A COMPARATIVE STUDY OF MULTI-CHANNEL PROCESSING METHODS FOR NOISY AUTOMATIC SPEECH RECOGNITION IN URBAN ENVIRONMENTS

Tran Huy Dat, Jonathan Dennis, Leng Yi Ren, Ng Wen Zheng Terence

Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632

ABSTRACT

For the distant speech recognition, the multi-channel processing has been proven to significantly improve the ASR performances compared to the single channel approaches. However, there is very little work has done to provide a comparative evaluation of the approaches, particularly with the modern Deep Neural Network (DNN) recognizers. In this paper, we address the above problem by evaluating the most recently reported mutti-channel methods for the distant speech recognition under urban environments using the 3rd CHiME Challenge database. Particularly, we analyse the effects of each stage of processing of beamforming, adaptive noise cancellation and dereverberation. The back-end processing components are also investigated. We further describe in details our best performing system which combines a harmonic to subharmonic ratio (SHR) voice activity detection, and correlative beamforming with adaptive channel selection in the from-end; semi-supervised DNN adaptation and RNN language model rescoring in the back-end. The system achieved impressive 60% and 55% relative WER reductions on the development set, as well as 65% and 60% of the same on the test set, for real and simulated data sets, respectively.

Index Terms— CHiME Challenge, distant speech recognition, correlative beamforming, semi-supervised speaker adaptation

1. INTRODUCTION

Distant speech recognition in realistic urban environments is a new and very challenging task due to high level of noise and interferences, echoes and reverberations, as well as attenuations and distortions. The 3rd CHiME Challenge [1] aims to test the performance of automatic speech recognition in a real-world, commercially motivated scenario: a person talking to a tablet device that has been fitted with a six-channel microphone array. The CHiME-3 scenario is ASR for a multi-microphone tablet device being used in everyday environments. For the challenge, four varied environments have been selected: cafe (CAF), street junction (STR), public transport (BUS) and pedestrian area (PED). For each environment, two types of noisy speech data have been provided, real and simulated. The real data consists of new 6-channel recordings of sentences from the WSJ0 corpus spoken in noisy environments. The simulated data was constructed by mixing clean utterances from that corpus into background recordings made in the four CHiME-3 environments. For ASR evaluation, the data is divided into official training, development and test sets.

Recently, different multi-channel approaches have been reported in the Reverb challenge [2] which is more focused on simulated room environments. The problem is that all the approaches had been evaluated with different back-end recognizers and hence it is impossible to compare them to make conclusions, particularly under this new urban environment conditions.

The aim of this paper is hence to perform comprehensive evaluations of the multi-channel approaches on the 3rd CHiME Challenge database with the same advanced DNN back-end recognizer in order to recommend the best design for the task. Since the multi-condition training has been approved to be superior to the clean training, we shall focus on that training method only. Finally, based up on the results from evaluations, we develop our best performing system which includes harmonic-to-subharmonic ratio (SHR) voice activity detection, correlative beamforming with adaptive channel selection in the front-end, semi-supervised DNN adaptation and RNN language model rescoring in the back-end.

2. MULTI-CHANNEL PROCESSING METHODS

In this section we highlight the most important multi-channel processing for distant ASR which have been reported in the literature.

2.1. Beamforming

Beamforming is a linear spatial filter which makes use of the known or estimated location of the sound source to utilize a sound directivity beam to enhance the target signal while reducing the background noise. This has been proven to be the most effective among speech enhancement methods for ASR due to its low distortion level while gradually improving the signal to noise ratio (SNR).

2.1.1. Delay-and-sum

The delay-and-sum (DS) method enhances the signal by aligning channels to the sound source before adding them up:

$$y\left(t\right)=\sum_{i=1}^{N}w_{i}x_{i}\left(t-\tau_{i}\right), \tag{1}$$

where the weights w_i are normally set to equal one, τ_i denotes the aligning time of the channels, $y\left(t\right)$ denotes the output beamforming, $x_i\left(t\right)$ -the channel signals. Since the sound source location is unknown, the channels are aligned to a reference channel and hence it is called the time difference of arrival (TDOA) which is estimated by maximizing the Generalized Cross Correlation Phase Transform (GCC-PHAT) denoted as follows:

$$\tau_{i}=\arg\max_{\Delta}\left[F^{-1}\left\{\frac{X_{i}\left(\omega\right)\bar{X}_{ref}\left(\omega\right)}{\left|X_{i}\left(\omega\right)\left[\bar{X}_{ref}\left(\omega\right)\right]\right|}\right\}\right], \tag{2}$$

where $X\left(\omega\right)$ is the complex spectrum of the signal $\mathbf{x}\left(t\right)$, Δ is the time argument of the inverse Fourier transform function which is denoted by F^{-1}. The reference channel is initialized as the channel with the highest estimated SNR from an initial segment on each channel and then iteratively tracking to the lowest negative TDOAs until they turn positive.

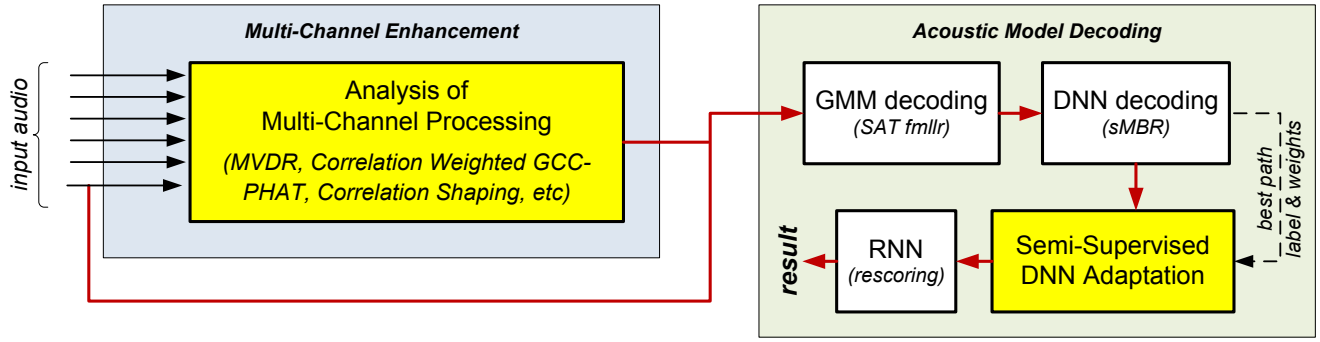

Fig. 1. Overview of our CHiME-3 speech recognition system for noisy reverberant conditions.

2.1.2. *Minimum Variance Distortion Response (MVDR)*

The MVDR-processed waveforms were provided by the organizers [1] as the baseline enhancement technique. Unlike the delay-and-sum, MVDR is a spectrum-based beamforming method designed to minimize the variance of the output subject to a constraint of gain, noted as

$$\mathbf{w} = \underset{\mathbf{w}^T\mathbf{H}_d=1}{\arg\min}\, \mathbf{w}^T\mathbf{Q}\mathbf{w} \tag{3}$$

Assuming an uncorrelated noise field, it yields a tractable form solution:

$$\mathbf{w} = \frac{\mathbf{Q}^{-1}\mathbf{H}_d}{{\mathbf{H}_d}^T\mathbf{Q}^{-1}\mathbf{H}_d}, \tag{4}$$

where $\mathbf{Q}^{-1}$ is the estimated noise covariance matrix [3] and $\mathbf{H}_d$ denotes the direct path of the sound source propagation which can be expressed as

$$\mathbf{H}_d\left(\omega\right) = \left[\alpha_1 e^{-i\omega\tau_1}, ..., \alpha_N e^{-i\omega\tau_N}\right]^T. \tag{5}$$

Assuming omnidirectional microphones, we can set the gains α_i to 1 and the time delay τ_i is replaced by TDOA estimated above. The MVDR is well known to be effective in diffused room noise conditions [2] but may get problems over unexpected directional interferences [4] due to its noise field uncorrelation assumption. The above will be later reaffirmed in our experiments.

2.1.3. *Adaptive correlative beamforming*

Unlike delay-and-sum beamforming, the adaptive correlative beamforming method takes into consideration possible changes in the sound source location and hence, allows updating of weighted coefficients. In this work, we build our system following the idea of updating the weighted coefficients using cross correlation coefficient segment-by-segment [5]. Our processing also includes start point detection, reference channel selection, TDOA alignment, and adaptive channel removal. The start point detection roughly detects when to start the processing. Our system uses the harmonic to sub-harmonic ratio (SHR) [6] [7] as a feature for voiced speech detection. To improve the robustness under mismatched conditions, we normalized the SHR by the sub-harmonic energy [8]

The reference channel selection is done by comparing the summary cross-correlation peaks between each channel and the rest of channels within a one second window:

$$\overline{xcorr_i} = \sum_{j=1,j\neq i}^{N} xcorr[i,j] \tag{6}$$

where $N = 6$ is the number of channels, and $xcorr[i,j]$ is the standard cross- correlation coefficients between channels i and j over the speech region of the utterance. The channel i with the highest average cross-correlation is chosen as the reference channel.

After alignment, the final step before output signal generation is an adaptive weighting and channel elimination strategy based on the standard cross-correlation $xcorr[i,j]$ of the aligned channels.

To reduce the effects of unexpected interferences, a smoothing is applied on the adaptive weight calculation: starting from an even weighting in the first analysis window, $w_m[c=1] = \frac{1}{M}$, the weights are updated continuously as follows:

$$w_i[t] = (1-\alpha)w_i[t-1] + \alpha\frac{\overline{xcorr}_i[t]}{\sum_{m=1}^{N}\overline{xcorr}_i[t]} \tag{7}$$

where $\alpha = 0.05$ and $\overline{xcorr}_m[t]$ is now calculated from the TDOA aligned frame t.

The adaptive channel elimination strategy is applied to remove noise and distortions affecting a small number of microphones, and also acts to remove microphones that have failed completely. As such, the value of $\overline{xcorr}_m[t]$ is used as a measure of the channel quality, and a threshold is applied to remove poor frames:

$$\overline{xcorr}_m[t] < \frac{1}{M}\sum_{i=1}^{N}(\overline{xcorr}_i[t]) - \beta \tag{8}$$

where segments matching this criteria have weights $w_m[t] = 0$ with the subsequent weights normalized to sum to 1. The parameter β controls the strength of the channel rejection, and was empirically set to 0.04 in our experiments. Overlap-and-add summation is then performed with a triangular window to smooth any discontinuities between overlapping frames. A further channel rejection strategy was also used, whereby if any channel repeatedly met the elimination criteria over the utterance, it would be removed completely and the beamforming reinitialized without this channel. Particularly, the channel is removed if quarter of frames were rejected. The rejection method was found to perform well in the case of microphone failure or distortion, or to automatically remove the rear-facing noise channel if the noise level was too high.

2.2. Adaptive noise cancellation

Adaptive noise cancellation is used as the second stage of beamforming for further cancellation of the directional noise components remaining from the initial beamforming output. This can also be understood as cancelling side lopes in the beamforming directivity and

is often named as Generalized Sidelobe Canceller(GSC) [9]. The basic idea of GSC here is to generate the noise references by multiplying the alighted multi-channel signals with a blocking matrix, the simplest being the averaging of subtractions of the reference from the remaining channels. In our implementation, the subtractions are done across the non-rejected channels. The rejected channels are also used without subtraction. Finally, the noise references are given by averaging across the above. An adaptive filter algorithm is used to map the reference noises channels to the one remaining in the output beamforming before subtracting the latter out. The normalized least mean square (NLMS) is then used in the adaptive filter [9].

2.3. De-reverberation

De-reverberation has been shown to be effective for reverberant speech in room conditions [2], but most of proposed methods rely on the reverse modelling of room impulse responses (RIR). In contrast to the REVERB challenge, Chime-3 focuses on urban outdoor environments hence RIR reversal is not suitable. We investigate a general approach for acoustic channel equalization, particularly the correlation shaping (CS) method proposed in [10]. The idea of CS method is to employ an linear adaptive filter to map the autocorrelation function on each channel (after normal alignment) to its , that of clean speech. The filter coefficients are updated using stochastic gradient decent. To reduce the effects of varying vocal tract filters, the adaptive mapping is performed on LPC residuals and some specific area of lags value is introduced. In our experiments, the CS equalization performed better than the inverse RIR approaches [2] as well as other dereverberation methods such as kurtosis equalization and the spatial temporal averaging (SMERSH) [11].

3. ACOUSTIC MODELLING AND DECODING

The following subsections describe the back-end processing in the proposed ASR system. The Kaldi toolkit is used for all experiments [12]. The following components are detailed: (1) feature extraction and GMM-HMM, (2) DNN acoustic modelling, (3) semi-supervised DNN adaptation, and (4) language modelling and rescoring.

3.1. Feature Extraction and Auxiliary GMM-HMM

The feature extraction and auxiliary GMM-HMM system used is trained using the CHiME-3 baseline script[1] without modifications. The GMM-HMM is required in an auxiliary role to provide speaker adaptive transforms (SAT) and the initial alignments for training the subsequent DNN system by forced alignment, which inherits the same tied-state structure. The SAT approach uses feature-space maximum likelihood linear regression (fMLLR) transforms, with speech segments extracted from each conversation assumed to come from the same speaker. For training, the fMLLR transforms are computed from forced-alignments, while for testing, the fMLLR transforms are computed from lattices by using 2 passes of decoding.

3.2. DNN Acoustic Modelling

The DNN acoustic model is trained with fMLLR (SAT) features from the GMM-HMM system that are spliced ± 5 frames and rescaled to have zero mean and unit variance. The DNN has 5 hidden layers, where each hidden layer has 2k sigmoid neurons, and has 1995 dimensions in the softmax output layer, taken from the GMM-HMM model. The hidden layer weights are initialised using layer-wise restricted Boltzmann machine (RBM) pretraining, and after this fine-tuning is performed to minimize the per-frame cross-entropy between the labels and network output. Finally, the DNN is re-trained by sequence-discriminative training to optimise the state minimum Bayes risk (sMBR) objective. Two rounds with four iterations each are performed, with realignment carried out in-between, always with a fixed learning rate of 1e-5.

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/software.html

3.3. Semi-supervised DNN adaptation

While the multi-conditional training helps to reduce mismatch between training and set, a semi-supervised DNN adaptation technique is utilised, on each individual speaker separately, to further reduce the mismatch between training and testing conditions [13, 14]. Additional iterations of fine-tuning of the DNN requires a frame-level label, and potentially also a confidence measure, and these are generated based on the initial output of the system after RNN rescoring, as shown in Figure 1. The frame-level confidence c_{frame_i} is extracted from the lattice posteriors $\gamma(i, s)$, which express the probability of being in state s at time i. The decoding output gives us the best path state sequence, $s_{i,1best}$, and the confidence values are the posteriors under this sequence, as follows [14]:

$$c_{frame_i} = \gamma\left(i, s_{i,1best}\right) \tag{9}$$

The best path state sequence and confidence measures are then used as the target labels and weightings respectively for additional iterations of DNN fine-tuning. In our experiments, all weights in the network are updated separately for each speaker. A learning rate of 0.0008, which was optimized through development data, is used, with halving performed each iteration.

3.4. Language Model and Rescoring

We utilise the language model provided by the hallenge for lattice generation during the decoding. However, we additional train an RNN language model using the "RNNLM-0.3e" package [15], with 20k words, 300 hidden units, 300 classes, and 2000m direct connections. The RNN language model can significantly reduce the perplexity, hence is used to rescore the output decoding lattice, with interpolation weight 0.3 against uses the basic LM.

4. CHALLENGE RESULTS & DISCUSSIONS

All the discussed methods were evaluated on the development set and the selected ones are chosen for the test evaluation. The enhancement methods are only applied to the test data as we found that applying enhancement to the training data degrades performances of the ASR systems with multi-condition training. The results presented in Table 1 all use single channel speeches from the official data in training. The baseline DNN-SMBR system uses Mel-filterbank features without speaker adapive training (SAT).

4.1. Comparison of enhancement methods

It is clear that all of the enhancement methods improve upon the baseline provided by **Noisy**. The only exception is **MVDR** which shows a significant loss in performance in the real test set. However, its simulated test performance is the best among all the methods compared. This large discrepancy can be explained by that the noise in the simulated signals are simply added to the reverberant channels hence created diffused noise field which is suitable

Acoustic Model	Enhancement	Development Set		Test Set	
		Real	Sim	Real	Sim
Baseline DNN-SMBR	Noisy	16.56	14.42	32.99	21.67
	MVDR	22.77	8.36	53.38	11.36
	CorrBF	12.29	11.43	23.58	18.11
DNN-SAT-SMBR	Noisy	12.11	10.94	22.13	13.76
	MVDR	13.54	5.48	23.35	6.39
	DS	8.79	8.64	16.45	14.31
	CorrBF	8.45	8.24	15.66	11.79
	CorrBF-GSC	9.20	8.52	-	-
	CS	9.83	10.52	-	-
DNN-SAT-SMBR + SSSA	Noisy	10.55	9.40	19.74	10.66
	MVDR	12.15	4.74	20.34	5.17
	DS	7.37	7.33	14.20	10.97
	CorrBF	7.10	7.01	13.34	9.23
	CorrBF-GSC	7.68	7.11	-	-
	CS	8.21	8.43	-	-
DNN-SAT-SMBR + SSSA+RNN	Noisy	9.56	8.72	18.12	9.67
	MVDR	11.48	**4.28**	18.55	**4.58**
	DS	6.65	6.79	13.01	10.03
	CorrBF	**6.52**	6.38	**11.46**	8.55
	CorrBF-GSC	7.00	6.37	-	-
	CS	7.54	7.68	-	-

Table 1. Challenge results comparing the WER performance of each step in the proposed system. Note: SSSA = Semi Supervised Speaker Adaptation, RNN = Recurrent Neural Network Language Model Rescoring. **Noisy** refers the original noisy utterances provided in the challenge; **MVDR** refers to the baseline enhanced data, using MVDR beamforming, provided in the challenge; **DS** is the simple fixed beamforming technique described in "Delay-and-sum"; **CorrBF** is the adaptive correlative beamforming; **CorrBF-GSC** applies adaptive noise cancellation (GSC) on top of the output of CorrBF; **CS** refers to channel equalization using correlation shaping.

for the MVDR beamforming methods. In contrast, the noise in urban environments in 3rd CHiME Challenge is the multi-source type which contains many directional interference components generating spikes in the covariance matrix estimation and hence may not be suitable for MVDR beamforming methods.

The **DS** is among the simplest forms of beamforming possible yet it shows a visible improvement over the baseline. It is explained by the low distortion level in the output DS beamforming. The **CorrBF** improves on this even further and is eventually found to be the best enhancement method with consistent around 60% relative improvements on WER for all the datasets. A simple explanation is that variations in the speakers' head positions and the recording tablet is better captured by the adaptive weights compared to fixed beamforming. It seems that applying adaptive noise cancellation (GSC) did not improve the ASR performances, in most of cases, as expected. The reason could be the distortions introduced in the output signals which ASR performance is especially sensitive to. The best performance on real test data is at **11.46%** WER and further improved to **10.01%** WER using multi-channel data in training.

The **CS** method, which produced cleanest output speech based on our own listening, is still noticeably worse than the **CorrBF** method. This may be due to the fact that, the reverberation effects in CHiME-3 challenge is not significant compared to the noise effects due to the short distance between speaker head to the tablet microphones. The inverse RIR dereverberations which were successfully applied in the Reverb challenge were more than 5% WER worse than the **CorrBF** method and hence not reported in details here.

Condition	Development Set		Test Set	
	Real	Simulated	Real	Simulated
BUS	7.45	5.55	17.89	6.43
CAF	6.61	8.11	9.43	8.54
PED	4.94	4.97	9.19	8.24
STR	7.08	6.87	9.34	11.00

Table 2. Challenge results from our best performing DNN + SSSA + RNN system. The WER performance of each condition is shown for comparison with other participants.

Processing Step	Dev WER Imp.	Test WER Imp.
Corr. Beamforming	3%	6%
Semi-supervised DNN	1%	2%
RNN Rescoring	0.6%	1%

Table 3. Comparison of the approximate WER improvements given by the key components of the system.

4.2. Comparison of backend system

For the back-end, the semi-supervised DNN adaptations have found to be very useful and it can be explained by the fact that the non-stationary noise conditions may introduce mismatch between training and testing. Furthermore,the speaker information can also be adapted through this semi-supervised adaptation. The RNN language rescoring which performs some kind of adaptation in language model (LM)is also helpful for the task with the unknown domain of speaking. These two components, together with the conventional speaker adaptive transform and discriminative learning are essential to deliver good performances of noisy reverberant ASR tasks.

It is worth noting that applying enhancement on the training data, did not help to improve the ASR performances. In our experiments, it consistently yields 5-7% degradation on WER. It seems that the multi-condition training is sensitive to distortions hence the enhanced audio could not help to improve.

4.3. Analysis of Word Error Rate Improvements

The results for our best performing system in terms of WER are summarized in Table 2. A summary of the contribution of each processing step to the final WER result is shown in Table 3. It can be seen that our correlation-weighted beamforming algorithm gives the most significant and consistent improvement in performance, which highlights the performance improvement gained when multiple microphones are available. In addition, the semi-supervised DNN adaptations and RNN language rescoring also helps to further reduce any mismatch between training and testing conditions.

5. CONCLUSION

This paper performs comprehensive comparisons of multi-channel approaches for distant speech recognition under urban environments using multi-condition training which was previously proven to be the most effective training method. The simple adaptive correlative beamforming is found to be the most useful in front-end due to its low level of distortion and ability to adapt the change of the head to microphone positions. At the back-end, semi-supervised DNN adaptation and RNN language rescoring are helpful to reduce the mismatch between training and testing.

6. REFERENCES

[1] Emmanuel Vincent Jon Barker, Ricard Marxer and Shinji Wanatabe, "The third 'chime' speech separation and recognition challemge: Dataset, task and baselines," in *ASRU 2015*. December 2015, IEEE.

[2] Keizo Kinoshita, Marc Delcroix, Takashi Yoshioka, Takeshi Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas, "The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[3] Xavier Mestre and Miguel A. Lagunas, "On diagonal loading for minimum variance beamformers," in *ISSPIT*. 2003, pp. 459–462, IEEE.

[4] J. Wang C.Y. Sun Liu, F and R.Du, "Robust mdvr beamformer for nulling level control via multi-parametric quadratic programming," in *Progress in Electromagnetic Research C*, 2011, vol. 20, pp. 239–254.

[5] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.

[6] Xuejing Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–333.

[7] Thomas Drugman and Abeer Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics.," in *Interspeech*, 2011, pp. 1973–1976.

[8] Jonathan W.D. and H.D. Tran, "Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: i2r's system description for the aspire challenge," in *ASRU 2015*. December 2015, IEEE.

[9] Osamu Hoshuyama and Akihiko Sugiyama, "Robust adaptive beamforming," in *Microphone Arrays*, Michael Brandstein and Darren Ward, Eds., Digital Signal Processing, pp. 87–109. Springer Berlin Heidelberg, 2001.

[10] Bradford W. Gillespie and L.E. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, April 2003, vol. 1, pp. I–676–I–679 vol.1.

[11] Patrick A. Naylor and Nikolay D. Gaubitch, *Speech Dereverberation*, Springer Publishing Company, Incorporated, 1st edition, 2010.

[12] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding (ASRU)*. 2011, IEEE.

[13] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7947–7951.

[14] Karel Vesely, Mirko Hannemann, and Lukas Burget, "Semi-supervised training of deep neural networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 267–272.

[15] Tomáš Mikolov, "Statistical language models based on neural networks," 2012.