The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments

Felix Weninger, Jürgen Geiger, Martin Wöllmer, Björn Schuller, and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Germany

[weninger,geiger,woellmer,schuller,rigoll]@tum.de

Abstract

We present the Munich contribution to the PASCAL 'CHiME' Speech Separation and Recognition Challenge: Our approach combines source separation by supervised convolutive non-negative matrix factorisation (NMF) with our tandem recogniser that augments acoustic features by word predictions of a Long Short-Term Memory recurrent neural network in a multi-stream Hidden Markov Model. The performance of our source separation approach is demonstrated in a sequence of gradually refined speech recognisers. While NMF drastically improves performance for all investigated recognisers, best results are obtained with the multi-stream approach along with a novel adaptation technique for noise dictionaries in supervised NMF. On the final Challenge test set, the proposed system delivers an average keyword recognition accuracy of 87.86 % across SNRs ranging from -6 to 9 dB, reducing the error rate from 44 % to 12 % compared to the Challenge baseline.

Index Terms: Non-Negative Matrix Factorisation, Tandem Speech Recognition

1. Introduction

Automatic speech recognition (ASR) over distant microphones in a noisy environment is a challenging problem leading to a multitude of research on signal enhancement (the front-end) and robust recognisers (the back-end). In fact, these fields are closely coupled: Given imperfect signal enhancement by source separation, robust speech recognition is required to cope with remaining interferences or even distortions induced by the separation. On the other hand, it is still commonly observed that speech recognition performance degrades at low signal-to-noise ratios (SNRs), so the need for source separation algorithms remains.

In the last decade, monaural source separation techniques by non-negative matrix factorisation (NMF) have emerged as a promising technique that is portable across application scenarios and acoustic conditions [1–4]. For instance, the previous CHiME challenge [5] featured an NMF-based approach for cross-talk separation [6] that used speaker models (speech dictionaries) in a supervised NMF framework. In this contribution, we use a convolutive extension of NMF that has delivered promising results for speech denoising [2], and use the increased modelling power to model whole words in the speech dictionaries.

On the other hand, regarding the back-end of our recogniser, we use a multi-stream recogniser employing word predictions of a bidirectional Long Short-Term Memory (BLSTM) recurrent neural network. In our previous studies, tandem architectures using BLSTMs have delivered excellent results in challenging speech recognition scenarios [7, 8].

The architecture of our system is depicted in Figure 1. The (noisy) speech signal is enhanced by convolutive NMF (as described in Section 2). Then, acoustic features are delivered to the BLSTM net, generating a word prediction in a secondary feature stream that is decoded along with the acoustic features in a Hidden Markov Model (HMM) framework (cf. Section 3). Thereby confusions of the BLSTM can be modelled by the HMMs, so that information from the BLSTM is exploited in a complementary way, and additional techniques for noise-robustness such as maximum-a-posteriori (MAP) adaptation or multi-condition training can be seamlessly integrated. The parameterisation of the system components is described in detail in Section 4.

We tuned and evaluated our system on the CHiME corpus, which contains 24 200 utterances (34 speakers) of the Grid corpus (17 000 in the training and 3 600 in each of the development and test set), convolved with different impulse responses. The development and test set are overlaid with stationary and non-stationary noise at six different SNRs from -6 to 9 dB. Along with noisy speech, 4 hours of pure background noise are provided. The corpus is described in detail in [9]. We present our experimental results in Section 5, and conclude in Section 6.

To increase clarity of the following section, we introduce the following notations: for a matrix $\mathbf{A}$, $\mathbf{A}_{i,j}$ denotes the element at row i and column j. The notation $\mathbf{A}_{:,j}$, resembling Matlab syntax, symbolises the j-th column of $\mathbf{A}$ (as a column vector). We write $\mathbf{A} \otimes \mathbf{B}$ for the elementwise product of matrices $\mathbf{A}$ and $\mathbf{B}$; division of matrices is always to be understood as elementwise. Column-wise concatenation of matrices $\mathbf{A}$ and $\mathbf{B}$ is written as $[\mathbf{A}\ \mathbf{B}]$. Finally, for a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $p \geq 0$, we define $\overset{p\rightarrow}{\mathbf{A}} \in \mathbb{R}^{M\times N}$ as a 'shifted' version of $\mathbf{A}$ where the entries of $\mathbf{A}$ are shifted p spots to the right, filling with zeros from the left.

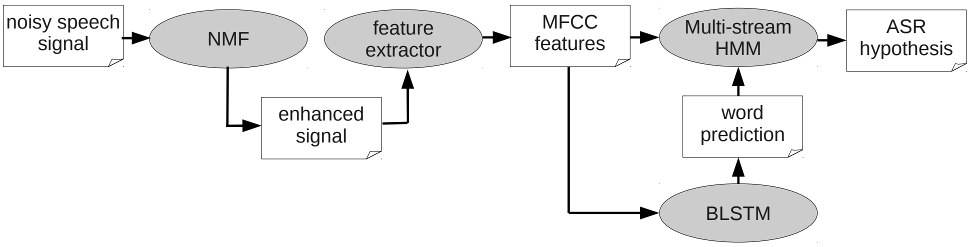
2. Speech Enhancement by Convolutive NMF

Our speech enhancement approach is based on the assumption that speech is corrupted by additive noise:

$$\mathbf{V} = \mathbf{V}^{(s)} + \mathbf{V}^{(n)}, \tag{1}$$

where $\mathbf{V} \in \mathbb{R}_+^{M\times N}$ is an observed magnitude spectrogram of noisy speech, $\mathbf{V}^{(s)}$ is the (true) spectrogram of the speech signal, and $\mathbf{V}^{(s)}$ is the (true) noise spectrogram. Furthermore, we assume that both, the speech and noise spectrograms, can be modelled as convolutions of base spectrograms (dictionaries) $\mathbf{X}^{(s)}(j) \in \mathbb{R}_+^{M\times P}$, $j = 1, \ldots, R^{(s)}$, respectively $\mathbf{X}^{(n)}(j)$, $j = 1, \ldots, R^{(n)}$, with non-negative activations $\mathbf{H}^{(s)} \in \mathbb{R}_+^{R^{(s)}\times N}$,

Figure 1: *Block diagram of the proposed system: Speech enhancement by NMF and multi-stream BLSTM-HMM decoding.*



$\mathbf{H}^{(n)} \in \mathbb{R}_+^{R^{(n)} \times N}$:

$$\mathbf{V}^{(s)}_{:,t} \approx \sum_{j=1}^{R^{(s)}} \sum_{p=1}^{\min\{P,t\}} \mathbf{H}^{(s)}_{j,t-p+1} \mathbf{X}^{(s)}_{:,p}(j), \quad (2)$$

$$\mathbf{V}^{(n)}_{:,t} \approx \sum_{j=1}^{R^{(n)}} \sum_{p=1}^{\min\{P,t\}} \mathbf{H}^{(n)}_{j,t-p+1} \mathbf{X}^{(n)}_{:,p}(j), \quad (3)$$

for $1 \leq t \leq N$. Defining

$$\mathbf{W}^{(s)}(p) = [\mathbf{X}^{(s)}_{:,p+1}(1) \ \cdots \ \mathbf{X}^{(s)}_{:,p+1}(R^{(s)})], \quad (4)$$

$p = 0, \ldots, P-1$ and $\mathbf{W}^{(n)}(p)$ analogously, one obtains an NMF-alike notation of this signal model, denoting the approximation of $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(n)}$ by $\mathbf{\Lambda}^{(s)}$ and $\mathbf{\Lambda}^{(n)}$:

$$\begin{aligned} \mathbf{V} &\approx \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)} \\ &= \sum_{p=0}^{P-1} \mathbf{W}^{(s)}(p) \overset{p\rightarrow}{\mathbf{H}}{}^{(s)} + \sum_{p=0}^{P-1} \mathbf{W}^{(n)}(p) \overset{p\rightarrow}{\mathbf{H}}{}^{(n)} \quad (5) \end{aligned}$$

In the remainder of this paper, we assume that both, $\mathbf{W}^{(s)}(p)$ and $\mathbf{W}^{(n)}(p)$ can be estimated from training data, as shown in Sections 4.2 and 4.3. The speech enhancement problem is thus reduced to finding suitable non-negative coefficients (activations) $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$ – then, the estimated clean speech spectrogram $\widehat{\mathbf{V}}^{(s)}$ is obtained by filtering the observed spectrogram $\mathbf{V}$:

$$\widehat{\mathbf{V}}^{(s)} = \frac{\mathbf{\Lambda}^{(s)}}{\mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)}} \otimes \mathbf{V}. \quad (6)$$

To jointly determine a solution for $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$, we iteratively minimise the element-wise sum of the β-divergence d_β between the observed spectrogram $\mathbf{V}$ and the approximation $\mathbf{\Lambda} := \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)}$:

$$d_\beta(\mathbf{V}|\mathbf{\Lambda}) = \sum_{i=1}^{N} \sum_{j=1}^{M} d_\beta(\mathbf{V}_{i,j}|\mathbf{\Lambda}_{i,j}), \quad (7)$$

starting from a (Gaussian) random solution. In NMF-based speech enhancement, using d_1 (equivalent to Kullback-Leibler divergence) is very popular [2, 3, 10], since it seems to provide a good compromise between separation quality and computational effort.

The minimisation of d_1 (7) is performed by the multiplicative update algorithm for convolutive NMF proposed in [2, 11], which can be very efficiently implemented using linear algebra routines employing vectorisation. Note that the asymptotic complexity of this algorithm is polynomial ($O(RMNP)$), and linear in each of $R := R^{(s)} + R^{(n)}$, M, N, and P. All experiments for this paper were performed with the NMF implementations found in our open-source toolkit openBliSSART [12] to enforce reproducibility of our results.

3. Multi-Stream BLSTM-HMM Recogniser

To enhance recognition accuracies, we apply our recently introduced multi-stream BLSTM-HMM recogniser [8], which was shown to prevail over conventional single-stream HMM-based recognition in challenging ASR scenarios. Our multi-stream system decodes both, low-level MFCC features, and frame-wise word/keyword estimates generated by a bidirectional Long Short-Term Memory recurrent neural network (RNN). Long Short-Term Memory (LSTM) networks were introduced in [13] and can be seen as an extension of conventional recurrent neural networks that enables the modelling of long-range temporal context for improved sequence labelling. They 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). Since these gates allow for write, read, and reset operations within a memory block, an LSTM block can be interpreted as (differentiable) memory chip in a digital computer. Further details on the LSTM principle can be found in [14].

In recent years, the LSTM technique has been successfully applied for a variety of speech-based pattern recognition tasks, including phoneme classification [14], keyword spotting [15], and emotion recognition [16].

Another shortcoming of standard RNNs is that they have access to past but not to future context. This can be overcome by using *bidirectional* RNNs, where two separate recurrent hidden layers scan the input sequences in opposite directions. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions.

For our multi-stream BLSTM-HMM recogniser, we use a combination of the principle of bidirectional networks and the LSTM technique (i e., bidirectional LSTM). In every time frame t the multi-stream HMM uses two independent observations: the MFCC features $\mathbf{x}_t$ and the BLSTM word prediction feature b_t. The vector $\mathbf{x}_t$ also serves as input for the BLSTM, whereas the size of the BLSTM input layer corresponds to the dimensionality of the acoustic feature vector. The vector of BLSTM output activations o_t contains one probability score for each word in the vocabulary at each time step (vocabulary size V).

b_t is the index of the most likely word:

$$b_t = \arg\max_w (o_{t,1}, ..., o_{t,w}, ..., o_{t,V}). \quad (8)$$

In every time step the BLSTM generates a word prediction according to Equation 8, and the HMM models $\mathbf{x}_{1:T}$ and $\mathbf{b}_{1:T}$ as two independent data streams. With $\mathbf{y}_t = [\mathbf{x}_t \; b_t]$ being the joint feature vector consisting of continuous MFCC and discrete BLSTM observations and the variable a denoting the stream weight of the first stream (i. e., the MFCC stream), the multi-stream HMM emission probability while being in a certain state s_t can be written as

$$p(\mathbf{y}_t|s_t) = \left[\sum_{m=1}^{M} c_{s_t m} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{s_t m}, \boldsymbol{\Sigma}_{s_t m}) \right]^a \times p(b_t|s_t)^{2-a}. \quad (9)$$

Thus, the continuous MFCC observations are modeled via a mixture of M Gaussians per state while the BLSTM prediction is modeled using a discrete probability distribution $p(b_t|s_t)$. 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 $\boldsymbol{\Sigma}$. The distribution $p(b_t|s_t)$ is trained to model typical phoneme confusions that occur in the BLSTM network.

4. System Parameterisation

4.1. Preprocessing and Fourier Transform

For NMF speech enhancement, audio signals were down-mixed from stereo to mono by averaging channels and transformed to the spectral domain by short-time Fourier Transformation using a window size of 64 ms (corresponding to 1024 samples at a sample rate of 16 kHz) and 75 % overlap, i. e., 16 ms frame shift. This kind of parameterisation has been proven to deliver excellent results in speech enhancement [2, 3] at an acceptable computational effort. We use the square root of the Hann function for windowing both in forward and backward transformation in order to reduce artifacts, as proposed e. g., in [1].

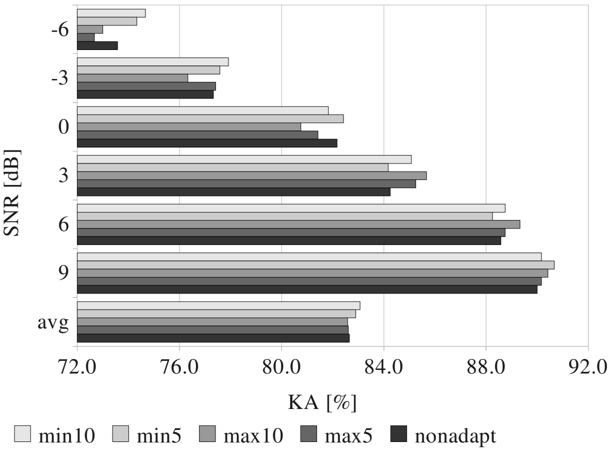
For extraction of MFCC features, we use the common parameterisation of 25 ms Hamming windows at 10 ms frame shift.

4.2. Speech Dictionaries

In this contribution, we computed speech dictionaries for supervised NMF by an algorithm that is particularly suited to speaker-dependent small vocabulary speech recognition tasks, as featured in the Challenge: The task is to decode utterances from a reverberated and noisy version of the Grid corpus containing voice command utterances [9]. Our approach is based on the observation that convolutive NMF is very well suited to modelling spectral sequences corresponding to words [17]. Thus, in our approach each dictionary entry corresponds to a 'characteristic' spectrogram of a certain word ($R^{(s)} = 51$). Furthermore, speaker-dependent dictionaries can be used for the separation since the speaker identity is assumed to be known.

Consequently, the characteristic spectrograms are obtained from the training set by convolutive NMF as follows. For each of the 34 speakers, we used the forced alignments, obtained by the baseline HMM-MFCC recogniser on the noise-free training set of the CHiME corpus [9], to extract all occurrences of each word occurring in the training data (51 words in total). Then, for each speaker $k \in \{1, \ldots, 34\}$ and word $w \in \{1, \ldots, 51\}$, we concatenated the magnitude spectra into a matrix $\mathbf{T}^{(s,k,w)}$, which was reduced to a characteristic spectrogram $\mathbf{w}^{(s,k,w)}(p)$ by a 1-component convolutive NMF,

Figure 2: *Optimisation of adaptive NMF on development set: Keyword recognition accuracies (KA) by SNR and on average (avg), for min-/max-adaptation strategies with $T = 10$ / $T = 5$ versus non-adaptive NMF. Single-stream multi-condition trained recogniser with MAP adaptation (cf. Section 4.5).*



$$\mathbf{T}^{(s,k,w)} \approx \sum_{p=0}^{P-1} \mathbf{w}^{(s,k,w)}(p) \overset{p\rightarrow}{\mathbf{h}^{(s,k,w)}}, \quad (10)$$

and formed a speaker-dependent dictionary

$$\mathbf{W}^{(s,k)}(p) = [\mathbf{w}^{(s,k,1)}(p) \; \cdots \; \mathbf{w}^{(s,k,51)}(p)]. \quad (11)$$

Thereby 100 NMF iterations were used. In our experiments, the parameter P was set to 13 based on inspection of the training set: This corresponds to a spectrogram of a 256 ms signal segment at 64 ms window size and 16 ms frame shift, which is enough to model single words in most utterances, and thus provides a good compromise between modelling power and computational complexity.

4.3. Noise Dictionaries

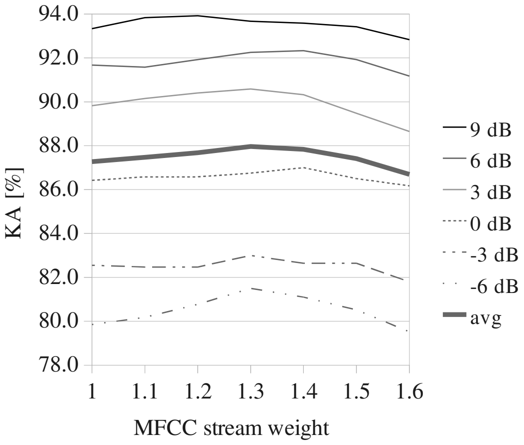
In contrast to the speech, the background noise is assumed to be highly variable among test conditions. Thus, it is desirable to create a noise dictionary as general as possible. To this end, we sub-sampled the set of training noise (approx. 4 hours) available for the Challenge, selecting 4 000 random segments of 256 ms length, concatenated them into a spectrogram $\mathbf{T}^{(n)}$, and reduced them to a dictionary $\mathbf{W}^{(n)}(p)$. Analogous to the speech dictionary, it contains 51 characteristic noise spectrograms ($R^{(n)} = 51$).

We considered a factorisation of the whole training noise provided for the Challenge as not feasible taking into account space and time complexity. Note that in contrast to noise, speech dictionaries can be constructed from all available data, since the training set is subdivided by words and speakers.

4.4. Adaptation of Noise Dictionaries

To gain an upper performance benchmark for the proposed denoising approach, assuming that the particular type of back-

Figure 3: *Multi-stream recogniser: Optimisation of MFCC stream weight a on development set by average keyword recognition accuracy (KA) across SNRs from -6 to 9 dB. Preprocessing by non-adaptive NMF + bandpass filter.*



ground noise is known, we adapt the general noise dictionaries by an exemplar-based approach. It is somewhat similar to the concepts in [18], yet we extend it to convolutive NMF and use the voice activity ground truth provided in the annotation of development and test data. Precisely, in the separation of test utterances j we use an adaptation dictionary consisting of T spectrograms with P frames, calculated from the background noise encountered before the utterance.

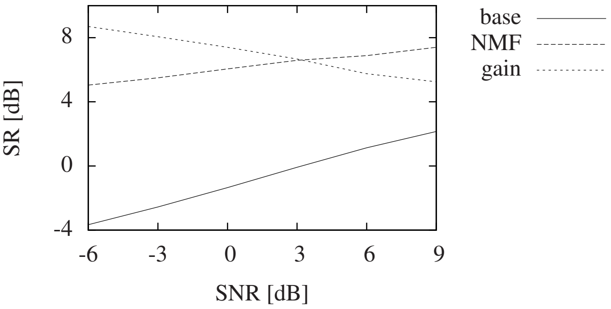
Then, we replace entries of the general noise dictionary according to the (sum of the element-wise) d_1 divergence of the adaptation spectrograms given the dictionary entries. From that, two adaptation methods can be derived, replacing according to maximum ('max-adaptation') or minimum divergence ('min-adaptation'). Best results on the development set were obtained by min-adaptation with $T = 10$ (see Figure 2), corresponding to a context size of 2.56 s before each utterance. Overall, keyword accuracy improvements by min-adaptation are most visible for lower SNRs (-6 and -3 dB) while max-adaptation performs slightly better for higher SNRs (3 dB and above). Still, none of these differences are significant according to a one-tailed z-test ($p > 0.05$, sample size 7 200). In the ongoing, adaptive NMF refers to min-adaptation with $T = 10$.

4.5. Single-Stream Recognisers

As we strive to evaluate source separation and speech recognition separately, we integrated speech enhancement into a sequence of increasingly complex speech recognisers, from the Challenge baseline towards our full-featured multi-stream BLSTM-HMM recogniser. The Challenge baseline uses standard 39-dimensional cepstral mean normalised MFCC features including delta and acceleration coefficients and word-level HMMs with varying number of HMM states (4–10) and 7-component Gaussian mixtures per state. Speaker dependent models are created by additional EM iterations using the training utterances for each speaker.

As a first step to improve the baseline recogniser, we modified the Mel filter bank for MFCC feature extraction with a cut-off frequency of 5 000 Hz, which – along with the preemphasis applied by default (coefficient 0.97) – results in bandpass (BP) filtering. Second, we opted for MAP adaptation (with factor $\tau = 5$) to create the speaker dependent models instead of additional EM iterations. Third, we included multi-condition training for increased noise robustness, by mixing all 17 000 training utterances with random segments of the 4 hours of training noise, to enforce broad coverage of SNRs and noise characteristics. The complete training cycle, including MAP adaptation, is then performed using clean and noisy training data instead of just using clean training data. Note that using the adaptation set provided by the Challenge did not further improve our results on the development set, probably due to its small size compared to the full training set.

Figure 4: *Development set: Average speaker ratio (SR) before (base) and after NMF, and SR gain.*



4.6. Multi-Stream BLSTM-HMM

In order to form the multi-stream BLSTM-HMM recogniser, the multi-condition, MAP adapted single-stream recogniser was enhanced by a BLSTM feature stream. The BLSTM network was trained on framewise word targets obtained via HMM-based forced alignment of the training set. As network input x_t we used cepstral mean normalised MFCC features enhanced by band pass filtering. Similar to the network configuration used in [8], the BLSTM network consisted of three hidden LSTM layers (per input direction) with a size of 78, 150, and 51 hidden units, respectively. Each LSTM memory block contained one memory cell. For training we used a learning rate of 10^{-5} and a momentum of 0.9. Zero mean Gaussian noise with standard deviation 0.6 was added to the inputs during training in order to improve generalisation. Prior to training, all weights were randomly initialised in the range from -0.1 to 0.1. Input and output gates used $\tanh$ activation functions, while the forget gates had logistic activation functions. The network was trained on the 51 different word targets. Training was aborted as soon as no improvement on the development set could be observed. For the multi-stream decoder, we applied a stream weight variable $a = 1.3$ that delivered best performance on the development set (see Figure 3). While the network was trained using the same (partly noisy) data used for building the multi-condition single-stream recogniser, validation and decoding was performed using features from NMF-enhanced signals.

5. Results

5.1. Development Set

Before turning to automatic speech recognition results, we evaluate the noise reduction by our NMF-based source separation

Table 1: *Development set: Keyword recognition accuracies [%]. Recognisers: Baseline, MAP speaker adaptation, and multi-condition training (MCT); multi-stream (MS) BLSTM-HMM with MAP and MCT. Preprocessing by non-adaptive / adaptive NMF (ANMF, min-adaptation, $T = 10$) and/or bandpass filtering (BP).*

Recogniser	**Preprocessing**	**SNR** [dB]						**Mean**
		-6	-3	0	3	6	9	
Baseline	-	31.08	36.75	49.08	64.00	73.83	83.08	56.30
Baseline	BP	34.42	40.42	51.42	65.25	75.25	83.75	58.42
Baseline	NMF + BP	62.17	67.67	73.17	78.50	83.75	86.08	75.22
Baseline	ANMF + BP	61.83	67.50	74.42	78.58	84.17	86.92	75.57
MAP	-	45.58	48.67	62.08	73.92	82.75	88.25	66.88
MAP	BP	46.58	52.08	63.83	74.58	82.25	89.00	68.05
MAP	NMF + BP	71.33	73.92	79.00	83.17	87.00	89.00	80.57
MAP	ANMF + BP	70.75	73.33	78.42	84.00	87.67	89.92	80.68
MCT + MAP	-	56.92	61.08	71.50	81.25	88.92	92.75	75.40
MCT + MAP	BP	54.83	62.42	72.00	80.50	87.00	90.75	74.58
MCT + MAP	NMF + BP	73.58	77.33	82.17	84.25	88.58	90.00	82.65
MCT + MAP	ANMF + BP	74.67	77.92	81.83	85.08	88.75	90.17	83.07
MS + MCT + MAP	BP	69.83	75.83	83.67	88.75	92.58	**94.83**	84.25
MS + MCT + MAP	NMF + BP	81.50	**83.00**	86.75	90.58	92.25	93.67	87.96
MS + MCT + MAP	ANMF + BP	**81.67**	82.67	**87.92**	**91.42**	**92.92**	94.08	**88.45**

on the development set. As evaluation measure, we chose the gain in speaker ratio (SR) proposed in [2] for evaluation of speech de-noising:

$$\mathrm{SR}(f) = 10 \log_{10} \frac{r(f(t), s(t))}{r(f(t), n(t))} \qquad (12)$$

where r denotes correlation, $f(t)$ is the signal to be evaluated, $s(t)$ is the ground truth speech signal, and $n(t)$ the ground truth noise signal. $n(t)$ was estimated by (monophonic) subtraction of the clean speech signal from the noisy signal. The average SR of the mixed signals per SNR (baseline), as well as the average SR of the noisy signals enhanced by convolutive NMF, and the corresponding SR gain are shown in Figure 4. Note that while the baseline SR appears to be linear in SNR, these measures are not equivalent. It turns out that the gain by NMF preprocessing is especially high for low SNRs – for -6 dB, an average SR gain of 8.7 dB is obtained.

Keyword recognition accuracies on the development set are shown in Table 1. Gradually refined speech recognisers deliver better and better performance on average across SNRs from -6 to 9 dB: 68.05 % KA with MAP adaptation, 74.58 % with multi-condition training and MAP, and 84.25 % with the multi-stream BLSTM-HMM recogniser (baseline: 56.30 %). On the other hand, NMF is able to boost the performance of all recogniser types, including the highly robust multi-condition, multi-stream BLSTM-HMM recogniser.

As expected, the strongest improvements by (adaptive) NMF are observed for the least robust recognisers, and vice versa. Precisely, the relative gain in keyword accuracy compared to simple bandpass filtering is 29.3 % for the baseline, 18.6 % for the MAP, 11.4 % for the multi-condition + MAP, and 5.0 % for the multi-stream BLSTM-HMM recogniser. All of these improvements are highly significant according to a one-tailed z-test ($p < 0.001$, sample size 7 200). Furthermore, the gain by signal enhancement (bandpass filtering and NMF) decreases with increasing SNR; notably, for 6 and 9 dB, performance of the multi-condition trained recogniser is lowered by employing NMF and/or bandpass filtering. Overall, ASR results are correlated with the observation of the gain in speaker ratio (Figure 4).

In summary, however, the combination of BLSTM-HMM recognition with NMF-based signal enhancement, enhanced by adaptive noise dictionaries, delivers best accuracy on the development set. Thus, this system will be used for creating our Challenge results on the final test set.

5.2. Final Test Set

Evaluation of source separation using the various recognisers on the final test set (Table 2) reveals the same trends found on the development set, despite the differences in acoustic conditions: Note that for the development and test sets, different room impulse responses have been used for articificial reverberation. The best single-stream recogniser, along with adaptive NMF, delivers a remarkable keyword accuracy of 84.35 % (50.8 % relative improvement over the baseline), while our full-featured ANMF-BLSTM-HMM system delivers our final Challenge result of 87.86 % accuracy, reducing error rate by 72 % relative compared to the baseline. In a fully realistic setting, that is, without adapting NMF using voice activity ground truth, 87.28 % average keyword accuracy are obtained.

6. Conclusion

We have introduced the Munich system for the CHiME Challenge, which integrates NMF-based source separation into a tandem BLSTM-HMM speech recogniser. On the test set, we were able to outperform the Challenge baseline by 55 % relative (31 % absolute) across six different SNRs from -6 to 9 dB. Thereby we have enforced reproducibility of our results by using open-source software.

By evaluating source separation with a sequence of gradually refined speech recognisers, we could show that robust speech recognition architectures and source separation contribute to recognition performance in a complementary way. Furthermore, we have been able to demonstrate the portability of our system across acoustic conditions, including different types of reverberation, stationary and non-stationary noise.

Table 2: *Final test set: Keyword recognition accuracies [%]. Recognisers: Baseline, MAP speaker adaptation, and multi-condition training (MCT); multi-stream (MS) BLSTM-HMM with MAP and MCT. Preprocessing by non-adaptive / adaptive NMF (ANMF) and/or bandpass filtering (BP), cf. Table 1.*

Recogniser	**Preprocessing**	**SNR** [dB]						**Mean**
		-6	-3	0	3	6	9	
Baseline	-	30.33	35.42	49.50	62.92	75.00	82.42	55.93
Baseline	BP	34.08	37.67	53.58	64.25	76.58	83.08	58.21
Baseline	NMF + BP	64.17	69.17	76.42	80.00	84.17	87.67	76.93
Baseline	ANMF + BP	63.50	68.33	77.42	79.50	83.50	87.33	76.60
MAP	-	41.08	47.08	61.67	73.83	81.75	89.83	65.87
MAP	BP	44.50	49.00	64.42	71.17	81.17	88.58	66.47
MAP	NMF + BP	71.83	76.17	82.33	85.50	87.75	89.17	82.13
MAP	ANMF + BP	71.92	75.08	81.75	86.42	87.00	90.33	82.08
MCT + MAP	-	55.08	61.17	71.17	81.67	87.42	92.50	74.84
MCT + MAP	BP	54.50	61.08	72.75	81.67	86.83	91.25	74.68
MCT + MAP	NMF + BP	75.58	79.25	84.08	87.67	88.33	90.58	84.25
MCT + MAP	ANMF + BP	74.67	80.42	83.92	88.67	87.92	90.50	84.35
MS + MCT + MAP	BP	68.50	75.58	82.17	88.33	90.58	**93.92**	83.18
MS + MCT + MAP	NMF + BP	**80.33**	83.50	86.67	90.00	90.25	92.92	87.28
MS + MCT + MAP	ANMF + BP	79.83	**84.00**	**87.92**	**90.67**	**91.83**	92.92	**87.86**

Future work should focus on integration of NMF activation features in a third feature stream, and extension of the convolutive model to allow different lengths of dictionary entries, in order to further enhance portability of the system.

7. Acknowledgements

The research leading to these results has received funding from the Federal Republic of Germany through the German Research Foundation (DFG) under grant nos. SCHU 2508/2-1 (Non-Negative Matrix Factorisation for Noise-Robust Feature Extraction in Speech Processing) and SCHU 2508/4-1 (Context-Sensitive Automatic Recognition of Spontaneous Speech with BLSTM Networks). This work was partially supported by the project AAL-2009-2-049 "Adaptable Ambient Living Assistant" (ALIAS) co-funded by the European Commission and the German Federal Ministry of Education (BMBF) in the Ambient Assisted Living (AAL) programme.

8. References

[1] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proc. of EUSIPCO*, Antalya, Turkey, 2005.

[2] 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.

[3] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. of Interspeech*, Makuhari, Japan, 2010.

[4] G. Evangelista, S. Marchand, M. Plumbley, and E. Vincent, "Sound source separation," in *DAFX - Digital Audio Effects, 2nd Edition*, U. Zölzer, Ed. Wiley, 2011.

[5] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, vol. 24, pp. 1–15, 2010.

[6] 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.

[7] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition - A Tandem BLSTM-HMM approach," in *Proc. of Interspeech*, Brighton, UK, 2009.

[8] 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.

[9] 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. of Interspeech*, Makuhari, Japan, 2010.

[10] 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.

[11] W. Wang, A. Cichocki, and J. A. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2858–2864, July 2009.

[12] F. Weninger, A. Lehmann, and B. Schuller, "openBliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks," in *Proc. of ICASSP*, Prague, Czech Republic, 2011.

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

[14] 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.

[15] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation*, vol. 2, no. 3, pp. 180–190, 2010.

[16] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[17] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Proc. of SAPA*, Jeju, Korea, 2004.

[18] J. Gemmeke and T. Virtanen, "Artificial and online acquired noise dictionaries for noise robust ASR," in *Proc. of Interspeech*, Makuhari, Japan, 2010.