COMPACT LONG CONTEXT SPECTRAL FACTORISATION MODELS FOR NOISE ROBUST RECOGNITION OF MEDIUM VOCABULARY SPEECH

Antti Hurmalainen[⋆] *Jort F. Gemmeke*[†] *Tuomas Virtanen*[⋆]

[⋆] Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[†] Department ESAT, Katholieke Universiteit Leuven, Belgium

ABSTRACT

In environments containing multiple non-stationary sound sources, it becomes increasingly difficult to recognise speech from its short-time spectra alone. Long-context speech and noise models, where phonetic patterns and noise events may span hundreds of milliseconds, have been found beneficial in such separation tasks. Thus far the majority of work employing non-negative matrix factorisation to long-context spectrogram separation has been conducted on small vocabulary tasks by exploiting large speech and noise dictionaries containing thousands of atoms. In this work we study whether the previously proposed factorisation methods are applicable to more natural speech and limited noise context while keeping the model sizes practically feasible. Results are evaluated on the WSJ0 5k -based 2nd CHiME Challenge Track 2 corpus, where we achieve approximately 4% absolute improvement in speech recognition rates compared to baseline using the proposed enhancement framework.

Index Terms— Spectral factorisation, speech recognition, noise robustness

1. INTRODUCTION

In conventional automatic speech recognition (ASR) it is common to employ short-term spectral features as the input for back-end recognition. A typical choice is computing mel-frequency cepstral coefficients (MFCCs) from 25 ms frames with a 10 ms shift. Hidden Markov models (HMMs), used to model temporal progression of speech, search for most likely paths by observing transition probabilities between two consecutive frames. Such short-term evaluation has been found sufficient for clearly spoken speech in optimal conditions. However, real-world speech recognition tasks rarely meet these expectations.

Apart from the linguistic variation taking place in casual speech, a major challenge for practical ASR is coping with signals corrupted by recording hardware, transmission channels, and environmental noise. The latter can be divided further into competing sources and acoustic phenomena such as reverberation. Whereas many kinds of constant channel errors and the effect of acoustic environment can be addressed with static compensation methods, additive noise from varying sources forms a greater obstacle. There is almost infinite variation in the sounds encountered in everyday situations, including semi-stationary background noise, sudden impacts, longer noise events, and competing speech. Especially the last example illustrates how the spectro-temporal behaviour of noise sources can be very close to actual target speech. Furthermore, in conditions falling below a 0 dB signal-to-noise ratio (SNR), noise sources start to dominate several spectral regions, making the short-time spectrum unreliable as a feature space for classification.

It has been demonstrated that increasing the temporal context of modelling units and observation windows is beneficial for discovering spectro-temporal regions dominated by speech or noise. Context of a few hundred milliseconds has been found relevant for speech modelling and perception in statistical speech analysis [1], intelligibility measurement [2] and direct observation of the auditory cortex [3]. The significance of temporal context for robust ASR has received further support in additive multi-source modelling with spectrogram factorisation, where the best results have been achieved by using observation windows spanning 200–300 ms [4, 5, 6].

However, an inherent downside of context expansion is that the modelling units become more specialised, and more units are required to cover the same event space than using a shorter context. In the previously referred experiments and related work, separation and classification quality were found to improve by using thousands of atoms even for small vocabulary tasks like 11-word Aurora-2 [7] and 51-word GRID/CHiME [6]. While early experiments have been conducted on large vocabulary, it is not clear whether the approach is viable for such tasks and eventually real world use.

To address this concern, we propose incorporating refined modelling methods to our non-negative matrix factorisation (NMF) framework. We apply long-context NMF to WSJ0-based 2nd CHiME Challenge Track 2 data, where medium vocabulary speech must be recognised from noisy mixtures ranging from +9 to -6 dB SNR. The identity of the target speaker is also unknown, which was not the case in the 1st CHiME Challenge involving difficult noise conditions [8]. New methods aiming at considerable basis reduction are compared to baseline results and large basis factorisation. In Section 2 we give the basics of spectrogram factorisation. Section 3 introduces recent methods which help in model size reduction. The experimental set-up is described in Section 4, whereafter results are listed and discussed in Section 5. Finally we present conclusions and ideas for future work in Section 6.

2. SPECTROGRAM FACTORISATION

By spectrogram factorisation we refer to techniques, where sound sources are separated in spectral domain by factoring a spectrogram matrix into its constituent parts. Furthermore, we concentrate on algorithms which take into account the temporal continuity of signals, that is, observe a context larger than individual frames. In earlier work, promising results have been achieved by using *non-negative* modelling. The motivation is that DFT resolution spectral magnitudes and features derived from them are mostly additive, thus non-negative additive models produce a good estimate of source component contribution.

Tuomas Virtanen has been funded by the Academy of Finland, grant #258708. The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049.

A common characteristic in previously proposed work is that spectral modelling units and observation windows consist of T consecutive frames. A single spectrogram model, *atom*, is a $B \times T$ matrix, where B is the number of spectral bands in the feature space. Within a similarly sized observation window, the observed spectrogram $\mathbf{Y}$ is modelled as a sum

$$\mathbf{\Psi} = \sum_{l=1}^{L} x_l \mathbf{A}_l, \tag{1}$$

where $\mathbf{\Psi}$ is the estimate of $\mathbf{Y}$, L is the number of atoms (indexed by l), $\mathbf{A}$s are atom spectrograms, and xs are their *activation weights*. All spectral features and activation weights are non-negative. By assigning atoms into individual sources, in this case speech and noise, it is possible to derive single source estimates such as $\mathbf{\Psi}^{\mathrm{s}}$ for speech and $\mathbf{\Psi}^{\mathrm{n}}$ for noise by only including the chosen set's atoms in summing. These estimates are then employed to separate the original spectrogram into its components.

As the duration of an utterance, here denoted by T_{utt} frames, is generally longer than an atom, we need a model to represent the whole $B \times T_{\mathrm{utt}}$ spectrogram as atom activations over time. Two alternative models have been used extensively in earlier work:

1. A 'sliding window' method, where $W = T_{\mathrm{utt}} - T + 1$ overlapping $B \times T$ windows are extracted from $\mathbf{Y}$ in 1 frame steps, and factored individually [4]. The utterance spectrogram estimate $\mathbf{\Psi}$ is produced by averaging over window estimates, hence as an average of up to T single-window factorisations per frame. As atom and observation spectrograms can be vectorised and $\mathbf{X}$ solved from equation $\mathbf{\Psi} = \mathbf{A}\mathbf{X}$, where $\mathbf{\Psi}$ is $BT \times W$, $\mathbf{A}$ is $BT \times L$ and $\mathbf{X}$ is $L \times W$, we call the method simply non-negative matrix factorisation (NMF) for short.
2. Non-negative matrix deconvolution (NMD), alternatively called convolutive NMF (CNMF), where the crucial difference to previously described NMF is that the utterance spectrogram estimate $\mathbf{\Psi}$ is produced jointly by all $\mathbf{X}$ entries via convolutive reconstruction. No averaging takes place as the overall spectrogram is a direct sum of timed activations.

Iterative update rules for determining $\mathbf{X}$ and $\mathbf{A}$ matrices are presented in detail in literature [9] and earlier work [4, 6]. Previous experiments suggest that sliding window NMF has inherent robustness against occasional mismatches and incorrect classification due to its averaging, whereas NMD is better suited for small atom count factorisation as its temporal model requires fewer shifted variants of each sound event than NMF. Both models are considered in this work with the focus being on NMD model reduction.

3. MODEL SIZE REDUCTION METHODS FOR FACTORISATION OF NOISY SPEECH

The basis generation algorithms in previously cited works have often relied on pseudo-random sampling of large amounts of *exemplars* from training material or from the noise neighbourhood of utterances to be recognised. The assumption is that given enough examples of sources, most observed events can be modelled as their linear combination. For abundant training data and model size, random sampling was found as good as initial attempts of refined selection. Later we have proposed informed speech basis reduction, replacing exemplars with state-centric templates, and noise basis reduction by NMD modelling [6]. Still, constraints such as small vocabulary, simplified grammar, or plentiful noise context were typically exploited in the experiments. In this section we present alternative speech and noise modelling methods, which produce compact bases for medium vocabulary speech separation in difficult conditions.

3.1. Variable length atoms

The first recently introduced model extension allows the length of atoms to vary within a basis. While in sliding window NMF the atom duration T is practically forced by design to be a constant in any single factorisation task, the same restriction does not apply to NMD. By using variable atom length it is possible to exploit long context and its benefits in separation whenever suitable, while also maintaining shorter units which also appear in natural speech and noise. Early experiments have been conducted on variable length bases for two-speaker separation [10] and robust ASR for small vocabulary [11], but the work presented here is among the first examples of variable length NMD modelling in semi-realistic ASR.

The convolutive utterance re-estimation formula for variable atom length T_l becomes

$$\mathbf{\Psi} = \sum_{l=1}^{L} \sum_{t=1}^{T_l} \mathbf{A}_{l,t} \overrightarrow{\mathbf{X}_l}^{(t-1)}. \tag{2}$$

$\mathbf{A}_{l,t}$ is the t^{th} frame column vector of atom l, $\mathbf{X}_l$ is the l^{th} row vector of $\mathbf{X}$, and operator $\rightarrow$ shifts it right by $t-1$ columns.

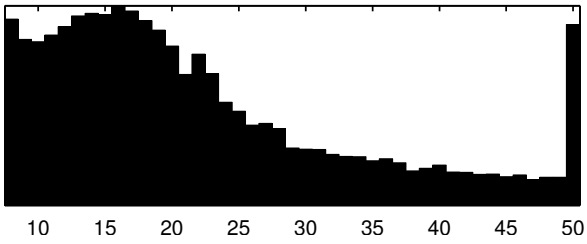
In this work we use strongly variable-length speech bases by employing a basis acquisition algorithm similar to the one presented for CHiME/GRID speech data [11]. The algorithm starts from the longest permitted atom length $T = T_{max}$, inspects the speech training data, and attempts to find length T segments matching to each other. The measure used for match-finding is a combination of spectral data and monophone annotations to take into account both spectral and linguistic similarity. If a sufficiently large group of matching segments (here called a *cluster*) is found, a speech atom is formed by averaging the matching spectrograms. The corresponding areas of training data are flagged as taken. Thereafter the algorithm continues searching for clusters, reducing the segment length by one whenever the minimum cluster size requirement can no longer be met at current length T. Consequently a basis of template atoms is generated in a decreasing order of atom length and frequency of occurrence in the training data.

3.2. Multi-stage factorisation with speaker-dependent bases

In WSJ0-based CHiME Track 2 data, training and test speaker identities form disjoint sets. In other words, no exactly matching speaker model can be chosen for test factorisation, and no clues about test speaker characteristics are initially available. However, it is obvious that factorisation with a closely matching speaker model has a better chance of capturing correct speech features among noises which may include competing non-target speakers. Earlier it has been illustrated how NMD can act as a speaker identifier, when multiple speaker-dependent bases are used for factorisation and the relative activation weights of each speaker's atoms are observed [12].

Based on these findings, we propose a method which allows approximate speaker identification and basis selection by using multi-stage factorisation. In the initial stage, a small number of atoms from all training speakers are used, and relatively few NMD iterations are computed. In each subsequent stage, speaker activity weights are used for selecting the best matching bases, while more atoms from the chosen speakers are introduced to factorisation. Eventually the system will converge to a small set of training speakers, whose speech profiles match best to the target speaker. The details for the presented set-up are given in Section 4.4. By dynamic management

Figure 1: Histogram of speech atom lengths from variable length basis acquisition, ranging from 8 to 50 frames.



of the number and size of bases, it is possible to perform multi-speaker modelling and semi-matched final factorisation, while the amount of simultaneously active atoms remains low. For accelerated basis set reduction, we use a group sparsity constraint, which favours solutions where activations come from a small number of bases [12].

3.3. Pre- and online-adaptation of noise atoms

For acquisition of noise models, there are three important sources whose availability and significance depends on the recognition task. First, we may have fixed training material for pure noise. Second, there is a varying amount of noise context surrounding target speech. Finally, noise can be estimated from the utterance itself by capturing features which do not match to any speech models. In previous work, all three methods have been exploited [6, 7, 13, 14] with occasional further extensions such as artificial noise atoms [15].

Previously we have achieved the best results by sampling large exemplar bases randomly from training data [4] or semi-randomly from the local context [6] according to availability. However, both methods are prone to including a lot of redundancy or unnecessary, near-silent spectral data. Furthermore, exemplars sampled from additive multi-source mixtures cannot model accurately the same events appearing alone or in different combinations. Therefore in this work we use methods based on NMD learning to acquire smaller noise models with a higher efficiency.

Regardless of which data is used for noise learning, we apply iterative NMD atom update rules described in literature [9, 16]. For CHiME Track 2 data, we use two sources for noise atoms: first, background training data which is first reduced to its loudest sections, and second, the 'embedded' utterances with 5 seconds of noise context before and after. It has been found that to prevent overfitting and fragmentation of learnt atoms into unusably small spectro-temporal units, adaptation should be terminated earlier than the commonly employed amount of factorisation iterations for fixed bases. Computationally the simplest way to implement this is to reduce the number of iterations to approximately 20–30 (compared to 200–400 of earlier work), which can be achieved in long semi-supervised factorisation by only performing a basis update after a certain interval of activation update iterations.

4. EXPERIMENTAL SET-UP

A factorisation framework was designed for the 2nd CHiME Challenge medium vocabulary (Track 2) dataset [17]. Its speech data consists of WSJ0 5k vocabulary utterances and is divided as follows:

- 7138 training utterances jointly from 83 speakers, both 'clean' (without additive noise) and mixed at a random SNR
- 409 development test utterances jointly from other 10 speakers and repeated at 6 SNRs
- 330 evaluation test utterances from other 8 speakers, 6 SNRs

Noisy utterances are mixed with non-stationary multi-source household noise at SNRs ranging from +9 to -6 dB in 3 dB steps. Noise data contains natural room reverberation. For speech data, similar impulse responses are simulated. All utterances are available with 5 seconds of noise context before and after the utterance. Approximately seven hours of pure noise data is also available for training. Recognition is measured by HTK toolkit's 'Err' word error rate.

Table 1: Statistics of speech bases used during multi-stage factorisation of the CHiME Track 2 evaluation set. For each stage, the number of active speaker bases and their combined atom count is reported as minimum, mean and maximum values.

Stage	Speakers			Atoms		
	min	mean	max	min	mean	max
1	83	83	83	4150	4150	4150
2	9	24.3	36	900	2427	3600
3	2	8.9	17	612	3023	5754
4	1	3.8	9	304	1305	3246

4.1. Feature space

All factorisation experiments were conducted in monaural 40-band mel-spectral magnitude space. Features were extracted from binaural input signals with a frame length of 25 ms and frame shift of 10 ms, and averaged in absolute magnitude value domain. Mel bands were reweighted by a fixed equalisation curve derived from 2-normalisation of noisy 0 dB training utterances.

4.2. Speech bases

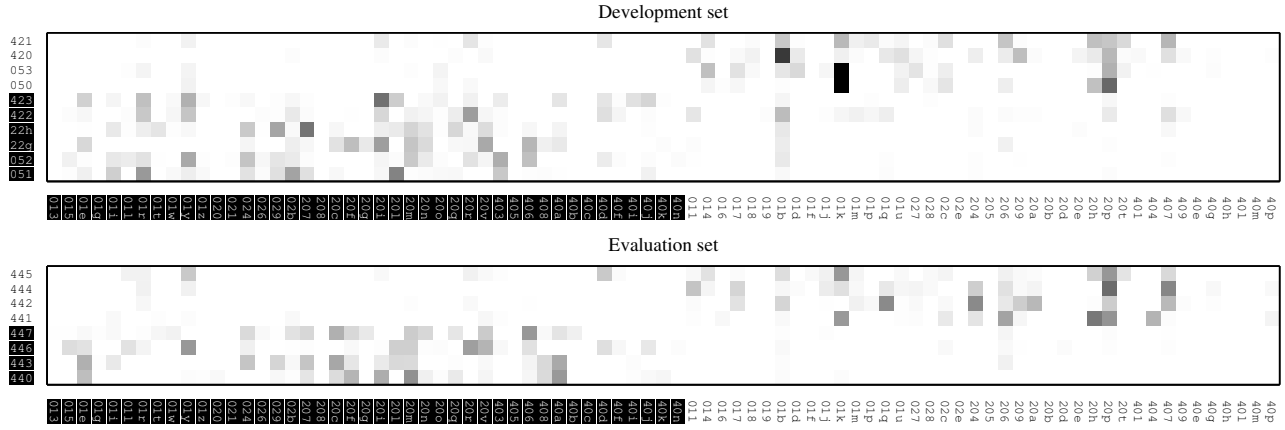
A variable-length speech basis was generated for each training speaker similarly to the algorithm described for 1st CHiME challenge data [11]. The similarity measure for frame vectors consisted of dot product between normalised, square root compressed mel magnitudes augmented with delta features, and monophone labels acquired from forced alignment using the baseline recogniser. Similarity between frames i and j was computed as

$$c_m(i,j) = c_\mathrm{s}(i,j) + c_l(i,j), \tag{3}$$

where the merged similarity c_m is the sum of spectral vector dot product c_s and correlation of monophone labels c_l, the latter ranging from 0 to 0.06 depending on how closely monophones and their substates matched in annotations. Sequences where all mutual frame pairs produced at least 0.92 total similarity were considered for clustering. A cluster was selected for atom construction if its source segments covered at least 0.15% of the speaker's noiseless training material. In other words, long segments were allowed to form atoms with fewer matches than short segments. Atom lengths ranged from 46 to 4 in clustering, whereafter the 2 preceding and following frames were added to atoms as their content is implied by delta features. Consequently the final length of speech atoms was between 8 and 50 frames (80–500 ms).

Figure 1 illustrates the distribution of speech atom lengths in combined speech bases. We notice that large variation takes place, reflecting the multitude of phonetic unit lengths appearing in natural speech. A large peak can be seen at length 50. Even longer correlating segments could be found, but their value for factorisation

Figure 2: Similarity of test speakers (y-axis) to training speakers (x-axis), measured as the amount of speaker-dependent basis activations in the last stage of 9 dB test set factorisation. For each test identity, the sum of activations is normalised to unity. Similarity increases toward black with the maximum intensity being 0.4. White-on-black identity names belong to male speakers, black-on-white to females.



becomes negligible thus they were truncated to the chosen maximum value. Mean atom length was 22.2 frames, approximately matching the previously favoured fixed contexts of 200–300 ms [4, 6, 13].

The number of atoms in the 83 speaker-dependent bases was from 276 to 397 with a mean value of 344 and combined atom count of 28579. Speakers with more variable pronunciation generated larger bases than very consistent ones. Because the cluster size was defined as a percentage of available data, no notable difference was present between speakers with fewer or more training utterances. By comparing the basis sizes to the 5000 word vocabulary, it is clear that the typical unit modelled was shorter than a complete word.

4.3. Noise bases

Two noise modelling methods were used: a fixed noise basis acquired by NMD learning from background training material, and online-adapted noise model from the embedded utterances.

For fixed basis acquisition, the seven-hour training material was first reduced to its loudest 20% frames, measured by spectral magnitudes. From the remaining material, segments shorter than 5 frames were removed, while the rest were padded by 10 frames before and 30 after, approximating the usual temporal decay profile of noise events. Thereafter the segments were faded in and out with a 10-frame transition, and concatenated into approximately five minute blocks of significant noise events. Each block was factored with 25 iterations of NMD basis adaptation to produce atoms with a joint duration of 10% of block length, that is, approximately 60 atoms of length 50 frames per block. While no attempt was made to force noise atoms into shorter or variable duration, in practice this often happened due to some of the atoms modelling short-duration noise events. The procedure as a whole generated 1729 fixed noise atoms.

Direct adaptation of noise atoms from embedded utterances followed mostly similar principles, yet employed significantly fewer atoms. The details are described in the next subsection.

4.4. Multi-stage factorisation

Training and test file factorisation was conducted using the 'embedded' files with 5 seconds of noise context to both directions. After feature extraction, the following bases were set up:

- Speech bases: for test files all 83 speaker-dependent bases, for training files all except self
- A variable amount of randomly initialised adaptive noise atoms, enough to cover 75% of embedded utterance duration
- Optionally, the fixed 1729-atom noise basis (See section 4.3.)

The motivation for given speech basis choices was to use a set of bases disjoint from the target identity. For development and evaluation sets this was automatically the case. For training utterances, the true matching identity was left out to prevent oracle modelling.

The adaptive noise atom count was left slightly below the amount required to cover all embedded utterance frames in order to promote discovery of recurrent features. These atoms were re-adapted from scratch for each utterance from its own context alone. Training and evaluation were run with and without the fixed noise basis to study whether the methods are applicable to entirely new situations where pre-training of noise models is not an option.

For factorisation, variable-length NMD was used with generalised Kullback-Leibler divergence as the spectral distance measure, and L_1 penalty as the sparsity constraint similarly to earlier work. L_1/L_2 group sparsity penalty was induced on speech activations as presented previously [12], with each speaker's atoms forming a group. Sparsity weights were defined by brief experimentation on development utterances and set to 0.07, 0.1, 0.1 and 0.11 for speech, groups, adaptive noise and fixed noise (respectively) when the latter was used, and 0.08, 0.1 and 0.1 for the rest when not. All sparsity values are proportional to the mean value of basis atom 1-norms.

Factorisation had four stages with basis pruning as follows:

1. All speech bases, 50 atoms per speaker, 50 iterations
2. Reduced set of bases, 100 atoms per speaker, 50 iterations
3. Further reduced set of bases, all atoms, 100 iterations
4. Final reduced set of bases, all atoms, 100 iterations

Each partial basis consisted of the first (longest) atoms of complete speaker-dependent bases. Between stages, activation matrix sums were calculated for each speaker dependent basis. A threshold value was set 10–20% from the geometric mean toward the largest value to remove all except the best matching identities. Activation weights of remaining speech atoms were left as is, whereas newly introduced atoms were given a small initial weight of 0.001. Noise atoms or their activation were not changed between stages.

Table 2: Results for CHiME Track 2 noise robust speech recognition, listed as word error rate ('Err') over SNRs. Tables on the left and the right show results for development and evaluation sets, respectively. First, the baseline results using provided 'noise' models are given. The next lines show results for proposed enhancement using adaptive noise atoms only, and then for both adaptive and fixed noise atoms. Finally reference results for large basis NMF are shown. Results are evaluated using provided and re-trained GMMs.

SNR (dB)		9	6	3	0	-3	-6	avg
Baseline ('noise')		44.34	49.05	55.71	59.89	67.43	73.17	58.27
Adapt. only	noise	44.03	48.91	55.04	58.35	65.97	71.19	57.25
	re-trained	42.93	48.24	53.76	57.53	64.71	70.92	56.35
Adapt. +fixed	noise	44.19	47.25	53.27	56.53	63.93	69.47	55.77
	re-trained	42.28	45.54	51.45	55.43	62.61	69.26	54.43
Large NMF	noise	43.33	46.75	51.66	56.51	64.61	69.28	55.36
	re-trained	39.13	44.18	47.65	52.29	60.56	66.23	51.67

(a) Development set

SNR (dB)		9	6	3	0	-3	-6	avg
Baseline ('noise')		41.73	45.32	51.06	58.42	63.09	70.43	55.01
Adapt. only	noise	42.59	45.19	49.71	56.53	61.76	66.75	53.76
	re-trained	40.30	44.44	48.70	54.04	60.34	66.90	52.45
Adapt. +fixed	noise	41.60	44.16	50.29	54.80	60.34	66.34	52.92
	re-trained	38.76	41.53	47.99	51.73	58.83	66.71	50.93
Large NMF	noise	42.35	44.35	48.81	54.01	60.17	65.18	52.48
	re-trained	37.40	39.14	43.51	50.94	55.58	61.85	48.07

(b) Evaluation set

Basic statistics of basis and atom counts in each stage are listed in Table 1 for the test set (with the fixed noise basis enabled). Notably, the simultaneous speech atom count never exceeded 5754, and the last stage employed on average 3.8 bases and 1305 atoms.

Figure 2 illustrates the convergence of different test speakers' (y-axis) factorisation toward matching training speaker bases (x-axis). 9 dB SNR experiments were used for the plot to minimise noise interference. We can observe that even though approximately 40 different utterances were factorised per test speaker, the algorithm generally converged toward a spiky distribution of only a few matching bases. The bases were also mostly from the same gender as the test speaker, and the set was unique for each individual speaker. Comparison by listening confirmed that approximately similar speaker profiles were generally found.

Speech and fixed noise activations were only permitted in the actual utterance area, whereas adaptive noise activations were permitted also in the noisy context to capture the immediate noise environment. As the adaptive basis size was generally below 30 atoms and only updated every 10 iterations (of total 300), factorisation effort was mostly concentrated on the noisy speech, and the overall complexity of the system remained comparable to previous small vocabulary experiments.

4.5. Enhancement and recognition

The activation matrices acquired from NMD were used to generate speech and noise spectrogram estimates as described in Sections 2 and 3.1. Mel spectrograms were mapped back to linear frequency domain and used as a time-varying filter defined as $\mathbf{\Psi}^{\mathrm{s}}/(\mathbf{\Psi}^{\mathrm{s}} + \mathbf{\Psi}^{\mathrm{n}})$ for the original noisy spectrograms [6].

Because the sparse NMD model with adaptive atoms occasionally produces rapidly changing spectro-temporal behaviour with heavy filtering in fully masked segments, it was found beneficial to apply a 0.1 minimum value to the filter weight value normally ranging from 0 to 1. Enhanced signals were recognised using the CHiME HTK tools, both with the multi-condition noise trained baseline models and models re-trained with enhanced training data.

For comparison, we also implemented a sliding window NMF system employing considerably larger exemplar bases similarly to earlier work. 10000 speech exemplars and 4000 noise exemplars were sampled randomly from training material, whereafter approximately 1000 noise exemplars were added from the context. Feature space, factorisation and enhancement followed generally similar principles to those presented for Aurora-2 and 1st CHiME data [4, 6], and for applicable parts they matched the NMD setup.

5. RESULTS AND DISCUSSION

Results for speech recognition experiments are given in Table 2 as word error rates (HTK 'Err') per SNR, separately for development and evaluation sets. The first row shows results using baseline 'noise' models and unenhanced waves. The next rows list results for proposed enhancement using adaptive noise only, and for adaptive+fixed noise. The last rows list results for reference NMF enhancement using large exemplar bases. Enhanced signals were evaluated using the baseline 'noise' models, and with GMMs re-trained from matching training set enhancement.

We observe that enhancement with the proposed approach generally yields improvement over the baseline already on the standard back-end models. Expectedly including a fixed noise basis acquired from background training material provides further improvement over just using noise adaptation from the embedded utterance. Without back-end re-training, the proposed system with both noise models is approximately comparable (2–3 % over baseline) to NMF with large exemplar bases. In re-training, the gap increases so that the improvements over unenhanced baseline are approximately 4% and 7% for proposed and NMF factorisation, respectively.

The proposed framework is our first attempt to develop a relatively lightweight factorisation and enhancement system for medium-vocabulary speech recognition in difficult conditions. Compared to the GRID-based 1st CHiME set [8], the new WSJ-based corpus introduced several new challenges. The 5000 word vocabulary with only limited training data available for each speaker requires a different approach to generating speaker-dependent speech bases. Furthermore, test identities coming from disjoint speaker sets prevented selecting a perfectly matching speech model.

We investigated using several small speaker-dependent bases, which complement each other concerning both vocabulary and speaker characteristics. A clear benefit of (approximate) identity matching is the ability to separate a target speaker from competing speakers, which is more difficult with a speaker-independent basis modelling all speakers simultaneously. From Figure 2 we see that at least at high SNRs the algorithm was able to find similar speaker profiles. An obvious problem of the method is that non-target speakers have a good chance of activating an alternative set of bases, and at < 0 dB even dominating the selection process. Currently this is only prevented by vocabulary matching via long context atoms. Further methods for correct selection could include spatial estimation and preliminary decoding during the selection process.

In noise modelling, initial results suggest that a noise model

adapted from a 5 second context only has a limited separation capability. Acquiring a comprehensive model beforehand improves results significantly. However, the obvious problem is applying the method to new noise environments. In a real-world system, continuous noise model updating during pauses in speech would be preferable in order to maintain a maximally good match. Such a system for continuous NMD recognition has already been proposed [18].

With respect to model complexity and the goal of achieving feasible basis sizes, we can observe that the proposed framework managed to improve average speech recognition rates by approximately 4% (absolute) compared to the unenhanced baseline with an average basis size of 1305 final stage speech atoms, 1729 fixed noise atoms, and generally less than 30 adaptive noise atoms – approximately $1/5^{th}$ of the reference NMF basis size. While more atoms were temporarily used for speaker selection, it must be noted that in these experiments we always started from all 83 candidates for each utterance. In practice, there is a lot of redundancy among the models with some of them barely activating at all, and in real world it rarely applies that speaker adaptation should be repeatedly started from scratch. Therefore we expect that the multi-speaker basis sizes could be easily reduced further. Regarding vocabulary size, already the current bases modelled sub-word units of a vocabulary 15 times larger than average atom count and covered a large part of common linguistic units, hence the requirements for truly large vocabulary should not be considerably greater.

6. CONCLUSIONS

We presented a spectrogram factorisation framework designed for medium vocabulary speech recognition using long temporal context yet compact bases. Several emerging or wholly novel ideas were proposed, including variable length modelling, multi-stage factorisation with basis pruning, and two noise models used in conjugation.

With refined bases, it was found feasible to separate unknown speaker's speech from very noisy mixtures with models smaller than were previously used for small vocabulary tasks with matching speaker identity. Approximately 4% absolute reduction was obtained in average word error rate in evaluation on the 2nd CHiME Challenge Track 2 corpus. As several novel aspects were introduced and combined for a new task with limited parameter tuning, we expect further improvements when their standalone and interoperation characteristics becomes better understood. Nevertheless, already the initial results appear promising regarding robust real-world speech recognition with practically applicable factorisation model sizes.

7. REFERENCES

[1] O. Räsänen and U. K. Laine, "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, vol. 45, no. 1, pp. 606–616, 2012.

[2] T. M. Elliott and F. E. Frédéric, "The Modulation Transfer Function for Speech Intelligibility," *PLoS Computational Biololgy*, vol. 5, no. 3, pp. e1000302, 2009.

[3] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing Speech from Human Auditory Cortex," *PLoS Biology*, vol. 10, no. 1, pp. e1001251, 2012.

[4] J. F. 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.

[5] 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 *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4681–4684.

[6] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition," *Computer Speech and Language*, vol. 27, no. 3, pp. 763–779, 2013.

[7] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun, "Toward a Practical Implementation of Exemplar-Based Noise Robust ASR," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, 2011, pp. 1490–1494.

[8] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[9] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations*, Wiley, 2009.

[10] D. Wang and J. Tejedor, "Heterogeneous Convolutive Non-Negative Sparse Coding," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012.

[11] A. Hurmalainen and T. Virtanen, "Acquiring Variable Length Speech Bases for Factorisation-Based Noise Robust Automatic Speech Recognition," (to be published).

[12] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012.

[13] 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 *Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME)*, Florence, Italy, 2011, pp. 24–29.

[14] R. Vipperla, S. Bozonnet, D. Wang, and N. Evans, "Robust Speech Recognition in Multi-Source Noise Environments using Convolutive Non-Negative Matrix Factorization," in *Proceedings of CHiME workshop*, Florence, Italy, 2011, pp. 74–79.

[15] J. F. Gemmeke and T. Virtanen, "Artificial and Online Acquired Noise Dictionaries for Noise Robust ASR," in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2012, pp. 2082–2085.

[16] 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.

[17] 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 *Proceedings of ICASSP*, Vancouver, Canada, 2013.

[18] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Detection, Separation and Recognition of Speech From Continuous Signals Using Spectral Factorisation," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 2649–2653.