Friedrich-Alexander University Erlangen-Nuremberg

Multimedia Communications and Signal Processing

Prof. Dr.-Ing. Walter Kellermann

Research internship

Spatial diffuseness features for robust speech recognition in everyday environments

Markus Bachmann

November 2016

Supervisors: Christian Hümmer

Hendrik Barfuss

Contents

List of Abbreviations

ASR	Automatic Speech Recognition
CDR	Coherent-to-Diffuse power Ratio
DDR	Direct-to-Diffuse Ratio
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DOA	Direction Of Arrival
DNN	Deep Neural Network
fMLLR	feature-space Maximum Likelihood Linear Regression
GCC-PHAT	Generalized Cross-Correlation with PHAse Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LDA	Linear Discriminant Analysis
MFCC	Mel Frequency Cepstral Coefficient
MLLT	Maximum Likelihood Linear Transform
RBM	Restricted Boltzmann Machine

SAT	Speaker Adaptive Training
STFT	Short Time Fourier Transform
sMBR	state-level Minimum Bayes Risk
TDOA	Time Difference Of Arrival
WER	Word Error Rate

Chapter 1

Introduction

Many devices of our everyday life like (toy) robots, smartphones, and TVs almost naturally employ Automatic Speech Recognition (ASR) for human machine interaction. However, robust ASR is still a challenge in everyday environments, especially when noise and reverberation are present in the recorded signals [DKNN13, YG15, HCE+15]. In order to compensate for the resulting reduction of the recognition accuracy, a variety of methods have been developed for front-end processing, e. g., spatial filtering methods [WWLR+14, HCE+15] or speech enhancement schemes [MMN+02, YG15]. The current trend in ASR removes the resulting separation between the recognizer and the front-end by replacing the explicit feature processing by implicit learning.

Following this trend, we incorporate spatial information about the diffuseness of the sound field into a Deep Neural Network (DNN)-based acoustic model by means of the so-called meldiffuseness features [SHMK15]. We calculate the diffuseness features from Coherent-to-Diffuse power Ratio (CDR) estimates that recently showed to be efficient for dereverberation and noise suppression [SK15]. As starting point for our work, we choose the ASR baseline of the CHiME-4 challenge, that targets state-of-the-art speech recognition in real-world environments.

This report proceeds with the presentation of the signal model followed by an introduction to meldiffuseness estimation and the CHiME-4 ASR baseline in Chap. 2. In Chap. 3, we give a brief overview over the CHiME-4 dataset. The performance of the proposed approach is evaluated in Chap. 4. Chap. 5 concludes this report.

Chapter 2

Automatic Speech Recognition System

Before introducing the meldiffuseness features and the CHiME-4 ASR baseline, we have a look at the signal model employed throughout this report.

2.1 Meldiffuseness Features

In every recording, the environment has an impact on the recorded version of the original utterance. If we record a single speaker with a single microphone, we often model these modifications to the original signal $u(t)$ by a single impulse response $h(t)$ of a linear system. Typically, the environment also adds some noise $n(t)$ to the signal including sensor, interference, and background noise as well as late reverberation. If we use multiple microphones for recording, every sensor indexed by m captures a different noise signal $n_m(t)$ and a different modification $h_m(t)$ of the original signal $u(t)$. So, the recorded signal at the microphone indexed by m is given by

$$x_m(t) = h_m(t) * u(t) + n_m(t) \tag{2.1}$$

in the continuous and

$$x_m[k] = h_m[k] * u[k] + n_m[k] \tag{2.2}$$

in the discrete time-domain, where $*$ denotes the convolution operator.

Assuming free-field propagation of sound waves, $h_m(t)$ corresponds to a shifted Dirac impulse, so that the desired signal

$$s_m(t) = h_m(t) * u(t) = u(t - t_{ref} - \tau_m) \tag{2.3}$$

is a delayed version of the original $u(t)$. These delays comprise the time t_{ref} that the sound waves take to travel from the speaker to the reference sensor of the microphone array and the Time Differences Of Arrival (TDOAs) (τ_m in the continuous and κ_m in the discrete time-domain), which are dependent on the position of the source relative to the array, between the microphones and the reference sensor. In the far-field, we can assume that plane waves arrive at the array so the TDOAs are only dependent on the Direction Of Arrival (DOA) of the sound waves.

In the following, we denote the corresponding Short Time Fourier Transform (STFT) domain signals by their uppercase letters $U(l,f)$, $S_m(l,f)$, $X_m(l,f)$, and $N_m(l,f)$, where l is the discrete time frame index and f the continuous frequency. In the STFT domain, the auto- and cross-power spectral densities are defined as

$$\Phi_{x,y}(l,f) = \mathcal{E}\left\{X(l,f)\, Y^*(l,f)\right\}, \tag{2.4}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation and $(\cdot)^*$ the complex conjugate of $(\cdot)$, respectively. In [DGTT+12], the authors define the diffuseness $D(l,f)$ as follows

$$D(l,f) = \frac{1}{DDR(l,f)+1}, \tag{2.5}$$

where $DDR(l,f)$ denotes the Direct-to-Diffuse Ratio (DDR). The DDR is equivalent to the CDR if we neglect early reflections and assume the direct sound and the perfectly diffuse reverberation to be mutually uncorrelated [SK15]. In this case, the diffuseness $D(l,f)$ can also be obtained by

$$D(l,f) = \frac{1}{CDR(l,f)+1}. \tag{2.6}$$

The CDR itself is the ratio between the direct-path signal and the diffuse noise. In [SK15], the authors derive the relationship

$$CDR\left(l,f\right)=\frac{\Gamma_{n}\left(l,f\right)-\Gamma_{x}\left(l,f\right)}{\Gamma_{x}\left(l,f\right)-\Gamma_{s}\left(l,f\right)} \tag{2.7}$$

between two omnidirectional microphones, where $\Gamma_{x}\left(l,f\right)$, $\Gamma_{s}\left(l,f\right)$, and $\Gamma_{n}\left(l,f\right)$ denote the coherence function of the observations, the desired direct-path signals, and the noise, respectively.

Assuming the direct-path signals to arrive as plane waves, their time-variant coherence function between two microphones indexed by p and q is given by

$$\Gamma_{s}\left(l,f\right)=\frac{\Phi_{s_{p},s_{q}}\left(l,f\right)}{\sqrt{\Phi_{s_{p},s_{p}}\left(l,f\right)\cdot\Phi_{s_{q},s_{q}}\left(l,f\right)}}=e^{j2\pi f\left(\tau_{q}\left(l\right)-\tau_{p}\left(l\right)\right)}. \tag{2.8}$$

For the coherence function of the diffuse noise, we consider a spherically isotropic sound field generated by the superposition of sound waves produced by an infinite number of uncorrelated noise sources, which are located on a sphere centered around the receiving sensors. In [CS62, SK15], the authors give the corresponding coherence function between two omnidirectional microphones by

$$\Gamma_{n}\left(l,f\right)=\frac{\sin\left(\frac{2\pi fd}{c}\right)}{\frac{2\pi fd}{c}}, \tag{2.9}$$

where d denotes the distance between the two sensor positions and c is the speed of sound.

We compute the short-time coherence function estimate of the observations

$$\widehat{\Gamma}_{x}\left(l,f\right)=\frac{\widehat{\Phi}_{x_{p},x_{q}}\left(l,f\right)}{\sqrt{\widehat{\Phi}_{x_{p},x_{p}}\left(l,f\right)\widehat{\Phi}_{x_{q},x_{q}}\left(l,f\right)}} \tag{2.10}$$

between two microphones indexed by p and q using the short-time estimate of the power spectral densities obtained by recursive averaging according to

$$\widehat{\Phi}_{x_{p},x_{q}}\left(l,f\right)=\lambda\widehat{\Phi}_{x_{p},x_{q}}\left(l,f\right)+\left(1-\lambda\right)X_{p}\left(l,f\right)X_{q}^{*}\left(l,f\right) \tag{2.11}$$

with forgetting factor λ between 0 and 1.

Inserting Eqs. (2.8), (2.9), and (2.10) into Eq. (2.7) leads to complex-valued results in general. However, the CDR and the diffuseness on the other hand are real-valued quantities. This contradiction results from the mismatch between the coherence models and the actual acoustic conditions. Thus, the direct application of Eq. (2.7) to estimate the CDR is not feasible. In order to overcome this problem, many CDR estimators have been proposed. In this report, we employ two CDR estimators which have been shown in [SK15] to be effective in dereverberation and noise suppression, namely a DOA-independent CDR estimator given by

$$CDR_{DOA-independent} = \frac{\Gamma_n \mathrm{Re}\left\{\widehat{\Gamma}_x\right\} - \left|\widehat{\Gamma}_x\right|^2 - \sqrt{\Gamma_n^2 \mathrm{Re}\left\{\widehat{\Gamma}_x\right\}^2 - \Gamma_n^2\left|\widehat{\Gamma}_x\right|^2 + \Gamma_n^2 - 2\Gamma_n \mathrm{Re}\left\{\widehat{\Gamma}_x\right\} + \left|\widehat{\Gamma}_x\right|^2}}{\left|\widehat{\Gamma}_x\right|^2 - 1} \tag{2.12}$$

and a DOA-dependent CDR estimator defined as

$$CDR_{DOA-dependent} = \frac{1 - \Gamma_n \cos\left(\arg\left\{\Gamma_s\right\}\right)}{\left|\Gamma_n - \Gamma_s\right|} \cdot \left|\frac{\Gamma_s^*\left(\Gamma_n - \widehat{\Gamma}_x\right)}{\mathrm{Re}\left\{\Gamma_s^* \widehat{\Gamma}_x\right\} - 1}\right|, \tag{2.13}$$

where $\mathrm{Re}\{\cdot\}$ and $|\cdot|$ denote the real-part and the magnitude of $(\cdot)$, respectively. Note that the dependencies on the time frame l and the frequency f have been omitted in Eqs. (2.12) and (2.13) for brevity.

If we employ more than two microphones for recording, we compute the CDR for every pair of sensors according to Eq. (2.12) or (2.13). After calculating the corresponding diffuseness values according to Eq. (2.6), we obtain the overall diffuseness $\overline{D}(l, f)$ by averaging the pairwisely computed diffuseness values. In order to obtain the meldiffuseness, we filter the overall diffuseness value $\overline{D}(l, f)$ with a Mel-scale filter bank.

2.2 CHiME-4 ASR Baseline

As the starting point for our research, we use the CHiME-4 ASR baseline [VWN+], which is based on [HCE+15].

This ASR system utilizes the BeamformIt toolkit [AWH07] by Xavier Anguera for front-end enhancement of the microphone signals: after Wiener filtering each individual channel signal $x_m[k]$ and choosing a reference channel, the tool localizes possible sources by extracting multiple TDOAs from each channel signal using the Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) method. It eliminates unreliable TDOA values by comparing the corresponding GCC-PHAT values with a noise threshold and selects those TDOAs $\kappa_m[c]$ that most likely belong to one speaker by a dual-step Viterbi postprocessing. The weights $w_m[c]$ for each channel m and for each processing block c are computed dependent on the dynamic range of the input signal and the cross-correlation, which is averaged over all blocks and over the total number of microphones M. After this procedure, we obtain weights that eliminate bad quality segments in the input signals. As the weighted-delay&sum equation

$$y[cS + k_S] = \sum_{m=1}^{M} w_m[c]\, x_m[cS + k_S - \kappa_m[c]] \tag{2.14}$$

could lead to discontinuities in the output signal, a triangular window smooths the calculated weights. The beamformed signal is then obtained as

$$\begin{aligned} y[cS + k_S] = \alpha[k_S] \sum_{m=1}^{M} w_m[c]\, x_m[cS + k_S - \kappa_m[c]] \\ + (1 - \alpha[k_S]) \sum_{m=1}^{M} w_m[c-1]\, x_m[cS + k_S - \kappa_m[c-1]], \end{aligned} \tag{2.15}$$

where S denotes the length of a processing block and k_S is the sample within the block currently processed. For a detailed and complete description of the BeamformIt toolkit, we refer to [AWH07].

For the speech recognition task, the CHiME-4 baseline ASR system employs the Kaldi toolkit [PGB+11]. The back-end is a hybrid Hidden Markov Model (HMM)-based system that utilizes Gaussian Mixture Model (GMM)-based acoustic models to extract the initial alignment for the DNN-based acoustic model, which performs the final recognition. We adapt all systems using multi-condition data.

The baseline trains multiple GMM-HMM models using various modifications of the standard 13 dimensional Mel Frequency Cepstral Coefficients (MFCCs). In a first step, a monophone GMM-HMM model is adapted using MFCC + Δ + $\Delta\Delta$ features. After the training of a triphone model with MFCC + Δ + $\Delta\Delta$ features, a ± 3 frame splicing is applied to the MFCCs and a Linear Discriminant Analysis (LDA) reduces the dimensionality of the resulting 91 dimensional vector to 40 dimensions. These features are further processed using a Maximum Likelihood Linear Transform (MLLT) and then employed to adapt a triphone GMM-HMM model. Finally, the baseline performs Speaker Adaptive Training (SAT), i. e., training on feature-space Maximum Likelihood Linear Regression (fMLLR)-adapted features, of a triphone GMM-HMM model on top of the LDA-MLLT features. Every training, with exception of the first one, uses the alignment extracted in the previous step. For the first decoding, the initial alignment is equally spaced.

The DNN-HMM acoustic model based on "Karel's DNN implementation" uses a ± 5 frame splicing of the same MFCC-LDA-MLLT-fMLLR features chosen as input for the lastly trained GMM-HMM acoustic model at its input layer. The DNN consists of 2048 neurons in each of the 7 hidden layers. Each layer is pretrained as Restricted Boltzmann Machine (RBM) by the Contrastive Divergence algorithm. The stack of these RBMs builds the first version of the DNN. Regular cross-validation prevents over-fitting of this DNN during the cross-entropy training. Repeated sequence-discriminative training using the state-level Minimum Bayes Risk (sMBR) criterion optimizes the DNN for whole sentences.

As the final step, the baseline rescores the previous 3-gram KN language model based results using a 5-gram KN language model and a recurrent neural network based language model [VWN$^+$].

Chapter 3

Dataset

We conduct all our experiments on the dataset provided by the CHiME-4 challenge [VWN+].

This dataset contains sample-synchronized 6-channel recordings of read WSJ prompts recorded in 24 bits at 48 kHz by a TASCAM DR-680 multi-track field recorder. The audio recordings were downsampled to 16 bit at 16 kHz for distribution. The microphone configuration is shown in Fig. 3.1. Here, the microphones are numbered according to their channel index in the recordings. All sensors, except microphone 2, face forward and are mounted flush with the front of the 1 cm thick frame. Microphone 2 faces backwards and is mounted flush with the back of the frame. The sensors are audio-technica ATR3350 omnidirectional lavalier microphones. Like commercial devices, some of the sensors failed to record some of the utterances properly. This happens in about 12 % of the data due to hardware issues or coverage of the microphones by the hands or clothes of the speaker.

In addition to the six-channel recordings, the dataset offers a mono and a stereo track, where the microphones are chosen in such a way that none of the above mentioned issues appear.

The dataset can be split into real and simulated recordings. The real data is recorded in a bus, a cafe, a pedestrian area, and at street junctions. Additionally, recordings are also done in a booth. The latter are then mixed with noise recorded at the already

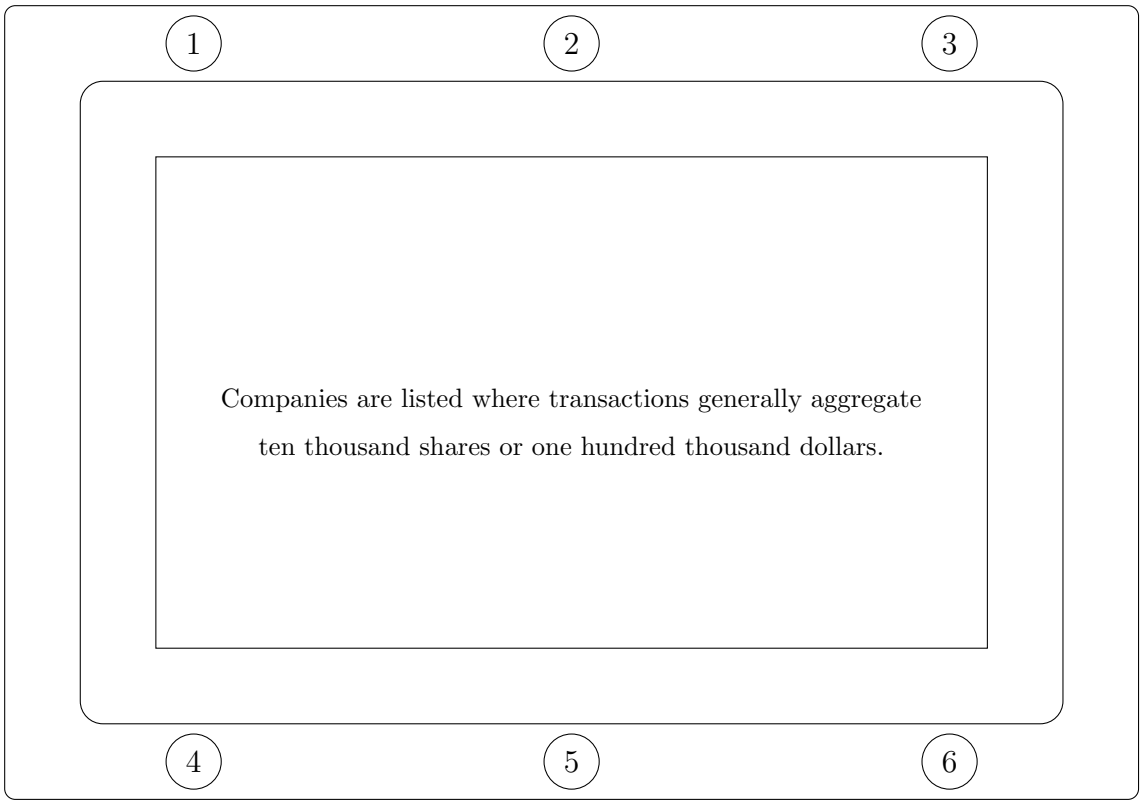

Figure 3.1: Microphone configuration for the CHiME-4 recordings. Microphones 1 and 3 – 6 face forward and are mounted flush with the front of the frame. Microphone 2 faces backwards and is mounted flush with the back of the 1 cm thick frame [VWBM].

mentioned locations to create the simulated recordings.

The dataset comprises three subsets, namely a training, a development and an evaluation set. The first one contains 1600 real utterances from four speakers and 7138 simulated ones obtained by mixing the WSJ0 SI-84 training set from 83 speakers with the above mentioned noise recordings. The development set includes 410 real and 410 simulated utterances in each of the four noisy environments from four other speakers. The evaluation set again comprises 330 real and 330 simulated utterances in each of the four noisy environments from four other speakers.

As mentioned above, some of the microphones sometimes failed to record the utterances properly. Assuming that these issues only affect the minority of the channels, we can distinguish clean signals from the distorted ones by means of their similarity to each other. We measure the similarity in terms of the cross-correlation coefficient

$$\rho_{p,q} = \frac{\mathrm{Cov}\left(x_p\left[k\right], x_q\left[k\right]\right)}{\sqrt{\mathrm{Var}\left(x_p\left[k\right]\right)\mathrm{Var}\left(x_q\left[k\right]\right)}} = \rho_{q,p} \tag{3.1}$$

for each channel pair indexed by p and q. In order to obtain a rating how distorted a signal is, we average the cross-correlation coefficients for each channel over all other available channels resulting in the averaged cross-correlation coefficient

$$\bar{\rho}_p = \frac{1}{M-1}\sum_{\substack{q=1 \\ q\neq p}}^{M} \rho_{p,q} \tag{3.2}$$

for the channel indexed by p, where M denotes the number of available channels. This estimator grants lower values to more distorted signals. Thus, we can detect distorted signals by comparing the corresponding averaged cross-correlation coefficient with a threshold.

Chapter 4

Experiments

In [BHSK], the authors propose appending meldiffuseness features to the input of the DNN-based acoustic model like already done in [SHMK15]. In the latter, the authors employ logmelspec + Δ + $\Delta\Delta$ features as input to the DNN in their base evaluation instead of MFCC-LDA-MLLT-fMLLR features. Thus, we evaluate the performance of the CHiME-4 baseline with both feature combinations in terms of Word Error Rate (WER) scores.

Like in [SHMK15], we compute the logmelspec features without dithering, preemphasis, and removal of the DC offset from the time-domain signals sampled with 16 kHz using a 25 ms Hann window with a frame shift of 10 ms. A 512-point Discrete Fourier Transform (DFT) transforms the blocks into the STFT domain resulting in 257 subbands. The subsequent 24-bin Mel-scale filter bank covers a frequency range from 64 to 8000 Hz. In contrast to this, the CHiME-4 baseline computes the MFCCs with dithering, preemphasis, and removal of the DC offset using a 25 ms Povey window with the same frame shift as above. A Mel-scale filter bank then reduces the 257 STFT subbands resulting from the same 512-point DFT as above to 23 Mel-scale frequency bins covering a range from 20 to 8000 Hz. After a subsequent Discrete Cosine Transform (DCT), the baseline shortens the feature vector down to 13 dimensions. We train and decode both feature combinations with the noisy fifth channel signal of the CHiME-4 challenge training and development set without using the BeamformIt toolkit.

		Development Set		
Preprocessing	Feature	Real	Simulated	Average
none	logmelspec + Δ + ΔΔ	12.21 %	10.59 %	11.40 %
none	**MFCC-LDA-MLLT-fMLLR**	**9.65 %**	**8.77 %**	**9.21 %**

Table 4.1: ASR WER for the noisy 5$^{\text{th}}$ channel recordings of the CHiME-4 challenge development set

From the results in Tab. 4.1, we can conclude that the baseline MFCC-LDA-MLLT-fMLLR features perform better by about 2.19 % on average. Thus, we employ the baseline features in our further experiments instead of the logmelspec+Δ+ΔΔ features utilized in [SHMK15].

In the next step, we evaluate the impact of the proposed meldiffuseness features onto the performance of the CHiME-4 ASR baseline. Therefore, we remove the preprocessing using the beamformer of the BeamformIt toolkit and instead append meldiffuseness features to the input of the DNN in the ASR system resulting in MFCC-LDA-MLLT-fMLLR + meldiffuseness features. We adopt the parameter values from the previous evaluation of the MFCC-LDA-MLLT-fMLLR features. The forgetting factor λ in Eq. (2.11) is set to 0.68 as proposed in [SHMK15]. We calculate the meldiffuseness features from the DOA-independent and the DOA-dependent CDR estimates denoted as $\text{meldiffuseness}_{\text{DOA-independent}}$ and $\text{meldiffuseness}_{\text{DOA-dependent}}$, respectively. We extract the TDOAs required by Eq. (2.8) from the output of the BeamformIt toolkit.

In the training phase, we calculate the MFCCs from the noisy fifth channel of the CHiME-4 challenge training set. In order to compute the meldiffuseness features, we randomly select another channel per utterance from the training set excluding the second channel. In the decoding phase, we estimate the MFCCs from one channel and the meldiffuseness features from both channels of the CHiME-4 challenge stereo development set if we evaluate the system without the beamformer. If we employ the BeamformIt toolkit, we use both channels of the stereo development set as input to the beamformer and compute the MFCCs from its output.

We list the results in Tab. 4.2. In our evaluation, the combination of the BeamformIt

Preprocessing	Feature	Development Set		
		Real	Simulated	Average
none	MFCC-LDA-MLLT-fMLLR	10.65 %	13.76 %	12.21 %
BeamformIt	**MFCC-LDA-MLLT-fMLLR**	**8.27 %**	**9.33 %**	**8.80 %**
none	MFCC-LDA-MLLT-fMLLR + meldiffuseness$_{\text{DOA-independent}}$	10.31 %	12.83 %	11.57 %
none	MFCC-LDA-MLLT-fMLLR + meldiffuseness$_{\text{DOA-dependent}}$	9.57 %	12.04 %	10.80 %

Table 4.2: ASR WER for the two-channel track of the CHiME-4 challenge development set

toolkit and the MFCC-LDA-MLLT-fMLLR features yield the best results with 8.80 % WER reducing the average WER by about 3.41 % compared to the baseline evaluation without beamformer. On the other hand, the appended meldiffuseness features at the input of the DNN also lead to a decrease of the average WER by about 0.64 % for the DOA-independent and about 1.41 % for the DOA-dependent estimate.

Comparing the averaged WERs in the second row of Tab. 4.1 and in the first row of Tab. 4.2, we can recognize that changing the channel for the computation of the MFCC-LDA-MLLT-fMLLR features from the fifth channel to one channel of the stereo development set drops the average WER by about 3 %. Thus, we repeat the previous evaluations with a different dataset. In the cases where we append the meldiffuseness features to the input of the DNN, we compute the MFCC-LDA-MLLT-fMLLR features from the noisy fifth channel of the training and development dataset. For the computation of the meldiffuseness features, we use another random channel excluding the second channel from the training set and the channel with the highest averaged cross-correlation coefficient from the development set in addition to the fifth channel of the corresponding datasets. We select the additional channels on a per utterance basis. If we employ the beamformer, we calculate the MFCC-LDA-MLLT-fMLLR features from the noisy fifth channel of the training set in the training phase. In the decoding phase, we use the same channels as input to the beamformer as for the meldiffuseness feature computations.

Tab. 4.3 lists the results. Again, the MFCC-LDA-MLLT-fMLLR features computed

		Development Set		
Preprocessing	Feature	Real	Simulated	Average
none	MFCC-LDA-MLLT-fMLLR	9.65 %	8.77 %	9.21 %
BeamformIt	**MFCC-LDA-MLLT-fMLLR**	**8.17 %**	**8.42 %**	**8.29 %**
none	MFCC-LDA-MLLT-fMLLR + meldiffuseness$_{\text{DOA-independent}}$	9.27 %	8.77 %	9.02 %
none	MFCC-LDA-MLLT-fMLLR + meldiffuseness$_{\text{DOA-dependent}}$	9.05 %	8.23 %	8.64 %

Table 4.3: ASR WER for our self-defined two-channel track of the CHiME-4 challenge development set

from the beamformed signals yield the lowest average WER of 8.29 % reducing the averaged WER by about 0.92 %. Appending meldiffuseness features to the input of the DNN here also leads to a decreased WER by about 0.19 % for the DOA-independent and about 0.57 % for the DOA-dependent estimate of the meldiffuseness features. The improvements are not as big as in the previous evaluation, but the corresponding WERs are lower. This suggests that the ASR system exhibits a better performance on our dataset in general.

Chapter 5

Conclusion

In this report, we replaced the preprocessing step in the CHiME-4 challenge ASR baseline using the BeamformIt toolkit by appending spatial diffuseness features to the input of the HMM-DNN acoustic model. In order to calculate the diffuseness features, we employed the DOA-independent and DOA-dependent CDR estimates proposed in [SK15]. We evaluated this system on the official stereo development set of the CHiME-4 challenge and on a two-channel dataset which was setup by ourselves and based on the official six-channel development set. In both cases, the appended diffuseness features improved the WERs slightly but were outperformed by the preprocessing step using the beamformer.

However, the evaluations on the two datasets showed that the performance of the whole system is still very sensitive to the performance of the baseline features. Thus, it might be beneficial to combine the beamforming preprocessor with diffuseness features appended to the acoustic model input although both methods exploit spatial information.

References

[AWH07] Xavier Anguera, Chuck Wooters, and Javier Hernando. Acoustic Beamforming for Speaker Diarization of Meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, Sept 2007.

[BHSK] Hendrik Barfuss, Christian Hümmer, Andreas Schwarz, and Walter Kellermann. Robust coherence-based spectral enhancement for speech recognition in adverse real-world environments. To appear, preprint available: `https://arxiv.org/pdf/1604.03393v2.pdf`.

[CS62] Benjamin Cron and Charles Sherman. Spatial-Correlation Functions for Various Noise Models. *The Journal of the Acoustical Society of America*, 34(11):1732–1736, 1962.

[DGTT+12] Giovanni Del Galdo, Maja Taseska, Oliver Thiergart, Jukka Ahonen, and Ville Pulkki. The diffuse sound field in energetic analysis. *The Journal of the Acoustical Society of America*, 131(3):2141–2151, March 2012.

[DKNN13] Marc Delcroix, Yotaro Kubo, Tomohiro Nakatani, and Atsushi Nakamura. Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling? In *INTERSPEECH*, pages 2992–2996, 2013.

[HCE+15] Takaaki Hori, Zhuo Chen, Hakan Erdogan, John R. Hershey, Jonathan Le Roux, Vikramjit Mitra, and Shinji Watanabe. The MERL/SRI system for the 3RD CHiME challenge using beamforming, robust feature

extraction, and advanced speech recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 475–481, Dec 2015.

[MMN+02] Duncan Macho, Laurent Mauuary, Bernhard Noé, Yan Ming Cheng, Douglas Ealey, Denis Jouvet, Holly Kelleher, David Pearce, and Fabien Saadoun. Evaluation of a noise-robust DSR front-end on Aurora databases. In *INTERSPEECH*, 2002.

[PGB+11] 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 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec 2011. IEEE Catalog No.: CFP11SRW-USB.

[SHMK15] Andreas Schwarz, Christian Hümmer, Roland Maas, and Walter Kellermann. Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4380–4384, April 2015.

[SK15] Andreas Schwarz and Walter Kellermann. Coherent-to-Diffuse Power Ratio Estimation for Dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6):1006–1018, June 2015.

[VWBM] Emmanuel Vincent, Shinji Watanabe, Jon Barker, and Ricard Marxer. The 4th CHiME Speech Separation and Recognition Challenge. `http://spandh.dcs.shef.ac.uk/chime_challenge/index.html`. Retrieved September 22, 2016.

[VWN+] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data

simulation mismatches in robust speech recognition. *Computer Speech & Language*. To appear.

[WWLR+14] Felix Weninger, Shinji Watanabe, Jonathan Le Roux, John R. Hershey, Yuuki Tachioka, Jürgen Geiger, Björn Schuller, and Gerhard Rigoll. The MERL/MELCO/TUM System for the REVERB Challenge Using Deep Recurrent Neural Network Feature Enhancement. In *Proc. REVERB Workshop*, 2014.

[YG15] Takuya Yoshioka and Mark JF Gales. Environmentally robust ASR front-end for deep neural network acoustic models. *Computer Speech & Language*, 31(1):65–86, 2015.