NON-NEGATIVE MATRIX DECONVOLUTION IN NOISE ROBUST SPEECH RECOGNITION

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

[⋆] Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland
[†] Radboud University, Nijmegen, P.O. Box 9102, 6500 HC Nijmegen, The Netherlands

ABSTRACT

High noise robustness has been achieved in speech recognition by using sparse exemplar-based methods with spectrogram windows spanning up to 300 ms. A downside is that a large exemplar dictionary is required to cover sufficiently many spectral patterns and their temporal alignments within windows. We propose a recognition system based on a shift-invariant convolutive model, where exemplar activations at all the possible temporal positions jointly reconstruct an utterance. Recognition rates are evaluated using the AURORA-2 database, containing spoken digits with noise ranging from clean speech to -5 dB SNR. We obtain results superior to those, where the activations were found independently for each overlapping window.

***Index Terms*—** Automatic speech recognition, noise robustness, deconvolution, sparsity, exemplar-based

1. INTRODUCTION

Widespread adoption of Automatic Speech Recognition (ASR) systems is still being hampered by insufficient robustness against background noise. Hidden Markov Model (HMM) based recognisers, where state likelihoods are estimated using Gaussian Mixture Models (GMM), have considerable problems when noisy frames no longer match to clean acoustic models. Various robustness methods have been suggested, including model compensation, missing data techniques and feature enhancement [1, 2, 3]. These approaches can typically achieve acceptable recognition rates in low to medium noise, but lose quality rapidly, when a large portion of spectral features is simultaneously corrupted by high noise levels.

In our previous work we have shown, that improved recognition rates can be achieved near or below 0 dB SNR by using an additive model of *exemplars* representing longer (100 – 300 ms) spectrogram segments [4]. Using a Non-negative Matrix Factorisation (NMF) algorithm, it is possible to separate the input signal to speech and noise. Furthermore, we have shown that speech content can be decoded directly from the labels of activated exemplars without reconstructing the separated speech signal [5].

In contrast to earlier exemplar-based methods, where the observation is compared to the nearest element in the dictionary, our framework reconstructs observations as a non-negative linear combination of exemplars. The number of simultaneously active exemplars is not limited by the design, although sparsity is enforced to improve the recognition quality. Similar methods have been used for source separation in image and music applications, among others. Common terminology for referring to such techniques includes *Sparse Classification* (SC) and *Sparse Representation based Classification* (SRC).

Tuomas Virtanen and Antti Hurmalainen have been funded by the Academy of Finland. The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program.

While the noise robustness of our algorithm improved by using longer exemplars, we also observed a decrease in clean speech recognition rates. The primary reason for this negative development is that the complexity of spectro-temporal features will increase in longer windows, thus requiring more exemplars to cover the larger variation in appearing patterns [6]. In addition, factorisation of individual analysis windows requires that correctly time-aligned exemplars are available in the dictionary, so the number of different temporal alignments to be covered also increases according to window length. However, simultaneous increasing of both exemplar count and length is not desirable due to computational constraints.

To improve the recognition accuracy of our system using a limited dictionary of long exemplars, we introduce a shift-invariant convolutive model. By reconstructing the whole observation at once as a convolution of exemplars and activations, we avoid the problem of temporal alignment of the exemplars in fixed windows. It is no longer necessary to include multiple shifted variants of features in the exemplars to represent the observation accurately. Consequently, better efficiency can be expected for similar dictionary size.

The content is organised as follows. Section 2 describes the key concepts of the paper: exemplar-based recognition, matrix deconvolution and differences to the previous model. In Section 3 we explain, how to obtain state likelihoods and the final recognition output from exemplar activations. The noisy spoken digit recognition test setup is given in Section 4. Results, discussion, and conclusions follow in Sections 5, 6 and 7, respectively.

2. EXEMPLAR-BASED DECONVOLUTION

2.1. Windowed exemplar model

The basis unit of our system, a speech or noise *exemplar*, is a $B \times T$ spectrogram matrix consisting of spectral magnitudes (square root of energy). B is the number of frequency bands and T the number of consecutive frames in each exemplar. Our observation matrix $\mathbf{Y}_{\text{utt}}$ is a $B \times T_{\text{utt}}$ spectrogram in the same domain, where T_{utt} is the total number of frames in the whole speech utterance.

The utterance is modelled as a linear weighted combination of exemplars in overlapping, exemplar-sized *windows*. The starting frame indices τ of windows range from 1 to $W = T_{\text{utt}} - T + 1$, and a window starting from frame τ covers frames $[\tau, \tau + T - 1]$. The linear combination is characterised by an $L \times W$ activation matrix $\mathbf{X}$, where each element X_{lw} represents the weight of exemplar l (from 1 to the total number L) activation in window w. The activation pattern can be determined for one window at a time as in our previous experiments [4, 5], or by generating joint activations for the whole utterance using a *deconvolution* algorithm.

2.2. Matrix deconvolution

The estimated model $\mathbf{\Psi}_{\text{utt}}$ for observation $\mathbf{Y}_{\text{utt}}$ using L exemplars can be written as

$$\mathbf{\Psi}_{\text{utt}} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}. \tag{1}$$

Each $\mathbf{A}_t$ is a $B \times L$ matrix representing frame t of the exemplars, thus the spectrogram of exemplar l can be found in columns l of $\mathbf{A}_1 \dots \mathbf{A}_T$. Here $\overset{\leftarrow i}{(\cdot)}$ and $\overset{\rightarrow i}{(\cdot)}$ are shift operators, moving the matrix entries left or right, respectively, by i units. In this case $\mathbf{\Psi}_{\text{utt}}$ is $T-1$ columns longer than the activation matrix $\mathbf{X}$, so shifting takes place in a T_{utt} wide zero-padded matrix, starting from its leftmost position. $T-1$ zero columns are added, no columns are discarded to generate the shifted matrix.

The exemplars and their activations are restricted to non-negative values. The exemplars are obtained from training data and fixed, whereafter the activations are estimated by minimising the generalised Kullback-Leibler divergence

$$d(\mathbf{Y}_{\text{utt}}, \mathbf{\Psi}_{\text{utt}}) = \sum y \log(\frac{y}{\psi}) - y + \psi \quad \forall (y, \psi) \in (\mathbf{Y}_{\text{utt}}, \mathbf{\Psi}_{\text{utt}}). \tag{2}$$

An L_1 norm penalty (sum of all elements) is applied to the activations, which has been found effective for magnitude spectrogram features [7].

As the approximated observation matrix $\mathbf{\Psi}_{\text{utt}}$ will be a temporal convolution between the basis and the activations, the algorithm is called Non-negative Matrix Deconvolution (NMD) [8]. In our previous work we called the method *convolutive sparse coding* [9]. NMD has already been used successfully for sound source separation in music and speech applications [10, 11].

The entries of the activation matrix are initialised to unity values, and the following update rule (based on [12]) is applied iteratively:

$$\mathbf{X} = \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t^T \cdot \overset{\leftarrow(t-1)}{[\frac{\mathbf{Y}_{\text{utt}}}{\mathbf{\Psi}_{\text{utt}}}]}}{\mathbf{\Lambda} + \sum_{t=1}^{T} \mathbf{A}_t^T \cdot \overset{\leftarrow(t-1)}{\mathbf{1}}}, \tag{3}$$

where $\otimes$ is elementwise multiplication, and all divisions are also elementwise. $\mathbf{\Lambda}$ is a sparsity matrix defining the penalty factor for each activation element, thus the total weighted penalty becomes $\sum x \cdot \lambda \quad \forall (x, \lambda) \in (\mathbf{X}, \mathbf{\Lambda})$. In our system, we set a different penalty weight for activations corresponding to speech and noise. The model $\mathbf{\Psi}_{\text{utt}}$ is evaluated before each update using (1).

2.3. Comparison to independent windows

In our previous work we used a sliding window approach, where all W overlapping $B \times T$ windows were factorised independently. Because the middle frames of the observation will be reconstructed several times in consecutive windows, averaging was applied in later steps to compensate for the effect. The implementation was somewhat simpler than in NMD — each window can be represented as a separate, concatenated observation vector, and the utterance can be processed as a factorisation between two matrices without shifting operations. However, it occasionally suffers from the fixed temporal positioning of its windows. An exemplar must match accurately to the temporal position of spectral features found in an individual window to be used there. When the window length is increased, it becomes less likely, that a matching exemplar is found in a limited dictionary. Each window must be factorised, and depending on its match to the dictionary, reconstruction quality may vary between windows. The effect of mismatches will be reduced during averaging, but not eliminated entirely. On the other hand, for NMD it suffices to find a single temporal position, where an exemplar matches the observed speech. The difference between the activation patterns is visualised in Figure 1.

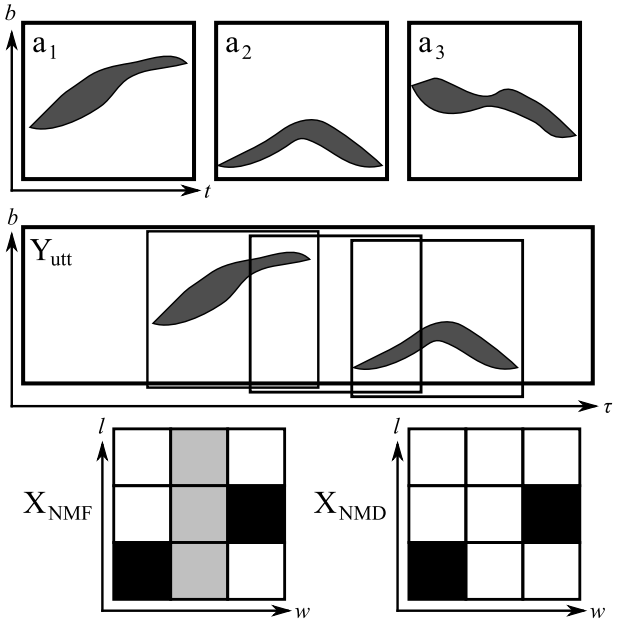

Figure 1: A stylised comparison of independent window (NMF for short) and deconvolution (NMD) methods. Utterance spectrogram $\mathbf{Y}_{\text{utt}}$ is represented using exemplars $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ in three windows. The first and last window match to exemplars 1 and 2, but in NMF the middle window must be reconstructed using inaccurate activations (bottom left matrix). In NMD, only enough exemplars to reconstruct the utterance are activated (bottom right matrix), thus the middle window remains empty.

3. DECODING

After determining the activation matrix $\mathbf{X}$, it is used to generate a state likelihood matrix $\mathbf{L}$. It consists of column vectors $\mathbf{l}_\tau$ for each frame in the utterance. These vectors, their length representing the total number of states in the system, describe the estimated likelihoods of states at time τ.

Each speech exemplar is labelled with a state sequence over its duration, so that in each frame it is assumed to be in exactly one state. When an exemplar is activated in window w, an update is made to T columns of $\mathbf{L}$ starting from w. A state label q in frame t of an exemplar will increment the element q of column $w+t-1$ by its activation weight. A formal description of this procedure is given in [5].

Even though silence states are also included in the labels, their activation is somewhat unpredictable. Because the magnitude of silent frames is zero in all bands, no exemplars are activated during true silence. Conversely, these states may appear within speech activity, when a speech-silence transition exemplar is used as a part of the sum. For these reasons, silence state likelihoods are reshaped according to a speech activity estimate derived from the total weight of active speech exemplars in each frame. The matter is discussed in

Table 1: Digit recognition rates for AURORA-2 test sets A and B at various window lengths and noise levels. The first three rows repeat the independent window factorisation ('NMF' for short) results given in [5]. The last three rows show the new deconvolution results ('NMD').

SNR	(dB)	clean	20	15	10	5	0	-5
NMF	T=10	96.2	95.3	94.4	92.1	84.7	71.2	39.6
	T=20	96.6	95.8	94.8	92.7	88.8	78.1	53.1
	T=30	94.7	93.4	93.3	92.2	89.9	79.5	56.7
NMD	T=20	96.7	96.3	95.4	93.9	90.1	78.5	57.5
	T=30	97.0	96.4	95.6	94.7	91.4	82.0	61.0
	T=40	93.5	94.4	94.2	91.5	88.6	78.3	55.2

(a) Test set A

SNR	(dB)	clean	20	15	10	5	0	-5
NMF	T=10	96.2	94.7	93.6	87.9	78.4	57.1	27.4
	T=20	96.6	95.3	93.7	89.9	82.7	63.1	35.7
	T=30	94.7	93.5	93.2	90.1	85.7	67.5	37.6
NMD	T=20	96.7	96.0	95.1	91.7	84.0	62.4	33.5
	T=30	97.0	95.6	94.7	92.1	86.4	68.1	36.4
	T=40	93.5	93.8	93.4	89.2	83.6	64.1	33.0

(b) Test set B

more detail in [4].

Finally, the summed likelihoods in each frame are normalised to unity, and the state likelihood matrix is decoded using the Viterbi algorithm.

4. EXPERIMENTS

The efficiency of deconvolution versus independent overlapping windows was studied using a test setup similar to the one described in [4] and [5]. AURORA-2 connected digit recognition test, which includes multiple noise types and noise levels, was used for evaluation. The same bases of 4000 speech and 4000 noise exemplars, generated by random selection from the multicondition training set in the earlier experiments, were used. In these bases, each exemplar is a $B \times T$ magnitude spectrogram consisting of 23 mel-scale spectral bands and T frames with 25 ms frame length and 10 ms frame shift. Window lengths 20 and 30 from the previous work were included, as this much temporal context has been found recommendable for sufficient noise robustness. In addition, a $T = 40$ basis was generated using a similar procedure to study the capability of deconvolution in even longer windows. State labels of speech exemplars were acquired via HMM-based forced alignment. All in all, 179 states were used: 16 for each digit ('zero', 'oh', 1–9) and 3 for silence.

We processed the same random subset of 100 utterances (10% of the complete test set) as in [5] for all four noise types in test set A and the four in test set B. Clean speech and all six noise levels, SNR 20, 15, 10, 5, 0 and -5 dB were included. Due to the different activation patterns between independent windows and deconvolution, the NMD sparsity parameters λ were reoptimised to 2.0 for speech and 1.5 for noise exemplars using the training set. The silence balancing algorithm was modified slightly to derive its SNR estimate from waveforms by comparing the mean power of the whole wave (signal+noise) to the lowest 20% of frame powers (only noise), because in NMD the exemplar activation levels were found to vary too much for this purpose. The silence parameters were retrained from the training set for each window length separately. 200 NMD iterations were used for the main experiment as before, although the computation was continued up to 250 iterations for further comparison.

5. RESULTS

The recognition rates of our test are summarised in Table 1. Previous results from our independent window experiments ('NMF') are shown first, sorted by window length [5]. The new convolutive model ('NMD') results follow.

In set A, convolutive $T = 30$ comes out uniformly superior to the alternative window lengths and also to our previous results. Convolutive $T = 20$ surpasses the NMF results and approximately equals convolutive $T = 30$ at high SNRs, but falls faster in the noisy end like it did in NMF. The newly introduced $T = 40$ (400 ms exemplars) is roughly comparable to the previous $T = 20/30$ NMF results. However, a decrease of approximately 3% from convolutive $T = 30$ is present already in the clean end, and it reflects to all the noisy rates. Overall, set A turns out to be a success for the convolutive algorithm.

In set B we observe mostly positive results, but also a few decrements. The improvements in clean speech recognition rates are also present here all the way until 0 dB, where convolutive $T = 20$ loses by a small margin to its NMF counterpart. For $T = 30$, this happens at -5 dB alone. $T = 40$ is again acceptable in comparison to the NMF results, but several percent below the new $T = 30$ rates.

The high contrast between set A (noise types matching to the basis) and set B (nonmatching noise) is still present and even emphasised in the convolutive approach. The possible reasons for this are discussed in Section 6.

Increasing the iteration count to 250 produced mixed results (not shown). Recognition rate changes between -1.4% and +3.7% (absolute) were observed. The largest and most systematic gains were in the noisy end of set A, all 0 dB rates increasing by $\geq 1.0\%$ and -5 dB by $\geq 2.2\%$. Elsewhere no regular trend was found.

In comparison to established methods, the current experimental setup does not yet achieve the clean speech recognition rates of carefully trained GMM-based implementations, which often exceed 99%. On the other hand, previous -5 dB rates achieved with noise-compensated or multi-condition trained GMMs include 17.1% [2], 24.6% [13] and 42.9% [4] for set A. All perform worse on set B, albeit by a smaller margin, when the methods do not utilise spectro-temporal features specific for each noise type. Uncompensated systems trained with clean speech typically fall below 10% at -5 dB.

6. DISCUSSION

Three main observations can be made from the results. First, in this test setup the convolutive method produces generally higher recognition rates than the independent window algorithm. Second, convolutive $T = 30$ achieves the highest clean speech recognition rate of all methods and windows presented here, improving significantly its earlier independent window performance. Third, test set B still turns out problematic, even more so than in NMF. Each of these observations deserves a brief analysis.

The improved overall rates are a positive outcome, and speak for the potential of NMD in exemplar-based recognition. However, the new algorithm also required some changes and retraining of parameters, which may play a role in the overall results. We still conclude, that significant gains were achieved by using NMD for the problem.

Because its joint, shift-invariant activation pattern appears inherently suitable for dictionary reduction and reverberation handling, we consider it the better candidate for further research within related topics, such as echoing noise and large vocabulary.

The second observation was the superiority of $T = 30$. Whereas in the previous independent window experiment it suffered from lower clean speech recognition rates, here it improves to the extent that it surpasses both of the $T = 20$ variants in all SNRs. It was our earlier assumption, that in such a long window the dictionary size becomes a limiting factor for independent windows, because several temporal alignments of features are required in the exemplars. We also assumed, that deconvolution might reduce the effect. The results support both of these theories. As the $T = 30$ basis was identical in both variants, and post-processing factors are negligible in clean speech recognition, we conclude that the nearly halved error rate in clean speech results from algorithmic differences. The other high percentages in set A follow the improved performance of clean speech throughout the noise levels. Window length 40 was found too large to be handled with this dictionary size, regardless of the use of convolution.

The primary problem of our current approach is highlighted by the third observation, namely the increasing quality gap between sets A and B. The noise types of set A are similar to those used in training and dictionary construction. Therefore the factorisation/deconvolution becomes a well defined separation problem, and generally plausible results can be achieved. The situation is notably different in test set B. Because the noise types do not match, especially in long windows we cannot expect to find good approximations for the observed noise in the dictionary. In NMF of independent windows, a lot of averaging will take place. Up to 30 different noise estimates from consecutive windows are mixed together. Therefore they are unlikely to form any major distracting features. In NMD, this kind of forced averaging is not present. The increased sparsity, which aided separation in set A, may become a hindrance instead. Sparse activations of nonmatching noise features are not suitable for representing the true noise in signals, thus the separation often fails. A telling detail is that in set A the noisy results improved further by increasing the iterations to 250. In set B this did not happen. Even a few decrements took place, suggesting that the algorithm had already reached an unstable peak level regarding separation quality.

It has been repeatedly seen that long temporal context is effective, or even required for handling high levels of background noise. We also found here additional support for the potential of exemplar-based sparse representation. However, while various speech patterns can be handled by a reasonably sized exemplar dictionary, the same cannot be said about all types of noise present in the real world. To cope with this issue, we have already taken initial steps towards adaptive and synthetic noise dictionaries [14]. Preliminary results show that even a simple synthetic dictionary can surpass the separation quality of a poorly matching sampled dictionary. Deconvolution should prove useful in such dictionary methods, because new patterns can be included as single entries without temporal repetition. The algorithm itself will take care of different temporal alignments.

7. CONCLUSIONS

A framework for an exemplar-based, deconvolutive speech recognition system was presented. Comparative results against an earlier setup with independent factorisation windows were shown using the AURORA-2 connected digit recognition test. Deconvolution with a window length of 30 frames (300 ms) surpassed the results of other window lengths and the previous approach almost uniformly. Recognition rates of $>80\%$ were observed at 0 dB SNR, and $>60\%$ at -5 dB. Improvements in clean speech recognition rates using long windows suggest, that deconvolution can overcome some of the dictionary size problems of independent windows. It turned out that the match between the dictionary and observed noise is crucial in deconvolution, even more so than in the independent window approach.

8. REFERENCES

[1] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proceedings of ICSLP*, 2000, pp. 869–872.

[2] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proceedings of ICASSP*, 2004, pp. 213–216.

[3] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, September 2005.

[4] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Accepted for publication in IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[5] T. Virtanen, J.F. Gemmeke, and A. Hurmalainen, "State-based labelling for a sparse representation of speech and its application to robust speech recognition," in *Proceedings of INTERSPEECH*, 2010.

[6] J. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proceedings of EUSIPCO*, 2009, pp. 1755–1759.

[7] 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, 2007.

[8] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, pp. 494–499. 2004.

[9] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[10] P.D. O'Grady and B.A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, pp. 88–101, 2008.

[11] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, 2007.

[12] T. Virtanen, *Sound source separation in monaural music signals*, Ph.D. thesis, Tampere University of Technology, 2006.

[13] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ISCA Tutorial and Research Workshop*, 2000, pp. 181–188.

[14] J.F. Gemmeke and T. Virtanen, "Artificial and online acquired noise dictionaries for noise robust ASR," in *Proceedings of INTERSPEECH*, 2010.