MODELLING SPECTRO-TEMPORAL DYNAMICS IN FACTORISATION-BASED NOISE-ROBUST AUTOMATIC SPEECH RECOGNITION

Antti Hurmalainen *Tuomas Virtanen*

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

ABSTRACT

Non-negative spectral factorisation has been used successfully for separation of speech and noise in automatic speech recognition, both in feature-enhancing front-ends and in direct classification. In this work, we propose employing spectro-temporal 2D filters to model dynamic properties of Mel-scale spectrogram patterns in addition to static magnitude features. The results are evaluated using an exemplar-based sparse classifier on the CHiME noisy speech database. After optimisation of static features and modelling of temporal dynamics with derivative features, we achieve 87.4% average score over SNRs from 9 to -6 dB, reducing the word error rate by 28.1% from our previous static-only features.

Index Terms— Automatic speech recognition, exemplar-based, spectral factorisation, noise robustness

1. INTRODUCTION

In its current state, automatic speech recognition (ASR) can achieve high phonetic classification quality in favourable conditions. However, the same cannot be said about noisy ASR. As the signal to noise ratio decreases towards zero or below, a majority of spectral features becomes corrupted, and traditional recognisers cannot match the observations to speech models reliably. Especially non-stationary noise is problematic for recogniser back-ends and difficult to counter with uniform compensation methods. Therefore detecting and removing non-speech artifacts becomes essential for noise-robust ASR.

To compare different robust ASR methods, PASCAL CHiME challenge was announced in 2010, and its results were gathered in a workshop in September 2011 [1]. As the test data includes very low SNRs, practically all challenge entries contained enhancement or separation steps for extracting real speech features from the noisy mixture [2]. Proposed approaches included beamforming, spatial uncertainty-of-observation, statistical speech-noise models and independent component analysis. Separation algorithms can thus be considered highly important for everyday ASR in general. What is less clear is how to select the algorithms and features for the task.

One significant group of separation methods consists of spectral factorisation. Due to the novelty of this branch, current work mostly focuses on modelling static spectrogram features. Nevertheless, we know that important characteristics of speech and noise can be found in spectral dynamics, that is, local changes in spectro-temporal patterns. In MFCC-based recognition, it has been found beneficial to augment the base features with *time derivatives*, also known as *delta coefficients* [3]. Another approach suggested for long temporal context modelling is using TRAP features, where the emphasis is on long term behaviour of a few spectral bands [4]. In our exemplar-based framework, spectrogram windows spanning up to 300 ms can capture a lot of temporal context [5], but some of the dynamic information is lost in the additive model. It has been suggested, that dynamics can be emphasised in factorisation-based recognition by including temporal and spectral derivatives in the feature vectors [6].

In this work, we inspect further the efficiency of derivative features on top of optimised Mel magnitudes to improve the robustness of factorisation-based recognition. The work is organised as follows. First, we introduce in Section 2 our exemplar-based factorisation framework and its recognition method known as *sparse classification* (SC). Then we describe the concept of derivative features in Section 3. The CHiME challenge data, our basic setup and feature space experiments are described in Section 4, whereafter we conclude in Section 5.

2. EXEMPLAR-BASED SPARSE CLASSIFICATION

While many separation methods are based on statistical speech and noise models, in our approach we make the models more explicit by representing the observed features as a combination of *exemplars* — spectrogram segments sampled directly from the training material or the local context [5].

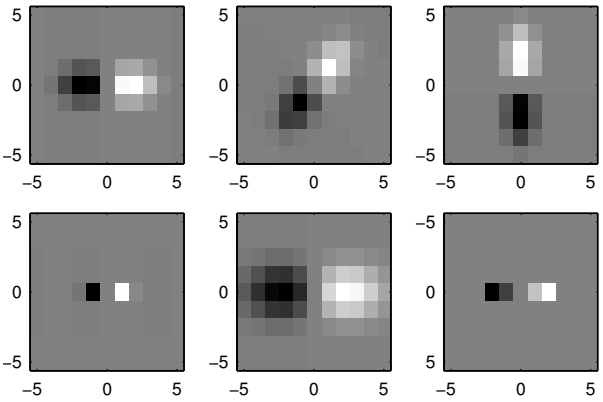
Each exemplar in our system is a $B \times T$ spectrogram matrix with B spectral bands and T consecutive frames. They are gathered to a *basis* or *dictionary*, which is used to model observed speech and noise features. Each observation window is represented as a linear combination of basis atoms. If we reshape the observation matrix to a vector $\mathbf{y}$ and each exemplar (basis atom) to a column vector $\mathbf{a}_i$, the problem becomes finding the activation weight vector $\mathbf{x}$ so that

$$\mathbf{y} \approx \sum_{i=1}^{m} \mathbf{a}_i x_i \tag{1}$$

where m is the number of exemplars in the basis. In matrix form the same equation can be given as $\mathbf{y} \approx \mathbf{A}\mathbf{x}$. Multiple observation windows can be given as parallel column vectors to solve the total activation matrix $\mathbf{X}$ ($m \times n$) for n windows at once. Finally, by assuming that basis and observation features are non-negative spectral magnitudes, and that activations should be non-negative too, finding $\mathbf{X}$ becomes a *non-negative matrix factorisation* (NMF) problem for a fixed basis. Enforcing additional sparsity on the solution ensures, that a few best fitting matches are favoured over unrealistically complex combination of multiple atoms. The iterative update rules used to find the $\mathbf{x}$ estimates are presented in [5].

To determine the utterance content from activations, each exemplar has a $Q \times T$ *label matrix*, describing the likelihood of each state $q \in [1, Q]$ over the exemplar's frames $[1, T]$. Label matrices are added together according to the corresponding exemplars' activation weights in temporal locations, where the activation was observed. This produces a $Q \times T_{\text{utt}}$ *likelihood matrix* for the whole utterance, which can be decoded using a standard Viterbi algorithm. The full procedure is described in earlier work [5, 7].

Figure 1: Spectro-temporal filters. Top row: 'Medium' length Gabor filters for temporal, diagonal and spectral direction. Bottom row: 'Short' and 'Long' Gabor filters, and length 2 HTK delta filter. Magnitudes are shown at a full greyscale range, thus not in scale.



As the decoding is based on activation weights and exemplar labels, there is no need to reconstruct the clean spectrogram or to synthesise the waveform for an external back-end. Even though spectrum or signal enhancement are also possible, in earlier work we have shown that direct classification performs better than the single-stream alternatives [5]. Multi-stream methods can improve the results significantly [8], but in this work we only use SC for simplicity and to eliminate the contribution of other components.

3. SPECTRO-TEMPORAL DERIVATIVE FEATURES

Current spectral factorisation algorithms are mostly employed in plain magnitude spaces, which model the activity in spectrogram bins, but not the dynamics over time and frequency. As in MFCC time derivatives, the NMF base features can be augmented by differential estimates. Because we are working in Mel spectrogram domain, it is possible to observe changes not only in time, but in any spectro-temporal direction by using 2-dimensional filters. The concept is similar to edge detection algorithms in image processing.

First, we construct a filter matrix in the spectro-temporal space. Then a derivative feature matrix is calculated by common 2-dimensional convolutive filtering of the static features, revealing the on- and offsets of spectrogram patterns. However, it should be noted that the differential estimates can have any sign, unlike the original non-negative magnitudes. To stay in the non-negative domain required by standard NMF algorithms, we must modify the features before factorisation.

The derivative feature matrix is reshaped to a vector, and represented by two vectors of the same size. The first contains the positive coefficients, and zeros where the vector was negative. Similarly, the second vector contains the absolute values of negative coefficients. If we denote the static features by a row vector $\mathbf{f}$ and its derivative by $d\mathbf{f}$, the augmented feature vector becomes

$$\hat{\mathbf{f}} = [\mathbf{f},\ d\mathbf{f}^{+},\ d\mathbf{f}^{-}] = [\mathbf{f},\ \max(d\mathbf{f}, \mathbf{0}),\ \max(-d\mathbf{f}, \mathbf{0})]. \quad (2)$$

If multiple derivatives are used, they are concatenated further to the vector as +/- pairs. Similar implementation was used in [6].

To learn the directions helpful for phonetic classification, we experimented with real-valued Gabor filters for multiple directions and sizes. Examples of filter matrices are shown in Figure 1, and they are described in more detail in Section 4.4.

4. EVALUATION

4.1. CHiME challenge data

The experiments were conducted using the PASCAL CHiME challenge database [1]. Its speech data consists of GRID corpus sentences, which follow a linear grammar of six word classes. The task is to recognise words belonging to the 'letter' and 'digit' classes, which contain 25 and 10 word options, respectively.

CHiME utterances are convolved with room response patterns, and mixed with household noises at six SNRs ranging from +9 to -6 dB. For training, there are 500 reverberated utterances for each of the 34 speakers, and six hours of plain background noise. The development and test sets consist of 600 utterances each, distributed between all speakers. Each set is repeated for all SNRs by mixing the utterances with different background segments containing an appropriate level of noise. All noisy utterances are presented within a long noise context as 'embedded' wave files. The development utterances are also available as 'clean' files with reverberation but no additive noise. Speaker identity is assumed to be known during recognition, while the target SNR is not.

4.2. Base setup

Our exemplar-based setup generally follows the one described in [7]. To reduce the number of parameters, we only use exemplar length of 20 frames (25 ms frame length, 10 ms shift), speaker-dependent speech bases and adaptively sampled noise bases in this work. The previous results for this setup and the GMM-based CHiME challenge baseline recogniser can be found in Table 3.

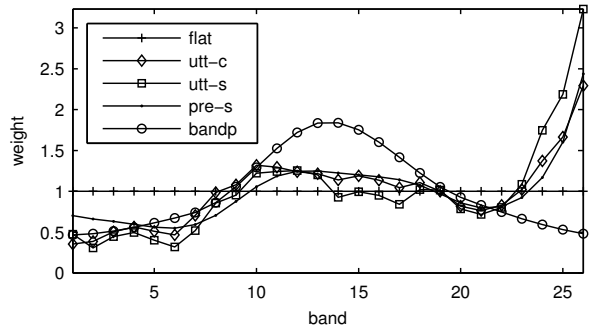
For each speaker, a speech basis is constructed by sampling 5000 exemplars from the 'clean' training speech semi-randomly. 5000 noise exemplars are also extracted for each test utterance by sampling the 'embedded' waveform files to both directions from the target utterance. In clean speech recognition, the noise basis is omitted. After converting all exemplars to Mel magnitudes and merging the speech and noise bases, a band weighting function is applied to define the contribution of each spectral band. Thereafter individual basis vectors are normalised to a Euclidean norm of 1.

Each test utterance is similarly converted into Mel magnitudes by extracting overlapping windows with a step of one frame. The band weights determined for the basis are applied to the observation as well. The observation windows are factorised to find out the activation vectors $\mathbf{x}$ as described in section 2. We initialise the activations to ones, and apply 300 rounds of an iterative update rule. The algorithm minimises the sum of estimation error (defined by KL-divergence) and a weighted L_1 penalty for non-zero activations.

As in earlier work, we used base sparsity values of 2.0 for speech and 1.7 for noise activations. However, the final sparsifying effect depends on the ratio between the penalty values and the 1-norms of basis vectors. The latter will increase by a factor of $\sqrt{R}$, if the length of 2-normed feature vectors is multiplied by R and their distribution remains similar. Therefore the $\sqrt{R}$ scaling is applied to the previously determined sparsity values, whenever the channel count, band number or derivative features change the feature vector length.

To avoid optimising for the test set, all parameter scans were performed on the development set. The 'clean' set was also used, although it does not belong to the final test set and is not included in any average values. The feature extractor was modified to use 512 FFT bins instead of the previous 256, producing small initial improvements over the earlier extraction. No changes were made to basis selection, factorisation or decoding algorithms. The learnt state mappings presented in [7] were not used in this work.

Figure 2: Mel band weighting curves for no adjustment ('flat'), online normalisation of the combined basis ('utt-c'), online speech basis normalisation ('utt-s'), precalculated normalisation from training speech ('pre-s') and bandpass filtering ('bandp').



4.3. Spectral band parameters

Before moving on to derivative features, we reoptimised the underlying static spectral magnitude space. In earlier work, we used 26 spectral bands calculated from 16 kHz signals as in the provided CHiME recogniser. The features were extracted separately for both channels, and the channel feature vectors were concatenated. These choices were re-evaluated as follows.

4.3.1. Band weighting

The Mel-scale distribution of speech and noise features is considerably uneven across bands. We can reweight the bands for two different goals; either to flatten the distribution for equal contribution of each band, or alternatively to emphasise certain bands for maximal classification quality. While the highpass filter commonly employed in MFCC extraction can improve clean speech recognition, we have found it too drastic for robust factorisation algorithms. Instead, five different weighting methods were tested:

1. No weighting ('flat')
2. Normalisation of the combined utterance basis bands ('utt-c')
3. Normalisation calculated from the speech basis only ('utt-s')
4. Precalculated normalisation of training speech bands ('pre-s')
5. Experimental bandpass filtering ('bandp')

Method 2 is our previous approach and depends on the adaptive noise basis of each utterance. Method 3 only depends on the current speech basis, that is, speaker identity. Methods 4 and 5 both produce fixed weighting, which simplifies the later steps. The bandpass weighting was included as an example of filter types, which emphasise the speech formant area and mostly discard frequencies over 4 kHz. All weighting methods are illustrated in Figure 2. For non-fixed weightings, means over all development data are shown.

The results are summarised in the first part of Table 1. We observe that 'do nothing' and online-computed speech weighting fare worse at certain SNRs than the other methods, which are approximately tied. Interestingly, the fixed weightings produce similar average rates, while bandpass filtering favour the clean end and precalculated speech normalisation the noisy one. The latter was chosen for further experiments due to its robustness, normalising effect and fixed shape. The differences between diverse weighting methods were generally small.

Table 1: Development set results for different spectral band parameter combinations. The format of experiment names is [band number] / [mono | stereo] / [weighting type].

SNR (dB)	clean	9	6	3	0	-3	-6	avg
26/s/flat	92.7	90.6	90.5	88.3	83.5	79.1	71.8	84.0
26/s/utt-c	93.7	91.8	91.8	89.8	83.5	78.5	72.2	84.6
26/s/utt-s	93.7	92.0	91.6	89.3	83.3	77.4	70.4	84.0
26/s/pre-s	93.6	91.4	90.8	89.3	84.7	78.9	72.7	84.6
26/s/bandp	93.7	92.0	91.7	89.8	83.8	78.8	71.8	84.6
26/m/pre-s	93.3	92.1	91.4	89.3	83.9	78.7	71.9	84.5
26/m/bandp	93.7	91.8	91.7	89.6	83.8	79.5	71.6	84.7
40/m/pre-s	93.6	92.3	91.6	89.8	85.0	79.7	72.7	85.2

4.3.2. Channel count

In our original parametrisation, binaural features were kept in separate entries of the feature vector, retaining some of the spatial information of the sound sources. To study whether it plays any role in recognition quality, the development set was also factorised using mono features by averaging the Mel magnitudes of channels. Apart from adjusting the sparsity value due to vector length halving, no other changes were made. Two fixed weighting curves, precalculated normalisation and bandpass filtering, were tested.

As can be seen from the results in Table 1 (rows 4–7), the accuracy of mono and stereo features is highly similar. Because mono features reduce the vector length and consequently computing costs by a half, they were used for further experiments.

4.3.3. Spectral band number

One fundamental question regarding feature selection is the number of Mel bands. To inspect this briefly, the band count was increased from 26 to 40. The results are shown on the last row of Table 1. We observe some $\sim$1% improvements and no decrements, suggesting that the gains may be worth the increased computational costs. While the next section was still evaluated using the original 26 bands, the final evaluation was performed on both values.

4.4. Spectro-temporal filters

After determining efficient base features, we tested three combinations of spectro-temporal Gabor filters: only temporal (forward and backwards), cardinal directions (temporal and spectral), and diagonal filters (45° angles). The prototype filter matrix was defined by

$$g(x, y) = \exp(-(\frac{x^2 + (\gamma y)^2}{2\sigma^2}) \sin(\frac{2\pi x}{\lambda}), \quad x, y \in [-5, 5] \quad (3)$$

with ellipticity γ set to 3, Gaussian envelope width factor σ to 2, and wavelength λ to 9, producing approximately one full sinusoid cycle. The prototype filter and two of its rotations are shown on the first row of Figure 1. The absolute sum of filter coefficients was set to 0.6 for each half of the filter. The results of augmenting directional filters to fixed-norm weighted mono features can be seen in the first part of Table 2. We notice that temporal direction improves the recognition rates, while including any of the spectral directions does not.

Settling for primarily temporal filtering, we tested the Gabor filter with its size increased and decreased by 50%, and in addition the delta filter employed by HTK using the default window length of 2 frames to both directions [3]. All were normalised to a 0.6

Table 2: Development set results for 2D filtering. Filter type is either Gabor [short | medium | long] in [temporal | cardinal | diagonal] directions, or HTK delta.

SNR (dB)	clean	9	6	3	0	-3	-6	avg
G/med/temp	92.8	92.0	91.7	89.5	83.9	80.3	73.2	85.1
G/med/card	92.9	91.8	90.3	89.3	83.8	78.1	71.7	84.1
G/med/diag	92.8	91.1	90.6	88.3	82.7	76.3	69.5	83.1
G/short/temp	93.3	92.2	92.2	90.3	85.6	81.4	73.5	85.9
G/long/temp	92.3	91.0	89.8	88.3	82.2	77.4	70.5	83.2
HTK delta	93.4	92.4	91.8	90.3	85.1	82.1	74.1	86.0

coefficient sum per side. The filters are shown on row 2 of Figure 1, and the results in the second part of Table 2. The best results were achieved using the shorter filters with little or no cross-band bleeding. The clean speech recognition rate does not improve over unfiltered base features, but the robustness against heavy noise increases. Changing the filter weight (not shown) did not produce any significant improvements.

4.5. Final test set evaluation

After optimisations, the test set was evaluated using the following parameter combination; mono features, precalculated speech-normalising band weights, and length 2 temporal delta filtering at weight 0.6. Both 26 and 40 spectral bands were used for determining their quality-cost tradeoff. The results are listed in Table 3. We notice significant improvements at each SNR in comparison to our earlier results. The word error rate is reduced by 13.9–32.8% at different SNRs, and the total error rate by up to 28.1%. Using 40 bands produces a large boost at -6 dB and modest gains elsewhere.

While the overall rates do not match the state-of-the-art results achieved in the CHiME workshop, where the best average score was 91.65% [9], it should be noted that the current highest ranking methods are relatively complex combinations of multiple techniques, whereas the approach presented here is a single stream classifier. Preliminary experiments suggests, that using sparse classification with complementary methods in multi-stream recognition can indeed achieve over 90% average recognition rate on the CHiME data already with the earlier, unoptimised features [8].

5. CONCLUSIONS

We studied alternative parametrisations of Mel features and their derivatives for factorisation-based speech recognition using CHiME challenge data and an exemplar-based sparse classifier.

First, we found out that the recognition algorithm is not particularly sensitive to band weighting, although some normalisation will improve the results over do-nothing. Mono features were found as effective as stereo for this data, allowing a 50% reduction in computational costs. Increasing the spectral band number from original 26 to 40 improved the results slightly.

Spectro-temporal filters were applied to the basis and observation features to model dynamic behaviour. Including temporal delta information produced significant improvements, while edge detection in spectral directions was found detrimental. The best temporal filters were relatively short with roughly 20ms temporal context to both directions, and no cross-band bleeding.

All in all, our feature space optimisation yielded 28.1% reduction in the total word error rate over all noisy conditions. Clean speech recognition rate remained at approximately 93–94%, which illustrates the difficulty of short word classification when no clues of word identity can be found from the neighbouring word context.

Table 3: Test set scores (%) for the CHiME baseline GMM recogniser, our previous SC features, and optimised features with their relative word error rate reductions (%) from the earlier results.

SNR (dB)	9	6	3	0	-3	-6	avg
GMM baseline	82.4	75.0	62.9	49.5	35.4	30.3	55.9
original SC, B=26	91.6	89.2	87.6	84.2	74.7	68.0	82.5
optimised SC, B=26	92.8	91.3	89.8	87.9	82.2	75.8	86.6
WER reduction	13.9	19.9	17.5	23.7	29.6	24.5	23.4
optimised SC, B=40	92.9	91.8	90.1	88.4	82.9	78.5	87.4
WER reduction	15.9	24.6	20.1	26.8	32.6	32.8	28.1

While the presented work was tested on the exemplar-based recogniser, it can be generalised to other algorithms based on non-negative spectral factorisation. The improved separation quality should prove useful both for feature-enhancing front-ends and for direct classifiers in standalone or combined recognition.

6. REFERENCES

[1] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1918–1921.

[2] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "Overview of the PASCAL CHiME Speech Separation and Recognition Challenge," in *Proc. Machine Listening in Multisource Environments (CHiME 2011), satellite workshop of INTERSPEECH 2011*, Florence, Italy, 2011.

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

[4] H. Hermansky and S. Sharma, "TRAPs – Classifiers of Temporal Patterns," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 1003–1006.

[5] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[6] M. Van Segbroeck and H. Van hamme, "Unsupervised Learning of Time-Frequency Patches as a Noise-Robust Representation of Speech," *Speech Communication*, vol. 51, no. 11, pp. 1124–1138, 2009.

[7] A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen, "Exemplar-based Recognition of Speech in Highly Variable Noise," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 1–5.

[8] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative Matrix Factorization for Highly Noise-Robust ASR: To Enhance or to Recognize?," in *Proc. ICASSP*, Kyoto, Japan, 2012.

[9] M. Delcroix et al., "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 *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 12–17.