Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition

Antti Hurmalainen[a,*], Jort F. Gemmeke[b], Tuomas Virtanen[a]

[a] *Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101, Tampere, Finland*
[b] *KU Leuven, Department ESAT-PSI, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium*

Abstract

Speech recognition systems intended for everyday use must be able to cope with a large variety of noise types and levels, including highly non-stationary multi-source mixtures. This study applies spectral factorisation algorithms and long temporal context for separating speech and noise from mixed signals. To adapt the system to varying environments, noise models are acquired from the context, or learnt from the mixture itself without prior information. We also propose methods for reducing the size of the bases used for speech and noise modelling by 20–40 times for better practical applicability. We evaluate the performance of the methods both as a standalone classifier and as a signal-enhancing front-end for external recognisers. For the CHiME noisy speech corpus containing non-stationary multi-source household noises at signal-to-noise ratios ranging from +9 to -6 dB, we report average keyword recognition rates up to 87.8% using a single-stream sparse classification algorithm.

Keywords: automatic speech recognition, noise robustness, non-stationary noise, non-negative spectral factorization, exemplar-based

1. Introduction

These days we are surrounded by devices and services, which could potentially use speech as their input. Possibly the largest hindrance to

[*] Corresponding author, tel. +358 41 482 3348, fax +358 3 3115 4989

Email addresses: `antti.hurmalainen@tut.fi` (Antti Hurmalainen), `jgemmeke@amadana.nl` (Jort F. Gemmeke), `tuomas.virtanen@tut.fi` (Tuomas Virtanen)

widespread adoption of automatic speech recognition (ASR) systems is their limited performance in noisy environments. In everyday situations, the presence of noise can be considered the norm rather than the exception. Therefore robustness against noise is a fundamental requirement for a recogniser intended for common use.

While current state-of-the-art speech recognition systems achieve near-perfect recognition rates on carefully pronounced speech recorded in clean conditions, their performance deteriorates quickly with decreasing signal-to-noise ratio (SNR). Many of the methods proposed for dealing with additive noise focus on increasing the system's sensitivity to the desired patterns over an undefined, roughly uniform noise floor. When the sound level of noise events becomes comparable to that of the target signal, it becomes increasingly important to model noise explicitly. This has been previously accomplished with, for example, model compensation techniques (Acero et al., 2000; Gales and Young, 1996) which allow modelling the interaction of speech and noise. Such techniques have been successfully used to recognise speech in mixtures of multiple speakers, given prior information on each speaker (Hershey et al., 2010).

Since non-negative matrix factorisation (NMF) algorithms were introduced for widespread use (Lee and Seung, 2001), they have been applied to numerous source separation problems. In audio signal processing, NMF has been successfully employed to separate signals consisting of multiple speakers, music, and environmental sounds by modelling a signal as a linear non-negative combination of spectral basis atoms (Heittola et al., 2011; O'Grady and Pearlmutter, 2007; Schmidth and Olsson, 2006; Smaragdis, 2007; Virtanen, 2007). Given a set of basis atoms (also known as *dictionary*) representing the expected sound sources — in robust ASR, speech and noise — observations can be modelled as a sparse linear combination of atoms. This representation can be used to do speech or feature enhancement, proved useful as a preprocessing step for robust speech recognition (Gemmeke et al., 2011c; Raj et al., 2010; Weninger et al., 2011). Alternatively, when speech atoms are associated with speech classes such as phones, the activations of atoms can provide noise robust likelihoods for hybrid decoding in an approach dubbed *sparse classification* (Gemmeke et al., 2011b; Hurmalainen et al., 2011b).

In the most straightforward approach of spectrograms factorisation, each frame is processed independently. However, in real-world situations, the short-term spectral characteristics of noise events can closely resemble actual speech, making the approach prone to misclassifications. Basic NMF methods have later been extended with prior models (Wilson et al., 2008b), smoothness constraints (Cichocki et al., 2006), temporal dynamic modelling and regularisation (Mysore and Smaragdis, 2011; Wilson et al., 2008a) and adding derivative features to the feature vectors (Van Segbroeck and Van hamme, 2009). Meanwhile, there has been an increasing interest in long context spectro-temporal templates for speech modelling. Example-based methods and longest matching segment searching have been proposed for large vocabulary speech recognition (Sundaram and Bellegarda, 2012; Wachter et al., 2003, 2007), dereverberation (Kinoshita et al., 2011) and denoising (Ming et al., 2011). Multi-frame atoms have also been combined with additive spectral modelling in NMF-based speech separation and enhancement (Smaragdis, 2007; Vipperla et al., 2011; Weninger et al., 2011). In our earlier work, we have found further support for the potential of multi-frame spectrograms as features for robust ASR (Gemmeke et al., 2011b; Hurmalainen et al., 2011b; Hurmalainen and Virtanen, 2012; Virtanen et al., 2010; Weninger et al., 2012). While the benefits of the model have been demonstrated in robust speech recognition, the problem of acquiring effective dictionaries — especially for non-stationary noise — has not been plausibly solved.

In this work, we have three goals. First, we propose a new method for acquiring speech basis atoms from a training set. Thus far, the best recognition accuracy in NMF-based recognition has been obtained by using a large number of atoms, which makes the approach computationally expensive. Therefore methods are needed for selecting smaller sets of atoms that still manage to model speech and noise accurately. The proposed algorithm yields sets of speech basis atoms that are much smaller than the previously employed exemplar sampling methods, which improves the practical applicability of the framework through reduced computational costs.

Second, we propose a method to learn noise basis atoms directly from noisy speech, rather than from pure noise sources. Previous studies show that impressive separation and recognition results can be obtained when accurate prior information on the noises is available. However, when the pre-generated noise model is inaccurate or mismatching, the performance of the methods degrades substantially (Gemmeke et al., 2011b). In our earlier work we employed a technique that samples noise basis atoms from the immediate context of an utterance, similar to the use of a voice activity detector (VAD) to estimate the characteristics of noise during speech inactivity as employed in other noise-robust ASR approaches (Demuynck et al., 2011). Nevertheless, in very noisy conditions a VAD will become unreliable, and for non-stationary noises the estimate acquired during speech inactivity may not match exactly to the noise observed during speech. It is also possible that no reliable source for noise-only segments is available in the first place. In order to overcome these obstacles, we propose to use spectrogram factorisation to learn the noise model using only the noisy speech observation itself as the source for the model. The factorisation algorithm will construct its own noise atoms during separation without prior information or assumptions on the noise events.

The final goal of the paper is to present the current state-of-the-art in spectral factorisation based, single-stream noise robust ASR through the use of spectrogram dynamics and binaural features. Temporal deltas and stereo features are added to the model for increased separation and recognition accuracy.

The rest of the paper is organised as follows: Section 2 describes the spectrogram factorisation tools that are used as the basis for the proposed methods. Section 3 proposes methods for speech and noise model acquisition and adaptation. In Section 4 we present an experimental set-up based on the CHiME noisy speech corpus (Barker et al., 2012) used for public evaluation in CHiME workshop in 2011 (Barker et al., 2011). In Section 5 we present our recognition results, obtained with both sparse classification and front-end speech enhancement based recognition. Discussion and conclusions follow in Sections 6 and 7, respectively.

2. Factorisation-based separation and recognition

2.1. Non-negative spectral modelling

NMF-based separation takes place in a spectro-temporal magnitude domain, where the temporal dimension consists of partially overlapping *frames*, and the spectral dimension of a number of frequency *bands*. In this work, the base unit used for additive modelling is a $B \times T$ spectrogram *window* of B Mel bands and T consecutive frames. These are the dimensions of each observation window in our system, and also of the *atoms*, which form the *basis* for modelling the observation.

We can represent noisy speech as a sum of two parts; a speech model $\hat{\mathbf{s}}$ consisting of speech atoms $\mathbf{a}^{\mathrm{s}}$ weighted by *activations* x^{s},

$$\hat{\mathbf{s}} = \sum_{j=1}^{J} x_j^{\mathrm{s}} \mathbf{a}_j^{\mathrm{s}}, \tag{1}$$

and a noise model $\hat{\mathbf{n}}$ using atoms $\mathbf{a}^{\mathrm{n}}$ and activation weights x^{n},

$$\hat{\mathbf{n}} = \sum_{k=1}^{K} x_k^{\mathrm{n}} \mathbf{a}_k^{\mathrm{n}}. \tag{2}$$

The model uses J atoms for speech and K for noise. The total speech-noise model for noisy observation $\mathbf{y}$ thus becomes

$$\mathbf{y} \approx \hat{\mathbf{s}} + \hat{\mathbf{n}} \tag{3}$$

and the estimated noisy observation

$$\hat{\mathbf{y}} = \sum_{j=1}^{J} x_j^{\mathrm{s}} \mathbf{a}_j^{\mathrm{s}} + \sum_{k=1}^{K} x_k^{\mathrm{n}} \mathbf{a}_k^{\mathrm{n}}, \tag{4}$$

using all in all $L = J + K$ atoms and weight coefficients. For now, we treat basis atoms and the observation as generic feature vectors and ignore their true spectro-temporal ordering, assuming only that they match. All variables are assumed to be strictly non-negative.

The fundamental task is to find the *activation vectors* $\mathbf{x}^{\mathrm{s}}$ (length J) and $\mathbf{x}^{\mathrm{n}}$ (length K), or together simply $\mathbf{x}$, which optimise the model under a chosen quality function. We optimise a cost function consisting of a sum of two factors; first, the generalised Kullback-Leibler (KL) divergence between the observation $\mathbf{y}$ and its approximation $\hat{\mathbf{y}}$

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{(y_i, \hat{y}_i) \in (\mathbf{y}, \hat{\mathbf{y}})} y_i \log \frac{y_i}{\hat{y}_i} - y_i + \hat{y}_i \tag{5}$$

and second, a penalty term for non-zero activations weighted elementwise by a sparsity vector $\boldsymbol{\lambda}$

$$f(\mathbf{x}) = ||\boldsymbol{\lambda} \otimes \mathbf{x}||_1 = \sum_{l=1}^{L} \lambda_l x_l. \tag{6}$$

The total cost function to be minimised becomes $d(\mathbf{y}, \hat{\mathbf{y}}) + f(\mathbf{x})$. The first factor measures spectral representation accuracy by generalised KL-divergence, which has been found to perform better than e.g. Euclidean distance or other tested error measures in source separation (Virtanen, 2007).

The second factor induces sparsity to the activation vectors, optionally using a customisable weight for each individual basis atom or group of atoms.

2.2. Sliding window factorisation

In this work we have used two different approaches for processing utterances longer than the window length T. The first is factorising the utterance in overlapping, independent windows (Gemmeke et al., 2011b). To process an utterance consisting of T_{utt} frames, we advance through it with a step of one frame so that the first window covers frames $[1 \ldots T]$, and the last $[T_{\mathrm{utt}} - T + 1 \ldots T_{\mathrm{utt}}]$. Consequently, we have $W = T_{\mathrm{utt}} - T + 1$ overlapping observation windows over the utterance. Each observation window spectrogram is reshaped to a vector $\mathbf{y}_w$ ($w \in [1, W]$). Similarly, each basis atom is reshaped to a vector $\mathbf{a}_l$ ($l \in [1, L]$). The vectorised observations are collected in a matrix $\mathbf{Y} = [\mathbf{y}_1 \ldots \mathbf{y}_W]$, and the atoms in a *basis matrix* $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_L]$. Then we solve for the $L \times W$ *activation matrix* $\mathbf{X}$ so that

$$\mathbf{Y} \approx \mathbf{AX}, \tag{7}$$

while minimising the cost function defined by equations (5) and (6). This can be achieved by applying iteratively the update rule

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{Y}/(\mathbf{AX}))}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda}}, \tag{8}$$

where $\otimes$ is elementiwse multiplication and all divisions are also elementwise. $\mathbf{1}$ is a $\mathbf{Y}$-sized all-ones matrix. $\mathbf{\Lambda}$ is a sparsity penalty matrix defined elementwise for each entry of $\mathbf{X}$, consisting of a $\boldsymbol{\lambda}$ vector for each observation window.

2.3. Convolutive factorisation

An alternative for handling temporal continuity over multi-window observations is *non-negative matrix deconvolution* (NMD), also known as *convolutive non-negative matrix factorisation* (Smaragdis, 2007) or *convolutive sparse coding* (Wang et al., 2011; Wang, 2008). Whereas in the sliding window approach (herefrom called simply 'NMF') each observation window and its corresponding activation vector is an independent entity, in NMD the whole utterance spectrogram $\mathbf{Y}_{\mathrm{utt}}$ is estimated jointly by all activations via convolutive reconstruction. It has been applied earlier to speech separation (O'Grady and Pearlmutter, 2007; Smaragdis, 2007), and to noise-robust speech recognition (Hurmalainen et al., 2011a,b; Vipperla et al., 2011; Weninger et al., 2011).

In this work, we use NMD as in (Hurmalainen et al., 2011a,b). In particular, we only use windows completely within the utterance spectrogram, not ones with their last frames extending beyond T_{utt} as in some implementations. Therefore the activation matrix size is $L \times W$ like in sliding window NMF. The update rule used for activations is

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow (t-1)}{[\frac{\mathbf{Y}_{\text{utt}}}{\mathbf{\Psi}_{\text{utt}}}]}}{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow (t-1)}{\mathbf{1}} + \mathbf{\Lambda}}, \tag{9}$$

where each $\mathbf{A}_t$ is a $B \times L$ matrix containing frame t of all basis atoms, and the estimated utterance spectrogram $\mathbf{\Psi}_{\text{utt}}$ is calculated by

$$\mathbf{\Psi}_{\text{utt}} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow (t-1)}{\mathbf{X}}. \tag{10}$$

Operators $\overset{\leftarrow i}{(\cdot)}$ and $\overset{\rightarrow i}{(\cdot)}$ denote a matrix shift, where the entries are moved left or right by i columns, respectively.

2.4. Speech enhancement

Spectrogram factorisation methods can be used to enhance the input signal before it is passed to a conventional recogniser back-end. Signal enhancement is performed by computing the estimated utterance spectrogram $\mathbf{\Psi}_{\text{utt}}$ as in Equation (10) using the final $\mathbf{X}$ and $\mathbf{A}$ matrices. We also compute an estimated speech spectrogram $\mathbf{\Psi}^{\text{s}}_{\text{utt}}$ by only using the basis atoms and activation rows corresponding to speech. In sliding window NMF the model is similar, except that we average the overlapping window estimates by dividing the frame columns of $\mathbf{\Psi}_{\text{utt}}$ and $\mathbf{\Psi}^{\text{s}}_{\text{utt}}$ by the number of windows contributing to each utterance frame, varying from 1 at the begin and end, to T in the midmost frames.

The clean speech spectrogram estimate is obtained by filtering it in the FFT domain. Because the factorisation model uses Mel-scale spectral resolution, we map the estimates to FFT resolution by inverting the Mel filterbank transform. Denoting the original FFT → Mel scale transform matrix by $\mathbf{M}$, we determine its pseudoinverse $\mathbf{M}^{+}$, and multiply the estimated Mel spectrograms by it from the left. A complex FFT-resolution spectrogram $\tilde{\mathbf{Y}}_{\text{utt}}$ of the original noisy utterance is computed at the temporal resolution of the system. It is then filtered elementwise by the estimated speech/total ratio to get complex speech spectrogram estimate $\tilde{\mathbf{Y}}^{\text{s}}_{\text{utt}}$ as

$$\tilde{\mathbf{Y}}^{\mathrm{s}}_{\mathrm{utt}} = \tilde{\mathbf{Y}}_{\mathrm{utt}} \otimes \frac{\mathbf{M}^{+}\mathbf{\Psi}^{\mathrm{s}}_{\mathrm{utt}}}{\mathbf{M}^{+}\mathbf{\Psi}_{\mathrm{utt}}}. \tag{11}$$

Finally, an enhanced signal is generated with overlap-add synthesis, which inverts the spectrogram derivation.

2.5. Recognition via sparse classification

Instead of using factorisation for signal enhancement, the activations can also be used directly for classification (Virtanen et al., 2010). In this approach, dubbed *sparse classification* (SC), speech basis atoms are associated with sequences of speech labels such as HMM-states. The activations of speech basis atoms serve directly as evidence for the associated speech labels, and the combined speech activations yield a state likelihood matrix, which is used in a hybrid HMM-based recogniser. In previous work it was observed that recognition of noisy speech using sparse classification leads to more accurate results than enhancement-based recognition (Gemmeke et al., 2011b). We have also found the performance of SC to improve in some scenarios by replacing the canonical HMM-based labelling of exemplars with atom-state mapping learnt from training set factorisation (Mahkonen et al., 2011).

3. Speech and noise modelling

3.1. Overview

To separate sound mixtures, we need atoms to model the contained single source components. In noise robust ASR this means models for pure speech and pure noise. In this section we describe on a general level our methods for generating speech and noise bases from training data, and propose methods for generating noise bases adaptively from the context or from the noisy utterance itself.

3.2. Pre-sampled exemplar bases

Both speech and noise bases can be acquired by sampling *exemplars*, instances of spectrograms extracted from the training material as demonstrated in our previous work (Gemmeke et al., 2011b; Hurmalainen et al., 2011b). For speech, this can produce plausible models with high classification capability. For noise, it is not guaranteed that similar sound events will be encountered in actual use cases. In our work on AURORA-2, we saw error rates increasing by up to 60% for mismatched noises (Gemmeke et al., 2011b; Hurmalainen et al., 2011a). Because a noise mismatch degrades the effectiveness of speech-noise separation, and keeping a generic database for all possible noise types would be infeasible, methods for context-sensitive noise modelling are needed for practical applications.

3.3. Context-based noise sampling

To reduce the mismatch between observed noise events and the noise basis, we can switch from using a generic noise database to sampling noise exemplars from the nearby context of the utterances to be recognised. It is generally plausible to assume that in ASR the input is continuous, and that there are moments when the target voice is not active. Since exemplars sampled from the immediate noise neighbourhood of utterances are likely to contain sources similar to those in the noisy speech, we exploit these moments without speech activity to update our noise model.

During development of our recognition system, we managed to reduce the error rates by 10–20% by switching from random to context-based noise sampling. The difference depends on the level of mismatch between training data and observed noise. Sampling the local noise context allows more compact bases, lower computational costs, and generally a better match to the noise encountered during speech. The context-based set-up uses annotated 'oracle' endpointing to sample its atoms from known noise segments, and exploits both preceding and following temporal context. Although in this work oracle endpointing was used in this work to reduce the number of factors affecting the results and to keep correspondence to earlier work, in (Hurmalainen et al., 2012) preliminary experiments are reported on VAD-based noise segment selection and dynamic basis management for continuous inputs.

3.4. Compact speech bases

Previously we have employed large, semi-randomly sampled speech bases, which typically consist of 4000–5000 exemplars per speaker (Gemmeke et al., 2011b; Hurmalainen et al., 2011b). Experiments have also shown, that further gains in recognition accuracy can be achieved by increasing the number of exemplars. Conversely, a small basis sampled in this manner does not model speech sufficiently well for sparse classification (Gemmeke et al., 2011a). While the large, partially redundant exemplar bases allow accurate modelling of observed speech, they may become difficult to acquire and manage for ASR tasks employing a larger vocabulary.

It is possible to use factorisation algorithms to *learn* the speech bases from training material. This has been previously used for speech separation (Smaragdis, 2007) and speech modelling for denoising (Vipperla et al., 2011; Weninger et al., 2011). Unsupervised learning from diverse speech data will ideally discover recurrent phonetic patterns, which can be used for speech modelling. However, NMF-based algorithms may also separate the spectra of speech patterns into multiple overlapping atoms, or learn short-term events lacking the long temporal context preferred in sparse classification and robust separation. In our preliminary experiments, too much fragmentation has typically taken place in large training set learning for its application to speech basis generation.

To address the issue of basis sizes, in this work we propose modelling speech using *template atoms* with more controlled acquisition and less redundancy. The method is based on constructing an atom for each HMM state in the recognition system, including its typical context. According to HMM state labelling acquired via forced alignment, spectrograms of training data instances corresponding to the chosen state are gathered together, and a characteristic template of the state and its neighbourhood is constructed by averaging. The exact procedure for the CHiME database used in this study is described in Section 4.3.

The variant presented in (Weninger et al., 2011) learns a single basis atom from concatenated instances of one word at a time, making it conceptually similar to the templates used in this work. The main difference lies in our algorithm's capability to model words longer than a single window. By using multiple templates centered around one sub-word state at a time, the system is able to model words of arbitrary length. The partially redundant, state-centered templates can also model speed variations in long word pronunciation by combining multiple activations of sub-word atoms over time.

3.5. Learnt noise bases

Whereas speech training data is generally single-source and can be used as-is to model atomic speech events, noise training data and observations often contain multiple overlapping sources. Therefore learning the noise bases either from noise-only segments or noisy mixtures by applying factorisation algorithms may help us to discover recurrent single-source noise components from mixed signals. In the previously mentioned NMD experiments (Vipperla et al., 2011; Weninger et al., 2011), bases were learnt from segments known to contain only noise. The difference between sampled and learnt atoms primarily depends on the nature of the data. If the co-occurrence of noise sources is low, we can expect the bases to become fairly similar. Some fragmentation of noise events may take place in NMD learning if too many atoms are trained with insufficient sparsity constraints on activation. For strongly multi-source inputs, learning will become more favourable due to its ability to discover atomic sources from mixtures.

A different kind of scenario arises, if no source of pure background noise is available. In this case, we still have an option to learn and separate likely noise artefacts from the noisy utterance itself. Given a sufficiently accurate speech basis, we can factorise a noisy utterance by including self-learning noise atoms in the basis. In this approach, the speech basis is kept fixed, and only the noise part is updated on the fly.

Applying learning to sliding window NMF has some theoretical pitfalls, primarily due to having to learn multiple shifted versions of all noise events. A large learnt basis would be required, which in turn increases the risk of modelling speech with it as well. Preliminary experiments have not produced any promising results on this variant. The NMD model, on the other hand, is well suited for noise learning. Sparsity and a small number of noise atoms act as the restricting factors for isolating new noise events.

Basis learning can be included in the procedure given in Section 2.3. After each iteration of the activation update (9), $\mathbf{\Psi}_{\text{utt}}$ is re-estimated using Equation (10), and the basis is in turn updated by

$$\mathbf{A}_t \leftarrow \mathbf{A}_t \otimes \frac{\frac{\mathbf{Y}_{\text{utt}}}{\mathbf{\Psi}_{\text{utt}}} \overset{\rightarrow(t-1)}{\mathbf{X}}^T}{\mathbf{1} \cdot \overset{\rightarrow(t-1)}{\mathbf{X}}^T} \quad \forall t \in [1, T]. \tag{12}$$

Learning can be performed for all atoms in the basis or only for a subset of it. In the latter, only the entries of basis and activation matrices corresponding to the atoms to be updated are included in the equation arrays. Afterwards all modified atoms are reweighted to unitary 2-norm.

Ideally, any parts of the spectrogram which cannot be accurately explained with speech exemplars will be captured by the online-learnt noise atoms. This requires some careful calibration to ensure that co-occurring speech features are not captured together with the noise. The primary tool for this is the sparsity weight vector $\boldsymbol{\lambda}$ described in Section 2.1. However, we assume that even cautiously applied noise learning can detect and remove the largest instances of noise, thus filtering out the most harmful artefacts. This is a highly desirable goal for newly encountered noise events, for which we have no prior information.

4. Experimental set-up

4.1. CHiME corpus

For our experiments, we use the CHiME noisy speech database, published in 2010 to address the challenges posed by non-stationary multi-source noise environments (Barker et al., 2012). For its speech content, the database uses the GRID corpus, where 34 different speakers read simple six word command sentences with linear grammar (Cooke et al., 2006). Each utterance follows the syntax *verb-colour-preposition-letter-digit-adverb*. The word classes have cardinality of 4/4/4/25/10/4, respectively. Recognition performance is scored by the percentage of correctly classified *letter* and *digit* keywords. A baseline recogniser employing HTK binaries (Young et al., 2005) with acoustic models trained on clean speech is provided.

The database consists of following sets:

1. Training speech: 500 clean utterances per speaker
2. Training noise: 6+ hours of pure background noise
3. Development set: in total 600 utterances from all speakers, repeated over six SNRs ranging from +9 to -6 dB at 3 dB steps.
4. Test set: As development set, but with different utterances

Test and development utterances are provided in a long noise context as 'embedded' files with the utterance locations annotated. Development utterances are also available as clean speech. By 'clean' we denote audio without additive noise. All CHiME data is convolved with a room reverberation response, so none of the utterances are truly clean like their original GRID counterparts. All audio is binaural and sampled at 16 kHz.

Additive noise consists of actual household sounds, including appliances, family members, impacts and other sound events. Most of the events are momentary and highly varied, in many cases unique. Different SNRs have been generated by selecting noise segments which produce the desired dB ratio by themselves without scaling. Therefore all SNR-versions of the same development/test utterance contain different noise events.

4.2. Feature space

The feature space used in our experiments consists of magnitude spectrogram segments as described in Section 2.1. The Mel filterbank covers frequencies from 64 to 8000 Hz, divided evenly on a Mel scale with B bands. For the temporal resolution of frames, lengths between 8 and 256 ms have been previously studied (Smaragdis, 2007), and window shift usually varies between 10 and 32 ms. Often a longer frame is used for enhancement than for classification. However, we fix the frame parameters to 25/10 ms for compatibility with CHiME default models and sufficient resolution for sparse classification. In separation and enhancement, it appears that the total duration of atoms, measured in physical time, is more important than temporal resolution within the window (Smaragdis, 2007).

We have previously found repeated evidence for the optimality of window length of 20–30 frames (215–315 ms) for robust enhancement and recognition (Gemmeke et al., 2011b; Hurmalainen et al., 2011a,b). Durations used in other work include 70 ms (Vipperla et al., 2011), 80 ms (Wang, 2008), 176 ms (Smaragdis, 2007), 224 ms (O'Grady and Pearlmutter, 2007) and 256 ms (Weninger et al., 2011, 2012). Based on previous results and a grid search on the development data over a range of T values, we set the NMF window length to 20 frames (215 ms), but use 25 frames (265 ms) for NMD, which appears to favour slightly longer context (Hurmalainen et al., 2011a).

We have achieved improvements by increasing the number of Mel bands from 26 (Hurmalainen et al., 2011b) to 40 (Hurmalainen and Virtanen, 2012). For even larger numbers of frequency bands, the gains were negligible. Therefore B was set to 40 for these experiments.

The factorisation algorithms support processing signals using stereo features by concatenating the features pertaining to each individual channel. In previous work we observed that the use of stereo features only has a minor impact on the separation quality, while it doubles the data size and computational costs (Hurmalainen and Virtanen, 2012). Therefore the results were mostly computed using mono features averaged in the spectral magnitude domain. However, in the same study we found out that augmenting the static features with temporal derivatives ('deltas') similarly as in conventional GMM-based modelling (Young et al., 2005) does improve the recognition rates. Even though the long temporal context of atoms manages to model spectral behaviour over time to some extent by itself, adding explicit delta features will emphasise modulations, which contain significant information on speech and noise events. To generate enhanced signals and recognition results reflecting the current best performance of our framework, stereo features and temporal dynamics as in (Hurmalainen and Virtanen, 2012) were included in the final experiment of this work.

4.3. Basis generation

4.3.1. Speech

In this work, all our speech bases are speaker-dependent, and the knowledge of test speaker identity is exploited by selecting the corresponding speaker's basis. We use two variants; sampling large bases from the training material as described in Section 3.2, and using compact template bases introduced in Section 3.4.

The first method is to sample training utterances semi-randomly (Hurmalainen et al., 2011b). For each speaker, the 500 training utterances are split into 300 for basis generation and 200 for learning the mapping between speech exemplars and speech labels. The utterances selected for basis generation are sampled by extracting windows with a random step of 4–8 frames. The resulting, densely sampled sets of more than 10000 exemplars per speaker are reduced to 5000 while maximising the flatness of included word distribution. This mainly reduces the amount of exemplars from the originally overrepresented non-keyword classes that contain only four word options each. However, no attempt is made to control the exact positioning of exemplars within utterances. They may cover word boundaries, thus modelling specific word transitions.

The second method is based on constructing compact bases of state-centric speech templates. As in the provided CHiME recogniser models, our framework uses 250 speech states (4–10 states per word) to label speech basis atoms. For each state in the system, we select all instances of the word, which contains the chosen state. Based on a forced alignment by the CHiME recogniser, the words are positioned in a length T window with the target state in its midmost frame. We then take the median within each single spectrogram bin over all word instances to generate a prototype of each state and its immediate context. The process is illustrated in Figure 1, where template construction is shown for the third state (out of six) of the word ‘green’.

The midmost frames, always representing the nominal state, are most likely to match each other in the spectral domain. Therefore the spectral model is also most consistent in the middle of a template. As the temporal distance increases towards template edges, there is higher variation in the spectrogram content due to differences in pronunciation style, speed and coarticulation. Consequently, the edges fade out when a median is taken over instances. Especially, multiple neighbouring words candidates all have different spectrogram profiles. Consequently the median template model

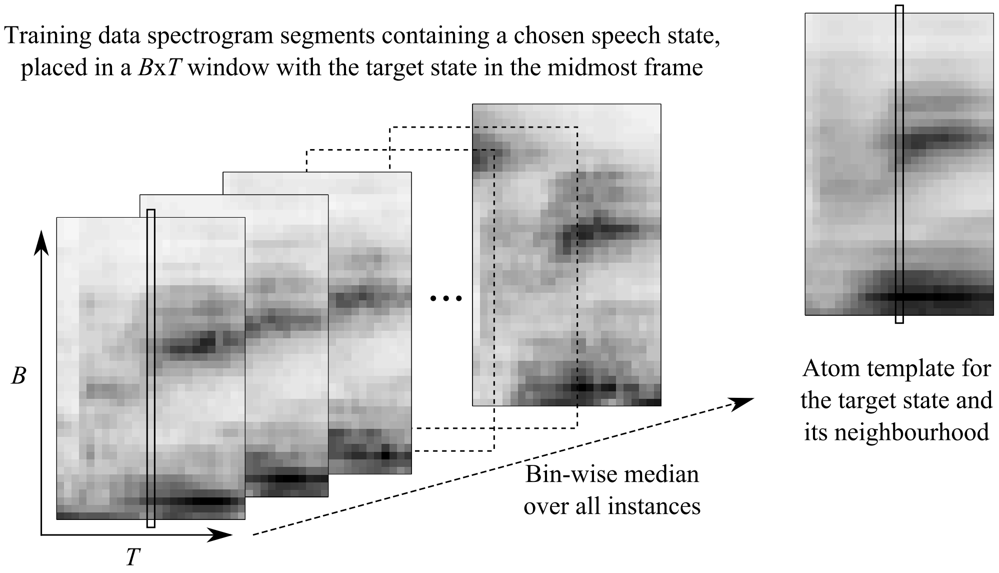
Training data spectrogram segments containing a chosen speech state,
placed in a BxT window with the target state in the midmost frame
B
T
...
Bin-wise median
over all instances
Atom template for
the target state and
its neighbourhood

Figure 1: Forming an atom template for a speech state and its neighbourhood. Training data spectrograms containing the state are placed in a $B \times T$ window, and bin-wise median is taken over the instances. In this example, the third state of word 'green' is modelled with a 40×25 template. In addition to the state itself, a large part of the word is captured as well, thus increasing the temporal context being modelled.

will generally remove the fragments of other words and continuity over word boundaries. For example, in the last training data instance in Figure 1 we can see a high-pitched fricative from a preceding word, whereas very little spectral activity remains in the first frames of the resulting template.

The compact bases cannot model all possible temporal alignments required by independent NMF windows, but they are suited for NMD's temporal model, which can find the best locations for a few temporally sparse activations. By losing word transition modelling and replacing redundant exemplars with median templates, the basis size is reduced to 1/20th of the large NMF bases.

4.3.2. Noise

In this work, we employ three different methods for modelling the additive, non-stationary noise in CHiME data:

1. Context-based sampling of the utterance's noise neighbourhood as presented in Section 3.3 and our earlier CHiME experiments (Hurmalainen et al., 2011b). The 'embedded' wave files are sampled to

both directions from the target utterance, and exemplars are extracted at random intervals of 4–7 frames from segments containing only noise. As before, we use 5000 noise exemplars for the NMF experiments. With these parameter settings, approximately 4.5 minutes of noise context got sampled into the basis (from 5–7 minutes of overall audio context with the skipped neighbouring utterances included). The nearest available noise segments were used so the amount of forward and backwards context was roughly symmetric, except at the ends of embedded recording sessions where only one direction is available.

2. The same algorithm, but used to generate a small noise basis of 250 exemplars for NMD. Because less temporal redundancy is required in the NMD model, the sampling interval is increased to 10–15 frames. Still, the overall context covered is reduced to approximately a tenth in terms of physical time span ($\sim$ 30 seconds of pure noise data).
3. Finally, we study noise modelling using neither context nor prior knowledge. Instead of passing a pre-generated basis to the factorisation algorithm, we randomly initialise $\lceil T_{\text{utt}}/T \rceil$ noise basis atoms — just enough to cover every frame of an utterance once — and update them in the NMD iteration loop as described in Section 3.5. The on-line updated atoms will adapt themselves to spectrogram patterns not matching to the speech basis, thus learning and modelling noise events found in the mixture.

The generic background training material was not used in any of these experiments. While potentially a sound option in some scenarios, it is debatable if a universal noise basis can be modelled for real world use. For reasons pointed out in Section 3.3, we favour context-aware noise modelling to improve the adaptivity to new noise environments.

4.3.3. Basis weighting

Earlier we have been using two-way normalisation of the basis. Each vectorised atom spectrogram was scaled to unitary Euclidean norm. In addition, the Mel band weights of the full basis were scaled so that the Euclidean norms over all spectral content within each band were equal. To satisfy both conditions together, ten alternating normalisation rounds were performed iteratively for an approximate solution. (Gemmeke et al., 2011b). In this work, we still normalise individual atoms as is preferable for the NMF update rules. However, fixed weights are acquired for Mel bands by gathering all training speech spectrograms, and computing weights which equalise the Euclidean norms over their Mel band content. Using a fixed band weighting profile stabilises and simplifies the model, because the two-way normalisation step can be omitted, and the weighting no longer changes in every noise basis update. When various band weighting methods were compared, the fixed, speech-normalising profile was found to perform comparably to two-way normalisation (Hurmalainen and Virtanen, 2012).

4.4. Factorisation

Activation matrices were computed using the update rules described in Section 2. We used CUDA GPU hardware, MATLAB and the GPUmat toolbox (The GP-You Group, 2010) for computation. Single precision variables and 300 iterative updates were used in all experiments.

In many previously reported implementations, the sparsity parameter λ has been set to a fixed value. However, its sparsifying effect is related to the 1-norms of the basis atoms, which will vary as a function of the dimensionality of the feature space. To make the level of sparsity more independent of the window parameters that determine the dimensionality, the penalty weights were set proportionally to the mean of the 1-norms of basis atoms. By conversion from the fixed parameters used in earlier experiments (Hurmalainen et al., 2011b; Hurmalainen and Virtanen, 2012), the sparsity value governing speech basis atoms was set to 0.1 of the mean of norms, and sparsity of noise basis atoms to 0.085. In basis-learning NMD, noise sparsity was increased after brief development data experiments to 0.1 to avoid bias toward the freely adapting atoms and consequently modelling speech with them as well.

4.5. Decoding

All our recognition methods are fundamentally based on the CHiME baseline recogniser and its language model. Variants for enhancement and sparse classification are employed as follows.

4.5.1. Signal enhancement

In signal enhancement, we synthesise the filtered spectrogram as described in Section 2.4. The enhanced wave files are recognised using HVite and two models with different training. First, we use the default CHiME models trained on reverberated, 'clean' training files to produce results compatible with the baseline system. The second system is trained on multicondition data consisting of the 17 000 clean utterances and the same utterances mixed with random training noise. Mean-only maximum-a-posteriori

(MAP) adaptation is used for generating the speaker-dependent models. These models are exactly the same as used in (Weninger et al., 2011) and later in our multi-stream recognition experiments (Weninger et al., 2012).

Neither of these models is retrained on speech data processed with our enhancement framework. Such a task would be laborious, considering that the enhanced output will differ slightly for all factorisation parameters, and that there is no standard training material with noise context as required by our adaptive algorithms. Therefore we only employ generic clean- and multi-condition trained models. A benefit of this choice is that earlier results exist for both models, allowing direct comparison.

In closely integrated recognition systems with matching spectral parameters, it would be possible to use the enhanced Mel scale spectrogram by itself for deriving the MFCC features. However, our separation framework and the two external recognisers all use slightly different parametrisation for their spectral features (e.g. Mel band count and preprocessing filters). Therefore enhanced speech was passed as time domain signals, which are universally accepted by all external recognisers regardless of their internal spectral representation.

4.5.2. Sparse classification

For direct classification via speech basis atom weights, we use label matrices representing the probabilities of different speech states over atom duration (Virtanen et al., 2010). In *canonical labelling*, labels are acquired directly from a forced alignment, and the matrices are binary so that for each frame of a speech basis atom only the nominal state is active with weight 1.

However, especially when using speech templates without transition context, some basis atoms may in practise match several different words in the CHiME state model. While phonetically similar, the words are denoted by different states in the system. For example, the first phones of “please” and “place” appear essentially the same. In order to reduce the risk of misclassification due to incorrect or overly strict label associations, we learn the mapping from activations to states by factorising the 200 training utterances not used for the basis, and calculating the mapping matrices using ordinary least squares (OLS) regression (Mahkonen et al., 2011). The non-binary conversion matrices acquired this way are able to model the multiple word associations of some speech atoms, improving the results in scenarios with more phonetic ambiguity (Hurmalainen et al., 2011b).

Preliminary experiments showed that OLS mapping improved the results of small basis experiments, thus this technique was used for the final sparse classification results. For large bases with static features, the results were mixed, with a small overall decrement in average score. With dynamic features included, the results of mapping were uniformly detrimental. Therefore no learnt mapping was used for large basis NMF experiments. The varying benefits of OLS are explained by the accuracy of canonical labels, and the amount of training data. For the large bases with full coarticulation context, the canonical labelling is already reasonably accurate, and no improvements were achieved by learning the mapping from limited training material (200 speaker-dependent utterances). Conversely, the templates constructed from multiple instances have indefinite labels to begin with, and better mapping can be learnt via training factorisation.

The utterances are decoded as described in (Hurmalainen et al., 2011b). In NMF decoding, we normalise the activation vectors of all windows to unitary sum. In NMD's temporal model, the activity levels may vary greatly across windows so no normalisation is applied to the basis activations. The resulting likelihood matrix is passed to a modified CHiME baseline recogniser, which performs the final recognition using the generated likelihoods and the default CHiME language model.

5. Evaluation

5.1. Modelling, factorisation and decoding methods

To compare the different methods for modelling speech and noise, the test set was factorised using three models:

1. Sliding window NMF, 5000 speech and 5000 sampled noise exemplars, $T = 20$ ('Large basis NMF')
2. NMD, 250 speech atoms, 250 sampled noise exemplars, $T = 25$ ('Small basis NMD, sampling')
3. NMD, 250 speech atoms, online-learnt noise model, $T = 25$ ('Small basis NMD, learning')

The 5000-atom sampled speech bases (used in model 1) and 250-atom template bases (models 2 and 3) are described in Section 4.3.1. The three noise models correspond to those described in Section 4.3.2.

Previously, we have got mixed results for applying NMD to large bases (Hurmalainen et al., 2011a,b). For CHiME data, no improvements were seen, while the computational complexity increases significantly. The large bases seem to contain sufficient temporal redundancy for NMF, which in

Table 1: Test set results for different factorisation configurations: large basis NMF, small basis NMD with sampled noise, and small basis NMD with online-learnt noise. All are decoded using feature enhancement (FE) with clean-trained (CHiME) and multi-condition trained (MC) models described in Section 4.5, and sparse classification (SC). Unenhanced baseline scores and two alternative enhancement systems are also shown.

SNR (dB)	9	6	3	0	-3	-6	avg
Baseline scores of FE recognisers (unenhanced)							
CHiME	82.4	75.0	62.9	49.5	35.4	30.3	55.9
MC	91.3	86.8	81.7	72.8	61.1	54.5	74.7
Large basis NMF							
FE, CHiME	92.2	88.8	85.8	80.5	73.3	61.4	80.3
FE, MC	92.8	92.3	90.7	87.6	82.2	75.7	86.9
SC	92.4	90.4	90.0	88.0	79.8	73.8	85.8
Small basis NMD, sampling							
FE, CHiME	91.3	87.0	83.5	76.2	68.2	56.3	77.1
FE, MC	93.0	91.2	90.0	85.2	79.0	72.9	85.2
SC	89.8	89.0	84.3	81.8	73.9	65.8	80.8
Small basis NMD, learning							
FE, CHiME	87.7	83.2	77.2	68.8	60.0	55.4	72.0
FE, MC	91.3	89.8	86.2	80.0	74.2	72.0	82.2
SC	87.8	83.5	79.8	75.0	66.4	60.6	75.5
Alternative NMD enhancement results							
EURECOM[a]	84.6	79.3	69.4	61.8	50.4	43.2	64.8
TUM[b]	90.6	88.3	87.7	84.1	79.2	75.6	84.2

[a] Vipperla et al. (2011)

[b] Weninger et al. (2011)

turn produces better results via multiple averaged estimates. Regarding compact bases, the 250+250 atom set-up was tested using both sliding window NMF and NMD. The scores were uniformly worse for NMF than for NMD (0.4–3.5% absolute, 2–20% relative decrement in recognition rates), confirming that the sliding window model is not as well suited for small bases with insufficient temporal alignment variants over the atoms.

All activation matrices acquired from different factorisation types were used for enhancement and recognition with the two GMM-based recognisers;

clean-trained original CHiME models ('CHiME') and the multi-condition trained model ('MC'), and also recognised using sparse classification ('SC'). The results are shown in Table 1. The unenhanced baseline performance of the external recognisers is shown on the first rows. Two alternative implementations for NMD enhancement, 'EURECOM' (Vipperla et al., 2011) and 'TUM' (Weninger et al., 2011) are also included on the last rows for comparison.

Table 2: Large basis NMF results for static-only mono features, and features with temporal dynamics and stereo channels included. Feature space extensions are applied individually as well as together. Results are shown for multi-condition trained feature enhancement (FE, MC), sparse classification (SC), and three external system combinations reflecting state-of-the-art results on CHiME data.

SNR (dB)	9	6	3	0	-3	-6	avg
Mono, static features only							
FE, MC	92.8	92.3	90.7	87.6	82.2	75.7	86.9
SC	92.4	90.4	90.0	88.0	79.8	73.8	85.8
Stereo, static features only							
FE, MC	93.2	92.2	91.0	87.8	82.4	76.3	87.1
SC	92.4	90.4	90.2	88.4	80.7	73.5	85.9
Mono, static and dynamic features							
FE, MC	93.3	92.1	90.0	87.7	83.1	76.6	87.1
SC	93.0	91.5	90.8	89.2	82.2	76.3	87.2
Stereo, static and dynamic features							
FE, MC	92.9	92.3	90.7	88.2	83.4	77.3	87.5
SC	92.8	91.7	91.1	89.3	83.4	78.6	87.8
Alternative systems for CHiME data							
FAU[a]	95.1	92.6	92.8	88.3	83.3	79.8	88.7
NTT[b]	95.8	94.2	93.7	92.3	88.3	85.6	91.7
TUM/TUT[c]	96.4	95.7	93.9	92.1	88.3	84.8	91.9

[a] Maas et al. (2011)

[b] Delcroix et al. (2011)

[c] Weninger et al. (2012)

5.2. Derivative and stereo features

As an additional evaluation, we recomputed the large basis NMF results while including binaural features and temporal dynamics as in (Hurmalainen and Virtanen, 2012). In stereo processing, features were extracted for both channels separately and treated like another set of spectral bands in feature vectors. Temporal dynamics were modelled by applying a delta filter, spanning two frames forward and backwards, to the static magnitude spectrograms. The newly acquired difference spectrogram was split into two parts, one containing positive delta values and another the absolute values of negative entries in order to keep the features non-negative. In other words, the two derivative spectrograms captured event on- and offsets, respectively. Both were concatenated with the static magnitude features of atoms and observations for separation. However, after acquiring the activation weights, only static magnitudes were used for generating the enhanced spectrogram and signals.

The results for multi-condition trained enhancement ('FE, MC') and sparse classification ('SC') using extended feature spaces are shown in Table 2. Stereo features and temporal dynamics are first applied each alone and then together. The scores are compared to static-only mono features, and three alternative systems presented in recent literature (Delcroix et al., 2011; Maas et al., 2011; Weninger et al., 2012).

6. Discussion

6.1. Findings

From the results in Table 1, showing the evaluation of different speech and noise modelling methods, we can make the general observation that larger bases and more context information produce better results. This is theoretically sound — the more information available, the better models for individual sources can be constructed. In sparse classification, there is approximately a 5% drop (absolute) in average recognition rate from large basis NMF to small basis NMD, and further to no-prior noise learning. Lower accuracy can already be observed in the cleanest conditions, suggesting that the small bases cannot classify words as accurately as the large bases. However, even the last SC variant performs at least 31%, and on average 44% better than the original CHiME recogniser, measured by relative word error rate reduction.

Interesting results can also be seen in the recognition rate differences between SC and the enhanced signal recognisers. We notice that SC nearly always exceeds the clean-trained CHiME recogniser, while the MC recogniser is mostly better than SC. Especially the small speech basis experiments favour the GMM-based recogniser with robust training. Only 1.7% reduction can be observed in the average score from the 10 000 atom NMF basis to 500 atom NMD. Another 3.0% decrement takes place, when all prior information on noise is removed. Still, enhancement using compact speech modelling and blind noise learning is able to reduce the error rate by up to 38% (relative) in the noisiest end, and by 30% on average in comparison to the same recogniser with unenhanced signals.

The results are also compared to other NMD-based enhancement systems tested on CHiME data. We observe that all our denoising algorithms perform better than the EURECOM approach, where noisy speech was modelled using 100 speech atoms, 100–200 noise atoms from the background data, and 20 atoms from the local context (Vipperla et al., 2011). The results were scored using the standard CHiME recogniser, which therefore should be used as the point of comparison. It is likely that a part of the difference in recognition rates arises from the temporal context, which in our experiments is 20–25 frames (215–265 ms) in comparison to EURECOM's 4 frames (70 ms).

The NMD enhancer used in TUM's CHiME experiments (Weninger et al., 2011) and in our joint work (Weninger et al., 2012) employed 51 speech atoms, 51 noise atoms learnt from the general background, and 256 ms window length. The temporal resolution was 64/16 ms, and the spectral resolution full 1024 FFT bins. The recogniser was the same as the MC model used in this work. We notice that the large basis NMF enhancer performs better than the TUM set-up. Small basis NMD with sampled noise works better in all but the lowest SNRs, and NMD without a noise model only at the highest SNRs. Especially the second case gives some insight to the two systems, which are in many ways similar but also differ in their parametrisation and modelling, primarily in spectral and temporal resolution. It should be inspected further, whether the resolution or the basis generation method plays a larger role in enhancement quality. Differences in the level of sparsity may affect the quality as well.

The final experiment (Table 2) on extended NMF feature spaces reveals more aspects regarding the choice between sparse classification and signal enhancement. Whereas in both static-only set-ups (mono and stereo) features enhancement works better, we notice that including dynamic information improves the SC quality more, making it in turn slightly better.

However, the differences are small, so the true order probably depends on implementation details such as external back-end training and the accuracy of atom-to-state mapping in SC. Nevertheless, both recognition methods benefit from dynamic features in separation, especially in the noisy end. The contribution of magnitude-domain stereo information is significantly smaller.

Three alternative recognition systems were also included in Table 2 for comparison. The first one (Maas et al., 2011) is a binaural signal enhancement front-end for a robust Sphinx-4 recogniser employing triphone HMMs. Its noise robustness is generally similar to the proposed system, while its initial clean end recognition rate appears better, probably due to more sophisticated back-end modelling. The NTT approach (Delcroix et al., 2011) combines multiple enhancement and model compensation steps to simultaneously exploit spectro-temporal and spatial information for separation. The TUM/TUT system, also combining multiple streams, consists of GMM recognition, a BLSTM network and a word-spotting version of our sparse classifier (Weninger et al., 2012). This multi-stream system managed to surpass all of its individual streams, and produced the best known average results on CHiME data at the time of writing. We can conclude that system, feature and stream combinations are currently producing state-of-the-art results in noise robust ASR. Factorisation-based methods are well suited for use in such combinations, but other features such as spatial information should also be considered in an efficient overall solution.

6.2. Computational complexity and costs

Regarding the computational complexity of factorisation-based speech and noise modelling, we can consider three aspects:

1. Training data requirements
2. Memory allocation
3. Computational costs

We have observed that a large basis of exemplars provides the best accuracy in modelling, and consequently the best recognition results. However, constructing a 5000+5000 atom basis using the approach taken in our NMF experiments requires significant amounts of training data, and for a larger vocabulary the requirements for similar coverage would increase further. Explicit modelling of large word segments and word transitions would require even larger bases, which would only be feasible with dynamic basis management. Fortunately, we have shown that both speech and noise bases can be reduced to a fraction of this size with only a modest decrement in recognition rates. On the other hand, the best results (smallest decrements) were observed using signal enhancement, where some of the training and modelling complexity is shifted on an external back-end.

The memory requirement for NMF bases is $B \cdot T \cdot L$ scalars, which for 10 000 single-precision 40×25 atoms is 40 megabytes. The amount can be reduced significantly by more efficient basis construction, phonetic modelling, and shifting the classification to a conventional recogniser. For example, our 500-atom bases only require 2 megabytes, and with learnt noise atoms even less. Therefore the memory requirements of exemplar-based factorisation are not unbearable for modern devices, including mobile ones.

The computational costs of NMF depend on data sizes, algorithms and naturally the hardware platform. On a dual core E8400 1333 MHz CPU, MATLAB implementation of the large basis (5000+5000 atoms) factorisation takes on average 80.8 seconds per utterance (46× audio duration). On a consumer-grade GeForce GTX260 graphics card, the same computation takes 7.0 seconds per utterance (4.0×). When the basis is reduced to 500 fixed atoms (16× reduction on data size, taking into account the increased window length), NMF execution times become 5.5 seconds (3.1× audio duration) and 0.62 seconds (0.35×) for the described CPU and GPU platforms, respectively. In CPU computing, the speedup factor is close to linear, whereas GPU computing scales better to large arrays due to heavy parallelisation.

Using NMD for factorisation complicates the comparisons. While fixed basis NMF can be computed trivially with elementary matrix operations which also parallelise directly, the NMD speed is highly dependent on algorithm design. The current small basis NMD implementation takes 3–6 seconds per utterance on a GPU, depending on whether basis learning is included. However, the same algorithm for a large basis takes approximately 10 seconds. This highly nonlinear correspondence to problem size illustrates, how the increased computing costs of NMD arise primarily from the overhead of additional algorithm steps. Code optimisation and possibly low-level implementation instead of interpreted MATLAB code would be beneficial in finding out the true performance of NMF and NMD. Nevertheless, it appears ultimately feasible to run the proposed set-ups in real time on parallel platforms.

7. Conclusions and future work

We presented several alternative methods for modelling speech and noise in factorisation-based speech recognition. Local context was used for adaptive noise modelling instead of acquiring a universal noise model from generic training data. The best results were achieved using a large exemplar-based basis consisting of actual instances of training and observation data. Meanwhile, we also demonstrated how significantly smaller bases can be employed for the task with only small losses in quality compared to the reduction factor in model size. Furthermore, we managed to model non-stationary multi-source noise using online-updated atoms without any prior information or context for the noise.

We found additional support for the optimality of 200–250 millisecond window length for both of our recognition methods; signal enhancement for external back-ends and sparse classification based on exemplar labels. When using large bases and dynamic features in addition to static spectra, we achieved better results by sparse classification than by enhancement. However, if the speech bases are reduced to generic templates without word transitions or pronunciation variance, signal enhancement for a multi-condition trained GMM recogniser performed better.

It appears that the current factorisation framework can produce plausible separation results for well-modelled data. Therefore even more effort should be spent on learning compact yet accurate speech and noise models for diverse use cases. The different noise acquisition methods (universal, local context, in-place learning) should be combined to maximise the model accuracy. Preliminary experiments suggest that such combination is indeed feasible, and a noise basis can be updated adaptively in continuous recognition using voice activity detection to locate noise-only segments. Recognition rates comparable to informed noise segment sampling have been achieved by using VAD-based basis adaptation without exploiting any look-forward context (Hurmalainen et al., 2012). For speech, the variations in pronunciation can be possibly handled via clustering or other techniques, which are able to represent the spectro-temporal space volumes with a small number of atoms per phonetic pattern. Switching from word-based to phonetic state models will be eventually needed for large vocabulary recognition.

One important feature type not exploited in this work is the spatial information available in binaural signals. It alone can act as a powerful separation method. Thereby introducing time-domain phase information to the framework might give significant improvements in multichannel recognition.

Regarding final recognition accuracy, there is a lot of potential in multistream algorithms, which combine enhancement, sparse classification, and complementary methods (Weninger et al., 2012). Different system combinations should be tested for better joint recognition rates. Especially the clean speech recognition rate, which in our standalone sparse classification is still suboptimal, can be improved by introducing alternative streams to the recogniser. Finally, it would be beneficial to optimise the practical implementation of NMF/NMD algorithms to best exploit current hardware, and thus allow actual deployment of separation-based robust ASR to everyday applications.

8. Acknowledgements

The research of Antti Hurmalainen has been funded by the Academy of Finland. The research of Jort F. Gemmeke was supported by IWT-SBO project ALADIN contract 100049.

References

Acero, A., Deng, L., Kristjansson, T., Zhang, J., 2000. HMM Adaptation using Vector Taylor Series for Noise Speech Recognition, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China. pp. 869–872.

Barker, J., Vincent, E., Ma, N., Christensen, C., Green, P., 2012. The PASCAL CHiME Speech Separation and Recognition Challenge. Computer Speech and Language (submitted) .

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2011. Overview of the PASCAL CHiME Speech Separation and Recognition Challenge, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy.

Cichocki, A., Zdunek, R., Amari, S., 2006. New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Tolouse, France. pp. V–621–624.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition. Journal of the Acoustical Society of America 120, 2421–2424.

Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S., Nakamura, A., 2011. 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 International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 12–17.

Demuynck, K., Zhang, X., Van Compernolle, D., Van hamme, H., 2011. Feature versus Model Based Noise Robustness, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 721–724.

Gales, M.J.F., Young, S.J., 1996. Robust Continuous Speech Recognition Using Parallel Model Combination. IEEE Transactions on Speech and Audio Processing 4, 352–359.

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

Gemmeke, J.F., Virtanen, T., Hurmalainen, A., 2011b. Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing 19, 2067–2080.

Gemmeke, J.F., Virtanen, T., Hurmalainen, A., 2011c. Exemplar-based Speech Enhancement and Its Application to Noise-robust Automatic Speech Recognition, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 53–57.

Heittola, T., Mesaros, A., Virtanen, T., Eronen, A., 2011. Sound Event Detection in Multisource Environments Using Source Separation, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 36–40.

Hershey, J.R., Rennie, S.J., Olsen, P.A., Kristjansson, T.T., 2010. Super-Human Multi-Talker Speech Recognition: A Graphical Modeling Approach. Computer Speech and Language 24, 45–66.

Hurmalainen, A., Gemmeke, J.F., Virtanen, T., 2011a. Non-negative Matrix Deconvolution in Noise Robust Speech Recognition, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 4588–4591.

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

Hurmalainen, A., Mahkonen, K., Gemmeke, J.F., Virtanen, T., 2011b. Exemplar-based Recognition of Speech in Highly Variable Noise, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 1–5.

Hurmalainen, A., Virtanen, T., 2012. Modelling Spectro-Temporal Dynamics in Factorisation-Based Noise-Robust Automatic Speech Recognition, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan. pp. 4113–4116.

Kinoshita, K., Souden, M., Delcroix, M., Nakatani, T., 2011. Single Channel Dereverberation Using Example-Based Speech Enhancement with Uncertainty Decoding Technique, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 197–200.

Lee, D.D., Seung, H.S., 2001. Algorithms for Non-negative Matrix Factorization, in: Advances in Neural Information Processing Systems 13, pp. 556–562.

Maas, R., Schwarz, A., Zheng, Y., Reindl, K., Meier, S., Sehr, A., Kellermann, W., 2011. A Two-Channel Acoustic Front-End for Robust Automatic Speech Recognition in Noisy and Rerverberant Environments, in: Proceedings of International Workshop on

Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 41–46.

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

Ming, J., Srinivasan, R., Crookes, D., 2011. A Corpus-Based Approach to Speech Enhancement From Nonstationary Noise. IEEE Transactions on Audio, Speech, and Language Processing 19, 822–836.

Mysore, G.J., Smaragdis, P., 2011. A Non-negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 17–20.

O'Grady, P.D., Pearlmutter, B.A., 2007. Discovering Convolutive Speech Phones using Sparseness and Non-Negativity Constraints, in: Proceedings of ICA, London, UK. pp. 520–527.

Raj, B., Virtanen, T., Chaudhure, S., Singh, R., 2010. Non-negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition, in: Proceedings of INTERSPEECH, Makuhari, Japan. pp. 717–720.

Schmidth, M.N., Olsson, R.K., 2006. Single-channel Speech Separation using Sparse Non-negative Matrix Factorization, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Pittsburgh, Pennsylvania, USA. pp. 2614–2617.

Smaragdis, P., 2007. Convolutive Speech Bases and their Application to Supervised Speech Separation. IEEE Transactions on Audio, Speech, and Language Processing 15, 1–14.

Sundaram, S., Bellegarda, J., 2012. Latent Perceptual Mapping with Data-Driven Variable-Length Acoustic Units for Template-Based Speech Recognition, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan. pp. 4125–4128.

The GP-You Group, 2010. GPUmat User Guide, version 0.27.

Van Segbroeck, M., Van hamme, H., 2009. Unsupervised Learning of Time-Frequency Patches as a Noise-robust Representation of Speech. Speech Communication 51, 1124–1138.

Vipperla, R., Bozonnet, S., Wang, D., Evans, N., 2011. Robust Speech Recognition in Multi-Source Noise Environments using Convolutive Non-Negative Matrix Factorization, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 74–79.

Virtanen, T., 2007. Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria. IEEE Transactions on Audio, Speech, and Language Processing 15, 1066–1074.

Virtanen, T., Gemmeke, J., Hurmalainen, A., 2010. State-based Labelling for a Sparse Representation of Speech and Its Application to Robust Speech Recognition, in: Proceedings of INTERSPEECH, Makuhari, Japan. pp. 893–896.

Wachter, M.D., Demuynck, K., Compernolle, D.V., Wambacq, P., 2003. Data-Driven Example Based Continuous Speech Recognition, in: Proceedings of EUROSPEECH, Geneva, Switzerland. pp. 1133–1136.

Wachter, M.D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., Compernolle, D.V.,

2007. Template-based Continuous Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing 15, 1377–1390.

Wang, D., Vipperla, R., Evans, N., 2011. Online Pattern Learning for Non-Negative Convolutive Sparse Coding, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 65–68.

Wang, W., 2008. Convolutive Non-Negative Sparse Coding, in: Proceedings of IJCNN, Hong Kong. pp. 3681–3684.

Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., Rigoll, G., 2011. 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. pp. 24–29.

Weninger, F., Wöllmer, M., Geiger, J., Schuller, B., Gemmeke, J.F., Hurmalainen, A., Virtanen, T., Rigoll, G., 2012. 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. pp. 4681–4684.

Wilson, K.W., Raj, B., Smaragdis, P., 2008a. Regularized Non-Negative Matrix Factorization with Temporal Dependencies for Speech Denoising, in: Proceedings of INTERSPEECH, Brisbane, Australia. pp. 411–414.

Wilson, K.W., Raj, B., Smaragdis, P., Divakaran, A., 2008b. Speech Denoising Using Nonnegative Matrix Factorization with Priors, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, USA. pp. 4029–4032.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2005. The HTK Book Version 3.3. Cambridge University Press.