THE TUM+TUT+KUL APPROACH TO THE 2ND CHIME CHALLENGE: MULTI-STREAM ASR EXPLOITING BLSTM NETWORKS AND SPARSE NMF

Jürgen T. Geiger[1], *Felix Weninger*[1], *Antti Hurmalainen*[2], *Jort F. Gemmeke*[3], *Martin Wöllmer*[1,4],
Björn Schuller[1], *Gerhard Rigoll*[1], *Tuomas Virtanen*[2]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Department of Signal Processing, Tampere University of Technology, Finland
[3]Department ESAT, KU Leuven, Belgium
[4]BMW Group, Munich, Germany
geiger@tum.de

ABSTRACT

We present our joint contribution to the 2nd CHiME Speech Separation and Recognition Challenge. Our system combines speech enhancement by supervised sparse non-negative matrix factorisation (NMF) with a multi-stream speech recognition system. In addition to a conventional MFCC HMM recogniser, predictions by a bidirectional Long Short-Term Memory recurrent neural network (BLSTM-RNN) and from non-negative sparse classification (NSC) are integrated into a triple-stream recogniser. Experiments are carried out on the small vocabulary and the medium vocabulary recognition tasks of the Challenge. Consistent improvements over the Challenge baselines demonstrate the efficacy of the proposed system, resulting in an average word accuracy of 92.8 % in the small vocabulary task and an average word error rate of 41.42 % in the medium vocabulary task.

Index Terms— Long Short-Term Memory, recurrent neural networks, non-negative matrix factorisation, dynamic Bayesian networks

1. INTRODUCTION

Automatic speech recognition (ASR) in reverberated environments with interfering noise sources is typically addressed by a combination of front-end enhancement, such as by speech source separation, and robust back-ends, involving model adaptation and improved ASR architectures. Speech source separation can be achieved through microphone array signal processing [1, 2]; alternatively, if only one microphone is available, monaural separation techniques such as non-negative matrix factorisation (NMF) [3–8] can be used. The latter is especially useful for use cases such as multimedia information retrieval, where multi-channel audio with specified microphone placement is usually not available. On the back-end side, improved recognition architectures are often based on system fusion, for example in tandem ASR. In this context, multi-stream systems have been introduced that fuse traditional Hidden Markov Models (HMMs) with neural networks [9, 10] and sparse coding techniques [7, 11]. An advantage of multi-stream ASR is that additional sources of information can be integrated without re-training the base system.

The benefits of NMF-based speech separation and multi-stream recognition have been successfully combined in our system for the previous 2011 CHiME Challenge [12], which featured a small vocabulary ASR task. Speech separation by convolutive NMF [4] was used in the front-end, and in the back-end an HMM recogniser was combined with a bidirectional Long Short-Term Memory (BLSTM) recurrent neural network in a multi-stream system. In [7], we refined this system by adding a third stream employing speech recognition by non-negative sparse classification [13].

In this contribution, we apply a similar system to both the small vocabulary and medium vocabulary tasks of the 2nd CHiME Speech Separation and Recognition Challenge[1] [14]. Thus, it is of interest whether the previously introduced methods generalise well to the increased vocabulary size, and in particular the recognition of phonemes instead of whole words. As in [7], the system presented in this work uses a multi-stream HMM to combine MFCCs with the word or phoneme predictions of NSC and/or a BLSTM-RNN. A flow-chart of the ASR system is depicted in Figure 1. The BLSTM predictions correspond to the discrete index of the word/phoneme with the highest activation. MFCCs as well as BLSTM predictions can be computed from enhanced speech signals, applying NMF as pre-processing. Through the multi-stream HMM framework, systematic errors of the BLSTM-RNN as well as NSC can be modelled by the HMMs in a conditional probability table (observed prediction given HMM state).

In the following, we will shortly describe the evaluation database of the 2nd CHiME Challenge and our employed methods before turning to the experimental setup and presenting results.

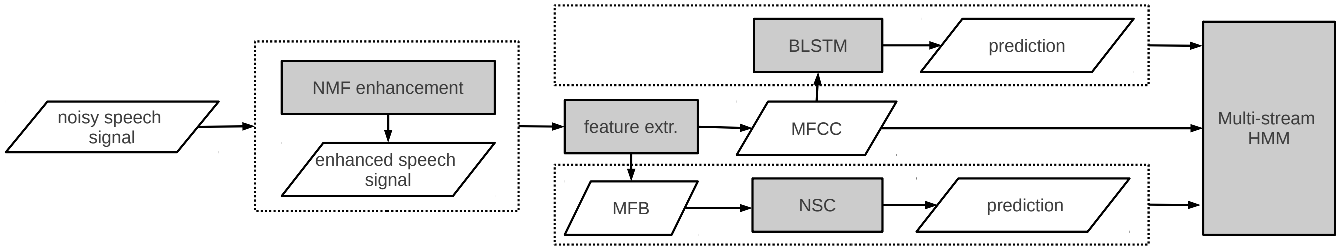
2. EVALUATION DATABASE

The small vocabulary task, as described in [15], consists of reverberated and noisy utterances from the Grid corpus [16] resembling command-and-control utterances with a fixed grammar and a vocabulary size of 51. Utterances have been convolved with real room impulse responses measured in a domestic environment, and overlaid with realistic noise recorded from the same environment at signal-to-noise ratios (SNRs) from -6 to 9 dB, in steps of 3 dB. A closed set of 34 speakers is used for training, development, and testing in the small vocabulary task. The medium vocabulary task is created in a similar way, using the same noise corpus but the speaker independent development and evaluation test sets of the Wall Street Journal corpus (WSJ-0) with 5 k vocabulary size and disjoint sets of 84, 10, and 8 training, development, and test speakers. For both tasks,

This research has been supported by the German Research Foundation (DFG) through grant nos. SCHU 2508/4 and /2. Antti Hurmalainen is funded by the Academy of Finland. The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049.

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

Fig. 1: *Block diagram of the proposed system: The central component is a multi-stream HMM fusing MFCC with optional word predictions by NSC (operating on Mel frequency bands, MFB) and/or the BLSTM-RNN (processing MFCC features). The MFCC feature extraction can optionally by performed on an enhanced speech signal, applying convolutive NMF as pre-processing.*



the same utterances are used at all SNRs in the development and test sets. The training sets comprise a randomly selected subset of utterances for each SNR. In the small vocabulary task, the training set has 17 000 utterances while the development and test sets consist of $6 \times 600 = 3\,600$ utterances. In the medium vocabulary task, there are 7 138 training, $6 \times 409 = 2\,454$ development, and $6 \times 330 = 1\,980$ test utterances.

3. METHODOLOGY

3.1. NMF Speech Enhancement

The speech enhancement component of our system uses exemplar-based spectrogram factorisation algorithms previously employed in noise robust ASR experiments on Aurora-2 and CHiME/GRID datasets [13, 17]. Speech and noise bases are generated by sampling a large amount of $B \times T$ mel spectrogram segments, *exemplars* from training data and the local context of test utterances. B is the number of spectral bands, and T the number of consecutive frames in an exemplar spectrogram. Thereafter the mel-spectral representation of an utterance is factorised using a sliding window method, where $B \times T$ observation windows are extracted with a shift of one frame, and represented as an additive combination of exemplar spectrograms,

$$\mathbf{V} \approx \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)} = \sum_{j=1}^{J} \mathbf{W}_j^{(s)} h_j + \sum_{k=1}^{K} \mathbf{W}_k^{(n)} h_k, \qquad (1)$$

where $\mathbf{V}$ is the true spectrogram, $\mathbf{\Lambda}^{(s)}$ and $\mathbf{\Lambda}^{(n)}$ are estimates for its speech and noise content, respectively, $\mathbf{W}$ are exemplar spectrograms and h their *activation weights*. The coefficients h_j and h_k are obtained through supervised NMF with a sparsity constraint on the activations. We denote the number of speech exemplars (*basis size*) by J and similarly noise basis size with K.

For enhancement, speech and noise spectrogram estimates are generated for full utterances by averaging the frame estimates of overlapping windows. These are mapped back to the linear frequency domain, and act as a time-varying filter for the original spectrogram $\mathbf{V}$, defined by

$$\widehat{\mathbf{V}}^{(s)} = \frac{\mathbf{\Lambda}^{(s)}}{\mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)}} \otimes \mathbf{V}, \qquad (2)$$

where the division and multiplication denoted by $\otimes$ are elementwise.

In summary, the main difference to the enhancement methodology used in our previous work [7] is the choice of exemplars as basis, rather than learnt representations of speech and noise. In preliminary experiments, the exemplar-based enhancement outperformed the method presented in [7] on the CHiME development sets.

3.2. Speech Recognition by Non-Negative Sparse Classification

The NMF setup described in Section 3.1 is also used for *non-negative sparse classification* as proposed in earlier work [7, 13, 17]. Each speech exemplar is equipped with a $Q \times T$ *label matrix*, which represents the likelihoods of Q phonetic states over the exemplar's T frames. State likelihood matrices are generated for utterances similarly to spectrogram estimation by summing the label matrices according to their activation weights, and averaging the frame likelihood vectors from overlapping windows. Consequently a matrix representing state likelihoods in every frame of an utterance is acquired. For the medium vocabulary system, we *learnt* the label matrices which map from exemplars to phonetic states. To learn the mapping we used Ordinary Least Squares (OLS) [18], using as source features the activations of speech exemplars augmented with an all-ones feature to model the intercept, and as target features the monophone identities. The training data for OLS consisted of the reverberated training set. For the small vocabulary task, state likelihoods are summed up per word, and the index of the most likely word is used as a discrete feature n_t, as in [7]. Accordingly, for the medium vocabulary task, state likelihoods are summed up per phoneme to obtain phoneme predictions.

3.3. BLSTM-based Speech Recognition

As additional source of information besides NSC predictions, a BLSTM network is used in the multi-stream framework. More precisely, the BLSTM network is used to generate framewise word / phoneme estimates, as introduced in [10]. Long Short-Term Memory (LSTM) networks were introduced in [19]; the underlying principle can be seen as an extension of conventional recurrent neural networks that enables the modelling of long-range temporal context for improved sequence labelling. LSTM networks are able to store information in linear memory cells over a longer period of time and can learn the optimal amount of contextual information relevant for the classification task. An LSTM hidden layer is composed of multiple recurrently connected subnets (so-called *memory blocks*). Every memory block consists of self-connected *memory cells* and three multiplicative *gate* units (input, output, and forget gates). Further details on the LSTM principle can be found in [20]. We use *bidirectional* LSTM networks (BLSTM) which have access to both, past and future context via forward and backward processing of the speech sequence. A BLSTM network with input units corresponding to MFCC features and one output unit per word (in the small vocabulary task) or phoneme (in the medium vocabulary task) is used to generate a discrete word/phoneme prediction feature b_t for each time step t.

3.4. Multi-Stream Decoding

Using the baseline HMM recogniser, the BLSTM-based system and NSC predictions, a multi-stream HMM system is built. In every frame t, the multistream HMM has access to up to three independent observations: the MFCC features $\mathbf{x}_t$, the BLSTM word/phoneme prediction b_t, and the NSC word/phoneme prediction n_t. The MFCC features are calculated either from the original noisy signal or from the one enhanced by NMF. With $\mathbf{y}_t$ being the joint feature vector and the variables λ^1, λ^2 and λ^3 denoting the stream weights of the MFCC, BLSTM and NSC streams, respectively, the multi-stream HMM emission probability in a certain state s_t can be written as

$$p(\mathbf{y}_t|s_t) = \left[\sum_{m=1}^{M} c_{s_t m}\mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_{s_t m},\boldsymbol{\Sigma}_{s_t m})\right]^{\lambda^1} \times p(b_t|s_t)^{\lambda^2} \times p(n_t|s_t)^{\lambda^3}. \tag{3}$$

More precisely, the continuous MFCC observations are modeled via a mixture of M Gaussians per state while the BLSTM and NSC predictions are modeled using conditional probability tables (CPTs) $p(b_t|s_t)$ and $p(n_t|s_t)$. These are simply obtained as the row-normalised confusion matrices of the BLSTM and NSC on the noisy development set. Note that in the medium vocabulary task, s_t denotes a context-dependent triphone state, while NSC and BLSTM provide predictions for context-indepenent units. In this case, the phoneme predictions from BLSTM and NSC are used for all corresponding triphones. The index m denotes the mixture component, $c_{s_t m}$ is the weight of the m'th Gaussian associated with state s_t, and $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\lambda^i > 0$ indicates presence of a stream.

4. EXPERIMENTS

This section describes the baseline recogniser, the system configuration and experimental setup.

4.1. HMM recogniser

As a baseline system, we use the HMM-based speech recognition system provided by the Challenge organisers. For the small vocabulary recognition task, this system is modified by using mean-only maximum-a-posteriori (MAP) adaptation ($\tau = 1.0$) to estimate speaker-dependent models. Furthermore, reverberated training data are combined with reverberated and noisy training data to increase the robustness of the recogniser. We create our own noisy training data by mixing all 17 000 training utterances with random segments of each of the seven different provided background noise recordings. Thus, the complete extended multicondition training (ext. MCT) set consists of 136 000 utterances. As we assume the SNR conditions in the test data to be unknown, we do not scale the noise or speech levels to obtain specific SNRs for training. In order to follow the Challenge guidelines, we furthermore perform experiments with the baseline noisy training set, or the combination with the reverberated set (denoted as MCT).

For all multi-stream systems, stream weights $\lambda = 1.0$ are used, except for the medium vocabulary recognition task with the HMM + BLSTM system, where values of $\lambda^1 = 1.1$ and $\lambda^2 = 0.9$ are used. These were obtained in earlier work on the CHiME 2011 development data [12].

For the medium vocabulary track, we additionally consider model re-training to adapt to potential distortions in the NMF enhanced signals. This is done by re-training the noisy baseline model using the maximum likelihood training procedure provided by the Challenge baseline, yet using features extracted from NMF enhanced reverberated and noisy training data. For the small vocabulary track, we do not consider model re-training, according to our finding in [12] that performance on the noisy Grid corpus cannot be further enhanced by this, probably due to the distinctive phonetic properties being preserved by the enhancement method.

4.2. Parameterisation

4.2.1. Spectrogram Factorisation

The speech enhancement and NSC setup of the small vocabulary track employs spectrogram factorisation methods as first described in the 2011 CHiME workshop [8] and later refined to a form which is also used in this work [17]. A speech basis comprising 5 000 exemplars is generated for each speaker by pseudo-random sampling of training data and selective reduction with word frequency equalisation. A matching speaker-dependent basis is always used for factorisation as the identity of target speakers was known. For a noise model, 5 000 exemplars are sampled from the noise context of each test utterance individually. All factorisation takes place in a 40-band monaural mel magnitude domain, where the bands were normalised by applying an equalisation curve acquired from training speech. Temporally, the model used 25 ms frames with 10 ms shift, and exemplar size (window length) of 20 frames. Utterances are factorised using sliding window NMF, where the cost function comprises generalised Kullback-Leibler divergence for spectral distance between the true spectrogram and its estimate, and weighted L_1 penalty for nonzero activations to induce sparsity. Iteration count, sparsity weights and other factorisation parameters are set as reported in earlier work [17].

For the medium vocabulary system, we used the same spectral features and factorisation method as for the small vocabulary track. From the reverberated isolated utterances in the training data, 10 000 speech exemplars were extracted by random sampling. Two noise dictionaries were used: a fixed noise dictionary of 4 000 exemplars randomly extracted from the embedded utterances in the noisy training set, and noise dictionary extracted from the 10 seconds of embedding noise in the noisy utterance that is being decoded. This second noise dictionary consists of all exemplars that can be extracted from the 1 000 frames of noise: $2 \cdot 500 - T + 1 = 981$ exemplars. This brings the total number of exemplars in the dictionary to 14 981. As for the small vocabulary task, the exemplar size is 20 frames. The sparsity for the speech was set at 0.075 times the average L_1 norm of the fixed part of the dictionary (speech and noise jointly). The noise sparsity was set at 0.5 times the speech sparsity. The number of iterations was kept constant at 400. These values were tuned using a small random subset of the AURORA-4 corpus.

4.2.2. BLSTM Configuration and Training

The BLSTM network which is used to generate predictions for the multi-stream recogniser was trained on framewise word targets for the small vocabulary task, and framewise phoneme targets for the medium vocabulary task. Therefore, HMM-based forced alignment of the training set was generated. As network input x_i, cepstral mean normalized MFCC features are used. In addition to the input and output layers, the BLSTM network is made of three hidden layers. For the small vocabulary task, these layers consist of 78, 150, and 51 hidden units, respectively. For the medium vocabulary task, 78, 128, and 90 hidden units are employed. Each LSTM memory block contains one memory cell. BLSTM topologies were chosen according to previously performed experiments on similar databases.

The networks are trained through gradient descent with a learning rate of 10^{-5} and a momentum of 0.9. The gradient descent algorithm minimises the root mean square error on the training data. For both tracks each, reverberated training data are used together with reverberated and noisy training data for network training. During training, zero mean Gaussian noise with standard deviation 0.6 is added to the inputs in order to further improve generalisation. All weights were randomly initialised in the range from -0.1 to 0.1. Input and output gates use hyperbolic tangent activation functions, while the forget gates have logistic activation functions. After every fifth epoch in the training phase, the overall root mean square error on the development set is evaluated. Using an early stopping strategy, training is aborted as soon as no improvement on the development set can be observed during 25 epochs.

In the small vocabulary recognition task, speaker identities are known which allows for creating a speaker-dependent system. The employed BLSTM network is adapted by performing additional training epochs using only the training utterances of the respective speaker.

4.3. Experimental Setup

In order to demonstrate the impact of our various system components on the recognition result, we perform a series of experiments. Thereby, we start from the baseline system trained on noisy data, and add one system component at a time. Subsequently, from the full system, we remove some system components.

In particular, for the small vocabulary task, the following experiments are performed: Starting from the baseline HMM system trained on noisy training data, we employ extended MCT by combining noise-free and noisy training data, generated by mixing noise-free data with seven different noise recordings. In addition, speaker dependent models are created using mean-only MAP adaptation. Next, we enhance test data for the BLSTM and MFCC streams with NMF; the NSC prediction is performed on unenhanced data following our previous findings [7]. Then, a double-stream system is created by incorporating the word predictions of the BLSTM. Finally, the full triple-stream system is obtained by adding the NSC prediction stream. Furthermore, we use this triple-stream system to decode unenhanced noisy data. Thereafter, in the triple stream system, we replace the optimised HMM recogniser (MAP + MCT) by the baseline noisy recognition system.

For the medium vocabulary task, we perform a similar series of experiments. The (noisy) baseline system is used to decode NMF-enhanced test data. In addition, we test the system which is re-trained with enhanced training data. The BLSTM and NSC streams are added to create double-stream systems, both evaluated with noisy and enhanced test data – again, enhancement is only used for the BLSTM and MFCC streams. Based on preliminary results with the development set, we did not perform experiments with the triple-stream system on the medium vocabulary task.

5. RESULTS

For the small vocabulary track, word accuracy (WA) is used as evaluation measure, while for the large vocabulary track, word error rate (WER) is employed. In line with the evaluation setup for the first track, word accuracy is only measured on the 35 letter and digit keywords; furthermore, note that in the first track there are no insertion or deletion errors due to the fixed grammar decoding.

5.1. Small Vocabulary ASR

Table 1 shows results on the development and test sets of the 2nd CHiME Speech Separation and Recognition Challenge, small vocabulary track. We report the mean accuracy across the six SNRs (-6 to 9 dB in 3 dB steps) on the development set, and the detailed accuracies per SNR on the test set. The baseline system, trained on noisy data, is improved by employing MCT and mean-only MAP adaptation for speaker adaptation, yielding an average WA of 80.0 % on the test set. Speech enhancement by NMF leads to further improvements, especially at lower SNRs, resulting in an average WA of 88.3 %. Note that this is significantly above the results reported in [12] using a smaller NMF base for enhancement. Consistent improvement accross all SNRs (average WA of 92.8 %) is achieved by incorporating word predictions from the BLSTM network recogniser. Adding the third stream with NSC predictions gives a small (but non-significant) improvement. While adding the NSC stream to the single-stream system leads to a significant[2] improvement (90.4 % average WA, not shown in the table), together with the BLSTM, it brings no further significant improvement. Having obtained substantial improvement compared to the baseline system (93.0 % average WA on test), the triple-stream system almost reaches human performance. It is, however, still significantly below human performance, mainly due to better performance of the human at low SNRs. Based on the triple-stream system, replacing the extended MCT set by a simpler version (by combining the official reverberated and noisy training sets), our best system which is compliant to the Challenge guidelines is obtained. With this system, an average WA of 92.8 % is obtained on the Challenge test set. Removing NMF speech enhancement from the triple-stream system notably results in a relatively small (but significant) performance degradation with an average WA of 91.2 % on the test set. Finally, using the baseline noisy models within the triple-stream system instead of the optimised ones leads to an average WA of 88.2 %.

5.2. Medium Vocabulary ASR

Results for the medium vocabulary track of the 2nd CHiME Challenge are displayed in Table 2. The baseline noisy models are used as a starting point for system improvements. Using NMF-enhanced test data improves the result to an average WER of 52.48 %. The improvement is highly significant[3] on the development set and on the test set. When the models are retrained to reduce the mismatch between the enhanced signals and noisy acoustic models, a substantial improvement is obtained on all SNRs, leading to an average WER of 48.07 %. Adding the BLSTM predictions to the noisy baseline system to form a double-stream system brings further significant improvements with an average WER of 41.76 % on the test set. Notably, when using the BLSTM, the inclusion of NMF enhancement only brings a small improvement to 41.42 % average WER, which is in contrast to the picture we observe in the small vocabulary track. This

[2] When we speak of significant differences, we mean statistical significance according to a simple z-test, using the significance level $\alpha = .05$. As a rule of thumb in the ranges of WA observed in our experiments on the small vocabulary task, results have to differ by 1.5 % absolute WA on average across SNRs, and by 4 % absolute WA per SNR to be significantly different.

[3] According to a z-test with $\alpha = .005$, treating the number of words as sample size.

Table 1: *CHiME 2013 development and test set (small vocabulary track): (Key)word accuracies (% WA) using single- and multi-stream ASR with MFCC, BLSTM, and/or non-negative sparse classification (NSC) streams. As a baseline, the HMM system trained on the noisy training set is used. MAP: mean-only MAP adaptation with $\tau = 1.0$; (ext.) MCT: (extended) multi-condition training. NMF Enh.: speech enhancement by NMF as pre-processing.* [1] *Performance of trained human on a subset of the CHiME 2011 test set (http://spandh.dcs.shef.ac.uk/projects/chime/PCC/results.html).*

WA [%]		**Devel.**			**Test**				**Test**
		Mean			**SNR** [dB]				**Mean**
Front end	Back end		-6	-3	0	3	6	9	
	Single Stream								
no Enh.	MFCC (noisy baseline)	68.8	49.3	58.7	67.5	75.1	78.8	82.9	68.7
no Enh.	MFCC (MAP, ext. MCT)	79.6	63.7	68.9	79.6	85.3	90.4	92.0	80.0
NMF Enh.	MFCC (MAP, ext. MCT)	87.4	78.3	84.3	89.3	91.3	93.3	93.3	88.3
	Double Stream								
NMF Enh.	MFCC (MAP, ext. MCT) + BLSTM	93.3	84.6	90.6	92.8	94.9	96.9	97.0	92.8
	Triple Stream								
NMF Enh.	MFCC (MAP, ext. MCT) + BLSTM + NSC	**94.2**	84.8	90.8	93.3	95.1	97.0	96.9	**93.0**
NMF Enh.	MFCC (MAP, MCT) + BLSTM + NSC	94.0	85.1	90.1	93.1	94.9	96.9	96.8	92.8
no Enh.	MFCC (MAP, ext. MCT) + BLSTM + NSC	92.9	81.7	87.9	91.9	93.9	95.8	96.2	91.2
no Enh.	MFCC + BLSTM + NSC	89.9	77.5	84.3	89.2	90.8	93.8	93.7	88.2
					Human[1]				
		–	*90.3*	*93.0*	*92.3*	*95.3*	*96.8*	*98.8*	*94.4*

probably indicates that the BLSTM would need to be re-trained using NMF-enhanced data to obtain optimal performance.

Let us briefly discuss the performance of NSC predictions by themselves. While these could significantly improve the baseline noisy system (average WER of 52.83 % on the test set), the system including NMF enhancement could not be improved with NSC predictions. Interestingly, the result by using the NSC stream on top of the noisy baseline is similar to the one obtained by using NMF enhancement without re-training of the MFCC stream (52.48 %, cf. above). Combining all three recognisers in a triple-stream system brought no improvement compared to the double-stream BLSTM system (results not shown in the table).

6. CONCLUSIONS

We have presented our approach to the 2nd CHiME Speech Separation and Recognition Challenge, employing speech enhancement by NMF and a multi-stream speech recogniser using predictions from a BLSTM-RNN system and NSC. We have shown that the general system architecture can be successfully applied to both small and medium vocabulary ASR. In the small vocabulary recognition task, the keyword accuracy (averaged over 6 SNRs) of the baseline system is improved by 35 % relative, resulting in an average WA of 92.8 % on the official Challenge test set. In the medium vocabulary track, the WER of the baseline system was improved by 25 % relative. On the test set, a WER of 41.42 % was obtained.

Conceptually, the design of the multi-stream architecture allows straightforward combination of multiple noise-robust ASR system. We found that system combination in a ‘plug-and-play’ fashion worked particularly well for the small vocabulary task, yet in the medium vocabulary task systems apparently have to be fine-tuned to each other for optimal performance, such as by re-training the HMMs to cope with NMF enhanced data. Thus, future work should address optimal adaptation of system components, alternative fusion strategies. Furthermore, we will consider discriminative training of the baseline MFCC stream to complement discriminatively trained neural networks.

7. REFERENCES

[1] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann, “A Two-Channel Acoustic Front-End for Robust Automatic Speech Recognition in Noisy and Reverberant Environments,” in *Proc. of CHiME Workshop*, 2011, pp. 41–46.

[2] D. Kolossa, R. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. da Silva Neto, and R. Martin, “Chime challenge: approaches to robustness using beamforming and uncertainty-of-observation techniques,” in *Proc. of CHiME Workshop*, Florence, Italy, 2011, pp. 6–11.

[3] M.N. Schmidt and R.K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in *Proc. of Interspeech*, Pittsburgh, PA, USA, 2006.

[4] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[5] K.W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” in *Proc. of ICASSP*, Las Vegas, NV, USA, 2008.

[6] G.J. Mysore and P. Smaragdis, “A Non-Negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics,” in *Proc. of ICASSP*, Prague, Czech Republic, 2011.

Table 2: *CHiME 2013 development and test set (medium vocabulary track): Word error rates (% WER) using single- and double-stream ASR with MFCC, BLSTM, and/or non-negative sparse classification (NSC) streams. As a baseline, the HMM system trained on the noisy training set is used. NMF Enh.: speech enhancement by NMF as pre-processing.*

WER [%]		**Devel. Mean**	**Test** **SNR** [dB]						**Test Mean**
Front end	Back end		-6	-3	0	3	6	9	
	Single Stream								
no Enh.	MFCC (noisy baseline)	58.27	70.43	63.09	58.42	51.06	45.32	41.73	55.01
NMF Enh.	MFCC	55.36	65.18	60.17	54.01	48.81	44.35	42.35	52.48
NMF Enh.	MFCC (re-trained)	51.67	61.85	55.58	50.94	43.51	39.14	37.40	48.07
	Double Stream (BLSTM)								
no Enh.	MFCC + BLSTM	45.32	58.57	50.07	43.94	37.06	32.67	28.25	41.76
NMF Enh.	MFCC (re-trained) + BLSTM	**44.85**	57.39	48.96	42.54	37.36	32.58	29.67	**41.42**
	Double Stream (NSC)								
no Enh.	MFCC + NSC	55.27	68.15	60.41	56.36	48.96	43.23	39.87	52.83
NMF Enh.	MFCC (re-trained) + NSC	52.58	64.49	57.22	52.29	45.30	40.05	37.38	49.45

[7] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J.F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-Negative Matrix Factorization for Highly Noise-Robust ASR: to Enhance or to Recognize?," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4681–4684.

[8] A. Hurmalainen, K. Mahkonen, J.F. Gemmeke, and T. Virtanen, "Exemplar-based Recognition of Speech in Highly Variable Noise," in *Proc. of CHiME Workshop*, Florence, Italy, 2011, pp. 1–5.

[9] A. Hagen and A. Morris, "Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR," *Computer Speech and Language*, vol. 19, no. 1, pp. 3–30, 2005.

[10] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream ASR framework for BLSTM modeling of conversational speech," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 4860–4863.

[11] Y. Sun, J.F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Using a DBN to integrate Sparse Classification and GMM-based ASR," in *Proc. of Interspeech*, Makuhari, Japan, 2010.

[12] M. Wöllmer, F. Weninger, J. Geiger, B. Schuller, and G. Rigoll, "Noise Robust ASR in Reverberated Multisource Environments Applying Convolutive NMF and Long Short-Term Memory," *Computer Speech and Language, Special Issue on Speech Separation and Recognition in Multisource Environments*, vol. 27, pp. 780–797, 2013.

[13] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[14] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. of ICASSP*, Vancouver, Canada, 2013.

[15] 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*, Makuhari, Japan, 2010, pp. 1918–1921.

[16] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[17] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition," *Computer Speech and Language*, 2012.

[18] K. Mahkonen, A. Hurmalainen, T. Virtanen, and J.F. Gemmeke, "Mapping sparse representation to state likelihoods in noise-robust automatic speech recognition," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 465–468.

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

[20] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.