LEARNING STATE LABELS FOR SPARSE CLASSIFICATION OF SPEECH WITH MATRIX DECONVOLUTION

Antti Hurmalainen, Tuomas Virtanen

Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland

ABSTRACT

Non-negative spectral factorisation with long temporal context has been successfully used for noise robust recognition of speech in multi-source environments. Sparse classification from activations of speech atoms can be employed instead of conventional GMMs to determine speech state likelihoods. For accurate classification, correct linguistic state labels must be assigned to speech atoms. We propose using non-negative matrix deconvolution for learning the labels with algorithms closely matching a framework that separates speech from additive noises. Experiments on the 1st CHiME Challenge corpus show improvement in recognition accuracy over labels acquired from original atom sources or previously used least squares regression. The new approach also circumvents numerical issues encountered in previous learning methods, and opens up possibilities for new speech basis generation algorithms.

Index Terms— Automatic speech recognition, noise robustness, non-negative matrix factorization, sparse classification

1. INTRODUCTION

Conventional automatic speech recognition (ASR) systems have typically been based on calculating the likelihoods of short-time frame spectra for phonetic state models derived from training speech. For clearly articulated speech in good conditions, representations such as mel-frequency cepstral coefficients (MFCCs) of short-term spectra suffice for classification and recognition. However, especially additive, non-stationary noise rapidly corrupts the immediate spectrum to the extent that simple compensation methods cannot restore the features adequately. With multiple active sources such as competing speakers and background noise, the problem often becomes ill-posed for a single-channel input.

In recent years, noise robust ASR has gained increasing attention due to its high relevance for everyday applications. Several alternative enhancement and recognition approaches and their combinations have been proposed [1]. In single-channel recognition, a common trend is increasing the temporal context of observation windows in order to recognise characteristic spectro-temporal patterns of sources from mixtures. A context of 100+ milliseconds has been used e.g. in longest segment matching [2], deep belief networks [3] and spectrogram separation via *non-negative matrix factorisation* (NMF) [4, 5, 6]. Combination of multiple long-context algorithms has recently produced state-of-the-art results in the 2nd CHiME separation and recognition challenge [7, 8].

Two different recognition methods have been demonstrated for NMF-based modelling. The common approach is using speech and noise spectrogram estimates as an enhancement filter for an external back-end. Alternatively, the activations of speech dictionary atoms themselves provide clues about the content of observed speech [9]. Deriving the phonetic state likelihoods directly from activations is dubbed (non-negative) *sparse classification* or *sparse coding*, (N)SC. The method could be considered an extension of early template matching algorithms. A major benefit of the approach is its ability to bypass the GMM block of conventional back-ends. A large part of recognition can be conducted already within the factorisation framework. We have observed ASR results favouring either feature enhancement (FE) or sparse classification depending on the task, dictionary size and back-end training [4, 6]. Furthermore, FE and SC streams have been found to complement each other [10] with further gains from combination to neural networks [5, 8]. Outside ASR, sparse classification has been used e.g. for face recognition [11], music genre classification [12] and speaker identification [13].

In this work we address a persistent issue in sparse classification, namely translating the activation weights of speech atoms into state likelihoods. We have previously seen that in some cases the correspondence between atoms and linguistic states is not entirely straightforward. As one solution, ordinary and partial least squares regression (OLS, PLS) were used to learn the mappings [14]. Especially for small template dictionaries, such mappings produced improvements in accuracy over original atom labels [6]. However, applying common linear regression methods to NMF/SC data involves numerical issues affecting both accuracy and complexity of learning and recognition. Therefore we propose a new learning algorithm based on the same non-negative modelling that is used for separation of spectrograms. Experiments on the 1st CHiME Challenge data show uniform improvements in recognition rates using SC, while the model also fits inherently better to the separation framework.

In Section 2 we introduce briefly the underlying factorisation model and formulae employed in the new algorithm. Previous and proposed learning methods are given in Section 3. The experimental set-up on CHiME data is described in Section 4. Results, discussion and conclusions follow in Sections 5 and 6.

2. CONVOLUTIVE NON-NEGATIVE MODELLING

The fundamental concept in non-negative spectrogram factorisation is modelling a $B \times T_{\text{utt}}$ spectrogram matrix $\mathbf{Y}$ as an additive combination of shorter, $B \times T$ atoms over the utterance's duration. Here B is the number of spectral bands in all data, T_{utt} is the total number of frames in an utterance, and T is the number of frames in an atom. We have used two methods for modelling observations where $T_{\text{utt}} > T$; a *sliding window* method where overlapping window spectrograms of length T are picked from $\mathbf{Y}$ and factored individually, and *non-negative matrix deconvolution* (NMD, or *convolutive NMF*, CNMF) where the whole observation is modelled jointly by all activations [15, 16]. The focus of this work is on NMD as it has shown more promise for compact and adaptive models required in real world applications [6].

T. Virtanen has been funded by the Academy of Finland, grant #258708.

Let us define a *matrix convolution operator* $\circledast$, which produces a $B \times T_{\text{utt}}$ spectrogram estimate $\mathbf{\Psi}$ from an $L \times W$ *activation matrix* $\mathbf{X}$ and a *basis array* $\mathbf{A}$ which contains the $B \times T$ spectrograms of L atoms. Its arrangement may vary depending on the implementation so we treat it as an abstract data container. $W = T_{\text{utt}} - T + 1$ is the number of *window indices* in a convention where all occurrences of atoms are expected to fit entirely within the output spectrogram matrix. A commonly used formulation [15, 16] is

$$\mathbf{X} \circledast \mathbf{A} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}} \tag{1}$$

where each $\mathbf{A}_t$ contains the t^{th} frames of atoms in a $B \times L$ matrix and operator $\rightarrow$ shifts $\mathbf{X}$ right in a $L \times T_{\text{utt}}$ zero-padded matrix starting from its leftmost position. An alternative formulation,

$$\mathbf{X} \circledast \mathbf{A} = \sum_{l=1}^{L} \sum_{t=1}^{T_l} \mathbf{A}_{l,t} \overset{\rightarrow(t-1)}{\mathbf{X}}_l \tag{2}$$

is used especially for variable-length atoms where atom duration T_l may vary between atom indices l [17, 18]. In this case $\mathbf{A}_{l,t}$ is the t^{th} frame column vector of atom l and $\mathbf{X}_l$ the l^{th} row vector of $\mathbf{X}$. The number of eligible window indices is set for each atom separately to $W_l = T_{\text{utt}} - T_l + 1$.

Given a fixed basis array $\mathbf{A}$, the activations are obtained as

$$\mathbf{X}_{\text{opt}} = \arg\min_{\mathbf{X}} \left[d(\mathbf{X} \circledast \mathbf{A}, \mathbf{Y}) + c(\mathbf{X}) \right] \tag{3}$$

for a spectral distance measure d and an (optional) activation cost function c. Generalised Kullback-Leibler divergence and L_1 sparsity cost have been commonly employed in speech separation. All matrices are assumed non-negative. Iterative algorithms for solving the problem are presented in literature [15, 16, 17]. In our separation task, the basis is concatenated from speech and noise atoms ($\mathbf{A}^{\text{s}}$, $\mathbf{A}^{\text{n}}$), which correspondingly produce speech and noise activations $\mathbf{X}^{\text{s}}$, $\mathbf{X}^{\text{n}}$. Single-source estimates from equations (1,2) can be used for spectral separation, or the speech half $\mathbf{X}^{\text{s}}$ of $\mathbf{X}_{\text{opt}}$ itself for sparse classification.

In our SC approach, speech is recognised by finding a $Q \times T_{\text{utt}}$ *state likelihood estimate matrix* $\hat{\mathbf{Z}}$ giving the likelihoods of all Q linguistic states of the system for each frame. The matrix is decoded with a pre-determined language model and Viterbi algorithm exactly as if the likelihoods were acquired from GMM evaluation of frame features. We generate the matrix from speech activations $\mathbf{X}^{\text{s}}$ as

$$\hat{\mathbf{Z}} = \mathbf{X}^{\text{s}} \circledast \mathbf{B} \tag{4}$$

where array $\mathbf{B}$ contains a $Q \times T$ *label matrix* for each atom, reflecting the atom's correspondence to linguistic states over its duration. Notably, the model is the same as in spectrogram estimation, only with linguistic states replacing spectral bands as the feature vector dimension. The problem to be solved is finding the ideal label content for $\mathbf{B}$ to maximise recognition accuracy.

3. FINDING LABEL MATRICES FOR ATOMS

All label acquisition methods rely on having transcribed training data. Forced alignment with a conventional GMM-HMM recogniser is used to find state sequences for training files, assigning each utterance frame index τ to one state q_τ of the linguistic model. We can represent the same state information as a binary state matrix $\mathbf{Z}$ ($Q \times T_{\text{utt}}$), whose each column vector $\mathbf{z}_\tau$ has a single '1' entry at index q_τ, while the rest is zeros. This is also the ideal likelihood matrix whose decoding would produce perfect recognition results.

3.1. Earlier methods

In speech basis generation methods where *exemplars* are sampled directly from training utterances, the easiest approach is to observe which spectral frames of $\mathbf{Y}$ were used for the atom, and to select the corresponding T columns of $\mathbf{Z}$ as the label matrix. The matrix will be binary, and we call the method *canonical* or *source* mapping. If atoms are acquired by averaging multiple training segments, the label matrix is similarly averaged from their $\mathbf{Z}$ ranges.

However, especially in word-based state systems, the same spectral features may match multiple states, and more generally the correspondence to states is not strictly binary. Therefore we have proposed using least squares regression for learning the label matrices, given the atom activations of training speech [14]. Each frame column $\mathbf{b}_t$ of label matrices in $\mathbf{B}$ is determined by factoring labelled training files and then solving a least squares problem between activations and columns of $\mathbf{Z}$ with a $t-1$ frame delay. Improvements over source mapping were observed [6, 14], but the method has its drawbacks. First, least squares is a free-signed algorithm which occasionally produces negative state likelihoods to $\mathbf{B}$, violating our non-negative activation and likelihood model. Second, its temporal model was accurate for sliding window factorisation but becomes inaccurate in convolutive modelling. Finally, for large bases and especially partial LS, large-scale numerical operations arise which become inconveniently slow. Therefore it would be favourable to find a more accurate and preferably computationally faster learning algorithm for the labels.

3.2. Proposed NMD learning

Interestingly, efficient mappings can be found by slightly unusual application of the same matrix deconvolution algorithm that is used for spectrogram factorisation. While in conventional NMF/NMD we are primarily learning the activations and possibly some of the atoms, label learning does the opposite. First, we factor clean training utterances using a fixed speech basis and thus gain activation matrices $\mathbf{X}$ for each utterance. Then, we switch the target observation from spectrogram matrix $\mathbf{Y}$ to the ideal state likelihood matrix $\mathbf{Z}$ and solve the new basis of atom labels alone as

$$\mathbf{B}_{\text{opt}} = \arg\min_{\mathbf{B}} d(\mathbf{X} \circledast \mathbf{B}, \mathbf{Z}) \tag{5}$$

over all $\mathbf{X}$/$\mathbf{Z}$ pairs for a single label array $\mathbf{B}_{\text{opt}}$, which will be our mapping data. The usual iterative solving methods apply, only with activations being fixed and the basis updated in minimisation. The label array can be first initialised e.g. to ones or randomly.

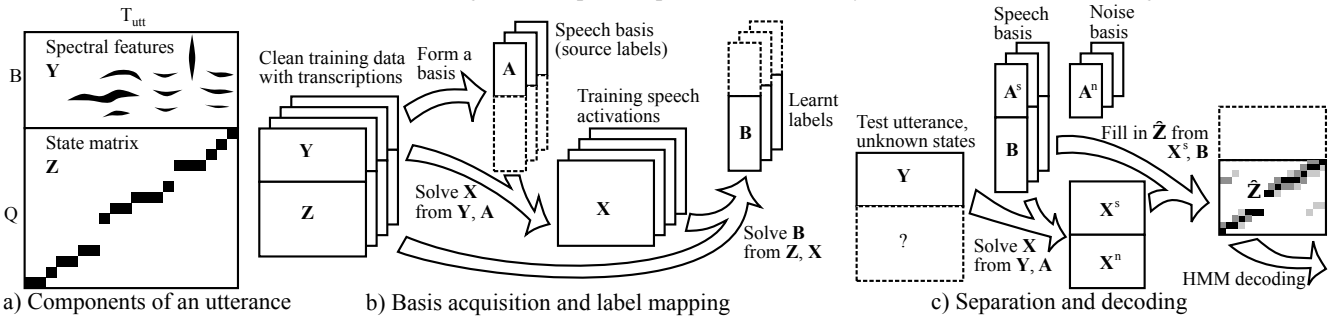
In conventional unsupervised or semi-supervised NMF/NMD learning tasks it is customary to normalise basis atoms between iterations to stabilise convergence speed and sparsity between atoms. In label learning there is better justification for not normalising the atoms, hence allowing the algorithm to allocate varying amounts of state content to label matrices. Atoms with highly ambiguous correspondence to states receive less classifying power, while atoms with consistent behaviour get more significance.

The distance measure d can be chosen equivalently to spectrogram factorisation, although it is not compulsory. Whereas KL-divergence has been found useful for emphasising small spectral details of sound [19], our target label data is strictly binary so the requirement does not apply. In addition, the KL measure

$$d_{\text{KL}}(z_i, \hat{z}_i) = z_i \log \frac{z_i}{\hat{z}_i} - z_i + \hat{z}_i \tag{6}$$

is not well defined for zeros which dominate $\mathbf{Z}$. Therefore another measure like Euclidean distance d_{Euc} may be preferable for states.

Fig. 1. Data structures and algorithm steps of a sparse classification system with NMD label learning.



The data structures and algorithm steps of the overall SC framework are illustrated in Figure 1. The leftmost segment a) shows how each utterance is represented by a spectrogram and a corresponding state content matrix. Segment b) shows how training files are used for basis acquisition and training factorisation, whereafter label matrices are solved by optimising $\mathbf{B}$. In segment c) state likelihoods are found for a test utterance by factoring it with speech and noise bases, and computing $\hat{\mathbf{Z}}$ from speech activations and atom labels. Final recognition is performed by decoding $\hat{\mathbf{Z}}$ with HMM probabilities of the language model.

4. EXPERIMENTS

4.1. Corpus and recognition task

Different label assignment methods were compared using the 1st CHiME Challenge task [20]. In its default language model, 250 subword states are used to represent a small 51-word vocabulary originating from the GRID corpus [21]. Utterances follow a linear *verb-colour-preposition-letter-digit-adverb* grammar with classes having cardinalities of 4, 4, 4, 25, 10 and 4, respectively. Performance is scored by correct recognition of 'letter' and 'digit' keywords.

Due to the word-based model, there are several states whose spectral features are effectively identical. For example, words 'please' and 'place' bear high similarity in most of their states. Even more crucially, many of the letter name keywords ('b', 'c', 'd' etc.) only differ in one phone, which produces several spectrally similar states across words. Consequently there is a considerable chance of misclassification if too strict assignment to states is used for atoms with partially ambiguous spectral content.

The 1st CHiME corpus contains 500 training utterances for each of its 34 speakers. These have no additive noise. Evaluation is carried out on 600-utterance development and test sets, which have additive, highly non-stationary room noise at SNRs ranging from +9 to -6 dB in 3 dB steps. The development set is also available as 'clean', that is, no added noise. All audio has room reverberation.

4.2. Speech bases

For comparison of labelling methods, we used factorisation results acquired previously with two different compact speech models. In the first one, a 250-atom basis is generated for each speaker by modelling spectral features of one state and its neighbouring context at a time. In this model, window length is fixed to 25 frames [6]. More recently, a heuristic clustering algorithm has been used to generate variable-length bases, where the number of atoms and the distribution of atom lengths is allowed to vary between speakers [18]. Here we use the basis variant where combined label and spectrum data was used for finding recurring segments, because it appeared to produce the best sparse classification results of studied variable-length methods [18]. The average size of these bases is 182 atoms, whose average length is 22.2 frames.

4.3. Factorisation model

The factorisation framework was otherwise identical for both basis types apart from the variable-length convolution model required by the second type. CHiME audio was converted into 40-band mel-spectral monaural features. Factorisation was performed with speaker-dependent speech bases matching each target speaker. For noisy development and test utterances, 250 fixed-length (25 frames) noise atoms were extracted from the noise context of 'embedded' utterances. All activation matrices used in these mapping experiments come directly from the previously presented studies with no re-factorisation. The details of factorisation parameters are given in the original articles [6, 18].

For each speaker, 300 training utterances were used to construct the basis, while the other 200 were factored to get activation matrices for learning the mappings.

4.4. Acquiring labels

Three label assignment methods were compared:

1. *Source mapping*, where each label matrix is averaged from the state matrices of training segments which were used to generate the atom.
2. *Ordinary least squares* (OLS) mapping, learnt with regression between activation matrices and transcriptions of 200 factored training utterances [14].
3. *Convolutive learning* with the proposed algorithm described in Section 3.2 and the same subsets of training utterances as in OLS.

The convolutive algorithm was further tested with Kullback-Leibler ('NMD/KL') and Euclidean ('NMD/Euc') criteria for similarity. For each basis and learning method, the mapping matrices $\mathbf{B}_l$ had identical length T_l to their corresponding atoms, fixed or variable depending on the experiment. The number of NMD iterations in convolutive learning was determined by observing performance on the development set. As in earlier work, final likelihoods for test utterances were generated by computing $\hat{\mathbf{Z}} = \mathbf{X}^{\mathrm{s}} \circledast \mathbf{B}$ matrices, which were then decoded using the default CHiME HMMs.

5. RESULTS AND DISCUSSION

5.1. Parameters for learning

Before moving on to final evaluation, two algorithmic choices regarding NMD learning were considered; first, which distance measure to use and second, how many iterations are required. These were grid-scanned by generating mappings for multiple parameter combinations and calculating keyword recognition rates for noisy development data. Results are shown in Figure 2. Average accuracy over the six SNRs is plotted as a function of learning iterations. Four combinations of fixed/variable length bases and KL/Euclidean distance are plotted as separate curves.

We notice that the highest recognition accuracy is reached in only a few iterations, in stark contrast to supervised spectral separation experiments where hundreds of iterations still improved the results [10]. A likely reason is that the binary target matrices lead to rapid convergence, whereafter overlearning may take place due to the small size of training sets, e.g. on average just eight instances of each 'letter' keyword in the 200 utterances. Similar trends have been observed in NMD-based speech and noise basis learning, where a low iteration count prevents overadaptation and yields better results than letting the minimisation algorithm converge completely [22, 23, 24].

Unfortunately more detailed conclusions are difficult to determine because the different parameters show varying and often contradictory behaviour. No truly consistent trend takes place in the iteration count. Eventually we settled for using four iterations for all methods. Further iterations only increased the computation time with no observable effect on the combined recognition accuracy. Regarding distance measures, for fixed length bases both choices perform equally well, but in variable length modelling KL-divergence falls below all other combinations. Exact reasons for this differing behaviour are not well known yet. Nevertheless, both basis types and distance measures were included in final evaluation.

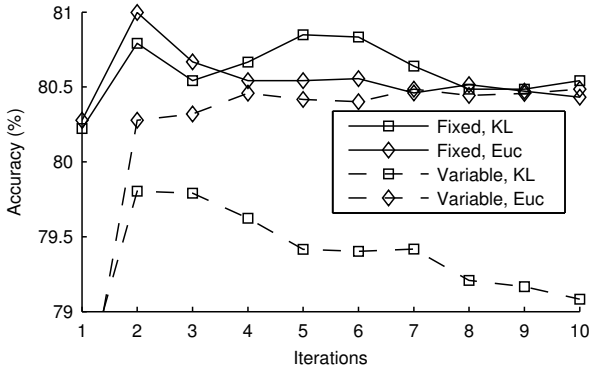
5.2. Test set results

The final results for different mapping algorithms are shown in Tables 1 and 2. Four labelling methods are compared: direct assignment from atom sources ('source'), ordinary least squares regression ('OLS') [14], and proposed NMD learning with KL and Euclidean distance measures ('NMD/KL', 'NMD/Euc'). Results for fixed-length speech modelling are given in Table 1, and for variable-length modelling in Table 2. In each case, recognition accuracies are listed for each SNR level of development and test sets, and as averages over noisy conditions. For each set and condition, the best result among different mapping methods is highlighted. However, note that clean and single-SNR sets comprise only 600 utterances and 1200 keywords each, thus each hit or miss accounts for roughly 0.1% difference. Noisy averages, containing six times more trials, can be considered more reliable.

5.3. Discussion

In general the proposed methods — especially with Euclidean distance — yield consistent improvement over using the atom source transcriptions as labels and in most cases also over OLS mapping. For a comparison, the average results using 20 times larger exemplar bases were 85.9 and 85.8% for development and test sets, respectively [6]. Therefore gaining up to 3.5% absolute improvement in test set accuracy with compact bases just by better interpretation of the same activations can be considered significant.

Fig. 2. Recognition accuracy (average of noisy development data) over the number of NMD iterations used for learning the mappings. Results are shown separately for fixed and variable length bases, and KL and Euclidean distance measures.



While determining strictly optimal labelling would be effectively impossible, we can speculate that the proposed methods are getting close to the limits set by factorisation output. The convolutive mapping model closely reflects the activation pattern of NMD factorisation, hence the significance of different atoms and atom-frames in separation and classification can be captured accurately. The development graphs in Figure 2 suggest that with the selected training data, the system is already prone to overlearning even though a closer absolute match to training transcriptions could be found by further iterations. Similar behaviour has also been seen in basis learning with NMD [22, 23, 24]. Because early halting of the descent does not produce repeatable results across different initialisation values and optimisation algorithms, it might be worth studying if overlearning could be prevented by regularisation of the label matrix content instead.

The remaining clean speech recognition errors mostly arise from difficult 'letter' keyword pairs like 'm'/'n' and 'b'/'v', which cannot be classified reliably with the employed mel-spectral features and speech bases regardless of mapping. Noisy conditions obviously introduce their additional errors, which must be solved primarily with better noise modelling. However, as the original 1st CHiME corpus had no noisy training data, mappings were learnt from clean speech factorisation. Learning labels from multi-condition factorisation might provide improved robustness due to capturing confusion of speech atoms with noise, which was not possible in the presented experiments.

The obvious benefit of the proposed method is that its convolutive structure is similar to our spectrogram modelling, and it produces inherently non-negative mapping data which is directly applicable for decoding. Its computational complexity depends on design aspects such as data dimensions, solving algorithm, iteration count and exploitation of sparsity. Nevertheless, the problem can be easily split and solved in a small memory space, unlike least squares which typically requires matrix inversion increasing in size along dimensionality. Learning time with MATLAB prototype code was about 5 minutes per speaker on a dual-core desktop PC.

Regarding future work on SC-based speech recognition, focus should be shifted back on feature spaces, basis acquisition methods, factorisation algorithms and possibly language and state models. Interestingly, the proposed algorithm may prove helpful in optimising these components. Because NMD learning is able to allocate state weight on atoms with most consistent classification capability, it can

Table 1. Keyword recognition rates (%) for the 1st CHiME Challenge corpus using sparse classification, fixed-length speech bases, and four methods for assigning state labels to atoms. Results are shown individually for each SNR level of development and test sets, and averages over noisy conditions. For each condition, the highest result across methods is highlighted.

labelling	development set								test set						
method	clean	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	n.avg	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	n.avg
source	91.3	88.7	86.8	83.7	76.5	69.3	62.9	78.0	87.2	86.4	81.5	80.3	70.3	63.9	78.3
OLS	**91.7**	88.7	86.8	83.7	76.5	69.3	62.9	78.0	89.8	**89.0**	84.3	81.8	73.9	65.8	80.8
NMD/KL	91.5	89.3	87.9	**86.5**	79.9	**73.9**	**66.4**	**80.7**	90.6	**89.0**	85.3	**83.0**	**75.6**	67.4	**81.8**
NMD/Euc	**91.7**	**90.1**	**88.4**	85.7	**80.4**	72.8	65.9	80.5	**90.8**	88.8	**85.5**	82.6	75.4	**67.5**	**81.8**

Table 2. 1st CHiME corpus keyword recognition rates (%) using variable-length speech bases. Results are formatted equivalently to Table 1.

labelling	development set								test set						
method	clean	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	n.avg	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	n.avg
source	90.3	88.3	87.0	85.4	78.1	72.3	64.9	79.4	88.2	87.3	83.3	80.3	72.4	65.1	79.4
OLS	90.7	88.3	87.2	84.7	78.4	72.9	66.8	79.7	88.7	87.1	83.3	81.7	74.0	67.2	80.3
NMD/KL	90.2	87.5	87.8	84.6	78.3	73.2	66.4	79.6	**89.2**	87.2	83.6	80.8	74.8	66.6	80.4
NMD/Euc	**91.0**	**88.8**	**88.0**	**85.8**	**79.0**	**74.0**	**67.3**	**80.5**	**89.2**	**87.4**	**84.0**	**82.2**	**75.3**	**67.3**	**80.9**

be used to determine redundant or unreliable atoms during basis generation, and to evaluate whether classification from acquired activation weights is possible in the first place. Therefore we can predict development of more closely integrated NMD systems handling both spectral and state content simultaneously in model construction and processing of input utterances.

6. CONCLUSIONS

A method based on convolutive non-negative matrix modelling was proposed for learning linguistic state content of atoms in sparse classification of speech. Acquired labels were evaluated on the 1st CHiME noisy speech recognition task and compared to alternative state mapping algorithms. Consistent improvements were achieved in speech recognition accuracy compared to using baseline atom identity or least squares regression for determining the correspondence of speech atoms to language model states. The proposed algorithm also circumvents earlier numerical issues in learning, and facilitates closer integration of spectral and state components in a factorisation-based speech recognition framework. We expect to employ the algorithm in further experiments on standalone sparse classification and joint methods, which have been found effective in noise robust speech recognition.

7. REFERENCES

[1] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[2] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S. Hahm, and A. Nakamura, "Speech Recognition in the Presence of Highly Non-stationary Noise Based on Spatial, Spectral and Temporal Speech/Noise Modeling Combined with Dynamic Variance Adaptation," in *Proceedings of the 1st CHiME workshop*, Florence, Italy, 2011, pp. 12–17.

[3] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 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] 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, pp. 126–130.

[8] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM+TUT+KUL Approach to the CHiME Challenge 2013: Multi-Stream ASR Exploiting BLSTM Networks and Sparse NMF," in *Proceedings of the 2nd CHiME workshop*, Vancouver, Canada, 2013, pp. 25–30.

[9] T.N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J.F. Gemmeke, J.R. Bellegarda, and S. Sundaram, "Exemplar-Based Processing for Speech Recognition: An Overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, 2012.

[10] J.F. Gemmeke and H. Van hamme, "Advances in Noise Robust Digit Recognition using Hybrid Exemplar-Based Techniques," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012, pp. 2134–2137.

[11] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[12] Y. Panagakis, C. Kotropoulos, and G.R. Arce, "Music Genre Classification via Sparse Representations of Auditory Temporal Modulations," in *Proceedings of EUSIPCO*, Glasgow, Scotland, 2009, pp. 1–5.

[13] R. Saeidi, A. Hurmalainen, T. Virtanen, and D.A. van Leeuwen, "Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification," in *Odyssey speaker and language recognition workshop*, Singapore, 2012.

[14] K. Mahkonen, A. Hurmalainen, T. Virtanen, and J. Gemmeke, "Mapping Sparse Representation to State Likelihoods in Noise-Robust Automatic Speech Recognition," in *Proceedings of INTERSPEECH*, Florence, Italy, 2011, pp. 465–468.

[15] 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.

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

[17] D. Wang and J. Tejedor, "Heterogeneous Convolutive Non-Negative Sparse Coding," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012, pp. 2150–2153.

[18] A. Hurmalainen and T. Virtanen, "Acquiring Variable Length Speech Bases for Factorisation-Based Noise Robust Speech Recognition," in *Proceedings of EUSIPCO*, Marrakech, Morocco, 2013.

[19] T. Virtanen, "Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[20] 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.

[21] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[22] F. Weninger, M. Wöllmer, and B. Schuller, "Sparse, Hierarchical and Semi-Supervised Base Learning for Monaural Enhancement of Conversational Speech," in *Proceedings of ITG Conference on Speech Communication*, Braunschweig, Germany, 2012.

[23] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A Comparative Study on Sparsity Penalties for NMF-based Speech Separation: Beyond LP-Norms," in *Proceedings of ICASSP*, Vancouver, Canada, 2013, pp. 858–862.

[24] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Compact Long Context Spectral Factorisation Models for Noise Robust Recognition of Medium Vocabulary Speech," in *Proceedings of the 2nd CHiME workshop*, Vancouver, Canada, 2013, pp. 13–18.