The STC System for the CHiME 2018 Challenge

Ivan Medennikov[1,2], *Ivan Sorokin*[1], *Aleksei Romanenko*[1,2], *Dmitry Popov*[1], *Yuri Khokhlov*[1], *Tatiana Prisyach*[3], *Nikolay Malkovskii*[3], *Vladimir Bataev*[1], *Sergei Astapov*[2], *Maxim Korenevsky*[1†], *Alexander Zatvornitskiy*[1,2,3]

[1] STC-innovations Ltd, St. Petersburg, Russia
[2] ITMO University, St. Petersburg, Russia
[3] Speech Technology Center Ltd, St. Petersburg, Russia

{medennikov,sorokin,romanenko,popov-d,khokhlov,prisyach,malkovskiy, bataev,astapov,zatvornitskiy}@speechpro.com, maxim.korenevski@emr.ai

Abstract

This paper describes the Speech Technology Center (STC) system for the 5th CHiME challenge. This challenge considers the problem of distant multi-microphone conversational speech recognition in everyday home environments. Our efforts were focused on the single-array track, however, we participated in the multiple-array track as well. The system is in the ranking A of the challenge: acoustic models remain frame-level tied phonetic targets, lexicon and language model are not changed compared to the conventional ASR baseline. Our system employs a combination of 4 acoustic models based on convolutional and recurrent neural networks. Speaker adaptation with target speaker masks and multi-channel speaker-aware acoustic model with neural network beamforming are two major features of the system. Moreover, various techniques for improving acoustic models are applied, including array synchronization, data cleanup, alignment transfer, mixup, speed perturbation data augmentation, room simulation, and backstitch training. Our system scored 3rd in the single-array track with Word Error Rate (WER) of 55.5% and 4th in the multiple-array track with WER of 55.6% on the evaluation data, achieving a substantial improvement over the baseline system.

Index Terms: CHiME-5 challenge, noise-robust ASR, acoustic models, speaker adaptation

1. Introduction

Significant progress in Automatic Speech Recognition (ASR) area was made in recent years. Many ASR tasks have been thoroughly studied, and human parity level was achieved or even outperformed for some of them [1, 2]. However, there are still many challenges for researchers in ASR. One of such challenges is Distant Speech Recognition (DSR). DSR specific factors such as reverberation, noisiness, simultaneous speech of several speakers, etc. degrade ASR system performance drastically.

To date many efforts are devoted to DSR [3–6]. One of these efforts is the 5th CHiME challenge [7] which considers the problem of distant multi-microphone conversational speech recognition in everyday home environments. The main features of the CHiME-5 challenge are:

- simultaneous recordings from multiple microphone arrays;
- real conversation, i.e. talkers speaking in a relaxed and unscripted fashion;
- a range of room acoustics from 20 different homes each with two or three separate recording areas;
- real domestic noise backgrounds, e.g., kitchen appliances, air conditioning, movement, etc.

Audio data are recorded in real dinner parties conditions. In each party, two participants act like hosts, and another two are guests. Moreover, each party is arbitrarily divided into three phases by location: kitchen, dining, and living. The minimum duration of each phase is 30 minutes. Conversational speech with a great amount of overlapped segments recorded in reverberant and noisy conditions is a main difficulty of CHiME-5. Details on the challenge can be found in [7].

A large number of methods were developed for improving multi-microphone DSR. First of all, dereverberation is able to reduce the effect of interference caused by multiple reflections of acoustic signal from different surfaces [5, 8, 9]. Besides dereverberation, signal arrival direction determination is exceptionally helpful in case of multiple microphones. Direction of arrival is then used in various beamforming approaches that allow to highlight target speech against the background of other acoustic events [10–12]. Also separation of overlapped speech is crucial for accurate speech recognition. There are traditional approaches for source separation [13–16] as well as various modern techniques based on neural networks [17–19]. Finally, noise level reduction significantly decreases WER in DSR conditions. Thus, there are many different approaches aimed at improving the ASR system for speech recorded on a distant microphone array. Moreover, all of them should act in concert.

This paper provides a description of the STC system for the 5th CHiME challenge. The first key feature of the system is multi-channel speaker-aware acoustic model which utilizes neural network beamformer similarly to [20]. Focusing on a target speaker is performed following the concept of auxiliary inputs [21]. We used this model to extract speaker-dependent bottleneck (SDBN) features, as this approach has shown its effectiveness for conversational speech recognition in our previous works [22–24]. The second key feature is speaker adaptation using target speaker masks which are per-frame probabilities of a current frame usefulness. Masks extractor is a binary classifier on top of the SDBN features. We present two simple and effective schemes for speaker adaptation with these masks.

The rest of the paper is organized as follows. Section 2 describes data preparation techniques used in the system, including array synchronization, signal enhancement, and data augmentation. Acoustic modeling is covered in Section 3. Section 4 presents experiments and results. Finally, Section 5 concludes the paper.

[†] Maxim Korenevsky is now with EMR.AI Inc., San Francisco, CA, USA

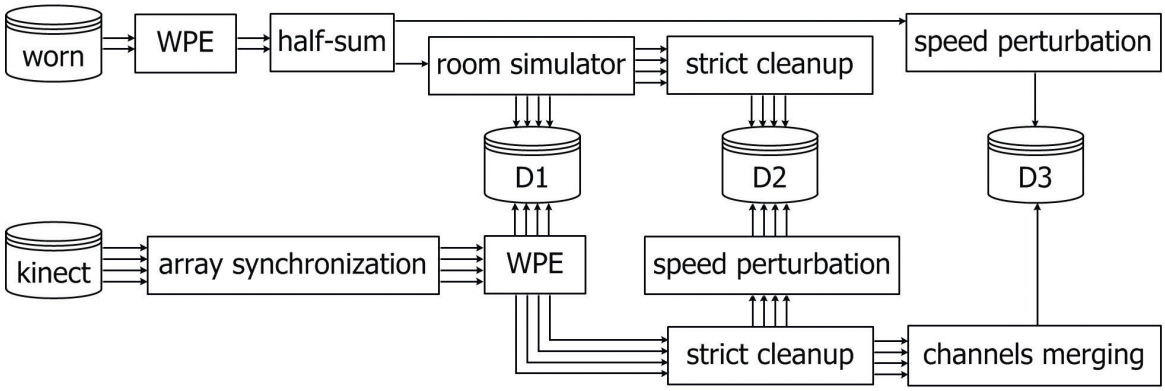

Figure 1: *Preparing of datasets*

2. Data preparation

2.1. Array synchronization

We assumed that the original training data are not synchronized perfectly. In order to compensate any asynchrony, we realigned original utterance segmentation on Kinects using the baseline GMM. The following scheme was used: extract per-frame features for a Kinect utterance, then compute likelihoods of the features sequence shifted by a few frames given the alignment obtained on the corresponding worn utterance. Finally, we modified the original json segment bounds using the shift value with maximum likelihood score.

More formally, if x_t is a feature vector corresponding to a frame t, $\mathcal{L}_M(x \mid p)$ is the likelihood of a feature vector x given the pdf-id p for model M, m is utterance length in frames, $p_0, \ldots, p_m$ is worn pdf-id alignment, then the shift value is calculated as

$$\tau = \underset{-5 \le l \le 5}{\operatorname{argmax}} \sum_{i=5}^{m-5} \log \mathcal{L}_M(x_{i+l} \mid p_i),$$

where summation is performed over speech frames only. So, new utterance bounds l_{new}, r_{new} are obtained from old utterance bounds l_{orig}, r_{orig} as

$$l_{new} = l_{orig} + \tau,$$
$$r_{new} = r_{orig} + \tau.$$

2.2. Signal enhancement and data augmentation

We found out that a half-sum of left and right worn channels (raw waveform mean, can also be considered as beamforming in a forward direction) provides moderate WER improvement. So, we added a half-sum of worn channels to the training set and also used it to obtain senone alignment. This alignment was used for target speaker masks generation (see Subsection 3.2) and was transferred to Kinect channels.

Both worn and Kinect recordings were dereverberated using the multi-channel version of the Weighted Prediction Error (WPE) algorithm [9]. We used open-source implementation[1] described in [25].

We slightly modified Kaldi cleanup procedure [6] used in the baseline setup and called it $strict\ cleanup$. All utterances changed by vanilla cleanup were excluded. Furthermore, utterances with WER higher than a predefined threshold were also excluded. As a result, the amount of training data has been cut in half.

In order to generate large-scale simulated data, we implemented the room acoustics simulator in a similar way as described in [26]. The simulation uses binaural signals of relatively clean utterances as the source. The noises used in the simulation were extracted from Kinect signals using a simple speaker-invariant model.

Additionally, we applied speed perturbation data augmentation technique [27]. Unlike the baseline recipe, we did not use volume perturbation.

Finally, recently proposed mixup data augmentation technique [28] was used for acoustic model training. This approach performs on-the-fly generation of virtual training examples by combining the existing ones. We used Kaldi-compatible implementation[2] described in our previous paper [29].

Figure 1 presents the preparation of 3 datasets D1, D2, and D3, used for training of the acoustic models. Note that mixup is not presented in the scheme, as it is performed on-the-fly during the training.

3. Acoustic modeling

3.1. Multi-channel speaker-aware acoustic model

Following the approach applied in Google Home [20], we built a multi-channel acoustic model that performs neural network beamforming. This model takes 4-channel complex spectra as an input and performs computations with both real and imaginary parts of complex numbers. Assuming that speaker-aware nature of the acoustic model is crucial for speech recognition accuracy, we investigated neural network speaker mask estimation. To focus on a target speaker, we applied the idea of auxiliary inputs used in [21].

First of all, we trained speaker embedding extractor depicted in Figure 2a using the triplet ranking loss [30]. In order to build a mask branches $mask_1$ and $mask_2$, we used Residual Attention Network [31]. Moreover, after the convolutional layers we applied the $attention$ network with a soft mask on the time dimension. Speaker embeddings were produced for each frame and then averaged into a single auxiliary vector of dimension 512 for each utterance. The averaging procedure was

[1] https://github.com/fgnt/nara_wpe

[2] https://github.com/Speechpro/Mixup

based on the speaker-adapted classifier designed to recognize 4 speakers, which was trained only on the target session recordings. This simple classifier was built on top of the embedding features using the LightGBM toolkit[3].

Once the speaker-aware auxiliary vector was defined for all utterances, we biased mask estimation and attention layers with auxiliary inputs as depicted in Figure 2b. Multi-channel speaker-aware acoustic model was trained to classify 3912 pdf-ids. Training of this model was performed in 12 epochs using the CNTK toolkit [32]. Then, we added a 128-dimensional linear bottleneck layer before the output layer by applying singular value decomposition. After that, one epoch of fine-tuning was carried out. This model was used for extraction of speaker-dependent bottleneck features.

D1 dataset was used for training of both speaker embedding extractor and speaker-aware acoustic model. In the case of the single-array track, we performed speaker adaptation only on a single test session from the reference Kinect, i.e. trained 12 different classifiers. For the multiple-array track, we trained classifier on reference utterances from test session, and used this model for non-reference device signals.

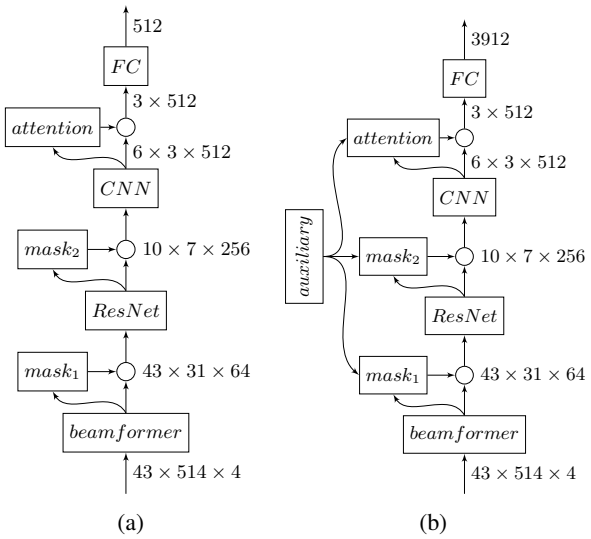

Figure 2: *Multi-channel (a) speaker embedding extractor and (b) speaker-aware acoustic model architectures.*

3.2. Adaptation using target speaker masks

We found that speaker-invariant acoustic models and even models adapted with i-vectors often recognize all speech in the input signal, including background speech and cross talk. To overcome this shortcoming, we explored adaptation using a frame-level mask of a target speaker.

The key point in building such masks is to train a classifier that predicts whether we should use that frame or not. The best option we found was to build a binary classifier on top of the SDBN features extracted from the speaker-aware acoustic model. Classifier was trained using the LightGBM toolkit.

Ideal targets for the classifier training were constructed in two steps. First, senone alignment for all training utterances using half-sum of two worn channels was obtained. Second, for all frames containing silence only or target speaker overlapped with no more than two additional speakers, 1 was assigned; in all other cases target was 0.

Two schemes of speaker adaptation using the obtained per-frame target speaker masks (classifier probabilities) were applied:

- Mask Filtering (MF): Classify each frame as target or non-target speech by a probability threshold, and then simply throw out non-target speech frames before the decoding. This method is easy to implement and also speeds up the decoding.
- Mask Appending (MA): Train acoustic model using a concatenation of input features and the soft-mask. Acoustic model architecture is modified by adding a gate with sigmoid layer to the input of the network.

3.3. Final acoustic models

In order to utilize the diversity of datasets, input features, model architectures, and speaker adaptation approaches, we trained a lot of various acoustic models. 4 of them were included in the final system:

- $\mathbf{AM}_1$: Time Delay Neural Network (TDNN) [33] trained with LF-MMI criterion [34] on D1 dataset and SDBN features. Training was performed with mixup data augmentation [29] and backstitch regularization [35].
- $\mathbf{AM}_2$: TDNN combined with unidirectional Long Short-Term Memory layers (TDNN-LSTM) trained with LF-MMI criterion on D2 dataset and 4-channel concatenation of 80-dimensional log mel filterbank features. MA speaker adaptation scheme was applied. Training was performed with mixup data augmentation.
- $\mathbf{AM}_3$: the same as $\mathbf{AM}_2$, but features were 40-dimensional Mel Frequency Cepstral Coefficients (MFCC).
- $\mathbf{AM}_4$: Bidirectional Long Short-Term Memory (BLSTM) network trained with cross-entropy criterion on D3 dataset and MFCC features appended with an i-vector. As in previous models, mixup data augmentation was applied. MF speaker adaptation scheme was used in the evaluation stage. Additionally, we applied a softmax temperature to a prior distribution for this model during the decoding.

4. Experimental evaluation

Table 1: *Overall WER (%) for individual models as well as their combination on the development set*

Track	Model	WER
Single	$\mathbf{AM}_1$	63.4
	$\mathbf{AM}_2$	63.3
	$\mathbf{AM}_3$	63.8
	$\mathbf{AM}_4$	66.0
	$\mathbf{AM}_1^\star$	63.8
	$\mathbf{AM}_1 + \mathbf{AM}_2 + \mathbf{AM}_3 + \mathbf{AM}_4$	59.4
Single+Dev	$\mathbf{AM}_1 + \mathbf{AM}_2 + \mathbf{AM}_3 + \mathbf{AM}_4$	56.6
Multiple	$\mathbf{AM}_1 + \mathbf{AM}_2 + \mathbf{AM}_3 + \mathbf{AM}_1^\star$	58.1
Multiple+Dev	$\mathbf{AM}_1 + \mathbf{AM}_2 + \mathbf{AM}_3 + \mathbf{AM}_1^\star$	53.5

In the single-array track, at first, posterior-level combination of 4 channels for the single-channel model $\mathbf{AM}_4$ was per-

[3] https://github.com/Microsoft/LightGBM

formed. Then, lattice-level combination of 4 systems (STC confusion networks based implementation described in [36]) was applied.

In the multiple-array track, we applied naive combination of word lattices for all 6 devices. Moreover, we also added lattices from the reference device to the combination. Single-channel model $\mathbf{AM}_4$ was excluded due to very high computational cost (unlike LF-MMI models, this model did not have frame subsampling, that led to 3 times longer decoding). Instead of $\mathbf{AM}_4$, we took model $\mathbf{AM}_1^*$ which is similar to $\mathbf{AM}_1$, but without per-utterance averaging of speaker embeddings. STC lattice combination was replaced with the Kaldi implementation due to computational reasons.

Table 1 shows the results of individual models as well as their combination on the development data. It should be noted that, in order to achieve maximum performance on the evaluation data, we augmented the training datasets with the development sessions and retrained our acoustic models. For clarity, we denote this scenario as *Single+Dev* and *Multiple+Dev* tracks. The results of the combination of these models were submitted as our final system. In Table 2, we report detailed results of the final system for each session and location. Finally, approximate contributions in WER reduction for the methods applied in the system are presented in Table 3.

Table 2: *WER (%) for the final system per session and location*

Track	Session		Kitchen	Dining	Living	Overall
Single+Dev	Dev	S02	65.5	56.2	52.4	56.6
		S09	55.7	56.8	51.9	
	Eval	S01	65.9	48.8	65.8	55.5
		S21	58.9	49.2	51.9	
Multiple+Dev	Dev	S02	62.1	52.2	50.2	53.5
		S09	51.2	51.6	51.4	
	Eval	S01	65.4	47.6	65.2	55.6
		S21	61.6	50.0	51.3	

Table 3: *Contributions in WER reduction for the methods applied in the system*

Method	WER reduction, % abs.
Array synchronization	0.9
Alignment transfer (worn half-sum to kinect)	1.3
WPE dereverberation	1.4
Strict cleanup	1.3
Room simulator	1.6
Speed perturbation	0.9
Mixup training	1.1
Backstitch training	0.5
Multi-channel models	2.2
Speaker adaptation (i-vectors)	2.4
Speaker adaptation (auxiliary)	4.1
Speaker adaptation (MF)	4.9
Speaker adaptation (MA)	7.0
Combination of 4 models	3.9

5. Conclusions

In this paper we presented our system for the 5th CHiME Challenge which scored 3rd in the single-array track and 4th in the multiple-array track.

As can be seen from Table 3, back-end approaches provided the main improvement in terms of WER. Data preparation techniques and models combination were also beneficial. On the other hand, our system almost does not use front-end approaches, except WPE dereverberation. We assume that significant performance gain can be achieved with strong front-end.

It is also interesting that there is no improvement on the evaluation data in the multiple-array track comparing to the single-array track. Probably, this is due to a large mismatch between development and evaluation data.

6. Acknowledgements

We would like to thank Aigul Nugmanova for valuable discussions on the speaker masks application. We also thank our former colleagues Valentin Mendelev and Alexey Prudnikov for their contribution to the development of our team and technologies.

This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.575.21.0178 (ID RFMEFI57518X0178).

7. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward Human Parity in Conversational Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, Dec 2017.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L. Lim, B. Roomi, and P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," *CoRR*, vol. abs/1703.02136, 2017. [Online]. Available: http://arxiv.org/abs/1703.02136

[3] S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, Eds., *New Era for Robust Speech Recognition, Exploiting Deep Learning*. Springer, 2017.

[4] J. P. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The chime challenges: Robust speech recognition in everyday environments," in *New Era for Robust Speech Recognition, Exploiting Deep Learning*. Springer, 2017, pp. 327–344.

[5] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "The REVERB challenge: A benchmark task for reverberation-robust ASR techniques," in *New Era for Robust Speech Recognition, Exploiting Deep Learning*. Springer, 2017, pp. 345–354.

[6] V. Peddinti, V. Manohar, Y. Wang, D. Povey, and S. Khudanpur, "Far-field ASR without parallel data," in *INTERSPEECH*, 2016.

[7] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.

[8] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[9] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing 20.10: 2707-2720.*, 2012.

[10] B. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine, Vol. 5 No. 2 pp. 424.*, 1988.

[11] D. H. Johnson and D. Dudgeon, "Array signal processing," *Englewood Cliffs, NJ: Prentice Hall*, 1993.

[12] H. Van Trees, "Optimum array processing," *New York, NY: Wiley-Interscience*, 2002.

[13] M. Togami, "Online speech source separation based on maximum likelihood of local gaussian modeling," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.213-216*, 22-27 May 2011.

[14] P. Comon, "Independent component analysis, a new concept?" *Signal Processing, vol. 36, no. 3, pp. 287-314*, 1994.

[15] D. Wang, "Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design," *Trends in Amplification, vol. 12, no. 4, pp. 332-353*, 2008.

[16] H. Tao, J.-Y. Zhang, and L. Yu, "Adaptive Blind Source Seperation Using Temporal Predictability," *Proceedings of 2006 International Conference on Communications, Circuits and Systems*, 2006.

[17] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *CoRR*, vol. abs/1508.04306, 2015. [Online]. Available: http://arxiv.org/abs/1508.04306

[18] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct 2017.

[19] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation," *CoRR*, vol. abs/1607.00325, 2016. [Online]. Available: http://arxiv.org/abs/1607.00325

[20] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. K. Chin, K. C. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic Modeling for Google Home," in *INTERSPEECH*, 2017.

[21] K. Zmolikova, M. Delcroix, K. Kinoshita, and T. Higuchi, "Optimization of speaker-aware multichannel speech extraction with asr criterion," in *ICASSP*, 2018.

[22] A. Prudnikov, I. Medennikov, V. Mendelev, M. Korenevsky, and Y. Y. Khokhlov, "Improving acoustic models for russian spontaneous speech recognition." in *SPECOM*, ser. Lecture Notes in Computer Science, vol. 9319. Springer, 2015, pp. 234–242.

[23] I. Medennikov, A. Prudnikov, and A. Zatvornitsky, "Improving English Conversational Telephone Speech Recognition," in *INTERSPEECH*, 2016.

[24] Y. Khokhlov, I. Medennikov, A. Romanenko, V. Mendelev, M. Korenevsky, A. Prudnikov, N. Tomashenko, and A. Zatvornitsky, "The STC Keyword Search System for OpenKWS 2016 Evaluation," in *INTERSPEECH*, 2017.

[25] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, Oct 2018.

[26] C. Kim, A. Misra, K. K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani, "Generation of Large-Scale Simulated Utterances in Virtual Rooms to Train Deep-Neural Networks for Far-Field Speech Recognition in Google Home," in *INTERSPEECH*, 2017.

[27] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *INTERSPEECH*, 2015.

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv:1710.09412*, 2017.

[29] I. Medennikov, Y. Khokhlov, A. Romanenko, D. Popov, N. Tomashenko, I. Sorokin, and A. Zatvornitskiy, "An Investigation of Mixup Training Strategies for Acoustic Models in ASR," in *INTERSPEECH*, 2018.

[30] M. Ye and Y. Guo, "Deep triplet ranking networks for one-shot recognition," *arXiv preprint arXiv:1804.07275*, 2018.

[31] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," *arXiv preprint arXiv:1704.06904*, 2017.

[32] F. Seide and A. Agarwal, "CNTK: Microsoft's Open-Source Deep-Learning Toolkit," in *KDD*, 2016.

[33] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[34] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *INTERSPEECH*, 2016, pp. 2751–2755. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-595

[35] Y. Wang, V. Peddinti, H. Xu, X. Zhang, D. Povey, and S. Khudanpur, "Backstitch: Counteracting Finite-Sample Bias via Negative Steps," in *INTERSPEECH*, 2017, pp. 1631–1635.

[36] Y. Khokhlov, I. Medennikov, A. Romanenko, V. Mendelev, M. Korenevsky, A. Prudnikov, N. Tomashenko, and A. Zatvornitsky, "The STC keyword search system for OpenKWS 2016 evaluation," in *INTERSPEECH*, 2017.