IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, July 2019.

TWO-STAGED ACOUSTIC MODELING ADAPTION FOR ROBUST SPEECH RECOGNITION BY THE EXAMPLE OF GERMAN ORAL HISTORY INTERVIEWS

Michael Gref [1,2], *Christoph Schmidt*[1], *Sven Behnke*[1,3], *Joachim Köhler*[1]

[1]Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Germany
[2]Institute for Pattern Recognition (iPattern), Niederrhein Univ. of Applied Sciences, Germany
[3]Autonomous Intelligent Systems (AIS), Computer Science Institute VI, Univ. of Bonn, Germany
{michael.gref, christoph.andreas.schmidt, sven.behnke, joachim.koehler}@iais.fraunhofer.de

ABSTRACT

In automatic speech recognition, often little training data is available for specific challenging tasks, but training of state-of-the-art automatic speech recognition systems requires large amounts of annotated speech. To address this issue, we propose a two-staged approach to acoustic modeling that combines noise and reverberation data augmentation with transfer learning to robustly address challenges such as difficult acoustic recording conditions, spontaneous speech, and speech of elderly people. We evaluate our approach using the example of German oral history interviews, where a relative average reduction of the word error rate by 19.3% is achieved.

Index Terms— Robust speech recognition, domain adaption, transfer learning, multi-condition training, data augmentation, oral history

1. INTRODUCTION

Automatic speech recognition (ASR) has undergone enormous improvements in recent years. Nowadays, it is successfully used in many applications, both in the commercial and industrial sectors. ASR not only enables the development of smart speech assistants but is also used for subtitling, information mining, analytics, and recommendation.

However, training state-of-the-art ASR systems requires large amounts of annotated speech. If training data is not available to a sufficient extent, only unsatisfactory results are achieved. Especially for challenging scenarios, often only little training data is available, and off-the-shelf ASR systems perform poorly. Such challenges can arise from different acoustic conditions such as noise and reverberation, but also varying recording equipment, spontaneous, fast speech, unclear pronunciations and dialects can be very challenging.

In this work, we propose an approach to tackle these challenges by combining multi-condition training via data augmentation and transfer learning on very little data in a two-staged acoustic modeling adaption. We evaluate our approach using the example of German oral history interviews, in which all aforementioned challenges occur to varying degrees.

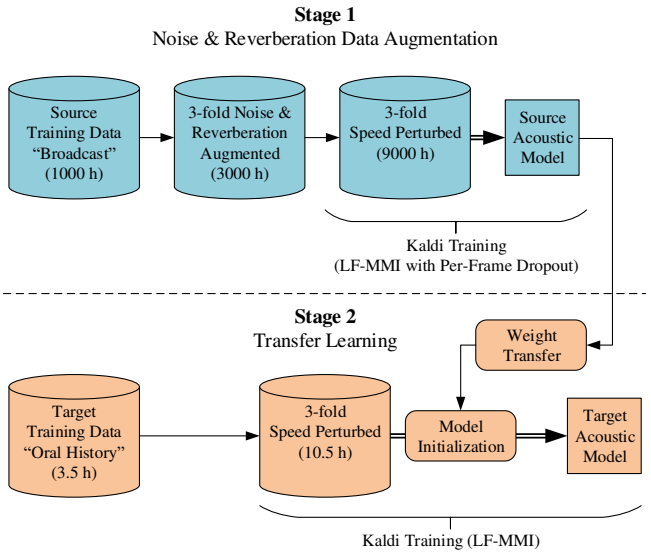

Fig. 1. Proposed approach. Noise and reverberation data augmentation is applied in Stage 1 to obtain a robust acoustic source model. In Stage 2, transfer learning is applied to tackle further challenges such as spontaneous speech.

2. RELATED WORK

ASR is a popular and highly researched area and new approaches are regularly proposed. Currently, lattice-free maximum mutual information (LF-MMI) trained models achieve state-of-the-art results on many different ASR tasks [1].

Oral history in historical research refers to conducting and analyzing interviews with contemporary witnesses. In Germany, this kind of research focused above all on the period of the Second World War and National Socialism. In the meantime, it has also come to include many other topics and historical periods. In our prior work [2, 3] we studied the ap-

plication and adaption of state-of-the-art ASR to German oral history interviews.

Applying data augmentation to training data is a common approach to increase the amount of training data in order to improve the robustness of a model. In ASR it can be used, e.g., to apply multi-condition training, when no real data in the desired condition is available. Data augmentation is, however, limited to acoustic effects that can be created in a sufficiently realistic manner - such as additive noise and reverberation. The data augmentation of reverberant speech for state-of-the-art LF-MMI models has been studied by Ko et al. [4]. Several speed perturbation techniques to increase the training data variance have been investigated by Ko et al. [5]. The proposed method in this work is to increase the data threefold by creating two additional versions of each signal using the constant speed factors 0.9 and 1.1—a method that is used in many recent Kaldi training routines by default.

Transfer learning is an approach used to transfer knowledge of a model trained in one scenario to train a model in another related scenario to improve generalization and performance [6]. It is particularly useful in scenarios where only little training data is available for the main task but a large amount of annotated speech is available for a similar or related task. A detailed overview of transfer learning in speech and language processing is given by Wang et al. [7]. Transfer learning for ASR systems using LF-MMI models has been studied by Ghahremani et al. [8] for many different common English speech recognition tasks.

However, most works in ASR, such as the aforementioned, studied transfer learning with a much greater amount of annotated speech than is available in the oral history task, for instance. In addition, most works focus on either data augmentation or transfer learning, usually to address a particular task or challenge, such as robustness to noise and not the robustness of an acoustic model as a whole.

3. PROPOSED APPROACH

We aim at improving the performance of robust ASR systems by performing a two-staged acoustic modeling adaption using a very little amount of target training data. An overview of the proposed method is given in Fig. 1. In the first stage, multi-condition training is applied using noise and reverberation data augmentation to obtain a robust acoustic source model. The second stage applies transfer learning to tackle the remaining challenges of the target data that could not be synthesized in the first stage, such as spontaneous speech, dialectics and pronunciations.

The first stage of the approach is based on our prior work [3], where multi-condition training using noise and reverberation data augmentation was used to decrease the acoustic mismatch of conventional clean training data and oral history interviews. This has been proven to significantly increase the performance of ASR systems on German oral history interviews. In contrast to the aforementioned work, where the amount of training data is kept to the same size, in our approach the data is increased 3-fold.

Defining discrete-time-signals as sequences of sample values, the applied augmentation can be described as

$$(x_n)_{n\in\mathbb{N}} := (s_n)_{n\in\mathbb{N}} * (h_n)_{n\in\mathbb{N}} + (w_n)_{n\in\mathbb{N}} * (\tilde{h}_n)_{n\in\mathbb{N}} \quad (1)$$

if both noise and reverberation inside a simulated room affects the speech signal. Here, $*$ is the convolution operation for sequences, $(s_n)_{n\in\mathbb{N}}$ the sequence of the clean speech signal, $(h_n)_{n\in\mathbb{N}}$, $(\tilde{h}_n)_{n\in\mathbb{N}}$ are room impulse responses modeling the reverberation of one room at different positions and $(w_n)_{n\in\mathbb{N}}$ describes the sequence of the noise signal. If only reverberation and no background noise affects the speech signal, $\forall n \in \mathbb{N} : w_n = 0$ applies and yields

$$(x_n)_{n\in\mathbb{N}} := (s_n)_{n\in\mathbb{N}} * (h_n)_{n\in\mathbb{N}}. \quad (2)$$

We use 266 room impulse responses of small and medium sized rooms along with 14.5 hours of real-life noise recordings collected from different sources - such as the Aachen Impulse Response database [9], other freely available and inhouse data. We create the following two artificially corrupted versions of our source training data and merge them with the original (clean) set to create a 3000 hour multi-condition source training set:

- **Reverb**: All signals are convolved according to Equation (2) with randomly selected room impulse responses of small or medium sized rooms. No noise is applied here.
- **Reverb+RealNoise**: Similar to **Reverb** but added noise recordings according to equation (1) applying a random signal-to-noise ratio between 10 and 20 dB. The noises have been randomly selected from real-life recordings, e.g. street noises, bus noises, police sirens, hairdryers. To avoid overfitting, we randomly selected and superposed up to three different noises for one audio file before applying the reverberation.

The transfer learning in Stage 2 is inspired by the work of Ghahremani et al. [8]. In our setup, a full weight transfer of the entire source model for initialization of the target model is applied without any layer freezing. In particular, the output layer is not replaced in contrast to some other transfer learning approaches in ASR, since the same set of phonemes and the same decision tree is used both in the source and target scenario. In the transfer learning stage, the i-vector extractor of the model trained in Stage 1 is used without any adaption.

The neural network training routine in Stage 2 is almost equal to the one used in Stage 1 with only slight adjustments. An overview of the parameters that are different in the transfer learning stage is given in Table 1. In Stage 1 we apply per-frame dropout according to Cheng et al. [10]. The training in

Table 1. Changed training parameters in Stage 1 and 2

Parameter	Stage 1	Stage 2
Init./final learn rate	1e-3 / 1e-4	1e-6 / 1e-7
Dropout Schedule	0, 0@0.2, 0.3@0.5, 0	0, 0

Stage 2 is performed without dropout. Our previous experiments with transfer learning showed that dropout seems to reduce the performance training on small data sets. In both stages, the training is performed for four epochs with a reducing learning rate. The initial and final learning rate in the second stage is lower than in the first stage due to the significantly smaller amount of training data. Note that 3-fold speed perturbation is applied in every setup, since we consider this technique to be a default procedure in the Kaldi training routines.

4. EXPERIMENTAL SETUP

4.1. Lexicon and language model

The lexicons needed for training in Stage 1, Stage 2 and decoding are all obtained using the same grapheme-to-phoneme pronunciation model trained with Sequitur G2P [11]. This model is created using the German pronunciation database Phonolex from the Bavarian Archive for Speech Signals.

For decoding, we use a 500,000 words 5-gram broadcast language model. This model is trained on broadcast text corpora consisting of 75 million words. Decoding parameters are kept the same for all experiments. In particular, the language model weight is kept to a fixed value for all experiments.

4.2. Acoustic model

4.2.1. Training data

For training the source system in the first stage, we utilize a 1000 hour large-scale corpus of German broadcast speech data *GerTV1000h* [12]. This data set can be considered to be out of domain for the oral history scenario, since the broadcast recordings differ from oral history in terms of the used recording technology, audio signal quality and speech characteristics.

As target data, we use the oral history data set proposed in our prior work [2]. It consists of 3.5 hours audio from 35 different speakers recorded in real oral history interviews. All audio signals are resampled to the sample frequency of the training data (16 kHz). The set contains 27,708 transcribed spoken words with a vocabulary of 4582 words. The recordings took place between 1980 and 2012, representing a wide range of recording technology, interview methodology, dialects and pronunciations. The set is manually transcribed and segmented and has an average segment length of 5.3 seconds with overall 2392 segments.

Table 2. Hidden Layer of the Acoustic Model

#	Type	Temporal Context
1	TDNN	$\{-2,-1,0,1,2\}$
2 & 3	TDNN	$\{-1,0,1\}$
4	LSTM	–
5 & 6	TDNN	$\{-3,0,3\}$
7	LSTM	–
8 & 9	TDNN	$\{-3,0,3\}$
10	LSTM	–

4.2.2. Acoustic model neural network topology

All acoustic model networks use a 300-dimensional input at each time-step consisting of five consecutive 40-dimensional MFCC features and a 100-dimensional i-vector [13]. We use a topology with ten hidden layers that was proposed and investigated by Cheng et al. [10]. The acoustic model neural network consists of seven TDNN layers [14, 15] and three LSTM layers [16] stacked in the order given in Table 2.

The applied implementation uses LSTM layers with forget gates [17], peephole connections [18] and projection layers [19]. The LSTM layers have a cell dimension of 1024 and a projection dimension of 256. The TDNN layers are 1024-dimensional.

4.2.3. Experiments

All experiments are carried out using the Kaldi ASR toolkit [20]. As part of the Kaldi training routines, the aforementioned speed perturbation [5] is applied on the entire training data to increase the amount of data three-fold before neural network training. All models are trained using the LF-MMI [1] criterion. Overall, four major types of setups are examined in our experiments:

1. **Baseline**: For comparison, we train a baseline acoustic model with the same setup as in Stage 1, excluding the noise and reverberation data augmentation and the entire transfer learning Stage 2.
2. **Stage 1 Only (Data Augmentation)**: Evaluating the performance of the source model trained in Stage 1 using the noise and reverberation data augmentation on the 35 speaker sets.
3. **Stage 2 Only (Transfer Learning)**: Applying the transfer learning experiments on the clean-trained baseline model.
4. **Proposed Approach**: Applying Stage 1 and Stage 2.

4.2.4. Leave-one-speaker-out evaluation

Since only very little data from the target domain is available, we apply a leave-one-speaker-out evaluation on the tar-

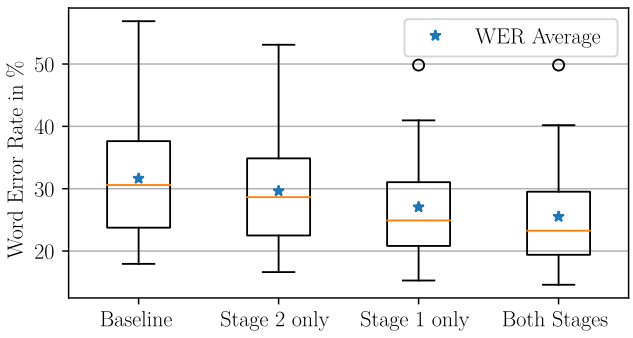

Fig. 2. Boxplot diagrams of the achieved WER in the 35 leave-one-speaker-out experiments for each setup. The star within the boxplots marks the average word error rate of all experiments w.r.t to the number of words in each of the sets.

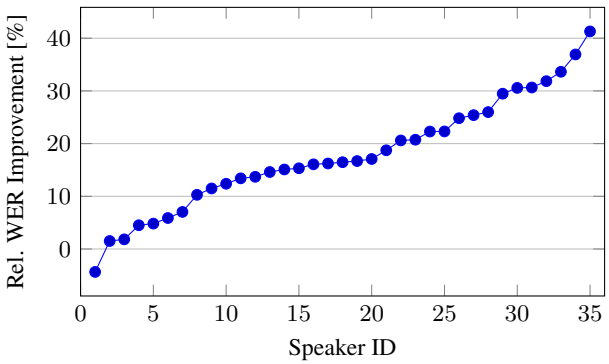

Fig. 3. Relative word error rate improvement of the proposed approach compared to the clean baseline for each leave-one-speaker-out experiment. IDs are sorted by the improvement for better visualization.

get data. This approach can be understood as a k-fold cross-validation where the data set is partitioned according to speakers. This means each subset consists of exactly one speaker. Then we loop over the data subsets and keep one speaker out of the training set for validation and train one model on the data of the remaining $k-1$ speakers. This way, we run k experiments in Stage 2 and evaluate each trained model on the speaker that was not present in the training data. We trained one model in Stage 1 and then used this model as the source model for all 35 different leave-one-speaker-out experiments in Stage 2.

5. RESULTS AND DISCUSSION

5.1. Leave-one-speaker-out experiments

The results of the 35 leave-one-speaker-out experiments for the four different setups are given in form of a boxplot diagram in Fig. 2. Our experiment shows that the word error rates (WER) significantly decreases when applying the proposed approach, compared to the baseline. In the clean setup, only one half of the experiments achieve a WER below 30%. However, in the proposed approach, this is the case for about 75% of the experiments. With the exception of one outlier, all experiments in the proposed approach have a WER below or near 40%. Half of the experiments achieve a WER below 24% in the proposed approach. On average, the WER decreases from 31.6% in the clean setup to 25.5% using the two-staged approach.

The relative WER improvements of each leave-one-speaker-out experiment using the proposed approach compared to the clean-trained baseline model are shown in Fig. 3. For 34 out of the 35 experiments the WER does decrease and only for one experiment the WER slightly increases. The speaker in this one experiment is recorded in a rather clean acoustic condition and has no noteworthy peculiarities in the nature of his speaking. For 27 out of the 35 experiments, the WER improves by more than 10% relative to the baseline.

The results using Stage 1 only are slightly worse than in the two-staged approach. On average, the WER achieved using only Stage 1 is 27.1%. This means removing Stage 2 decreases the speech recognition performance by 6.3% relative. Removing Stage 1 and applying the transfer learning stage on the clean baseline model gives an average WER of 29.6%. Thus, in the setup examined, transfer learning on a clean source model yields on average slightly worse results than the sole data augmentation in Stage 1—but is also a significant improvement to the baseline.

A more in-depth look at the 35 individual experiments is given in Fig. 4 where the relative WER improvement in comparison to the proposed two-staged approach is given when only one of the stages is applied. It is evident that the data augmentation has a large impact on the speech recognition performance in many of the experiments. However, for four

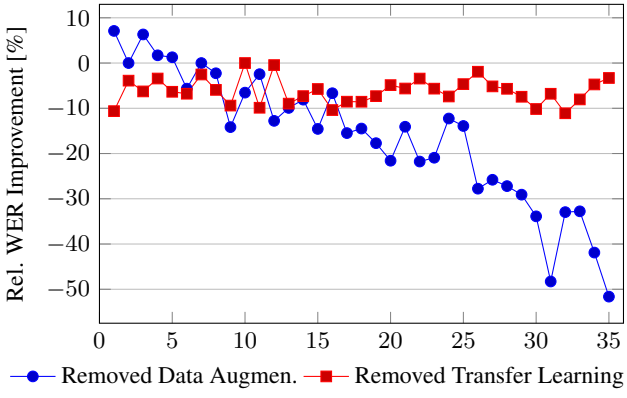

Fig. 4. Relative WER change for each leave-one-speaker-out experiments in case one of the stages is removed from the proposed approach. Speaker IDs are in the same order as in Fig. 3. Negative values indicate an increased WER.

Table 3. Word error rates on several in-house evaluation sets from different domains. Legend: y: yes; n: no, p: partly

Evaluation Set	**Size** [min.]	**Noise**	**Reverb.**	**Spont. Speech**	**Baseline**	**Stage 2 only**	**Stage 1 only**	**Two-Staged**
DiSCo Planned Clean	55	n	n	n	9.03	9.23	8.95	**8.89**
DiSCo Spontaneous Clean	115	n	n	y	10.25	10.06	**9.90**	9.94
DiSCo Planned Mix	87	y	n	n	11.64	11.67	**10.80**	10.83
DiSCo Spontaneous Mix	66	y	n	y	19.48	18.80	17.54	**17.41**
General German Broadcasts	61	p	p	p	12.31	11.87	11.49	**11.24**
Challenging Broadcast Radio	52	y	p	y	23.43	23.20	22.69	**22.02**
Challenging Broadcast TV	53	y	p	y	17.78	17.44	17.28	**17.02**
Spoken QALD-7	15	p	y	p	20.59	19.70	18.34	**17.72**
Humanities (Interaction)	49	y	y	y	66.50	64.37	47.81	**47.13**

experiments it even increases the WER. The improvement by the transfer learning on the other hand is quite consistent for the experiments.

5.2. Robustness with several evaluation sets

Finally, we investigate the robustness of the proposed approach by evaluating it on several different German in-house evaluation sets from other domains. Some of the sets partly share some challenges of oral history interviews, such as spontaneous speech or reverberation.

DiSCo [21] is a corpus for the German broadcast domain and is split in four evaluation sets: planned and spontaneous speech each in clean and mixed acoustic conditions. The two Challenging Broadcast evaluation sets are similar to the DiSCo Spontaneous Mix set and contain several challenging interviews and recordings with a lot of spontaneous speech, often in challenging acoustic conditions, and even some overlapping speech. The Spoken QALD-7 corpus contains in-house recorded questions for a question answering system based on [22] by several speakers using a web interface and their respective microphone—a headset or build-in laptop microphone for instance. The in-house Humanities evaluation set contains recordings of people informally talking to each other about different topics recorded in challenging acoustic conditions.

For this experiment, we use the entire oral history set in the second stage for transfer learning and no data is held out for evaluation. The word error rates of such a model on the evaluation sets are given in Table 3. Even though we used the two-staged acoustic modeling adaption to improve the performance on oral history interviews, the model performs better than the comparison models on many of the evaluation sets. The increase in performance is higher on rather challenging test sets while maintaining or even slightly increasing the good performance on the more clean tasks. Therefore, we conclude that the two-staged approach not only is useful for a specific task but also helps to increase the generalization of the acoustic model.

6. CONCLUSION

In this work, we proposed a two-staged acoustic modeling adaption for robust speech recognition and evaluated the approach on the challenging example of German oral history interviews. We evaluated the reliability of our approach with a leave-one-speaker-out evaluation method in which we perform 35 experiments for one setup. We showed that the proposed approach increases the speech recognition performance in 34 of the 35 experiments and performs better than using one of the methods alone. On average, the word error rate decreases relatively by 19.3%. Furthermore, we showed that our approach helps to increase the generalization of acoustic models and leads to increased recognition for challenging recordings while maintaining the good performance on clean tasks.

7. ACKNOWLEDGEMENTS

This research has been funded by the Federal Ministry of Education and Research of Germany (BMBF) in the project *KA*^{3} *- Kölner Zentrum für Analyse und Archivierung von AV-Daten* (Cologne center for the analysis and archiving of audiovisual data) (project number: 01UG1811B).

8. REFERENCES

[1] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *17th Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 2751–2755.

[2] Michael Gref, Joachim Köhler, and Almut Leh, "Improved transcription and indexing of oral history interviews for digital humanities research," in *Eleventh Inter-*

national Conference on Language Resources and Evaluation (LREC), 2018.

[3] Michael Gref, Christoph Schmidt, and Joachim Köhler, "Improving robust speech recognition for german oral history interviews using multi-condition training," in *13. ITG Symposium on Speech Communication*. 2018, pp. 256–260, IEEE.

[4] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[5] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *16th Annual Conference of the International Speech Communication Association (Interspeech)*, 2015, pp. 3586–3589.

[6] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville, *Deep Learning*, Adaptive computation and machine learning. MIT Press, 2016, pp. 526–528.

[7] Dong Wang and Thomas Fang Zheng, "Transfer learning for speech and language processing," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 1225–1237.

[8] Pegah Ghahremani, Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 279–286.

[9] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th International Conference on Digital Signal Processing*, 2009, pp. 1–5.

[10] Gaofeng Cheng, Vijayaditya Peddinti, Daniel Povey, Vimal Manohar, Sanjeev Khudanpur, and Yonghong Yan, "An exploration of dropout with lstms," in *18th Annual Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 1586–1590.

[11] Maximilian Bisani and Hermann Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[12] Michael Stadtschnitzer, Jochen Schwenninger, Daniel Stein, and Joachim Köhler, "Exploiting the large-scale german broadcast corpus to boost the fraunhofer IAIS speech recognition system," in *Ninth International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 3887–3890.

[13] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] Alexander H. Waibel, Toshiyuki Hanazawa, Geoffrey E. Hinton, Kiyohiro Shikano, and Kevin J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[15] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *16th Annual Conference of the International Speech Communication Association (Interspeech)*, 2015, pp. 3214–3218.

[16] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[18] Felix A. Gers and Jürgen Schmidhuber, "Recurrent nets that time and count," in *IJCNN (3)*, 2000, pp. 189–194.

[19] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *15th Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 338–342.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Dec. 2011, IEEE Signal Processing Society.

[21] Doris Baum, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim Köhler, "DiSCo - A german evaluation corpus for challenging problems in the broadcast domain," in *International Conference on Language Resources and Evaluation (LREC)*, 2010.

[22] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano, "7th open challenge on question answering over linked data (QALD-7)," in *Semantic Web Challenges - 4th SemWebEval Challenge at (ESWC)*, 2017, pp. 59–69.