test utterances will then be assumed to come from arbitrary microphone arrays and will be processed by a multichannel system. This causes a mismatch between the processing of enrollment and test recordings, and it can be perceived as a more difficult scenario[1]. Since enrollment files will remain untouched, the following analysis will be focused only on the test part of the trials.

Development test set

In the original set of test recordings, every utterance is uttered by only one speaker. Each speaker uttered at least 1 utterance and at most 6 utterances, while the average is 2.6 utterances per speaker. Every utterance was recorded in all noisy conditions – music, babble, television – and also *without* the presence of a distractor (*none* in Table 1). Eight microphones – 2, 4, 6, 8, 9, 10, 11, and 12 – were recording simultaneously. Positions and types of individual microphones are available in [12] and on the VOiCES website[2].

Table 1: *Statistics of the development and evaluation trail lists defined for the VOiCES challenge [19]. The term* clean *in the enrollment recordings section of the evaluation set denotes the fact that also original (not retransmitted) LibriSpeech utterances were used. Noise types are abbreviated as follows: B – babble, T – television, M – music, and N – none.*

		Enroll.	Test	Total
Dev. set	env	room 1	room 2	room 1, 2
	noise	N	N, B, T, M	N, B, T, M
	spkrs	103	189	196
	utts	128	489	617
	mics	2	8	2 + 8
	files	256	15,648	15,904
Eval. set	env	clean; room 3	room4	clean; room 3,4
	noise	–; N	N, B, T	N, B, T
	spkrs	44; 56	96	100
	utts	186; 70	336	592
	mics	–; 2	11	2 + 11
	files	186; 140	11,066	11,392

Based on the fact that the same content was recorded by eight microphones, we decided to randomly group 4 microphones. This grouping always resulted in the creation of 2 adhoc microphone arrays. Thus, the original 4,005,888 trials were reduced to 1,001,472 trials as 8 original trials were reduced to 2 (enrollment recordings remained unchanged). Overall, 996,448 trials are impostor and 5,024 are the target ones.

Evaluation test set

As well as in the set of development test recordings, every utterance is uttered by only one speaker. The minimum of utterances per speaker is 1, the maximum is 9 and the average is 3.5. Every utterance was recorded with babble and television distractor in the background as well as without any noise (note that the music noise was not used). Eleven microphones – 4, 6, 8, 9, 10, 11, 12, 16, 17, 18, and 19 – were recording simultaneously.

In order to keep the number of microphones in created adhoc arrays consistent with the development set, we randomly selected two pairs of 4 microphones. The last array consists of 3 remaining microphones and 1 microphone randomly selected out of the already used ones. Therefore, 3,607,516 original trials were reduced to 983,868. Overall, 973,929 trials are impostor and 9,939 are the target ones.

Multichannel training data

Our training corpus is based on a complete set of recordings from room 1 and room 2. This set is part of the first VOiCES release. We did not use any files from the second release for training because it contains recordings from room 3 and room 4 and we did not want to provide models with acoustic conditions that are present in the evaluation data during training. Because the development data constitute a subset of the first release, we filtered them out not to train models on them. We also removed files where we found an inconsistency in lengths of source LibriSpeech and retransmitted recordings which would cause problems in training. In the following step, recordings were grouped

[1] There is still room for SV performance improvement by closing the domain gap between enrollment and test recordings. An example is presented in [20]. The authors artificially reverberate clean enrollment recordings so that they become similar to test recordings and perform embedding fusion on top of simulated audio. We did not explore such direction as it would be more involved since some enrollment signals are reverberant. Moreover, clean enrollment recordings are only in the evaluation set.

[2] https://voices18.github.io/

to quartets based on speaker identity, chapter, segment, room, and distractor type ensuring that all four microphones recorded the same content. Microphones in microphone arrays were chosen randomly to enhance diversity of the set. The resulting training dataset consists of 57,800 examples (arrays comprising four microphones) spanning voices of 200 speakers. They are fully overlapped with speakers in the development set. The average duration of utterances is more than 15 seconds.

Unfortunately, this dataset has some drawbacks. It is similar to the development set in terms of speakers and acoustic conditions. It can also comprise the same content recorded by different microphones. This was tolerated for the sake of a larger set of data. Due to the mentioned disadvantages, results obtained on the development set should be treated carefully.

3. Multichannel speaker verification with beamforming

As discussed before, various approaches to multichannel SV have been proposed (including multichannel pre-processing or end-to-end models), and there are others yet to come. This paper aims at the utilization of the VOiCES dataset for training and evaluation and discovering its limits rather than at the proposal of a new multichannel SV approach.

One of the downsides of the corpus is its size and limited number of speakers. Therefore, it is hardly usable for state-of-the-art embedding extractor training. Instead, we will use the VOiCES data to train only pre-processing models which will output single-channel audio used by downstream embedding extractor. Specifically, we will examine training of neural networks that are used to estimate statistics for beamforming (spatial filtering).

3.1. X-vector extractor and speaker verification backend

In order to perform SV, we follow the standard approach employing a neural network embedding extractor followed by a PLDA backend [3].

X-vector extractor [1] is used to estimate single fixed-length embedding for every utterance of arbitrary length. We adopt a deeper architecture proposed in [21]. 30-dimensional MFCC features extracted from a single-channel audio are fed to 9 layers of TDNN/DNN – a frame-level part of the network. The resulting context is 11 frames to each side of the central frame. The frame-level block is followed by mean and standard deviation computation. The segment level part consists of 2 fully-connected layers followed by a layer with softmax to perform speaker classification.

The x-vector extractor was trained on 1.2 million speech segments from 7,146 speakers from the VoxCeleb 1 and 2 development sets plus additional 5 million segments obtained with data augmentation. All training segments were 200 frames long. The model was evaluated on the original trials of the VOiCES challenge – model 14 in [22].

PLDA backend involves two pre-processing steps: the x-vector dimension is reduced from 512 to 250 by LDA, length-normalization of embeddings is applied. For the backend training, we concatenated all segments from each session of the VoxCeleb 1 and 2 development data. Including augmentations, this resulted in 830K files.

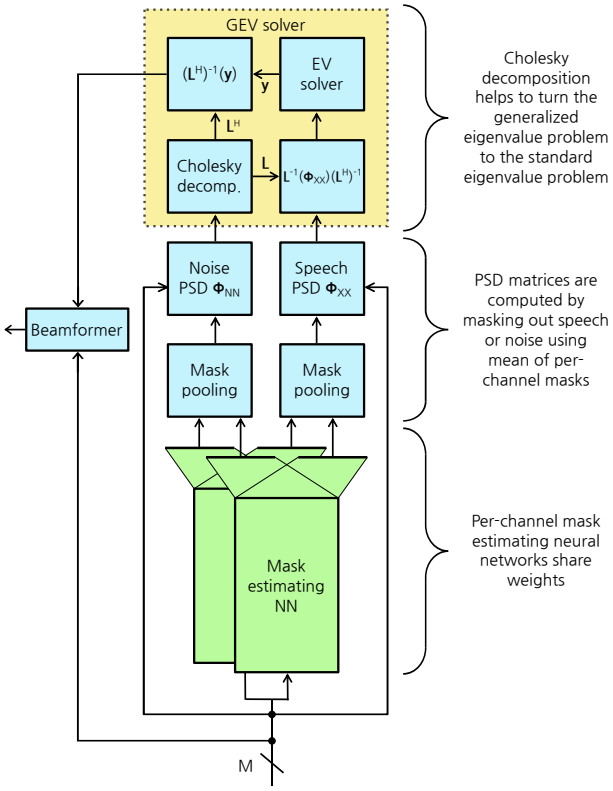

Figure 1: *GEV beamformer with integrated neural network. Green blocks represent trainable parts, blue ones are fixed. Computation flow is displayed for one frequency bin. M stands for the number of channels (4 in our case). $\mathbf{\Phi}_{NN}$ and $\mathbf{\Phi}_{XX}$ represent noise and speech power spectral density matrices, respectively. $\mathbf{\Phi}_{NN}$ is decomposed by means of Cholesky decomposition such that $\mathbf{\Phi}_{NN} = \mathbf{L}\mathbf{L}^H$. Eigenvalue (EV) solver computes a principal eigenvector $\mathbf{y}$ of the input matrix. Transformed principal eigenvector corresponds to a principal eigenvector of a generalized eigenvalue (GEV) problem and is used as a weight vector for beamforming.*

3.2. Beamforming frontend

Beamforming producing a single-channel spatially-focused audio is used as a pre-processing step for the x-vector extractor. As our baseline, we will use a fixed beamformer that requires no training, and therefore, does not use the VOiCES training data. Specifically, we will use the weighted delay-and-sum beamformer implemented by the BeamformIt tool [23].

The BeamformIt baseline will be compared with an adaptive generalized eigenvalue (GEV) beamformer [18] which estimates beamforming weights using the statistics of the input data. In [17], Heymann et al. proposed estimation of these statistics (power spectral density (PSD) matrices) by using neural networks. However, this approach cannot be followed directly with the VOiCES training dataset as this method requires the knowledge of the exact decomposition into the additive noise and clean audio.

The full processing chain of the GEV beamformer, along with neural network estimation, is displayed in Figure 1. Mask estimating NNs share weights, and each of them processes one channel (magnitude STFT representation). They predict two masks with values between zero and one according to the preva-

lence of noise or speech. Masks are subsequently mean-pooled and applied to STFT representation of recordings to mask out speech or noise while computing PSD matrices. The two PSD matrices are fed to a generalized eigenvalue solver, and the principal eigenvector is used as a beamformer weight vector.

Optimization of masks

The original NN-supported GEV work [17] suggested training the mask estimating NN directly by minimizing binary cross-entropy between outputs and ideal binary masks (IBMs), i.e. optimization of outputs of green blocks in Figure 1. However, in order to compute IBMs, knowledge of speech and additive noise is required. Hence, simulated data is required for training and the VOiCES training dataset cannot be directly used. We, therefore, prepared a simulated dataset comprising the same amount of data and equivalent utterances. We created data resembling the real ones by means of room simulation (image source method) and positional noise source addition. Original LibriSpeech recordings were used as inputs to the simulation. Reverberation time RT60 was drawn uniformly from the interval [0.3, 0.9] s. Shop, crowd, library, office, a real fan, and street noises were selected from the Freesound library[3] and added with SNRs from 3 dB to 20 dB.

It might be worth exploring variable simulation parameters along with various types of noise. However, this would require extensive experimentation. In real applications, it is not possible to expect certain types of noise, therefore, increased variability in training data is usually beneficial to enhance robustness. This is one of the reasons why we opted for diverse training data in terms of acoustic conditions.

Optimization of a beamformer output

We have recently proposed a solution to backpropagation through generalized eigenvalue decomposition by turning it to standard eigenvalue decomposition through Cholesky decomposition [24]. Block diagram of this approach is also displayed in Figure 1 and we refer an interested reader to [24] for details. This representation allows the generalized eigenvalue solver to be an integral part of the model that updates the weights of a mask estimating NN optimizing directly the output of the beamformer. Here the objective function is mean squared error (MSE) between magnitude spectra of clean speech and beamformer output.

It is convenient that this approach can be compared with the mask optimization since the network remains unchanged and the only difference is the way of training. Another benefit is that simulated data are no longer required and the VOiCES training corpus can be already used because it also contains source LibriSpeech recordings.

We use the following architecture for both NN-supported beamformers. The first layer that is supposed to capture temporal dependencies is an LSTM with 513 units. It is followed by two linear layers, each of which comprises 513 neurons with a sigmoid activation function. Finally, the last layer is divided into 2 branches of 513 neurons and sigmoid activation functions. A dropout of 50% is used during training.

4. Experiments

All the presented results are expressed in terms of equal error rate (EER [%]) and minimum detection cost (C_{det}) as defined for the VOiCES challenge in [19] (the prior probability of a target trial P_{tar} is set to 0.01).

4.1. Comparison of beamformers and training datasets

In this subsection, we compare the same beamformer utilizing the same neural network. As mentioned before, the difference is in the training procedure, where only the approach optimizing beamformer output can rely only on the VOiCES data without the need for simulation. For brevity, we will refer to the models as *BCE-model* and *MSE-model* according to the loss function that is being optimized during mask estimator training.

As per results on the development dataset in Table 2, the BCE-model reaches better performance than the MSE-model. This suggests the superiority of the mask optimizing approach. Training data for the MSE-model comprises the same speakers and the same acoustic conditions (rooms, noise types) as the development data. On the other hand, there is an overlap only in terms of speakers (as the rooms are simulated) regarding the training and development set for the BCE-model. We did not constrain the room simulation to model only the same rooms and microphone positions resembling those used when recording the VOiCES corpus. Despite the similarity of data associated with MSE-model, its performance is worse than that of BCE-model. We hypothesize that it is a more difficult task for the neural network to figure out that its outputs should mask out speech and noise only by optimizing MSE. However, direct mask optimization of BCE-model fits the underlying mathematical formulation of the GEV beamformer better.

Results on the development and evaluation datasets suggest that BCE-model generalizes better than MSE-model because relative performance degradation is more pronounced for MSE-model. We assume that this is where the diversity of the training data comes into play. Although BCE-model is trained on simulated data, they are more varied which can positively affect generalization. However, recordings for the MSE-model training come only from 12 microphones. Even though we randomly shuffle microphones while creating microphone arrays, variability is still limited[4].

In order to empirically support our assumption, we decided to augment the MSE-model training data to enlarge the variability and make the training more difficult. We decided to use a straightforward approach that resembles SpecAugment [25] for multiple reasons. It is possible to perform this type of augmentation online on a GPU, which would be complex for online augmentation by reverberation. By generating new training examples online, there is no requirement for additional storage space as in the case of offline augmentation. Also, no other external data is required and training of MSE-model can completely rely on the VOiCES dataset. Moreover, it was shown to be comparable to a standard augmentation by noise and reverberation in the context of SV [26]. We used only masking, no warping was applied. The difference between our approach and the original proposal is that we performed masking in magnitude STFT domain rather than in the mel-filter-bank domain. We started by applying a mild augmentation and gradually made it more aggressive. It turned out that harsher masking improves generalization as performance on the evaluation set gets better (especially for MSE-model). In the experiments

[3] http://www.freesound.org

[4] Note that restricting the simulation of the BCE-model training data to take into account only the rooms from the VOiCES dataset and the same microphone positions would result in even greater limitation of variability. Mask estimator processes channels separately and permutation of arrays would not have any effect.

Table 2: *Speaker verification results of a* single-channel *system (on original VOiCES trial set) and multichannel systems (on our modified trial set). Performance is displayed in terms of equal error rate (EER [%]) and minimum detection cost* C_{det}.

Method	Dev. set		Eval. set	
	EER	C_{det}	EER	C_{det}
Single-channel	2.03	0.261	5.51	0.459
BeamformIt	1.73	0.221	5.11	0.494
BCE-model	1.56	0.195	4.15	0.418
& SpecAugment	1.65	0.197	**4.05**	**0.415**
& removed mic 12	1.65	0.196	3.86	0.408
& removed mics 6, 12	1.72	0.195	3.62	0.398
MSE-model	1.81	0.196	5.11	0.514
& SpecAugment	1.82	0.205	**4.52**	**0.476**
& removed mic 12	1.77	0.207	4.39	0.469
& removed mics 6, 12	1.51	0.201	4.28	0.466

(presented in Table2), 2 frequency and 2 time masks were individually applied to all channels. Each time mask covers at most 5% of frames and the maximum range of frequency bins covered by each frequency mask is 15%. Based on our observation of the behavior on the evaluation set, it is reasonable to expect that we are not yet finished with tuning our SpecAugment parameters.

4.2. Microphone analysis

Motivated by a desire to identify poorly performing microphones that can, in turn, degrade the overall performance of the microphone array, a per-microphone analysis was performed. During these experiments, we used a single-channel x-vector extractor. For reference, the performance of the single-channel SV system on a complete set of the VOiCES trials is in Table 2 (the first row).

The original VOiCES development trial list was split into 8 parts, and the evaluation list to 11 parts corresponding to individual microphones. Each part contains the same enrollment recordings. The test recordings differ only in terms of microphone. Results of the analysis are in Table 3, where each part of the original trial list is denoted by the number of the corresponding microphone. Not surprisingly, microphones that are close to a loudspeaker (2, 4) provide the best results. Even though microphone 8 is located behind the loudspeaker, its proximity to the source ensures good performance. On the other hand, by far the worst-performing microphone is 12 which is placed on the wall and fully obstructed. The second worst microphone (6) is the omnidirectional condenser lavalier one placed far from the source.

In agreement with our aim, we removed the worst-performing microphone (12) from microphone arrays of both development and evaluation trial sets. In order to preserve the same number of trials as before removal, we made use of the fact that each enrolment recording is paired with 8 or 11 test recordings captured by different microphones in the original VOiCES trial definition. In our multichannel trial list we, therefore, removed microphone 12 from arrays and replaced it by a random microphone from a different array that recorded the same content. We followed the same procedure when removing two worst-performing microphones – 6, 12. As per evaluation results in Table 2, the negative effect of poor microphones in microphone arrays is clear. The BCE-trained model seems to be more prone to bad signals, as the overall performance improves more than for MSE-model after the removal of microphones 6 and 12. This may be attributed to a bad estimation of masks in conditions not seen in training (such as an obstructed microphone). The MSE-model training data include such situations.

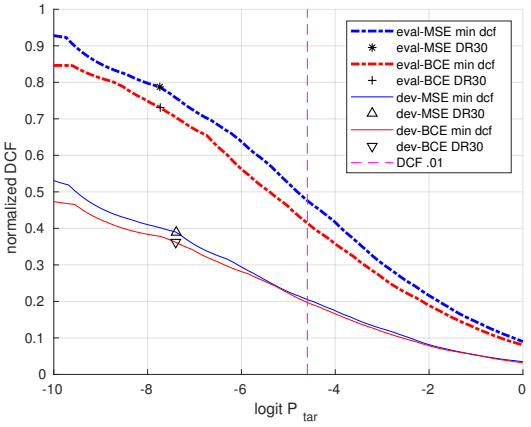

Figure 2: *Normalized DCF as a function of a logit of target speaker probability* P_{tar} *for MSE-model and BCE-model both with SpecAugment.*

4.3. Analysis with DCF plots

In the Figure 2, we analyze performance of two beamforming methods based on optimization of spectral masks (BCE) and on optimizing the output of the beamformer (MSE). We can observe that the performance of both methods is similar on the development set, while on the evaluation set, the BCE-model is consistently better across a wide range of operating points. We speculate that with MSE-model, we are still overtrained on the development set as we reach similar performance as BCE-model and at the same time we can observe large gains from data augmentation (see Table 2), while for BCE-model the effect of data augmentation is negligible. In our future work, we will certainly fine-tune the parameters of SpecAugment to find the limits of the method on this dataset. We can also observe that the selected DCF operating point sits reliably on the right side of the DR30 point (Doddington's rule of 30 [27]) which means that the measured results are statistically significant even with the substantially decreased amount of trials in the modified trial set.

5. Discussion

Our analysis of the VOiCES dataset and trials defined for the SV track of the VOiCES challenge revealed some properties of the corpus and suitability for multichannel experiments. Results drawn in this section may be considered as ideas that are worth taking into consideration while creating multichannel datasets.

For cross-site comparison and healthy competition of multichannel SV systems, the SV community will greatly benefit from having well designed public sets of development data, evaluation data, and trial lists. In this study, we adapted the exisiting VOiCES dataset originally intended for single-channel far-field systems and we created a recipe for development and evaluation of the multichannel SV systems. Even though we

Table 3: *Analysis of microphones in terms of SV performance. Bold numbers denote microphone labels defined by the VOiCES corpus. The first line of results represents EERs [%], the second line minimum detection cost C_{det} with omitted leading zero, and the third line assigns position labels to microphones. Meaning of the shortcuts is as follows:* clo *– closest to foreground speaker,* mid *– mid-distance to foreground speaker,* far *– farthest to foreground speaker,* beh *– behind foreground speaker,* tbo *– partially obstructed (on table),* cec *– overhead on ceiling (clear),* ceo *– overhead on ceiling (fully obstructed),*wal *– fully obstructed (wall).*

Development set								**Evaluation set**										
2	**4**	**6**	**8**	**9**	**10**	**11**	**12**	**4**	**6**	**8**	**9**	**10**	**11**	**12**	**16**	**17**	**18**	**19**
1.43	1.83	2.15	1.47	1.91	2.03	2.03	3.31	2.38	9.18	2.54	3.50	2.78	4.71	14.24	3.35	5.01	2.08	3.96
.221	.251	.274	.231	.248	.252	.261	.314	.304	.616	.312	.398	.343	.475	.839	.367	.499	.301	.484
clo	mid	far	beh	tbo	cec	ceo	wal	mid	far	beh	tbo	cec	ceo	wal	–	–	–	–

confirmed that it is indeed suitable for evaluaition purposes, it still has limitations when considering our scenario. Enrollment recordings were recorded only with two microphones. Therefore, potential beamforming could use only two microphones and the number of trials would reduce even more. To explore some realistic scenarios in consumer electronics, it would be useful to have a version of a dataset that could support trial definitions comprising multichannel enrollment recordings.

Variability in training data is convenient when designing robust models. The diversity of the dataset could be enhanced by incorporating more microphones. For instance, multichannel dataset [28] includes 31 microphones. The drawback of a huge number of simultaneously recording is a synchronization of all the channels. It calls for specialized hardware. As far as microphone arrays are considered, we grouped arbitrary microphones, so presumably, some sensors in the array are few meters apart and spatial aliasing may occur. It is also difficult to describe the properties of such arrays. Therefore, recordings from compact arrays would be certainly appreciated in a multichannel dataset.

We recognize that collecting datasets comprising far-field speech is difficult, costly, and time-demanding. Obviously, it cannot be compared with nowadays datasets such as Voxceleb [29, 30] in terms of the number of speakers and size. It is, therefore, difficult to use such a dataset for training of some more complex models. Therefore, we fully support the continuation of far-field multichannel data collection.

6. Conclusions

In this work, we have explored the potential of the VOiCES dataset to support training and evaluation of multichannel SV systems. We have identified several weak spots such as small amount of speakers and small variability in the acoustic environments and channels and we tackled these problems via data augmentation. In our set of experiments, we have confirmed that even with a dataset of this size and with the help of data augmentation, we can achieve interesting results and carry out research in the field of multichannel speaker verification. This was confirmed by successfully training our recently proposed method which directly optimizes the beamformer output via MSE. This method is especially appealing in the scenarios like here, when the database is being created by retransmitting relatively clean source data and recording with various noises through large number of different microphones.

In the future research, we would like to continue to explore the limits of data augmentation (and if possible use more data as well) with the MSE method and surpass the performance of the method based on ideal binary spectral masks. Eventually we would like to continue to research a SV-aware (end-to-end) beamforming as outlined in [24].

7. Acknowledgement

The work was supported by Czech Ministry of Interior project No. VI20152020025 “DRAPAK”, European Union’s Horizon 2020 grant no. 833635 “ROXANNE”, and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project “IT4Innovations” excellence in science - LQ1602.

8. References

[1] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep Neural Network Embeddings for Text-Independent Speaker Verification,” in *Proc. Interspeech 2017*, Aug 2017, pp. 999–1003.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011, ISSN: 15587916.

[3] S. J. D. Prince and J. H. Elder, “Probabilistic Linear Discriminant Analysis for Inferences About Identity,” in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep Hypersphere Embedding for Face Recognition,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.

[5] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive Margin Softmax for Face Verification,” *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[7] W. Cai, J. Chen, and M. Li, “Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System,” in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[8] Z. Huang, S. Wang, and K. Yu, “Angular Softmax for Short-Duration Text-independent Speaker Verification,” in *Proc. Interspeech 2018*, 2018, pp. 3623–3627.

[9] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, “BUT System Description to VoxCeleb Speaker Recognition Challenge 2019,” *arXiv e-prints*, p. arXiv:1910.08847, 2019.

[10] X. Qin, H. Bu, and M. Li, "HI-MIA : A Far-field Text-Dependent Speaker Verification Database and the Baselines," *arXiv e-prints*, p. arXiv:1912.01231, 2019.

[11] D. Garcia-Romero, D. Snyder, S. Watanabe, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker Recognition Benchmark Using the CHiME-5 Corpus," in *Proc. Interspeech 2019*, 2019, pp. 1506–1510.

[12] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, and J. van Hout, "Voices Obscured in Complex Environmental Settings (VOICES) corpus," *arXiv e-prints*, p. arXiv:1804.05053, Apr 2018.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[14] D. Cai, X. Qin, and M. Li, "Multi-Channel Training for End-to-End Speaker Recognition Under Reverberant and Noisy Environment," in *Proc. Interspeech 2019*, 2019, pp. 4365–4369.

[15] H. Taherian, Z.-Qiu Wang, and D. Wang, "Deep Learning Based Multi-Channel Speaker Recognition in Noisy and Reverberant Environments," in *Proc. Interspeech 2019*, 2019, pp. 4070–4074.

[16] J.-Y. Yang and J.-H. Chang, "Joint Optimization of Neural Acoustic Beamforming and Dereverberation with x-Vectors for Robust Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 4075–4079.

[17] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural Network Based Spectral Mask Estimation for Acoustic Beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[18] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[19] M. K. Nandwana, J. van Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The VOiCES from a Distance Challenge 2019 Evaluation Plan," *arXiv e-prints*, p. arXiv:1902.10828, Feb 2019.

[20] X. Qin, D. Cai, and M. Li, "Far-Field End-to-End Text-Dependent Speaker Verification Based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation," in *Proc. Interspeech 2019*, 2019, pp. 4045–4049.

[21] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker Recognition for Multi-speaker Conversations Using x-vectors," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5796–5800.

[22] P. Matějka, O. Plchot, H. Zeinali, L. Mošner, A. Silnova, L. Burget, O. Novotný, and O. Glembek, "Analysis of BUT Submission in Far-Field Scenarios of VOiCES 2019 Challenge," in *Proc. Interspeech 2019*, 2019, pp. 2448–2452.

[23] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2021, September 2007.

[24] L. Mošner, O. Plchot, A. J. Rohdin, L. Burget, and J. Černocký, "Speaker Verification with Application-Aware Beamforming," in *Proceedings of ASRU 2019*. 2019, pp. 411–418, IEEE Signal Processing Society.

[25] D. S. Park, W. Chan, Y. Zhang, Ch. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[26] S. Wang, J. Rohdin, O. Plchot, L. Burget, K. Yu, and J. Černocký, "Investigation of SpecAugment for Deep Speaker Embedding Learning," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[27] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST Speaker Recognition Evaluation – Overview, Methodology, Systems, Results, Perspective," *Speech Communication*, vol. 31, no. 2, pp. 225 – 254, 2000.

[28] I. Szőke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and Evaluation of a Real Room Impulse Response Dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.