NON-NEGATIVE MATRIX FACTORIZATION FOR HIGHLY NOISE-ROBUST ASR: TO ENHANCE OR TO RECOGNIZE?

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

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2] Department ESAT, Katholieke Universiteit Leuven, Belgium
[3] Department of Signal Processing, Tampere University of Technology, Finland

weninger@tum.de

ABSTRACT

This paper proposes a multi-stream speech recognition system that combines information from three complementary analysis methods in order to improve automatic speech recognition in highly noisy and reverberant environments, as featured in the 2011 PASCAL CHiME Challenge. We integrate word predictions by a bidirectional Long Short-Term Memory recurrent neural network and non-negative sparse classification (NSC) into a multi-stream Hidden Markov Model using convolutive non-negative matrix factorization (NMF) for speech enhancement. Our results suggest that NMF-based enhancement and NSC are complementary despite their overlap in methodology, reaching up to 91.9 % average keyword accuracy on the Challenge test set at signal-to-noise ratios from -6 to 9 dB—the best result reported so far on these data.

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

1. INTRODUCTION

In order to increase robustness of automatic speech recognition (ASR) against varying background noise in reverberant environments, efforts have been devoted to signal enhancement in the front-end on the one hand, and robust architectures of the back-end recognizer on the other hand. In the last decade, the use of non-negative matrix factorization (NMF) has led to impressive results both for front-end signal enhancement and for hybrid or tandem ASR backends. Treating speech enhancement as a source separation problem (speech and noise), NMF-based techniques can be used to factorize spectrograms into non-negative speech and noise dictionaries and their non-negative activations. On the one hand, a clean speech signal can be estimated from the product of speech dictionaries and their activations [1]. On the other hand, if the speech dictionaries are appropriately labelled—e. g., by correspondence to words, phonemes, or Hidden Markov Model (HMM) states—the activations of their entries directly reveal content of the utterance if sparsity constraints are followed (non-negative sparse classification, NSC) [2]. This has been successfully exploited for *exemplar-based* techniques in speech decoding [2, 3].

Recently, in the 2011 PASCAL CHiME Challenge [4], ASR systems were evaluated on noisy and reverberated voice command utterances of the Grid corpus [5], convolved with impulse responses and overlaid with background noise—both measured in a real domestic environment—at six SNRs from -6 to 9 dB. The HMM baseline provided by the organizers indicates the challenge of the data set, yielding 55.9 % average accuracy in recognizing 35 phonetically close keywords (letters and digits). Interestingly, both NMF speech enhancement and exemplar-based recognition were successful: Combining NMF-enhanced Mel frequency cepstral coefficients (MFCCs) with word predictions by a bidirectional Long Short-Term Memory (BLSTM) recurrent neural network (RNN) in a tandem approach yielded 87.3 % average accuracy [6]; 83.8 % accuracy was obtained in a hybrid system mapping exemplar activations computed by NMF to HMM state likelihoods [3].

In this paper, we propose to combine the NMF-based speech enhancement and sparse classification methods. Thus, we treat NSC and NMF enhancement as separate systems despite their overlap in methodology. A flow-chart of the ASR system is depicted in Figure 1. Similar to [6], our fusion strategy uses a multi-stream HMM to combine MFCCs with the word predictions of NSC and/or a BLSTM-RNN. These predictions correspond to the discrete index of the word with the highest activation, respectively. MFCCs as well as word predictions can be computed from enhanced speech signals, applying convolutive 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). Experiments on automatic recognition of noisy and reverberated speech are carried out in order to investigate the effect of combining NMF-based speech enhancement and sparse classification methods.

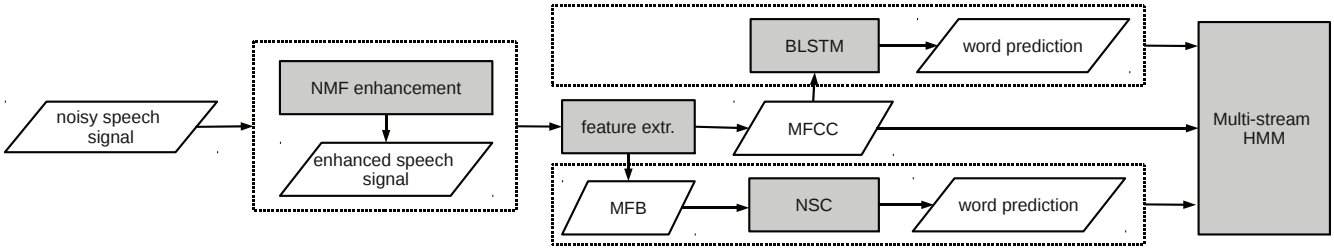
Starting from this broad picture, we now flesh out the details of the evaluation database and the proposed ASR system.

2. EVALUATION DATABASE

Our approaches for speech enhancement and ASR systems are evaluated on the official corpus provided for the 2011 PASCAL CHiME Challenge [4]. The speaker-dependent ASR task is to recognize voice commands of the form *command–color–preposition–letter–digit–adverb*, e. g., *"set white by U seven again"* in a noisy living room. For best comparability with the challenge results, we evaluate by the official challenge competition measure, which is keyword accuracy, i. e., the recognition rate of letters (25 spoken English letters excluding 'W') and digits (0–9). The corpus contains 24 200 utterances of 34 speakers from the Grid corpus [5], subdivided into a training (17 000 utterances), development (3 600), and

This research has been supported by the German Research Foundation (DFG) through grant nos. SCHU 2508/2 and /4, the project AAL-2009-2-049 ALIAS co-funded by the EC and the German BMBF, the Academy of Finland and IWT-SBO project ALADIN contract 100049.

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



test set (3 600). Each set has been convolved with a different binaural room impulse response (BRIR), corresponding to varying room configurations (e. g., doors open/closed, curtains drawn/undrawn). The development and test sets have been mixed with genuine binaural recordings from a domestic environment obtained over a period of several weeks. The noise is highly instationary due to abrupt changes such as appliances being turned on/off, impact noises such as banging doors, and interfering speakers [4]. The six signal-to-noise ratios (SNRs) employed in the development and test set range from 9 dB down to -6 dB in steps of 3 dB. Six hours of training noise (disjoint from development and test) are provided for noise modeling. More details of the domestic audio corpus and the mixing process can be found in [4]. All these data are publicly available at http://spandh.dcs.shef.ac.uk/projects/chime/PCC/datasets.html. For the experiments reported in this paper, all signals were downmixed to mono by averaging channels.

3. SPEECH ENHANCEMENT BY CONVOLUTIVE NMF

As a preprocessing step in the front-end, our multi-stream architecture uses speech enhancement by convolutive non-negative matrix factorization as in [6]. We assume that the observed magnitude spectrogram $\mathbf{V}$ of a noisy and reverberated speech signal can be approximated as the sum $\mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)}$ of (reverberated) speech and noise spectrograms. In turn, both of these are represented by non-negative convolutive bases (*dictionaries*) spanning P frames, denoted by $\mathbf{W}^{(s)}(p)$ and $\mathbf{W}^{(n)}(p)$, $p = 0, \ldots, P-1$, and their non-negative activations $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$:

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

Here $\rightarrow$ denotes a 'right shift' of matrix columns, filling with zeros from the left. We estimate both, $\mathbf{W}^{(s)}(p)$ and $\mathbf{W}^{(n)}(p)$ from training data as in [6]: For each of the 51 words and each of the 34 speakers, the corresponding segments of the noise-free CHiME training utterances are extracted according to an HMM forced alignment and concatenated into a single spectrogram which is reduced to a dictionary atom by convolutive NMF using the Kullback-Leibler (KL) divergence as cost function. From the 51 characteristic word spectrograms per speaker, speaker-dependent dictionaries are formed. A general noise dictionary is obtained by sub-sampling the 4 hours of training noise provided with the CHiME corpus and applying convolutive NMF with 51 components. In our experiments, we use $P = 13$, 64 ms frame size and 16 ms frame shift, and hence longer windows than those commonly used in speech recognition. This has been proven beneficial for the quality of NMF-enhanced signals [1].

Speech enhancement is performed by jointly determining a solution for $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$ using NMF with fixed dictionaries learned from training data. The estimated clean speech spectrogram $\widehat{\mathbf{V}}^{(s)}$ is then obtained by filtering the observed spectrogram $\mathbf{V}$ ($\otimes$ denotes the elementwise matrix product):

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

From $\widehat{\mathbf{V}}^{(s)}$ we resynthesize a time domain signal for further processing in the multi-stream recognizer.

4. WORD PREDICTION BY SPARSE NMF

While long context spectral factorization has been used successfully for separation and enhancement tasks [1, 6], its results can also be exploited more directly in speech recognition. By inspecting the activation weights $\mathbf{H}$ of dictionary atoms as determined by sparse NMF and knowing the identity of each atom, we can find out the sources, which most likely contribute to the mixed observation. In the case of speech, this identity information bears the phonetic content of atoms, thus allowing phone and word classification based on the activation weights without spectral reconstruction or waveform synthesis. In this approach, the atoms correspond to sampled spectrogram segments instead of learned convolutive patterns as in our speech enhancement strategy, and are consequently referred to as *exemplars*. Due to its inherent capability of capturing speech patterns from noisy mixtures, high robustness has been achieved in low SNRs using NSC [2, 3].

In NSC, it is crucial that different phones can be told apart already during the factorization. Therefore we use the same temporal resolution as in common MFCC recognition—25 ms frame size and 10 ms shift. As features, we use 26 Mel scale spectral magnitude bands, again derived from the number commonly used for calculating MFCCs. This resolution is believed to capture most of the information needed for direct classification, while keeping the computational complexity manageable. Both convolutive NMF (as for enhancement in this study) which factorizes the whole utterance jointly, and independent factorization of each window have been shown to enable modeling temporal context within a window [7]. In this work we use exemplar windows spanning 20 frames, and independent factorization of each window, based on our earlier results on CHiME data [3]. The other factorization options, including weighting of features, sparsity penalty values and the number of iterations were exactly set as in [3]. For the sparse classification task, 5 000 speaker-dependent speech exemplars and 5 000 noise exemplars are extracted from the training data. This combined speech-noise basis

is kept fixed during NMF iterations. After receiving the sparse activation weight vector for each window, the weights and the predetermined label sequences encoding the phonetic information of speech exemplars are used to construct a state likelihood matrix for the observation. The details of this NSC setup and its standalone recognition results in a hybrid ASR system are given in [3]. In this work, we determine the most likely word identity n_t for each frame t of the observation by summing state likelihoods corresponding to each word. The resulting sequence of word predictions is then used as a feature stream in the tandem decoder (cf. the next section).

5. MULTI-STREAM SPEECH RECOGNITION

The back-end of the proposed ASR system is based on a multi-stream HMM recognizer recently proposed as an efficient method to integrate Long Short-Term Memory (LSTM) modeling into speech decoding [8]. Long Short-Term Memory networks were introduced in [9] and can be seen as an extension of conventional recurrent neural networks that enables the modeling of long-range temporal context for improved sequence labeling. 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). Further details on the LSTM principle can be found in [10]. In the following, we will use *bidirectional* LSTM networks (BLSTM) which have access to both, past and future context via forward and backward processing of the speech sequence.

Employing a BLSTM network with input units corresponding to MFCC features and one output unit per word, we generate a discrete word prediction feature b_t for each time step t that is equivalent to the index of the output unit with maximum activation—in analogy to the NSC word prediction n_t. Thus, in every time frame t the multi-stream HMM has access to up to three independent observations: the MFCC features $\mathbf{x}_t$, the BLSTM word prediction b_t and the NSC word prediction n_t. $\mathbf{x}_t$ can be calculated from either the original noisy signal or from the one enhanced by convolutive NMF. With $\mathbf{y}_t$ being the concatenation of $\mathbf{x}_t$, b_t and n_t 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

$$p(\mathbf{y}_t|s_t) = \left[\sum_{m=1}^{M} c_{s_t m}\mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_{s_t m},\boldsymbol{\Sigma}_{s_t m})\right]^{\lambda_1} \times p(b_t|s_t)^{\lambda_2} \times p(n_t|s_t)^{\lambda_3}. \quad (3)$$

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 $p(n_t|s_t)$. 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.

6. EXPERIMENTS

We evaluate our system against the baseline provided by the 2011 CHiME Challenge [4] organizers. As basic techniques for increased robustness, we use mean-only maximum-a-posteriori (MAP) adaptation to estimate speaker-dependent GMs modeling the MFCC stream, and multi-condition training (MCT) for both the MFCC GMs and the weights within the BLSTM layers. The MCT training data is generated by mixing all 17 000 training utterances with random segments of the training noise in the CHiME corpus. Thus, the complete MCT (clean and noisy) training data consists of 34 000 utterances. Note that we intentionally do not scale the noise or speech levels to obtain specific SNRs for training, as we assume the SNR conditions in the test data to be unknown. Then, we evaluate the effect of integrating either or both the BLSTM and NSC word prediction streams, and for each case the impact of additional NMF enhancement.

The HMM topologies of the proposed system correspond to the baseline [4]. We employ left-to-right word-level Hidden Markov Models (HMMs) with 4–10 states and seven Gaussian mixtures (GM) per state ($M = 7$). In this study, we use 39-dimensional standard cepstral mean normalized MFCC features as in the baseline. The BLSTM network applied for generating the estimates b_t for the multi-stream system is trained on framewise word targets obtained via HMM-based forced alignment of the clean training set. Similar to the network configuration used in [8], the BLSTM network consists of three hidden LSTM layers (per input direction) with a size of 78, 150, and 51 hidden units, respectively. Each LSTM memory block contains one memory cell. The remaining training configurations are the same as in [8]. To create speaker-dependent networks, we adapt the BLSTM word predictor by performing additional training epochs using only the training utterances of the respective speaker. For each speaker, a network is generated by initializing with the weights of the speaker independent networks and training until no further improvement on the development data of the respective speaker can be observed. By using speaker-dependent BLSTMs, performance on the CHiME test set could be improved by about 3 % absolute with respect to our previous study [6]. The stream weights are optimized for the MFCC-BLSTM case as in [6] and are 1 or 0 (present / absent) otherwise.

7. RESULTS

Experimental results on the development and test set of the CHiME corpus are shown in Table 1. A noticeable improvement of almost 19 % absolute in keyword accuracy (on test) is gained by using MCT and mean-only MAP adaptation for the GM modeling of the MFCC stream, as detailed in [6]. When using only the MFCC stream in a MAP-adapted HMM with MCT, NMF enhancement delivers a gain of about 10 % absolute as reported in [6]. Still, this improvement is mostly visible for lower SNRs, while at 9 dB SNR there is a slight degradation. Considering a noise-robust back-end by additional BLSTM modeling without any front-end enhancement yields 86.3 % average accuracy; while the improvement by the BLSTM is slightly smaller than the one by NMF enhancement at low SNRs, we now observe significant gains even in 'almost clean' speech (4.5 % absolute at 9 dB, $p < .001$ according to a z-test). Using NMF enhancement in combination with BLSTM modeling gives an average accuracy of 90.5 %, again boosting the performance at low SNRs while inducing a slight degradation at 9 dB SNR: It appears that at 9 dB the BLSTM alone delivers predictions so robust that NMF separation artifacts outweigh the benefit of additional noise suppression.

Modeling the NSC word prediction in analogy to the BLSTM in a double-stream HMM, we obtain 83.7 % average accuracy without prior speech enhancement. Most notably, this accuracy is boosted to 87.4 % when using NMF enhancement in addition to NSC, indicat-

Table 1: *Keyword recognition accuracies [%] on the CHiME corpus using multi-stream HMMs with MFCC, BLSTM, and/or non-negative sparse classification (NSC) feature streams. –: not present ($\lambda_i = 0$), ✓: present, ✓+: computed from NMF enhanced signal.*

Streams			Devel	Test SNR [dB]						Test	
MFCC	BLSTM	NSC	Mean	-6	-3	0	3	6	9	Mean	
CHiME Challenge Baseline											
✓	–	–	56.3	30.3	35.4	49.5	62.9	75.0	82.4	55.9	[4]
Speaker adaptation / multi-condition training											
✓	–	–	74.6	54.5	61.1	72.8	81.7	86.8	91.3	74.7	[6]
✓+	–	–	82.7	75.6	79.2	84.1	87.7	88.3	90.6	84.2	[6]
✓	✓	–	86.5	72.8	79.0	85.4	90.8	93.8	95.8	86.3	
✓+	✓+	–	90.1	82.9	87.2	90.3	93.7	93.9	94.8	90.5	
✓	–	✓	85.3	67.2	75.1	85.0	89.8	92.0	93.4	83.7	
✓+	–	✓	89.3	79.1	82.8	88.7	91.2	92.7	93.5	88.0	
✓+	–	✓+	88.2	80.4	83.2	87.5	89.9	90.3	92.8	87.4	
✓	✓	✓	91.0	76.9	82.9	88.8	92.3	95.3	96.4	88.8	
✓+	✓+	✓	**92.6**	84.8	88.3	92.1	93.9	95.7	96.4	**91.9**	
✓+	✓+	✓+	92.1	84.5	87.9	91.0	93.5	95.0	95.6	91.3	

ing that our NMF and NSC approaches both contribute to robustness: We argue that although NMF is the basis of both, the input representation and dictionaries are considerably different in our approach (cf. Sections 3 and 4)—in fact, NSC was shown to produce better recognition of noisy speech than recognition of enhanced signals reconstructed from the same sparse representation [2]. Interestingly, we observe even higher performance (88.0 %) when using enhancement only for the MFCC stream: Enhancement degrades performance of the MFCC-NSC model starting from 0 dB SNR. This can probably be attributed to a mismatch of the NSC dictionaries, which are built from unprocessed speech and noise data, and the characteristics of the separated signal with separation artifacts and remaining interferences.

Finally, the overall best results are obtained by a triple-stream HMM fusing the enhanced MFCC stream with BLSTM and NSC word predictions, reaching 91.9 % keyword accuracy on the test set. Again, using NMF enhancement prior to NSC does not further improve performance; yet again, without NMF enhancement at all, performance is considerably lower (88.8 %). Notably, it can be seen that the triple-stream approach significantly ($p < .005$) outperforms both double-stream approaches, providing evidence for complementarity between the BLSTM and NSC streams.

8. CONCLUSIONS

We have successfully integrated non-negative sparse classification in a multi-stream BLSTM-HMM decoder using NMF speech enhancement, resulting in 91.9 % average keyword accuracy on the CHiME data set containing highly non-stationary noise at SNRs from -6 to 9 dB. This is the best result reported so far on these data. Our results suggest that NMF enhancement and recognition are complementary; it remains to investigate whether this is due to different parameterizations (features, dictionaries) or fundamental methodological differences. Furthermore, since the triple-stream approach delivers best results, we argue that robustness by flexible context modeling in the BLSTM is complementary to explicit noise modeling in NSC. Future work should address better integration of source separation and decoding by adapting the GMMs, the BLSTM as well as the NSC dictionaries to handle separated signals, and generalize the triple-stream approach to large vocabulary ASR.

9. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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, pp. 1918–1921.

[5] 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.

[6] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. of CHiME Workshop*, Florence, Italy, 2011, pp. 24–29.

[7] A. Hurmalainen, J. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 4588–4591.

[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, pp. 4860–4863.

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

[10] 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.