CHiME Challenge: Approaches to Robustness using Beamforming and Uncertainty-of-Observation Techniques

Dorothea Kolossa[1], *Ramón Fernandez Astudillo*[2], *Alberto Abad*[2], *Steffen Zeiler*[1], *Rahim Saeidi*[3], *Pejman Mowlaee*[1], *João Paulo da Silva Neto*[2], *Rainer Martin*[1]

[1]Institute of Communication Acoustics, Ruhr-Universität Bochum
[2] Spoken Language Laboratory, INESC-ID, Lisbon
[3] School of Computing, University of Eastern Finland

dorothea.kolossa@rub.de,ramon@astudillo.com,alberto.abad@l2f.inesc-id.pt,steffen.zeiler@gmx.de
rahim.saeidi@uef.fi,pejman.mowlaee@rub.de,joao.neto@inesc-id.pt,rainer.martin@rub.de

Abstract

While much progress has been made in designing robust automatic speech recognition (ASR) systems, the combination of high noise levels and reverberant room acoustics still poses a major challenge even to state-of-the-art systems. The following paper describes how robust automatic speech recognition in such difficult environments can be approached by combining beamforming and missing data techniques.

The combination of these two techniques is achieved by first estimating uncertainties of observation in the beamforming stage, either in the time or frequency domain, and subsequently transforming these observations with associated uncertainties to the domain of speech recognition. This strategy allows the use of reverberation-insensitive cepstral features, which can still be decoded robustly with the help of uncertainty information gained from the beamforming front end.

In this paper, we investigate a number of different preprocessing options with the somewhat surprising result that a simple fixed delay-and-sum beamformer and a null-steering beamformer, when combined with uncertainty decoding techniques, resulted in the most robust design among a much wider set of investigated techniques.

Index Terms: robustness, automatic speech recognition, beamforming, uncertainty decoding

1. Introduction

The goal of the CHiME challenge is to measure the progress that has been made in the last decade in distant microphone speech recognition and to establish a benchmark for further work in highly robust ASR [1]. For this purpose, the CHiME corpus covers natural environments by including various simultaneous audio sources in reverberant mixtures. As spatial cues are important for source separation, the corpus was recorded with a binaural microphone setup.

Many state-of-the-art speech separation or enhancement techniques turn out to be inefficient when used alone for the CHiME challenge, because of their inherent assumptions. For instance, many speech enhancement methods rely on noise estimates provided by noise estimation schemes. Such methods often assume that the noise signal shows less rapid changes than the speech, and are therefore limited in performance when the interfering noise signal has highly dynamic characteristics [2].

On the other hand, beamforming methods can cancel out non-stationary but directional interferers by incorporating spatial knowledge. Still, beamformers such as the Generalized Sidelobe Canceller (GSC) [3] have limitations in real life scenarios. For instance, the GSC is sensitive to direction-of-arrival (DOA) mismatch and suffers from signal leakage or low performance under environmental reverberation [4].

From the above discussion, it is plausible that standard speech enhancement or beamforming methods alone are insufficient for the CHiME corpus. In this paper, we investigate different approaches to provide robust speech recognition by combining standard beamforming techniques with uncertainty-of-observation techniques.

Uncertainty-of-observation techniques have proven beneficial in many contexts. They consider the speech features not as precisely known values, but use their time varying estimation error variances [5], or distinguish, in a binary fashion, between reliable and unreliable features [6]. With these approaches, noise, interfering speech and reverberation can all be treated as contributions to speech observation uncertainty, and decoding can then take place under consideration of these uncertainties, e.g. by uncertainty decoding or modified imputation [7].

However, since uncertainty estimation from beamforming is naturally given in the domain where the beamformer operates, i.e. in the time or time-frequency domain, the observation uncertainties need to undergo a transformation in order to serve as reliability information for the recognizer's MFCC features. To this end, we consider the speech features together with their uncertainties as random variables and calculate the impact that feature extraction has on their mean and variance. The mean value of the random variable output by uncertainty propagation can also be considered an MMSE estimator of the features [8], and the covariance output is used for more robust recognition by uncertainty decoding or modified imputation. Optionally, linear discriminant analysis (LDA) is used to reduce the dimensionality of MFCC features while maximizing class separability.

Finally, we employ Recognizer Output Voting Error Reduction (ROVER) [9] to combine the outputs of multiple speech recognition scores into a single one. The fusion enables us to achieve a lower error rate than any of the individual systems.

The organization of the paper is as follows. In the next section, we present the beamforming methods that we have used. In Section 3, the idea of uncertainty propagation as an interface between beamformer and uncertainty-based ASR is discussed. Section 4 discusses the model training and the experimental results on the CHiME corpus are reported in Section 5, and Section 6 concludes the work.

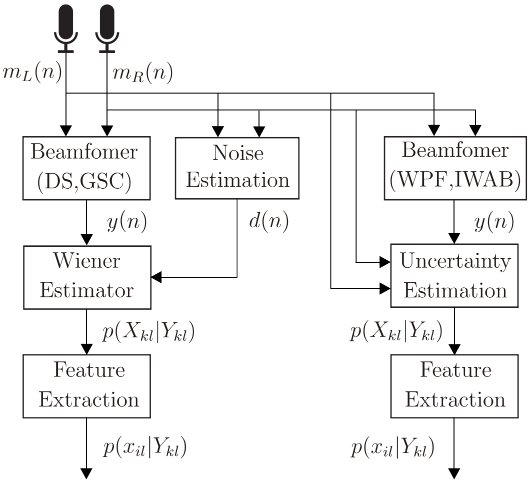

Figure 1: *Block diagram of the proposed approach. An initial step of beamforming is combined with MMSE post-filtering and uncertainty propagation.*

2. Beamforming

Microphone array processing [10] has been broadly used as a pre-processing stage to enhance distant recorded signals that might be used for any speech application, and in particular for speech recognition. Many different proposals exist for microphone array designs but most of them can be summarized into two major trends: fixed and adaptive beamforming. On one hand, fixed beamformers as the delay-and-sum (DS) [11] are quite simple solutions but are ineffective in reducing highly directive noise sources. On the other hand, adaptive beamformers, like the Generalized Sidelobe Canceller (GSC) [3], present a higher capability of interference cancellation but are much more sensitive to steering errors and suffer from signal leakage and degradation. In order to overcome some of the drawbacks of fixed and adaptive beamforming different robust solutions are used. Furthermore, a postprocessing Wiener filtering stage can be applied to the output of beamformers to improve the performance for diffuse noise fields [12]. To solve the problems of the adaptive beamforming, Hoshuyama et al. [13] propose using an adaptive blocking matrix (ABM) where coefficients are constrained to a determinate target error region.

In this work, the use of beamforming techniques was favored against alternative multi-microphone approaches (i.e. blind speech separation) due to the possibility to exploit knowledge of the fixed position of the speaker (broadside of the microphone pair). For this evaluation campaign we have developed and assessed several different beamforming configurations. The best performing beamformer candidates are described below.

2.1. Delay-and-sum beamformer (DS)

The delay-and-sum beamformer [11] aligns the different microphone signals to compensate for the different path lengths from the source to the various microphones. The combination of these aligned signals is

$$y(n) = \alpha_L m_L(n) + \alpha_R m_R(n - \tau_d) \tag{1}$$

where m_L and m_R are the left and right microphone channels, α_L and α_R are the microphone gains and τ_d is the delay that compensates the different propagation delays. In this particular case $\tau_d = 0$ and $\alpha_L = \alpha_R = 1$. The simplicity of the delay-and-sum beamformer is its most important strength, resulting in a convenient and practical choice for many microphone array applications. Thus, delay-and-sum beamforming is widely used despite its frequency dependent response and the weakness in reducing highly directive noise sources.

2.2. Robust Generalized Sidelobe Canceller (GSC)

A Generalized Sidelobe Canceller (GSC) beamformer basically consists of a fixed $y_f(n)$ and an adaptive $y_a(n)$ beamforming path. The adaptive path estimates the non-desired components $\mathbf{m_o}(n)$ through a spatial blocking matrix $\mathbf{B}$ that blocks target signal direction and allows all the other directions. These non-desired components are used for reducing the correlated noise components of the output of the fixed beamformer through a multiple input canceller stage with adaptive filters $\mathbf{w_a}$:

$$y(n) = y_f(n) - y_a(n) = \alpha^T \mathbf{m}(n) - \mathbf{w_a}^T \mathbf{m_o}(n) \tag{2}$$

$$\mathbf{m_o}(n) = \mathbf{B}\mathbf{m}(n) \tag{3}$$

where $\mathbf{m}(n) = [m_L(n), m_R(n)]^T$ is the vector formed by the two-channel inputs and $\alpha = [\alpha_L, \alpha_R]^T$ are the weights of fixed beamformer.

In this work, we have used a robust modification of the GSC structure like the one described in [13] named CCAF-NCAF (coefficient-constrained adaptive filters and norm-constrained adaptive filters) structure. The blocking matrix (BM) is adaptively designed to allow a concrete target-looking error region and to minimize the leakage of the desired signal to the beamformer noise estimate, while the filters of the multiple-input canceler are constrained to help guide their adaptation.

2.3. Wiener post-filtering for microphone arrays (WPF)

The use of an adaptive Wiener post-filter with a beamformer is known to allow effective frequency filtering of the signal by using spatial signal characteristics [12]. The general Wiener gain is formulated in the frequency domain as

$$H(k,l) = \frac{\Phi_X(k,l)}{\Phi_X(k,l) + \Phi_N(k,l)} \tag{4}$$

where k and l are the frequency and time-frame indices respectively and $\Phi_N(k,l)$ and $\Phi_X(k,l)$ account for the power-spectral densities of noise after the beamformer and the desired source respectively.

When multiple inputs are available, the Wiener filter can be computed by combining the cross-power spectral densities and the power spectral density of the different microphones of the array. Assuming that the received signal is an additive mixture of the desired signal and noise, that they are uncorrelated and that noise is uncorrelated also between microphones and have an equal power spectral density, then the gain of the filter can be approximated as

$$H(k,l) \approx \frac{2\max\{\Re\{E\{M_L(k,l)M_R(k,l)^*\}\}, 0\}}{E\{|M_L(k,l)|^2\} + E\{|M_R(k,l)|^2\}} \tag{5}$$

where $M_L(k,l)$ and $M_R(k,l)$ correspond to the STFT of the left and right microphone channels. The expectations are computed by smoothed periodograms and a flooring of the denominator at zero was used to prevent negative Wiener gains. $\Re$ denotes the real value.

It is clear that given the above assumptions the post-filter is particularly convenient in the presence of spatially white noise, however it is also useful in diffuse noise fields which reasonably approximate these conditions.

2.4. Integrated Wiener-filtering with Adaptive Beamformer (IWAB)

In [14], a beamformer is proposed consisting of the combination a robust GSC-like beamformer with Wiener post-filtering. The conventional delay-and-sum of the fixed beamformer path $y_f(n)$ is replaced by the Wiener beamformer, resulting in a filter-and-sum beamformer nested in a GSC-like robust structure with enhanced performance. In this evaluation, we have integrated the robust GSC-like beamformer and the Wiener postfilter described above in this section, where the filter is the one given in Eq. (5).

3. Single Channel Speech Enhancement and Robust Feature Extraction

Microphone array processing techniques are often complemented with a second step of single channel speech enhancement to eliminate residual noises. The efficiency of the such steps can be improved by integrating them with the ASR system through uncertainty propagation techniques. This leads to minimum mean square error (MMSE) estimates directly in the domain of recognition features [8] and provides estimation variances as well. Such variances can be utilized to improve the recognition furthermore by employing observation uncertainty techniques like modified imputation [15].

As described in [8], an MMSE-MFCC estimator can be attained by using the posterior distribution associated with a Wiener filter. Since a Wiener filter can be interpreted as a Bayesian estimator for Gaussian prior and likelihoods, the associated complex Gaussian posterior distribution has the form

$$p(X_{kl}|Y_{kl}) = \mathrm{N}_{\mathbb{C}}(X_{kl}; \hat{X}_{kl}, \lambda_{kl}) \qquad (6)$$

where $\hat{X}_{kl}$ is the estimation of the Wiener filter and λ_{kl} the corresponding estimate variance

$$\lambda_{kl} = \frac{\tilde{\Phi}_X(k,l)\tilde{\Phi}_D(k,l)}{\tilde{\Phi}_X(k,l) + \tilde{\Phi}_D(k,l)}. \qquad (7)$$

Here the parameters $\tilde{\Phi}_X(k,l)$ and $\tilde{\Phi}_D(k,l)$ are used to denote the power spectral densities of speech and residual noise used to derive the Wiener filter. Note that these can be different from the power spectral densities obtained for the WPF in Eq. (4) since they can be determined from other sources.

Two strategies were followed to determine the parameters of the posterior distribution.

3.1. Wiener Filter with Beamforming Based Noise Estimate

The first strategy, displayed in Fig. 1, left, simply applies a single channel Wiener estimator to the outputs of the DS and GSC beamformers and computes the associated posterior. However, rather than providing a noise variance estimate $\tilde{\Phi}_D(k,l)$ by using voice activity detection or minimum statistics, this estimate was obtained from the beamformer information.

Since the speaker is known to be positioned in front of the microphone array, any asymmetry between the microphones can be interpreted as either an interfering signal or the effect of asymmetric reverberation. Therefore, in the case of the DS beamformer, a very simple measure of the residual noise was attained from the subtraction of the two channel inputs as

$$d(n) = m_L(n) - m_R(n), \qquad (8)$$

from which the power spectral density $\tilde{\Phi}_D(k,l)$ was computed. In the case of the GSC a more elaborated estimate was derived from the blocking matrix. In both cases, the speech power spectral density $\tilde{\Phi}_X(k,l)$ was obtained using the well known decision directed method [16].

3.2. Approximate Wiener Post-Filtering Uncertainty

The second strategy, displayed in Fig. 1, was applied to the WPF and IWAB. This did not use any additional enhancement step but rather aimed at deriving a measure of uncertainty for the estimation obtained in the beamforming step.

In principle, since both WPF and IWAB employ Wiener filters, it should be possible to derive the associated posterior from the gain in Eq. (4) and directly determine the parameters of Eq. 6. Nevertheless, due to the particular form in which the gain is computed, the WPF is more aggressive than the conventional Wiener filter. Directly propagating the WBF posterior through the feature extraction resulted in poor results.

The impact of the artifacts induced by the WPF is mitigated when resynthesizing the signal back into a time domain signal $y(n)$. To take advantage of this fact, an equivalent gain of the Wiener filter after resynthesis was computed by comparing the STFT of the input to the WPF with the STFT of the output of the beamformer $y(n)$. The parameters of the posterior were then derived from this gain.

3.3. Robust Feature Extraction

For our setup we employed magnitude based Mel-cepstral coefficients as feature extraction with additional cepstral mean subtraction, delta and acceleration parameters. Linear discriminant analysis (LDA) was also used in some of the setups. Magnitude based cepstra proved to be consistently better than the conventional magnitude squared cepstra in all experiments. To derive the corresponding MMSE-MFCC estimator, we apply the recipes given in [17]. First the propagation of the Wiener posterior through the magnitude transformation is attained as

$$\begin{aligned}\mu_{kl}^{\mathrm{ABS}} &= \Gamma(1.5)\sqrt{\lambda}\exp\left(\frac{\nu}{2}\right) \\ &\cdot \left[(1-\nu)\, I_0\left(-\frac{\nu}{2}\right) - \nu I_1\left(-\frac{\nu}{2}\right)\right]\end{aligned} \qquad (9)$$

where Γ is the gamma function and I_0, I_1 are the modified Bessel functions of order zero and one respectively. The parameter $\nu = |\hat{X}_{kl}|^2/\lambda_{kl}$ is the signal to noise ratio of the associated Rice distribution. The propagation through the filterbank and logarithm can be greatly simplified by assuming the filterbank outputs to be uncorrelated and log-normal distributed, leading to

$$\Sigma_{jjl}^{\mathrm{LOG}} \approx \log\left(\frac{\sum_{k=1}^{K} W_{jk}^2 \left(|\hat{X}_{kl}|^2 + \lambda_{kl}\right)}{\left(\sum_{k=1}^{K} W_{jk}\mu_{kl}^{\mathrm{ABS}}\right)^2} + 1\right) \qquad (10)$$

with W_{jk} as the weights of the Mel-filterbank. The mean after the log-filterbank can be derived as

$$\mu_{jl}^{\mathrm{LOG}} \approx \log\left(\sum_{k=1}^{K} W_{jk}\mu_{kl}^{\mathrm{ABS}}\right) - \frac{1}{2}\Sigma_{jjl}^{\mathrm{LOG}}. \qquad (11)$$

Once the propagation through the logarithm has been attained, the pending transformations are the discrete cosine transform, delta and acceleration parameters and cepstral mean subtraction. Since these are all linear they pose no additional difficulty and thus the mean and variance of the recognition features can be computed.

3.4. Recognition with Observation Uncertainty Techniques

Three options were used for recognition. In the simplest case, the MMSE-MFCC estimate was directly passed to the recognizer (termed "no VC" for "no variance compensation"). When using Jasper for recognition, the available variances were also used to modify the recognizer to account for the observation uncertainty. For this purpose, modified imputation (MI) [15] and uncertainty decoding (UD) [18] were used.

4. Training

In all cases HMMs were trained using standard Baum-Welch re-estimation. For HTK the training and test scripts provided for the CHIME challenge were used. The only modification was lowering the mixture pruning threshold in speaker adaption. This allowed the use of MLLR adaptation while slightly reducing the performance of the unadapted models. For MLLR, one single global mean transformation was used for each speaker.

The differences between HTK- and Jasper-Training concern four aspects that will be described in the following section.

4.1. Jasper Training

Jasper is a Java-based recognition system for token passing in standard and coupled hidden Markov models [19]. Its core probability computation can be carried out in CUDA [20], which allows for fast training of full-covariance HMMs. The implications of this ability will be described in Sections 4.1.1 to 4.1.3. Also, the model structure used for JASPER was slightly different, which is detailed in Section 4.1.4.

4.1.1. Mixture splitting

A major shortcoming of Baum-Welch re-estimation is that its outcome is optimal only locally. Therefore, initial points are of high significance.

This is of interest also in selecting the directions for mixture splitting. However, a mixture-split can only follow the first eigenvector of the data covariance if the full covariance structure of the data is known. Therefore, in training mixture models, we opted for full-covariance matrices, and used the off-diagonals to inform mixture splitting.

4.1.2. Discriminative iteration control

Although it is typical to carry out a fixed number of Baum-Welch re-estimations after each mixture splitting, this may not give the maximally discriminative model set. Therefore, Jasper carries out re-estimations for each number of mixtures as many times, as performance on the development set continues to improve. Once a loss in accuracy is observed, a step-back takes place, so that the optimum performance model can be used. Since full-covariance models are trained, this is a computationally expensive approach, which is enabled by the massively parallel processing of log-likelihoods that CUDA can provide.

4.1.3. Linear Discriminant Analysis

The full-covariance models also support a linear discriminant analysis. We find the LDA matrix $\mathbf{W}$ by a generalized eigenvector decomposition. This leads to the transformed data

$$\mathbf{x}'_l = \mathbf{W}\mathbf{x}_l \qquad (12)$$

possessing the maximum ratio between inter- and intra-class covariance. In this context, *class* for us is equivalent to one GMM mixture component, so that we actually maximize discrimination between GMM components of the transformed data model. In all following experiments, this projection was onto 37-dimensional feature vectors $\mathbf{x}'_l$, where 37 was the optimum dimension for the development set using mixed training.

4.1.4. Model structure

The sentence model consists of a silence model at the beginning and the end, which is different from the standard setup. Between the silence models, a network for the sentence grammar is defined, which can be traversed by means of token passing and the forward-backward algorithm for recognition and training, respectively. All word models were strict left-right models without skips, using three states per phoneme.

5. Results

After establishing the baseline without signal processing or uncertainty-of-observation techniques in Sec. 5.1, we will show keyword accuracies for the isolated utterances of the development set first. These are organized in two sections: first, for models trained on clean data in Section 5.2, and secondly, for mixed training in Sec. 5.3. The best performing systems from the development set were finally evaluated on the isolated utterance test set, both stand-alone and in a Rover fusion of the three best systems, results for which can be found in Sec. 5.4.

5.1. Baseline results

The baseline results without signal processing are shown in Table 1. Whereas the first block gives official baseline results for the standard HTK configuration, the second block shows the Jasper baseline, obtained with clean training of speaker-dependent models. The final two blocks show results for mixed training, once with the HTK system that also reproduced the baseline results exactly, and once with Jasper.

method	-6dB	-3dB	0dB	3dB	6dB	9dB
clean						
HTK						
devel	31.08	36.75	49.08	64.00	73.83	83.08
test	30.33	35.42	49.50	62.92	75.00	82.42
Jasper						
devel	44.33	48.92	62.08	72.25	80.33	85.50
test	40.83	49.25	60.33	70.67	79.67	84.92
mixed						
HTK						
devel	63.83	70.92	78.50	85.17	89.58	92.42
test	63.00	72.67	79.50	85.25	89.75	93.58
Jasper						
devel	64.44	73.17	81.75	85.00	90.58	91.92
test	64.33	73.08	81.75	85.67	89.50	91.17

Table 1: Keyword recognition accuracy, no signal processing.

5.2. Clean Training

After beamforming with various strategies, the results of Jasper improve significantly, and best results are obtained once uncertainty propagation and, optionally, missing data recognition, are also applied. Both can be seen in Table 2. Here, the results of best averaging system (null-beamformer with uncertainty propagation and modified imputation) are shown in bold. Greek letters identify the systems that were later used in ROVER fusion.

The results for clean training of the HTK system can be seen in the left half of Table 4. As it can be seen here, clean training without adaptation is improved upon very notably by all systems using MLLR. As with the Jasper results, bold numbers indicate results of the system with best *average* performance for the considered condition.

method	-6dB	-3dB	0dB	3dB	6dB	9dB
WPF, no uncertainty propagation						
	50.33	59.67	72.17	80.25	87.25	91.25
Beamforming with uncertainty propagation						
DS						
no VC	54.42	60.83	71.67	80.67	86.00	89.92
UD	56.00	62.00	72.42	80.58	86.83	90.25
α : MI	**56.83**	**63.08**	**72.75**	**81.17**	**87.58**	**91.58**
WPF						
no VC	51.08	59.25	72.42	80.33	86.50	89.67
UD	53.42	61.33	73.58	81.33	87.58	90.00
γ : MI	54.33	63.00	74.08	81.75	88.00	89.50

Table 2: Jasper clean training results: keyword recognition accuracy with standalone beamforming (top) and with beamforming and uncertainty propagation.

5.3. Mixed Training

To reduce the mismatch between models and noisy data, a mixed trainig set was created by adding randomly selected samples from the noise-only development set to the entire clean training set at all SNR conditions. This improved the results notably as shown in the final block of Table 1.

As already for clean training, for mixed training beamforming also improves upon the baseline. But again, the best results, which are also marked in bold, are obtained using uncertainty propagation and missing data recognition, cf. Table 3 for the Jasper, and the right hand side of Table 4 for the HTK keyword recognition accuracies.

method	-6dB	-3dB	0dB	3dB	6dB	9dB
WPF, no uncertainty propagation						
ϵ : 39d	67.25	75.25	82.58	86.67	90.75	91.25
Beamforming (DS) with uncertainty propagation						
39d	74.00	79.25	84.25	88.25	90.50	92.83
LDA	71.50	77.50	86.00	89.25	92.50	93.25
UD						
39d	74.08	79.33	84.42	88.67	90.67	92.50
LDA	72.83	79.42	86.33	89.58	92.25	93.17
MI						
39d	75.58	79.67	84.42	88.67	90.92	92.67
δ :LDA	**75.00**	**79.92**	**86.58**	**90.08**	**92.92**	**93.17**

Table 3: Jasper mixed training results with best standalone beamforming (WPF) without uncertainty propagation (top), and with delay-and-sum beamforming (DS) and uncertainty propagation (MMSE MFCC). LDA-results were obtained with 37-dimensional features.

5.4. Final Test Results

The systems with the best performance on the development set were evaluated on the final test data results for which are shown in Table 6. The corresponding best methods have been marked in bold in Tables 2 and 3 for the Jasper, and in Table 4 for the HTK experiments. Generally speaking, the best-performing system was the delay-and-sum beamformer with uncertainty propagation, which is responsible for all entries in the table, with just the one exception of clean HTK training without MLLR, where the WPF gave best results.

In the last row of Table 6, finally, the results of ROVER fusion are shown. The three systems to be fused were selected based on best ROVER performance on the development set, and the fused system identifiers together with their development set results are shown in the following Table 5.

Systems	-6dB	-3dB	0dB	3dB	6dB	9dB
clean α, β, γ	57.75	64.92	74.08	82.67	89.42	91.58
mixed δ, ϵ, ζ	75.50	81.08	87.50	90.58	93.58	93.17

Table 5: Rover fusion results on development set.

6. Conclusions

Results of automatic speech recognition on reverberant and noisy data can be improved significantly by the combination of beamforming and missing data techniques. This combination can be achieved not only for frequency-domain but also for cepstrum domain recognition, if an appropriate transformation of observation uncertainties is used.

Alternatively to delay-and-sum beamforming, a Wiener beamformer has also given good results, but in the considered dataset, the combination with uncertainty-of-observation techniques was not competitive overall with a simple delay-and-sum beamformer. This indicates the need for further work on uncertainty estimation for beamformer output signals, which would be a promising route for further improvement.

We have tested all algorithms using both clean and matched training. It was observed that matched training leads to by far better recognition results, not only alone, but also in conjunction with all of the tested strategies for signal enhancement and uncertainty-based decoding, indicating both its wide applicability and also the ability of all preprocessing and robust recognition techniques to improve results even under well-matched conditions.

Among all experiments, the highest speech recognition results were obtained by ROVER fusion of multiple recognizer outputs. Among the single recognition systems, the ones showing best performance were generally those using a delay-and-sum beamformer for uncertainty estimation and propagation, with MLLR improving results for clean data, and Jasper with a precomputed LDA dimensionality reduction leading to the best overall performance for mixed training data.

7. References

[1] H. Christensen, J. Barker, N. Ma, and P. Green, “The CHiME corpus: a resource and a challenge for computational hearing in multisource environments,” in *Proc. Interspeech*, 2010.

[2] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007.

[3] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas and Propagation*, vol. 30, no. 1, pp. 27 – 34, 1982.

	Clean training						Mixed training					
method	-6dB	-3dB	0dB	3dB	6dB	9dB	-6dB	-3dB	0dB	3dB	6dB	9dB
NONR	31.08	36.75	49.08	64.00	73.83	83.08	63.83	70.92	78.50	85.17	89.58	92.42
unadapted												
ζ: DS^N	43.17	52.75	61.25	74.75	82.00	88.00	**69.58**	**77.00**	**81.83**	**87.67**	**91.42**	**92.75**
WPF^U	**44.83**	**53.33**	**64.00**	**75.83**	**84.92**	**88.92**	66.58	73.42	81.83	87.33	91.33	92.83
β: GSC^N	47.08	52.50	65.08	73.83	82.42	85.83	65.58	70.75	78.33	85.00	88.58	89.50
IWAB^U	44.75	53.58	65.58	74.83	85.58	87.58	64.75	73.08	79.08	84.33	90.83	91.67
MLLR												
DS^N	**57.42**	**65.83**	**74.00**	**82.33**	**87.50**	**89.75**	**70.17**	**78.67**	**82.58**	**88.25**	**91.58**	**92.58**
WPF^U	54.08	64.75	73.83	80.42	88.00	90.17	68.67	74.08	83.00	88.00	91.00	93.75
GSC^N	55.25	61.75	72.00	78.92	85.50	86.00	67.08	72.17	79.42	85.33	89.08	90.75
IWAB^U	54.00	63.00	72.42	79.75	87.00	88.25	66.67	74.67	79.92	85.50	91.50	91.42

Table 4: HTK keyword recognition accuracy without noise reduction (NONR) is shown in the first row. All other results were obtained with HTK and uncertainty propagation (MMSE-MFCC Estimation) and are shown for clean vs. mixed training data and with vs. w/o MLLR adaptation. A superset N indicates the use of noise estimation, a U that of uncertainty estimation.

	Clean training						Mixed training					
method	-6dB	-3dB	0dB	3dB	6dB	9dB	-6dB	-3dB	0dB	3dB	6dB	9dB
HTK	42.33	51.92	61.50	73.58	80.92	88.75	67.92	77.75	84.17	89.00	91.00	92.75
HTK +MLLR	54.83	65.17	74.25	82.67	87.25	91.33	68.25	79.75	84.67	89.58	91.25	92.92
Jasper	54.50	61.33	72.92	82.17	87.42	90.83	73.92	79.08	86.25	89.83	91.08	93.00
ROVER	57.58	64.42	76.75	86.17	88.58	92.75	74.58	80.58	87.92	90.83	92.75	94.17

Table 6: Keyword recognition accuracies on test set for best methods from development set.

[4] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 4, pp. 487 – 503, 2005.

[5] L. Deng, *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, to appear 2011, ch. Feature-Domain, Model-Domain, and Hybrid Approaches to Noise-Robust Speech Recognition.

[6] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[7] R. Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, to appear 2011, ch. Uncertainty Decoding and Conditional Bayesian Estimation.

[8] R. F. Astudillo and R. Orglmeister, "A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation," in *Proc. Interspeech*, 2010, pp. 713–716.

[9] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 1997, pp. 347 –354.

[10] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[11] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs: Prentice Hall, 1993.

[12] R. Zelinski, "A microphone array with adaptive postfiltering for noise reduction in reverberant rooms," in *Acoustics, Speech, and Signal Processing, International Conference on*, vol. 5, Apr. 1988, pp. 2578 –2581.

[13] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677 –2684, 1999.

[14] A. Abad and J. Hernando, "Speech enhancement and recognition by integrating adaptive beamforming and wiener filtering," in *Proc. 8th International Conference on Spoken Language Processing (ICSLP)*, 2004, pp. 2657–2660.

[15] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005, pp. 82–85.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.

[17] R. F. Astudillo, "Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition," Ph.D. dissertation, Technical University Berlin, 2010.

[18] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, May 2005.

[19] D. Kolossa, J. Chong, S. Zeiler, and K. Keutzer, "Efficient manycore CHMM speech recognition for audiovisual and multistream data," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 2698 – 2701.

[20] *NVIDIA CUDA Compute Unified Device Architecture Programming Guide*, NVIDIA Corporation, 2007.