Applying Non-Negative Matrix Factorization on Time-Frequency Reassignment Spectra for Missing Data Mask Estimation

Maarten Van Segbroeck*, Hugo Van hamme

Katholieke Universiteit Leuven, Department of Electrical Engineering – ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

{maarten.vansegbroeck,hugo.vanhamme}@esat.kuleuven.be

Abstract

The application of Missing Data Theory (MDT) has shown to improve the robustness of automatic speech recognition (ASR) systems. A crucial part in a MDT-based recognizer is the computation of the reliability masks from noisy data. To estimate accurate masks in environments with unknown, non-stationary noise statistics, we need to rely on a strong model for the speech. In this paper, an unsupervised technique using non-negative matrix factorization (NMF) discovers phone-sized time-frequency patches into which speech can be decomposed. The input matrix for the NMF is constructed using a high resolution and reassigned time-frequency representation. This representation facilitates an accurate detection of the patches that are active in unseen noisy speech. After further denoising of the patch activations, speech and noise can be reconstructed from which missing feature masks are estimated. Recognition experiments on the Aurora2 database demonstrate the effectiveness of this technique.

Index Terms: noise robust speech recognition, missing data techniques, mask estimation, speech separation, non-negative matrix factorization

1. Introduction

Additive noise results in a decrease in performance of ASR-systems due to the mismatch between the speech models obtained in clean conditions and the statistics of speech in the noisy test conditions. Missing Data Techniques have shown their effectiveness in reducing this mismatch. In the MDT approach, a missing data detector decides for each frame which features of the noisy speech are unreliable (missing) or reliable. The latter are treated as clean speech data in the acoustic models of the recognizer’s back-end. The missing features on the other hand are either marginalized or their value is estimated from the reliable data using the state distribution of the ASR as a prior (data imputation), see [1]. For reasons of accuracy, the ASR-system operates in a domain that is a linear transformation of log-spectra. Therefore, the data imputation technique needs to be extended to cover such linear transformations. An alternative MDT formulation was presented in [2] through the introduction of the ProSpect features and will be briefly restated in section 2.

One of the main advantages of missing data techniques over other noise reduction methods is that less assumptions need to be made about the noise type. Therefore, more constraints should be placed on the speech model. In this paper, the speech model consists of a collection of recurring acoustic patterns that are discovered and learned without supervision from clean

speech data. The learning algorithm involved makes use of non-negative matrix factorization (NMF) introduced by [3]. Thanks to the non-negativity constraints, NMF decomposes a matrix in additive (not subtractive) components, resulting in a parts-based representation of the data. NMF can therefore be seen as a learning algorithm that, when applied to an appropriate feature space, finds the parts or objects that the training data are built of. This approach was also used in a variety of other researches, see e.g. [4, 5, 6]. As will be explained in section 3, the patterns are discovered from a combination of two complementary feature representations that either reveal timing or frequency structure and which are derived from a reassigned time-frequency representation (RTFR) [7]. This allows the NMF to find patterns that reveal localized time-frequency structures in high resolution speech spectra. The advantage of learning patterns from a RTFR is that it facilitates the detection of these patterns in unseen noisy test utterances, e.g. the reassignment method improves the localization of the spectral energy in time and frequency, which already separates the speech to some extent from additive noise sources. The acoustic patterns will be referred to as *time-frequency patches* of speech and are highly interpretable on an acoustic and a phonetic basis. The patches can also model the dynamic spectro-temporal characteristics of speech, such as formant transitions.

The proposed missing data detector (MDD) will exploit this speech model by first retrieving the patches that are present in the high resolution representation of the noisy speech. Therefore, a noise model is also involved to explain the presence of the noise in the speech. This approach was also followed in [8] to improve the accuracy of the patch activations. The noise model can be trained offline or online during recognition. It suffices that this noise model roughly captures the spectral shape of the noise and it does not have to be representative for different levels of signal-to-noise ratio (SNR) in noisy speech. After a denoising procedure of the patch activations where only those patches are selected with highest activation energy, we can reconstruct the speech and the noise. Finally, these high resolution spectra are transformed into a low resolution representation in order to construct the missing feature masks for the MDT-based recognizer. A description of the proposed MDD will be outlined in section 4.

The experiments on the Aurora2 connected digit recognition task demonstrates the success of our system and are presented in section 5. Conclusions are given in section 6.

2. Missing Data Techniques

The speech recognizer is assumed to have a mainstream HMM-based architecture with Gaussian mixture models (GMM). In the front-end, a low resolution Mel-spectral representation is computed by a filter bank with D channels through windowing,

* This work was supported by “Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)”.

Figure 1: STFT representation (top), RTFR (middle) and enhanced RTFR (bottom) for the word “two” with focus on timing (left) and frequency (right) structure. In the bottom left panel, vertical lines correspond to plosive and vocal fold bursts, while the enhanced RTFR at the bottom right reveals the horizontal structure in the word, e.g. pitch and formant contours.

framing, FFT and filter bank integration. At a certain frame, let $\mathbf{s}$, $\mathbf{n}$ and $\mathbf{y}$ denote the vector of log-Mel spectral features for clean speech, noise and noisy signal respectively.

In a missing data approach, the components of $\mathbf{y}$ are divided into a reliable, $\mathbf{y}_r$, and an unreliable part, $\mathbf{y}_u$. The reliable components of $\mathbf{s}$ are approximated by $\mathbf{y}_r$. In the maximum likelihood per-Gaussian-based imputation [2], the missing part of $\mathbf{s}$ is estimated by minimizing the (negative) log-likelihood for each Gaussian mixture component over $\mathbf{s}$:

$$\frac{1}{2}(\mathbf{s} - \boldsymbol{\mu})' \mathbf{P}(\mathbf{s} - \boldsymbol{\mu}) \quad (1)$$

subject to the equality and inequality constraints:

$$\mathbf{s}_r = \mathbf{y}_r \text{ and } \mathbf{s}_u \leq \mathbf{y}_u \quad (2)$$

where $\boldsymbol{\mu}$ is the Gaussian mean and $\mathbf{P}$ is an inverse covariance or precision matrix, both estimated on clean training data. In most MDT systems, GMMs with diagonal covariance in the log-spectral domain are used, resulting in a diagonal structure for $\mathbf{P}$ and a tractable maximum likelihood estimation for $\mathbf{s}$. Higher accuracies are obtained with GMMs with a diagonal covariance in the cepstral domain, in which case $\mathbf{P}$ becomes non-diagonal. Imputation then becomes computationally more complex since the estimation of the unreliable part now requires the solution of a *Non-Negative Least Square* problem. In this paper, the ProSpect features defined in [2] will be used. Just like cepstra, they are computed by a linear transform of the logarithm of the filter bank energies. While these features can be applied in any speech recognition system, they show especially a clear benefit in MDT-based recognition since they reduce the computational requirements over cepstral MDT while the accuracy is maintained.

3. Discovery of time-frequency patches

The production of human speech can be regarded as a process of combining a small number of spectral patterns into many more different sequences. In this section, our goal is to find a set of patterns by analyzing continuous speech recordings. Therefore, we transform the speech data into a time-frequency reassignment spectrogram [7] on which we perform a non-negative matrix factorization (NMF). By imposing non-negativity constraints, NMF generates parts that are additive, unlike factorization techniques such as PCA or SVD. The finally obtained speech patterns are acoustic time-frequency patches of correlated energy.

3.1. Time-frequency reassignment

Time-frequency reassignment [7, 9] allows perfect localization of (well-separated) impulses, cosines and chirps, which constitute a reasonable model for speech. The corresponding re-assigned time-frequency representation (RTFR) has an increased sharpness of localization of the signal components without sacrificing the frequency resolution.

The RTFR is obtained by moving the spectral density value away from the point in the time-frequency plane where it was computed. It can be applied to different time-frequency representations each characterized by a different analysis kernel. In a RTFR, energy is reallocated from the geometric center of the spectral analysis kernel function to the center of gravity of the energy distribution. In this paper, the reassignment principle is applied to the short time Fourier transform (STFT). Let $H(t, \omega)$, $D(t, \omega)$ and $T(t, \omega)$ denote the STFT of the signal obtained with the window of choice $h(t)$, the derivative of $h(t)$ and the time weighted $th(t)$ respectively and let $\Re(X)$ and $\Im(X)$ be the real and imaginary part of X , then the energy at (t, ω) is reassigned to the center of gravity [7]:

$$(\hat{t}, \hat{\omega}) = \left(t + \Re \left[\frac{T(t, \omega)}{H(t, \omega)} \right], \omega - \Im \left[\frac{D(t, \omega)}{H(t, \omega)} \right] \right) \quad (3)$$

where the time and frequency offsets are computed from the ratios of the three STFTs.

To improve the visibility of acoustic events with short duration, we further enhance the localization of the energy along the time axis. Therefore, we first search for zero-crossing points in the time offsets of (3) and only those of them that are connected in the vertical direction (i.e. along the frequency axis) are retained. Finally, the corresponding energy of the RTFR is assigned to the retained points. When applied to speech with a sufficiently short analysis window, the enhanced RTFR clearly shows the vertical (i.e. well-localized in time) lines that are related to the burst of plosives and affricatives and energy releases by the vocal folds. By repeating the same procedure using the frequency offsets of (3), the tracks of time-varying spectral features such as pitch and formants can be clearly localized in frequency. We have found that formant structure is more apparent if we use shorter windows in the RTFR.

The different steps of the enhanced reassignment procedure are shown figure 1 for the word “two”. Firstly, a time-frequency representation is computed using a N -point STFT ($N = 128$). Subsequently, a RTFR is produced by reallocating the spectral energy to the gravity centers according to (3). The above mentioned enhancement steps are then applied to the RTFR to reveal either the timing or the frequency structure. Experiments have shown that an optimal choice for the window length is respectively 11 ms and 7 ms for male speakers and 6 ms and 4 ms for female speakers. The analysis window is shifted by 1 ms.

Experiments not included in this paper due to space constraints, have revealed that the enhanced RTFR is more robust than the STFT and produces better results for the approaches that will be described next.

3.2. Constructing the input matrix

In this section, we explain how the input matrix is created to which we will apply NMF for finding acoustic time-frequency patterns in speech signals. These patterns are discovered on clean training data. After pre-emphasizing the speech signals, we compute the enhanced RTFRs by the approach described in section 3.1. Both representations will be used to exploit the spectral information that is more apparent in either the vertical or the horizontal direction.

Figure 2: A collection of the discovered time-frequency patches with a duration of 60 ms. Some patches correspond to formant patterns and formant transitions (top row); others model inter-phone or silence-phone transitions, bursts and wideband spectra (bottom row).

Let us now define the spectral vector at a certain time t as $\mathbf{v}_t$ and $\mathbf{h}_t$, corresponding to the feature representation that reveals respectively the timing (vertical) and frequency (horizontal) structure. These representations will contain complementary information and therefore both are combined in the input matrix to discover the speech patches. Let us now define the real and non-negative column vector of dimension N as

$$\mathbf{c}_t = \max(\mathbf{v}_t, \mathbf{h}_t) \quad (4)$$

where the max-operator works element-wise. At each time step t , we take k consecutive columns of $\mathbf{c}_t$ representing the spectro-temporal structure of a short-time speech segment (i.e. of length k ms). These k columns are reshaped into a single column vector $\mathbf{C}_t$ of dimension kN , from which we construct a data matrix $\mathbf{C}$ of size $kN \times T$ by concatenating T columns from the clean training set.

3.3. Matrix factorization for unsupervised learning

By applying non-negative matrix factorization to the matrix $\mathbf{C}$, it is approximated by:

$$\mathbf{C} \approx \mathbf{B}\mathbf{A} \quad (5)$$

subject to the constraint that all matrices are non-negative and where the common dimension of $\mathbf{B}$ and $\mathbf{A}$ is much smaller than T and kN . Hence, equation (5) contains only additive linear combinations such that the factorization leads to a parts-based representation. These parts are found in the columns of $\mathbf{B}$ and their activation across time are given by the corresponding rows of $\mathbf{A}$. To retrieve a sparse representation of the data and to assure that each basic vector models a different time-frequency pattern, additional sparsity constraints must be enforced on $\mathbf{A}$. Therefore, we use sparse NMF [10] where the factorization is approximated by minimizing the objective function:

$$G(\mathbf{C}||\mathbf{B}\mathbf{A}, \lambda) = D(\mathbf{C}||\mathbf{B}\mathbf{A}) + \lambda \sum_{i,j} \mathbf{A}_{ij} \quad (6)$$

where the left term of (6) is the generalized version of the Kullback-Leibler divergence [3], while the right term in equation (6) enforces sparsity on $\mathbf{A}$ by minimizing the L_1 -norm of its columns. The trade-off between reconstruction accuracy and sparseness is controlled by the parameter λ . An algorithm for finding $\mathbf{B}$ and $\mathbf{A}$ given $\mathbf{C}$ based on multiplicative updates and with the additional sparseness constraint, can be found in [11]. To address scaling, the constraint that each column of $\mathbf{B}$ sums to unity is imposed. Experiments have shown that with the settings used in this paper, a good choice for λ is 1000.

Alternatively, convolutive NMF could be used to obtain a parts-based representation of the data [11]. Although this variant is appealing from a theoretical point of view, experiments, not reported in this paper, have shown that the activations of these patches are less robust to additional noise sources.

3.4. Time-frequency patches of speech

The columns of matrix $\mathbf{B}$ correspond to spectral patches which describe the recurrent time-varying spectra of speech. A set of 1000 time-frequency patches were discovered from the training set as described above with a patch length of 60 ms. A selection of these patches are shown in figure 2. Most patches describe formant movements over the duration of about a phone. A smaller set of time-frequency patches are modeling the beginning or ending of phones and phone-pair transitions. Others resemble wideband sounds and short-time energy bursts. Experiments have shown that a patch duration of 60 ms is optimal to obtain the recognition results of section 5.

4. NMF-based mask estimation

In the MDD, the high resolution RTFR of the noisy speech is described in terms of the discovered time-frequency patches. Patch activations are then computed and express to what extent each patch is present across time. To improve the accuracy of the activation matrix, some denoising steps are also performed during this process. From the final selected patches, speech and noise are reconstructed to estimate the spectral mask.

4.1. Denoising of patch activations during testing

To discover the patches that are present in test utterances of the Aurora2 database, the same procedure is used as in training except that we compute $\mathbf{A}$ in equation (5) by holding $\mathbf{B}$ fixed to the one obtained from training. However, to prevent that some of the speech patches are activated to explain the noise in the unseen noisy speech, some modifications are required during this process. Firstly, an extra noise model is included to explain the presence of additive noise sources. This noise model can be trained offline using the same method as described in section 3 or can be trained online. If we refer to the speech model with $\mathbf{B}_s$, to the noise model with $\mathbf{B}_n$ and use the subscript y to denote the noisy speech, then we can write:

$$\mathbf{C}_y \approx \begin{bmatrix} \mathbf{B}_s & \mathbf{B}_n \end{bmatrix} \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_n \end{bmatrix}. \quad (7)$$

A second necessary step to further denoise the speech activation matrix $\mathbf{A}_s$ is similar to a winner-takes-all strategy and is done during the iterative computation of eq. (7). Once convergence is reached, we select only those patches of $\mathbf{B}_s$ and $\mathbf{B}_n$ with highest activation. Therefore, we zero out, per column of $\mathbf{A}_s$ and $\mathbf{A}_n$, all the rows that do not correspond to these patches. Due to the property of the multiplicative update rules, these patches stay unactivated in further iterations. To this end, the total energy of speech and noise will be redistributed over the patches that are most probable to describe speech and noise.

In our experiments, 15 iterations are sufficient to reach convergence using a full activation matrix after which 5 iterations are performed using only the patches with highest activation energy.

4.2. Constructing the mask

Once equation (7) is solved using the additional denoising steps as mentioned in the previous section, we can reconstruct the data matrix of speech $\mathbf{C}_s \approx \mathbf{B}_s \mathbf{A}_s$ and noise $\mathbf{C}_n \approx \mathbf{B}_n \mathbf{A}_n$. After reshaping these matrices and a weighted summation to obtain N -dimensional high resolution spectra, a smoothing in time and frequency is performed to convert these spectra into a low resolution time-frequency representation. Time smoothing is performed by reframing with a sliding triangular window of 30 frames and a frameshift of 10 frames, after which a

mask type	SNR (dB)	Oracle masks				
		Subw.	Babble	Car	Exhib.	Avg.
binary	15	98.96	98.73	99.11	98.89	98.92
	10	97.70	98.19	97.82	96.73	97.61
	5	93.34	95.50	92.07	90.28	92.80
fuzzy	15	98.77	98.58	98.78	98.52	98.66
	10	97.67	97.94	98.42	97.13	97.79
	5	94.66	96.01	94.09	91.92	94.17

Table 1: Recognition accuracy (%) on Aurora2 test set A using ProSpect MDT with binary and fuzzy oracle masks.

mask type	SNR (dB)	Harmonicity masks				
		Subw.	Babble	Car	Exhib.	Avg.
binary	15	97.61	97.04	98.12	97.87	97.66
	10	94.87	93.29	95.02	93.86	94.26
	5	83.14	81.92	81.90	79.51	81.62
fuzzy	15	97.61	97.40	98.33	98.09	97.86
	10	95.12	94.50	95.47	94.14	94.81
	5	85.97	85.37	83.48	81.73	84.14

Table 2: Recognition accuracy (%) for harmonicity masks.

Mel-scaled filterbank is applied to smooth in the frequency domain. After logarithmic compression, the reconstructed spectra for speech and noise approximate the log-Mel feature vectors as used in section (2) and will be denoted by $\hat{s}$ and $\hat{n}$ respectively. Finally, the missing data masks are constructed by substituting these estimates into the binary or fuzzy mask decision criterion [12], i.e. respectively

$$\hat{s} < \hat{n} - \theta \quad \text{or} \quad \frac{1}{1 + \exp(-\rho(\hat{s} - \hat{n} - \theta))} \quad (8)$$

where the mask threshold θ and the slope parameter ρ are constants.

5. Recognition Experiments

Experiments are conducted on the TI-Digits speech database of Aurora-2 for the 4 noise types of test set A and SNR-levels between 15 dB and 5 dB. The MDT-recognizer is based on the HTK software package version 2.2 from Entropic. The digits are modelled as whole left-to-right HMMs with 16 states per digit and 20 Gaussians with diagonal covariance per state. The static features are modelled in the ProSpect domain and the velocity and acceleration features in the cepstral domain. Therefore, MDT will only be applied to the static features while the dynamic cepstra are left uncompensated.

As a reference, table 1 presents the accuracy obtained by the MDT-based recognizer using ideal masks constructed from the a priori knowledge of speech and noise, and table 2 presents the results from the (real) harmonicity masks of [13]. For all mask methods, the binary and fuzzy mask versions (see eq. (8)) are both evaluated. If the mask is fuzzy, the constrained optimization problem per Gaussian of [12] is used instead of the optimization problem described in section 2 which only holds for binary masks.

Finally, the recognition results of the proposed NMF-based mask estimator are shown in table 3. The noise model (section 4.1) was trained offline on a small part of the noise data containing the true noise type at 10 dB SNR. As can be seen from the table, this model holds equally well for other SNR-levels. Although the presence of the noise model has an important impact on the accuracy improvement compared to the harmonicity masks (especially for 5dB SNR), the main contribution to the performance should be explained by the strong speech model and the high resolution RTFR that is exploited to separate speech from noise.

mask type	SNR (dB)	NMF-based masks				
		Subw.	Babble	Car	Exhib.	Avg.
binary	15	97.88	97.22	98.39	97.66	97.79
	10	95.39	95.10	95.02	94.11	94.90
	5	88.95	87.22	86.88	86.15	87.30
fuzzy	15	97.97	97.61	98.54	98.24	98.10
	10	95.70	95.59	96.00	94.82	95.53
	5	89.13	88.97	88.25	87.47	88.46

Table 3: Recognition accuracy (%) for NMF-based masks.

6. Conclusions

We have presented a missing data mask estimation method that makes use of a speech model consisting of recurring acoustic patterns. These patterns are learned from a high resolution and reassigned time-frequency representation of the speech using NMF-techniques. Experiments conducted on the Aurora2 database have shown that these NMF-based masks perform uniformly better than those previously reported. Future work will include the application of our technique to large vocabulary continuous speech recognition using noise models that are adaptively retrained online.

7. References

- [1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," in *Speech Comm.*, vol. 34, 2001, pp. 267–285.
- [2] H. Van hamme, "PROSPECT features and their application to missing data techniques for robust speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, 2004, pp. 101–104.
- [3] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.
- [4] A. Park and J. Glass, "Towards unsupervised pattern discovery in speech," in *Proc. ASRU*, San Juan, Puerto Rico, Dec. 2005, pp. 53–58.
- [5] P. Smaragdis, "Convolutional speech bases and their application to speech separation," *IEEE Transactions of Speech and Audio Processing*, vol. 15, no. 1, pp. 1–12, Jan. 2007.
- [6] B. T. Meyer and B. Kollmeier, "Optimization and evaluation of gabor feature sets for ASR," in *Proc. ICSLP*, Brisbane, Australia, Sept. 2008, pp. 906–909.
- [7] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. on SP*, vol. 43, no. 5, pp. 1068–1089, 1995.
- [8] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. ICSLP*, Brisbane, Australia, Sept. 2008, pp. 411–414.
- [9] S. Hainsworth and M. Macleod, "Time-frequency reassignment: a review and analysis," Cambridge University Engineering Department, Tech. Rep. CUED/FINFENG/TR.459, 2003.
- [10] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
- [11] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutional non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, vol. 72, pp. 88–101, Dec. 2008.
- [12] M. Van Segbroeck and H. Van hamme, "Robust speech recognition using missing data techniques in the prospect domain and fuzzy masks," in *Proc. ICASSP*, Las Vegas, U.S.A., Apr. 2008, pp. 4393–4396.
- [13] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP*, Montreal, Canada, 2004, pp. 213–216.