Model Order Estimation Using Bayesian NMF for Discovering Phone Patterns in Spoken Utterances

Sayeh Mirzaei[1], Hugo Van hamme[1], Yaser Norouzi[2]

[1]Department of Electrical Engineering-ESAT, KU Leuven, Belgium
[2]Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran
smirzaei@esat.kuleuven.be, hugo.vanhamme@esat.kuleuven.be, y.norouzi@aut.ac.ir

10.21437/Interspeech.2013-37

Abstract

In earlier work, we have shown that vocabulary discovery from spoken utterances and subsequent recognition of the acquired vocabulary can be achieved through Non-negative Matrix Factorization (NMF). An open issue for this task is to determine automatically how many different word representations should be included in the model. In this paper, Bayesian NMF is applied to estimate the model order. The per-utterance word activations are given a gamma prior while the word models are assumed deterministic. Two Bayesian approaches are applied for obtaining optimal parameter values. First, the penalized joint log-likelihood of the parameters is considered as the objective function. Then, maximal marginal likelihood estimator (MMLE) is implemented which obtains the word models maximizing the likelihood after integration over the activations. The variational Bayesian algorithm, which maximizes a lower bound of the marginal log-likelihood, is applied to this optimization problem. The number of required latent components or basis vectors (model order) is estimated by evaluating likelihood metrics. The inferred model order is validated by observing error criteria on a test set. Experiments on synthetic data as well as real speech show that MMLE is more effective for the purpose of model order selection.

Index Terms: Model order estimation, Bayesian Non-negative Matrix Factorization (NMF), Maximum Joint Likelihood Estimation (MJLE), Maximum Marginal Likelihood Estimation (MMLE)

1. Introduction

Automatic Speech Recognition (ASR) systems have evolved a lot over the last decades and spoken human-machine interaction is finding its way into our daily lives: (car) navigation destination entry, directory enquiry, (email) dictation, voice search, etc. In all of these applications, the recognizer has a vocabulary that is static in the sense that it has been determined during application design. There are however applications where static vocabularies are not adequate because it is hard to predict which words a user will choose. Imagine giving instructions to a service robot. Most likely, you will use words that have very specific meanings in your home environment and you may use brand names that are local for your area and which are subject to change over time. In applications such as dictation software, this issue is addressed by offering a controlled procedure for extending the vocabulary, but it is difficult to imagine how this would work in applications where a *meaning* needs to be associated to the user-specific words.

Inspired by the way children *acquire* their ecologically relevant vocabularies, we have studied methods based on non-negative matrix factorization (NMF) [1] to find recurring acoustic patterns and relate these to events in other modalities [2][3], hence giving them a meaning. In applications such as home automation, *acquiring* the vocabulary from user interaction examples seems a viable approach, even for small vocabularies [4]. The task of the NMF is to find recurring acoustic patterns, *keywords*, that relate to actions on the user interface.

One of the open issues in this approach is to determine *how many* acoustic patterns are needed to properly model the acoustic data, i.e. how many different words is the speaker using? In earlier work, we have always assumed we know this in advance. Here, we evaluate to which extent *Bayesian NMF* can provide an answer to this question.

This paper is organized as follows: In section 2, the two Bayesian NMF approaches are briefly described. In section 3, the data construction procedure and training and test stages are explained. The results are presented in section 4. Finally, the conclusive remarks are stated in section 5.

2. Bayesian NMF

Non-negative matrix factorization has found vast popularity in several data analysis applications. The non-negative matrix $\mathbf{V}$ is approximated as the product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ of size $F \times K$ and $K \times N$ respectively. The elements of $\mathbf{W}$ and $\mathbf{H}$ are obtained by solving an optimization problem with the goal function which is defined based on the difference measure between the matrices $\mathbf{V}$ and $\mathbf{W}\times\mathbf{H}$. A particular choice for this measure is the generalized Kullback-Leibler (KL) divergence which is written as:

$$D_{KL}(\mathbf{A} \mid \mathbf{B}) = \sum_{f=1}^{F}\sum_{n=1}^{N}(a_{fn}\log\frac{a_{fn}}{b_{fn}} - a_{fn} + b_{fn}) \quad (1)$$

In [1], an algorithm with multiplicative updates is proposed which iteratively finds the parameters that minimize the divergence criterion. This approach is equivalent to its statistical interpretation, i.e. considering a Poisson generative model for data and maximizing the joint likelihood of data conditioned on the $\mathbf{W}$ and $\mathbf{H}$ parameters [5].

If a suitable prior distribution is presumed for the parameters of $\mathbf{W}$ and $\mathbf{H}$, we are involved in a Bayesian approach for non-negative matrix factorization. The Bayesian extension of the standard NMF solution provides more powerful modeling. Bayesian inference of the parameters allows us to adaptively perform model order selection based on the data. In [6], a full Bayesian approach has been applied for model selection. In this work, the Poisson distribution was assumed for the data and the elements of $\mathbf{W}$ and $\mathbf{H}$ are taken

with Gamma distribution as a prior. In [7], the dictionary elements of **W** are left deterministic while a Gamma prior is assumed for **H**. The generative model assumed for the data is:

$$v_{fn} \sim \mathrm{P}(v_{fn} \mid \sum_k w_{fk} h_{kn}) \tag{2}$$

where P denotes the Poisson distribution defined by:

$$\mathrm{P}(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{\Gamma(x+1)} \tag{3}$$

The activation coefficients h_{kn} are taken as random variables with Gamma prior, $h_{kn} \sim g(h_{kn} \mid \alpha_k, \beta_k)$ where

$$g(x \mid \alpha, \beta) = \left[\beta^{\alpha}\Gamma(\alpha)\right]^{-1} x^{\alpha-1} e^{\frac{-x}{\beta}},\ x \geq 0,\ \alpha > 0, \beta > 0$$

.

In our case, the data are histograms (see section 3), thus Poisson would be a suitable choice for the data distribution. Choosing a Gamma distribution as a prior can simplify the Bayesian inference procedure. Moreover, it is capable of imposing sparsity on the elements of **H**. In the following, the two different Bayesian methods which are implemented in this work are described.

2.1. MJLE

The first approach is maximum joint likelihood estimation (MJLE). The aim is to maximize the penalized joint log-likelihood of **W** and **H**.

$$C_{JL}(\mathbf{W}, \mathbf{H}) \triangleq \log P(\mathbf{V} \mid \mathbf{W}, \mathbf{H}) + \log P(\mathbf{H}) \tag{4}$$

where the term $\log P(\mathbf{H})$ can be regarded as a regularization term added to the joint likelihood function. A minorization-maximization (M-M) algorithm is presented in [7] for obtaining the updates of **W** and **H**. Here we apply the norm-constrained **W** update rules which impose the constraint that individual columns of **W** have unit l_1-norm. For the sake of conciseness, the rules for updating the **W** and **H** parameters are not repeated here. The penalized joint likelihood is evaluated in this case for the purpose of model order estimation.

2.2. MMLE

The MJLE method is prone to overfitting since the number of parameters is growing with the number of data points. To avoid this, the maximum marginal likelihood estimation (MMLE) scheme was proposed in [7] which has been shown to automatically prune out irrelevant components (columns) of the **W** matrix, hence being capable of estimating the proper model order. The log-likelihood is integrated over the **H** parameters and the obtained marginal log-likelihood is to be maximized

$$C_{ML}(\mathbf{W}) = \log P(\mathbf{V} \mid \mathbf{W}) = \log \int_{\mathbf{H}} P(\mathbf{V} \mid \mathbf{W}, \mathbf{H}) P(\mathbf{H}) d\mathbf{H} \tag{5}$$

Since this integral is intractable, a lower bound on the marginal log-likelihood is maximized in a variational Bayesian expectation maximization (VBEM) setting. It is based on the variational approximation of the exact posterior. The update relations can be found in [7]. For extracting the model order, the lower bound of the marginal log-likelihood is evaluated for different order values.

3. Experimental Procedure

In the present study we address the problem of finding recurring acoustic patterns *without supervision.* The underlying data representation that enables NMF to solve this problem is the *histogram of acoustic co-occurrences* (HAC) [2]. This is an utterance-level bag-of-features representation, i.e. a list of unordered (hence "bag") frequencies of ordered symbol pairs (hence "co-occurrence") with acoustic relevance. It is key that the HAC representation of an utterance will be the weighted addition of the HAC representations of the recurring patterns. Hence, the pattern HAC representations can be found without any supervision by NMF from multiple HAC utterance representations. The construction of the HAC vectors is described in detail in [8]. In short, speech is represented by a stream of MFCC vectors and their derivatives. Each frame is characterized by the posterior probabilities on a codebook of *M* Gaussians. Now consider the joint probability that the frame at time *t* is generated by Gaussian *i* while the frame at time $t+\tau$ is generated by Gaussian *j*. The $F = M^2$-dimensional HAC representation of an utterance is then this joint probability accumulated over the utterance. This process is repeated for *N* utterances, stacking the HAC-representations in a non-negative $F \times N$ data matrix **V**. Because the HAC representation of each utterance is approximately equal to the sum of the HAC representations of the *K* acoustic patterns that need to be discovered, **V** can be approximately decomposed in the matrix product $\mathbf{W} \times \mathbf{H}$, where **W** is a $F \times K$ matrix containing the HAC representations of the *K* patterns and the $K \times N$ matrix **H** contains the number of times the pattern was present in the respective utterances.

3.1. Training

In the training stage, the training data matrix is decomposed into non-negative matrices $\mathbf{W}_{train}$, with the columns corresponding to models for acoustic patterns, and $\mathbf{H}_{train}$ containing the activations for these patterns. Standard NMF as well as the two different Bayesian NMF approaches are applied for this purpose. Likelihood metrics corresponding to each scheme are utilized for model order estimation.

3.2. Testing

The obtained components in $\mathbf{W}_{train}$ are reused for decomposing the test data matrix, i.e. it is decomposed as $\mathbf{W}_{train} \times \mathbf{H}_{test}$ by minimizing the divergence criterion. To interpret and evaluate the quality of the result, the HAC-representations of the acoustic patterns obtained in the columns of $\mathbf{W}_{train}$ are related to the ground truth, which is represented in a matrix **G** with *m*,*n*-th entry equal to the number of times keyword *m* occurs in utterance *n*. Since the *k*,*n*-th entry of **H** represents the activation of acoustic pattern *k* in utterance *n*, we can estimate a nonnegative map-ping matrix **Q** such that $\mathbf{G} \approx \mathbf{Q} \times \mathbf{H}$ using NMF. The *m*,*k*-th entry of **Q** then represents the contribution of pattern *k* to word *m*. Notice that **Q** is estimated on the training data, *after* $\mathbf{H}_{train}$ is estimated from the acoustic evidence *without* using any ground truth information.

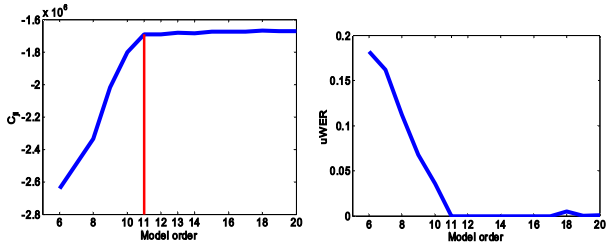

Figure 1: MJLE for ideal data. Left: Penalized joint likelihood. Right: unordered word error rate

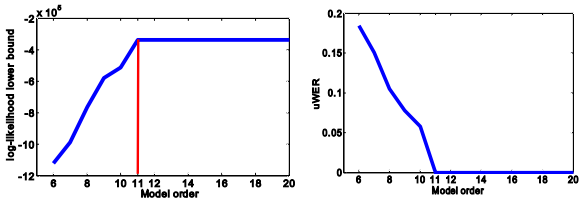

Figure 2: MMLE for ideal data. Left: Log-likelihood lower bound. Right: unordered word error rate

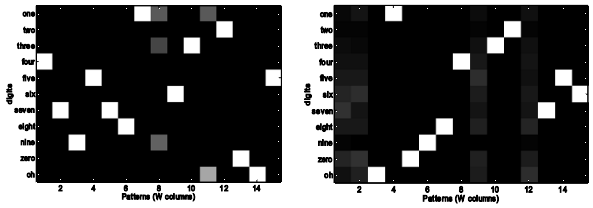

Figure 3: The **Q** matrix with order = 15 for ideal data. Left: MJLE. Right: MMLE

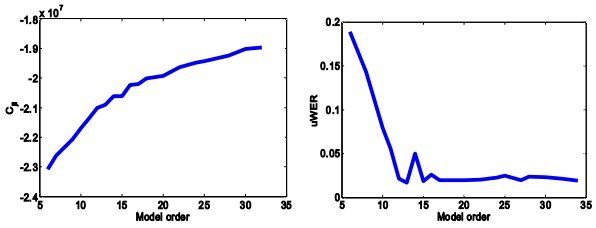

Figure 4: MJLE for female data. Left: penalized joint likelihood. Right: unordered word error rate.

Subsequently, the error metrics can be found by comparing the estimated grounding matrix $\mathbf{G}_{est} = \mathbf{Q} \times \mathbf{H}_{test}$ with the true grounding matrix, $\mathbf{G}_{test}$. This is expressed in the unordered word error rate (uWER) [2], which compares the identities of the top L_n words of column (utterance) n of $\mathbf{G}_{est}$ with the identities of the L_n nonzero entries of $\mathbf{G}_{test}$.

4. Results

The performance analysis is investigated by applying the algorithms on two data sets for which the vocabulary size is known. The first one is an idealized setup for which the NMF model holds exactly. The second one is real spoken data. Both data sets are balanced in terms of occurrence frequencies of the words.

4.1. Synthetic ideal digit strings

To obtain data for which the NMF model holds exactly, we start from written transcriptions of digit strings where each of the 11 digits has a unique transcription. The symbols in the HAC representation are taken to be the letters occurring in the transcription instead of the Gaussians used on real speech data. Hence the HAC representation is formed by counting the co-occurrence frequency of adjacent letter pairs (letter bigrams) in each utterance. Since we insert inter-word blanks, the HAC representation of an utterance can be written exactly as the sum of the HAC representation of each word occurring in the utterance. The dimension of the HAC-representation is 256, the number of different letter pairs. 3000 utterances were generated for training and the test data contains 1000 utterances. The number of digits for each utterance is chosen uniformly in the interval [2 5]. The gamma distribution hyper parameters are taken as $\alpha_k = 2$ and $\beta_k = 1$ for all of the **H** elements. The initial values of both **W** and **H** elements are taken as absolute value of a random normal variable (with mean 0 and variance 1) plus 1. The number of VBEM iterations is 1000.

First, the MJLE approach is implemented. The penalized joint likelihood is plotted for different number of columns assumed for **W** (model order) in fig.1. The inferred order evidently equals 11 which is the true number of digits. This figure also shows the unordered word error rate graph which implies the same model order.

The same graphs are plotted for MMLE approach in fig.2. The order can be implied the same way in this case. It should be noted that in the case of MJLE, when the order is taken more than 11, the extra columns of **W** are not near zero and contain repeated patterns due to overfitting. But for MMLE, the extra columns are approximately zero. This is an attractive property of MMLE which allows the proper order to be estimated by observing **W** columns as well. The mapping matrix **Q** for order 15 is shown in fig.3. It can be perceived from this plot that for MMLE, the extra columns of **W** are not mapped to the digits. However for MJLE, two columns are mapped to digits 5 and 7.

4.2. Spoken digit strings

The real data contains the utterances from the clean training set and the clean test utterances from set A of the AURORA-2 database [9]. This database contains utterances of 11 digits from male and female speakers, including leading and trailing silence, sometimes with inter word silence. The HAC representations are extracted based on 200 trained Gaussians and a single lag (lag = 50ms).

The window length for spectral analysis is taken 20ms and the frame shift is 10ms. The MFCC extraction uses 30 Mel-filter banks from which 12 MFCC coefficients are computed plus the frame's log-energy. Together with the first and second order delta's, a 39-dimensional feature vector is obtained. The initial gamma distribution hyper parameters as well as the initial values for **W** and **H** are taken the same as in section 4.1. The number of iterations for updating the parameters is 1000.

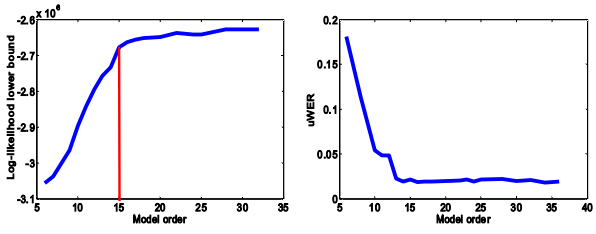

Figure 5: MMLE for female data. Left: VBEM lower bound of the log-likelihood. Right: unordered word error rate.

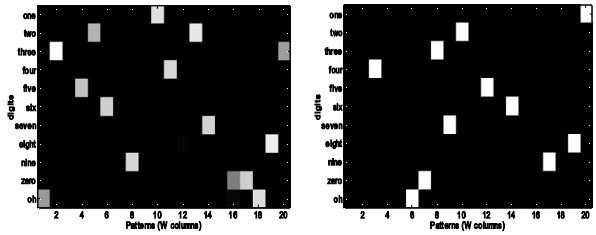

Figure 6: The **Q** matrix with order = 20 for female data. Left: MJLE. Right: MMLE

In this case, 3864 utterances are utilized in training and the test set contains all 4004 clean test utterances. First, the female utterances are considered. The MJLE likelihood and error metrics corresponding to female utterances are shown in fig.4. It is difficult to infer the order by lack of a knee point. However, applying MMLE for data decomposition leads to more implicative results, as illustrated in fig.5. The estimated order based on these graphs is 15. Obviously, the MMLE lower bound is a more powerful criterion for estimating the proper order since it discards irrelevant columns of **W**.

It seems that extra columns in **W** (more than 11) are required for modeling the silence sections that exist in the utterances. Also, some of the counts in the HAC representation stem from cross-word co-occurrences and don't follow a model of order 11. The **Q** matrix for order 20 is shown in fig.6. For MJLE, there is more than one column associated to a single digit. This is a consequence of overfitting as it was also observed in the form of repeated patterns in the case of ideal data. But for MMLE, the extra columns are not mapped to the digits and contain near zero elements. Furthermore, applying MMLE does not lead to cross patterns, i.e. an acoustic pattern is never used for two distinct words. However, in the case of ordinary NMF, cross-patterns do occur.

The performance metrics are also computed for complete data consisting of male and female utterances. The results are represented in fig.7 for MMLE. The estimated order is equal to 30 in this case, again leading to the best accuracy. This confirms the finding in [8] that the HAC representation requires gender-specific models for good accuracy.

The KL divergence for the ordinary NMF solution is also measured for different orders and plotted in fig.8. Like for MJLE, it is hard to infer a model order from these graphs. This reveals the usefulness of the Bayesian MMLE for estimating the order.

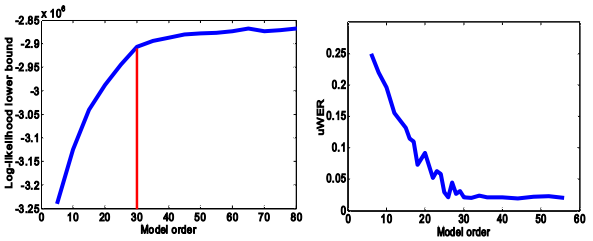

Figure 7: MMLE for complete data. Left: VBEM lower bound of the log-likelihood. Right: unordered word error rate.

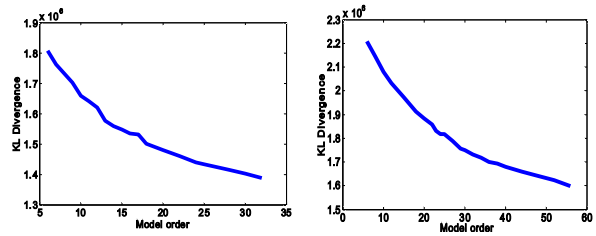

Figure 8: KL divergence of ordinary NMF. Left: female data. Right: complete data.

5. Conclusions

We proposed to apply the Bayesian NMF approach for discovering spoken words in speech utterances. The results show that this method is indeed capable of estimating the required number of patterns (model order) properly from the knee point in the likelihood function, and that the selected order leads to a low recognition error rate. It has been shown that the MMLE approach outperforms the MJLE for this purpose. We also observed that **W** columns for order values greater than the true number of words present in the data, are near zero. This is due to the ability of MMLE to discard irrelevant columns. As mentioned in the results section, some columns are required for modeling the silence and cross-word phenomena. The corresponding columns can be identified by observing the mapping matrix **Q**. We also presented that model order cannot be found by observing the KL divergence in the case of ordinary NMF.

6. Acknowledgements

This research was funded by the KU Leuven research grant OT/09/028(VASI).

7. References

[1] D. D. Lee and H. S.Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401,pp. 788-791, 1999.

[2] H. Van hamme, "HAC-models: a Novel Approach to Continuous Speech Recognition," *Proc. International Conference on Spoken Language Processing*, pp. 2554-2557, Brisbane, Australia, September 2008.

[3] J. Driesen, H. Van hamme and W. B. Kleijn, "Learning from Images and Speech with Non-negative Matrix Factorization Enhanced by Input Space Scaling," *In Proc. ARPA Spoken Language Technology Workshop*, Berkeley, USA, December 2010.

[4] Jort F. Gemmeke , J. van de Loo , G. De Pauw , J. Driesen , H. Van hamme and W. Daelemans, “A Self-Learning Assistive Vocal Interface Based on Vocabulary Learning and Grammar Induction,” *In Proc. Interspeech*, Portland, OR, USA, September 2012.

[5] T. Virtanen, A.T. Cemgil, S. Godsill, “Bayesian extension to non-negative matrix factorization for audio signal modeling,” *Proceedings of IEEE International Conference on Accoustic, Speech and Signal Processing (ICASSP 08)*, pp. 1825-1828, Las Vegas, Nev, USA, March-April 2008.

[6] A. T. Cemgil, “Bayesian inference in non-negative matrix factorisation models,” *Technical Report CUED/F-INFENG/TR.609*, University of Cambridge, July 2008.

[7] O. Dikmen, C. Fevotte, “Maximum Marginal Likelihood Estimation for Nonnegative Dictionary Learning in the Gamma-Poisson Model,” *IEEE Transactions on Signal Processing*, vol.60, no. 10, pp. 5163-5175, Oct. 2012.

[8] M. Sun and H. Van hamme. “Unsupervised Vocabulary Discovery Using Non-Negative Matrix Factorization With Graph Regularization,” *Proc. International Conference on Acoustics, Speech and Signal Processing*, pp. 5152-5155, Prague, Czech Republic, May 2011.

[9] H.-G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,". *In Proc. ISCA ITRW ASR2000 Workshop*, Paris, France, September 18-20, 2000