TRI-FACTORIZATION LEARNING OF SUB-WORD UNITS WITH APPLICATION TO VOCABULARY ACQUISITION

Meng Sun, Hugo Van hamme

Department of Electrical Engineering-ESAT, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium

mengsun@esat.kuleuven.be, hugo.vanhamme@esat.kuleuven.be

ABSTRACT

In prior work, we proposed a method for vocabulary acquisition based on a co-occurrence model and non-negative matrix factorization. The vocabulary is described in terms of co-occurrence statistics of frame-level acoustic descriptions and suffers from poor scalability to larger vocabularies. Much like whole-word HMM models, there is no reuse of a sub-word units such as phone models. In this paper, we apply the co-occurrence framework to learn a set of sub-word units unsupervisedly using a matrix tri-factorization and propose a method for computing their posteriorgram and finally show vocabulary acquisition from the posteriorgram. The method outperforms our prior work in that it can learn from a smaller set of labeled data and shows a better recognition accuracy.

Index Terms— semi-supervised learning, vocabulary acquisition, pattern discovery, spectral embedding

1. INTRODUCTION

Computational approaches to language acquisition have received increasing attention over the years [1, 2] because of their relevance in cognitive robots and in the modeling human language acquisition. In this research, a first task is to discover words in speech where the transcription is lacking. (Weak) Supervisory information may only occur from utterance-level labels stemming from detected events in other modalities like the video inputs [1], or the ground word tags [2].

Some models for unsupervised or weakly supervised spoken pattern discovery have been proposed, such as segmental dynamic time warping (DTW) [3] and DP-ngrams [4]. They search for recurrent acoustic patterns based on the *local alignment* of segments in training data. Those recurrent traces are subsequently associated to vocabulary patterns. In previous work based on co-occurrence statistics and non-negative matrix factorization (NMF), we have proposed an alternative model for unsupervised vocabulary discovery, which performed well on a small database [5]. The NMF model decomposes the data into recurrent parts, so it is supposed to have a *global view* of the data. The problem of this model is its restricted scalability to large vocabularies and the lack to view speech as a time series, which are actually two strengths of hidden Markov models (HMM). So the contribution of this paper is to discover hidden units and their transitions, i.e. an HMM-like probabilistic graph model, from the co-occurrence statistics.

The recently proposed spectral learning/embedding models of HMM show strong relations between co-occurrence statistics of observations and the hidden states [6, 7]. So we can discover sub-structures or hidden units from the discovered recurrent structures represented by the co-occurrence statistics with the idea of spectral embedding, i.e. matrix tri-factorization. It has two advantages. One is to find correct reusable intermediate units to improve the semi-supervised learning process (this paper) as in deep learning [8], an important step towards large vocabularies based on reusability of the units (future work). The other is to obtain an HMM-like probabilistic graph model but without the frame-level strict Markovian assumption for automatic speech recognition (ASR) (future work).

In this paper, we first extract repeated structures represented in the histogram of acoustic co-occurrences from partial or the complete training data. Then each extracted co-occurrence structure is modeled as an HMM-like probabilistic graph of the underlying hidden sub-word units using non-negative matrix tri-factorization (NMTF). A sequential labeling approach is subsequently applied to transform the representation of utterances from the acoustic level to the level of sub-word units. Experiments on vocabulary discovery are conducted to evaluate the two kinds of representations: acoustic co-occurrences and co-occurrences of sub-word units.

2. EXPLOITING AND UTILIZING HIDDEN UNITS

2.1. Tri-factorization learning of hidden units

Suppose an utterance is represented by its Gaussian posteriorgram, i.e. if x_t ($1 \leq x_t \leq M$, and M is the number of Gaussians) denotes the Gaussians that generated the signal analysis frame at time t, the posteriorgram is $\{P(\mathbf{x}_t),\ t =$

The research was funded by the K.U.Leuven research grant OT/09/028(VASI).

$1:T\}$. $P(\mathbf{x}_t)$ is the vector of Gaussian posterior probabilities of frame t, i.e. $(P(\mathbf{x}_t))_i = P(x_t = i)$. The potential hidden units it contains can be discovered from the following relations.

$$\begin{array}{rl} & P(x_t = i, x_{t+\tau} = j) \\ = & \sum_{k,l} P(x_t = i | y_t = k) \\ & P(y_t = k, y_{t+\tau} = l) P(x_{t+\tau} = j | y_{t+\tau} = l) \end{array} \quad (1)$$

where, i, j are the Gaussian indices, k, l are the indices of the hidden units, τ is the temporal parameter to define contextual dependencies between frames, and $\{P(\mathbf{y}_t),\ t = 1:T\}$ with $(P(\mathbf{y}_t))_i = P(y_t = i)$ is the posteriorgram of the utterance labeled by the hidden units. $P(x_t = i, x_{t+\tau} = j)$ and $P(y_t = k, y_{t+\tau} = l)$ are joint probabilites of Gaussians and the hidden units respectively.

The matrix formulation of Eq.(1) is $C^\tau = AB^\tau A^T$ where $C^\tau_{ij} = P(x_t = i, x_{t+\tau} = j)$ and $B^\tau_{kl} = P(y_t = k, y_{t+\tau} = l)$. We assume $A_{ik} = P(x_t = i | y_t = k) = P(x_{t+\tau} = i | y_{t+\tau} = k)$, i.e. the association matrix A between Gaussians and hidden units is time invariant because the dynamic behavior of hidden units is only reflected in B^τ. With multiple τ's, the learning algorithm of A, B^τ from C^τ is shown in Table 1.

2.2. Extraction of recurrent structures represented by the Gaussian co-occurrences

The quality of the hidden units obtained from the tri-factorization learning depends on the quality of the estimate of the joint probability C^τ. The learning model usually works well for a low-rank decomposition which corresponds to some HMM with a small number of states. Therefore R_1 recurrent structures represented by Gaussian co-occurrences $\{C^{r,\tau},\ r = 1, \ldots, R_1\}$ are first extracted from the training data by using an unsupervised NMF model in Eq.(2) [5].

$$V \approx WH \quad (2)$$

In the NMF model, each column of matrix V is the representation of a training utterance by its accumulated co-occurrence probabilities of Gaussians, which is obtained by flattening the matrix of accumulated co-occurrence probabilities of Gaussians, $\sum_{t=1}^{T-\tau} P(x_t = i, x_{t+\tau} = j)$, to a vector. Multiple contextual dependencies of τ are allowed, each of which produces a co-occurrence vector. The co-occurrence vectors of the same utterance are stacked to form a super vector as a column of V. With the factorization of Eq.(2), we obtain the recurring structures represented by Gaussian co-occurrences in the columns of W. Subsequently the sub-vector of the column $W_{:,r}$ with the same τ is rearranged in the $M \times M$ Gaussian co-occurrence matrix $C^{r,\tau}$, where τ is the contextual dependencies and r is the column index of W and $1 \le r \le R_1$.

Then $\{C^{r,\tau}, \tau = 1, 2, 3 \ldots\}$ are jointly factorized by the algorithm in Table 1 to obtain A^r and $B^{r,\tau}$. $B^{r,\tau}$ is initialized with a sub-band diagonal structure to generate a *left-to-right chain model* which will be called an *HMM-like probabilistic graph model*. The learned hidden units are stacked as $A = [A^1 A^2 \ldots A^{R_1}]$ and $B^\tau = \text{blkdiag}(B^{1,\tau}, B^{2,\tau}, \ldots, B^{R_1,\tau})$. The end units and head units of the $B^{r,\tau}$'s are connected in B^τ by filling small positive numbers at the corresponding locations. In this HMM-like probabilistic graph model, A performs as the observation matrix, and B^τ performs as the transition matrix.

2.3. Construction of the posteriorgram of the hidden units

Sequential labeling of the training and testing utterances are performed by using the learned HMM-like graph model from the training data. For the frame at time stamp t, we consider three probability contributions for the hidden units: observation, forward transition and backward transition.

The first probability estimate for the hidden units, $P_o(\mathbf{y}_t)$, comes from the observation at this time stamp $P(\mathbf{x}_t)$ which is a vector with Gaussian posterior probabilities. $P(\mathbf{x}_t | y_t = k)$ in Eq.(3) is a column of the matrix A, so $P_o(\mathbf{y}_t)$ can be estimated by NMF decoding.

$$P(\mathbf{x}_t) = \sum_k P(\mathbf{x}_t | y_t = k) P_o(y_t = k) \quad (3)$$

The second estimate of the probability on the hidden units is from the forward transition as is shown in Eq.(4).

$$(P_f^{(\tau)}(\mathbf{y}_t))^T = (P(\mathbf{y}_{t-\tau}))^T T^t_{t-\tau} \quad (4)$$

where $T^t_{t-\tau}$ is the local transition matrix from frame $t-\tau$ to frame t and $P(\mathbf{y}_{t-\tau})$ is the probability vector of hidden units at time stamp $t-\tau$. For each frame t, $T^t_{t-\tau}$ is estimated from the local co-occurrences of Gaussians, $C^t_{t-\tau}$, by the factorization $C^t_{t-\tau} = AT^t_{t-\tau}A^T$. $C^t_{t-\tau}$ is constructed from local information: the Gaussian co-occurrence matrix between $P(\mathbf{x}_{t-\tau})$ and $P(\mathbf{x}_t)$. Subsequently, the estimates from the forward transition with different τ's are summed and normalized in Eq.(5) to obtain its final estimate. The summation here means that the transition between hidden units is not strictly Markovian and the status of the current frame t can originate from any of its τ-nearest neighbor frames.

$$P_f(\mathbf{y}_t) = \frac{1}{\tau_0} \sum_{\tau=1}^{\tau_0} P_f^{(\tau)}(\mathbf{y}_t) \quad (5)$$

The third estimate of the probabilities on the hidden units is from the backward transition which is decoded from Eq.(6). $P_o(\mathbf{y}_{t+\tau})$ and $T_t^{t+\tau}$ are computed similarly as above by just using the information with the respective time stamps.

$$(P_o(\mathbf{y}_{t+\tau}))^T = (P_b^{(\tau)}(\mathbf{y}_t))^T T_t^{t+\tau} \quad (6)$$

Table 1. Tri-factorization learning of hidden units with multiple contextual dependencies

1	Initialization of A,$\{B^\tau\}$
2	**While** Stopping criteria is met
(1)	$P^\tau \leftarrow \mathbf{1}_{M\times 1} * \sum_i (A * (B^\tau + (B^\tau)^T))_{i,:}$;
(2)	$Q^\tau \leftarrow C^\tau \oslash (A * B^\tau * A^T)$;
(3)	$A \leftarrow A \odot (\sum_\tau Q^\tau * A * (B^\tau)^T + (Q^\tau)^T * A * B^\tau) \oslash (\sum_\tau P^\tau)$;
(4)	$A_{ik} \leftarrow A_{ik} / \sum_{i'} A_{i'k}$, $B^\tau_{kl} \leftarrow \sum_{i'} A_{i'k} * B^\tau_{kl} * \sum_{i'} A_{il}$;
(5)	$B^\tau \leftarrow B^\tau \odot (A^T * (C^\tau \oslash (A * B^\tau * A^T)) * A)$;

Then the estimates from the backward transition with different τ's are summed and normalized similarly in Eq.(7).

$$P_b(\mathbf{y}_t) = \frac{1}{\tau_0} \sum_{\tau=1}^{\tau_0} P_b^{(\tau)}(\mathbf{y}_t) \tag{7}$$

Finally, the estimate of $P(\mathbf{y}_t)$ is the product of the probabilities from observation, forward transition and backward transition in Eq.(8). The product implies that the activated hidden unit k of frame t should be both observable at this frame and be reachable from its τ-nearest neighbors.

$$P(y_t = k) = \frac{P_f(y_t = k) P_o(y_t = k) P_b(y_t = k)}{\sum_{k'} P_f(y_t = k') P_o(y_t = k') P_b(y_t = k')} \tag{8}$$

3. VOCABULARY DISCOVERY

Every utterance is represented by the Gaussian posteriorgram, $\{P(\mathbf{x}_t),\ t = 1, \ldots, T\}$, and the hidden-unit posteriorgram, $\{P(\mathbf{y}_t),\ t = 1, \ldots, T\}$. We now compare the two kinds of representations by performing weakly supervised spoken pattern discovery. The task is to discover vocabulary patterns with little human annotations. It is accomplished in the NMF framework of Eq.(9).

$$\begin{bmatrix} G_{:,1:N_1} & 0 \\ X_{:,1:N_1} & X_{:,N_1+1:N} \end{bmatrix} \approx \begin{bmatrix} Q \\ Y \end{bmatrix} Z \tag{9}$$

G is the ground truth matrix as supervision where its entry G_{sn} shows the frequency of appearance of the ground word s in the utterance n. N_1 is the number of labeled training utterances as supervision. During training, we always use all the training data, but only have an increasing number N_1 of utterances labeled. Cognitively, this process means that an agent or infant can hear a lot of utterances, but only a part of them are interpreted by a teacher or parent.

X is the data matrix, a column of which represents an utterance with its accumulated co-occurrence probabilities of Gaussians or hidden units. To represent long contextual dependencies, a *long-patch* method is utilized to define the co-occurrences of units when constructing the data matrix. Take the Gaussian posteriorgram $P(\mathbf{x}_t)$ as an example. A patch of length T_0=10 of the posteriorgram is the sum of the probabilities of units of the frames it contains: $P(\mathbf{x}_t) \leftarrow \sum_{t'=t}^{t+T_0} P(\mathbf{x}_{t'})$. Then τ=$\{$1,2,3$\}$ are utilized to compute co-occurrences between patches.

Y is the pattern matrix, each column of which is a learned vocabulary pattern. Q reflects the associations between the patterns and the ground words. Z is the coefficient matrix whose columns are the weights of the discovered patterns in the corresponding utterances. The columns of G, Y and X are ℓ_1 normalized to fit the probabilistic definition.

4. EXPERIMENTS AND RESULTS

For simplicity, we evaluate the model by discovering digits from the Aurora2/Clean database. The data set contains 11 English digits ("one" to "nine", "zero" and "oh") in 8438 training utterances and 1001 testing utterances from adult male and female speakers. Each utterance contains a digit string of length one through seven.

4.1. Acquire hidden units and posteriorgrams

The window length for spectral analysis was 20ms and the frame shift was 10ms. For each frame, 12 MFCC coefficients were computed plus the log-energy. MFCC and its Δ, $\Delta\Delta$ features were concatenated to a 39-dimensional feature vector on which a Gaussian mixture of M=1000 components were obtained by unsupervisedly training a Gaussian Mixture Model (GMM) with EM algorithm from the training data.

The number of recurrent structures of Gaussian co-occurrences in Section 2.2 is R_1=25. The contextual dependency parameter τ=1,2,3. For each $C^{r,\tau}$, 10 hidden units are extracted. So there are 10*R_1 hidden units in total.

4.2. Spoken pattern discovery

The common dimension between Y and Z, i.e. the number of vocabulary patterns, is R_2=12. The evaluation metric is unordered word error rate by only considering the appearance or not of digits without ordering them [5]. This metric can focus the evaluation on the representation of the vocabularies.

As is shown in Figure 1, the hidden units perform much better than the Gaussians with only a few labeled utterances, but not with a sufficiently large number of labeled utterances. This could be due to the lack of fine tuning of the hidden units. As pointed out in [8], a top-down fine tuning with supervision is important to help a multi-layer model beat its single-layer

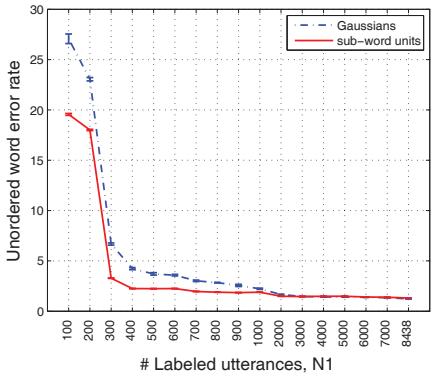

Fig. 1. The comparison of performance on vocabulary discovery between Gaussians and sub-word units.

Table 2. Error rates after fine tuning (%)

N_1	6000	7000	8438
Gaussians	1.38±0.00	1.35±0.00	1.24±0.00
hidden units	0.91±0.00	0.87±0.00	0.80±0.00

counterparts. Thus in the fine tuning stage, we use the ground truth information to classify the structures extracted in Section 2.2 into *digits* and *silence* and model silence with only 3 hidden units. Consistent improvement is observed in Table 2.

4.3. Visualization

Figure 2 shows the posteriorgram of the utterance "4625" with the extracted hidden units. Piece-wise constant traces or segments are observed from the figure. Some of the hidden units have long durations like 5 to 8 frames which probably corresponds to vowel parts, e.g. the segments of frames between 80 and 95, and the ones between 105 and 120. So it is appropriate to name the hidden units "sub-word units".

With the Viterbi alignment between an utterance and the HMM-like graph model with A, B^τ, further smooth segments are observed. One reason we don't apply Viterbi alignment here is that it is difficult to extend to two-dimensional data, e.g. images.

5. CONCLUSION

We have successfully extended the NMF-based vocabulary acquisition from co-occurrence statistics to now include a layer of sub-word units that are learned without supervision using a matrix tri-factorization. The method performs better at the acquisition of small vocabularies. The next steps are now to show that the sub-words allow to handle larger vocabularies and that it speeds up learning of new vocabularies. Furthermore, the proposed co-occurrence model and labeling approach only considers the τ-nearest neighbors of a data frame, making it a generic tool to extract patterns or topics from images.

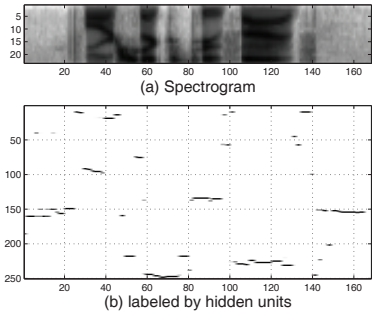

Fig. 2. Representations of utterance "4625".

Finally, space limitations have not allowed us to elaborate on the underlying graphical model, a tensor formulation of the method, nor on the relation to HMMs, which will be the topic of future publications.

6. REFERENCES

[1] D. Roy, "Grounded spoken language acquisition: Experiments in word learning," *IEEE Transactions on Multimedia*, vol. 5(2), pp. 197–209, 2003.

[2] L. ten Bosch, H. Van hamme, and L. Boves, "A computational model of language acquisition: focus on word discovery," in *INTERSPEECH*, 2008.

[3] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[4] G.Aimetti, L.ten Bosch, and R.K.Moore, "The emergence of words: Modelling early language acquisition with a dynamic systems perspective," in *INTERSPEECH*, 2009.

[5] M. Sun and H. Van hamme, "Unsupervised vocabulary discovery using non-negative matrix factorization with graph regularization," in *ICASSP*, 2011.

[6] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden markov models," *Journal of Computer and System Sciences*, p. to appear, 2011.

[7] B. Vanluyten, J. Willems, and B. De Moor, "Structured nonnegative matrix factorization with applications to hidden markov realization and clustering," *Linear Algebra and Its Applications*, vol. 429, pp. 1409–1424, 2008.

[8] G. E. Hinton, S. Osindero, and Y-W Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, July 2006.