LETTER

Unsupervised Learning of Continuous Density HMM for Variable-Length Spoken Unit Discovery

Meng SUN[†a)], *Member*, Hugo VAN HAMME[††], Yimin WANG[†], *and* Xiongwei ZHANG[†], *Nonmembers*

SUMMARY Unsupervised spoken unit discovery or zero-source speech recognition is an emerging research topic which is important for spoken document analysis of languages or dialects with little human annotation. In this paper, we extend our earlier joint training framework for unsupervised learning of *discrete* density HMM to *continuous* density HMM (CDHMM) and apply it to spoken unit discovery. In the proposed recipe, we first cluster a group of Gaussians which then act as initializations to the joint training framework of nonnegative matrix factorization and semi-continuous density HMM (SCDHMM). In SCDHMM, all the hidden states share the same group of Gaussians but with different mixture weights. A CDHMM is subsequently constructed by tying the top-N activated Gaussians to each hidden state. Baum-Welch training is finally conducted to update the parameters of the Gaussians, mixture weights and HMM transition probabilities. Experiments were conducted on word discovery from TIDIGITS and phone discovery from TIMIT. For TIDIGITS, units were modeled by 10 states which turn out to be strongly related to words; while for TIMIT, units were modeled by 3 states which are likely to be phonemes.

key words: *spoken unit discovery, unsupervised HMM learning, nonnegative matrix factorization, language acquisition*

1. Introduction

Unsupervised spoken unit discovery or zero-source speech recognition is an emerging research topic which is important for spoken document analysis of languages or dialects with little human annotation [1]. Supervised training of an automatic speech recognition (ASR) system normally requires a large collection of transcribed speech data, where continuous density hidden Markov models (CDHMM) with Gaussian mixture models (GMM) are conventional tools to represent each spoken unit. Forced alignment between acoustic features (e.g. MFCC vectors and their probabilities emitted by the states of CDHMM) and the word or phoneme transcriptions are conducted in conventional EM algorithms to estimate the observation and transition parameters in HMMs. However, in languages or dialects for which it is difficult to hire experts to make abundant annotations, one has to develop new techniques to train an ASR system with limited labeled data. This is the basic motivation for the recent work on unsupervised spoken unit discovery.

The key research objective is to discover the invariant speech structures in a weakly supervised way or in an unsupervised way in its extreme case, much like an infant learning a language [2]. Thanks to the repetition of the invariant speech structures, one can deploy the clustering methods from machine learning to discover them. One of the difficulties is that due to the lack of proper segmentation conventional vector-based clustering approaches cannot be implemented directly to deal with the variable length of the units in continuous speech. Thus, one has to consider the segmentation and clustering jointly. Segmental dynamic time warping (DTW) was applied to spectrograms and Gaussian posteriorgrams to find repeated structures in [3] and [4], respectively. Because the same unit uttered by different speakers can have very different spectral properties, a characterization of spoken units by a raw spectrogram will not generalize well over speakers. Therefore, in light of the successful implementation of GMMs in ASR, Gaussians were adopted to represent spoken units [4]. Besides using DTW, pre-segmentation was first implemented to extract short segments and spectral clustering of Gaussian components was subsequently utilized to cluster the mean vectors of the short segments into spoken units in [5]. However, a GMM itself cannot model the sequential nature of speech. Additional dynamic models should be introduced to work together with the GMM. In [6] an HMM was utilized to improve the clustering results of a GMM for spoken unit discovery, where ad hoc labels generated by the learned HMMs from the last iteration were treated as transcriptions to update the HMM parameters in the current iteration, as presented in [7].

Due to the large number of unknown parameters, the initialization of the GMM/HMM has a strong impact on its performance. DTW was utilized as initialization of HMMs to alleviate the problem of poor local optima in [8]. A Dirichlet process [9] and a hierarchical Dirichlet process [10] were integrated in the joint learning of GMM, HMM and speech units to model the Zipfian distribution of speech units. Without any prior segmentation, the algorithm was reported to be able to extract phone-like units which had strong correspondence with the English phones.

In this paper, we extend our recently proposed approach, joint training of nonnegative matrix factorization (NMF) and discrete density HMMs (DDHMM) [11], to the case with semi-continuous density (SCDHMM) and continuous density (CDHMM). Compared to DDHMM, SCDHMM and CDHMM have more powerful representation abilities. However, their unsupervised learning is more difficult due to the great number of unknown parameters involved, as the learning algorithm could be trapped into poor local optima.

Manuscript received August 9, 2015.
Manuscript revised September 27, 2015.
Manuscript publicized October 21, 2015.

[†] The authors are with the PLA University of Science and Technology, China.
[††] The author is with KU Leuven, Belgium.
a) E-mail: sunmengccjs@gmail.com

DOI: 10.1587/transinf.2015EDL8178

2. Preliminaries on HMM for Acoustic Modeling

In conventional ASR, the time-domain speech signal is mapped onto frames which are described by their spectral parameters, $\mathbf{O}_1, \ldots, \mathbf{O}_{T_n}$ where $\mathbf{O}_t$ denotes a frame vector such as MFCC+Δ+ΔΔ and T_n is the number of frames in the utterance n. An HMM for spoken unit discovery and speech recognition on sequences $\{\mathbf{O}_1, \ldots, \mathbf{O}_{T_n}\}$ (where $1 \leq n \leq N$ and N is the total number of sequences) is configured by connecting a number of left-to-right sub-HMMs with non-emitting beginning and ending states. Initially, we will assume that each unit is modeled by one sub-HMM. However, in view of pronunciation variation modeling in speech, it is conceivable that multiple sub-HMMs are used to model a single unit (e.g. word or phoneme). An HMM is characterized by the following elements [12].

- The hidden states of the r-th sub-HMM $\{S_{r,1}, \ldots, S_{r,L_r}\}$ where L_r is the number of states in the sub-HMM. So $K = \sum_{r=1}^{R} L_r$ is the total number of hidden states and R is the number of sub-HMMs. L_r is selected as a constant L in this paper.
- The sequence of hidden states $\{Q_1, Q_2, \ldots, Q_{T_n}\}$ which are aligned to $\{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_{T_n}\}$ of the n-th training sequence with length T_n.
- The transition matrix $\mathbf{T}_{K\times K}$ whose element $T_{k,k'}$ is the conditional probability of transition from S_k to $S_{k'}$: $\Pr(Q_{t+1} = S_{k'} | Q_t = S_k)$, $\forall t$. An HMM consisting of R sub-HMMs has a special structure with only non-zero probabilities on the diagonal/sub-diagonal locations $\{T_{(r-1)L+l',(r-1)L+l'}, T_{(r-1)L+l,(r-1)L+l+1}\}$ and cross-unit locations $\{T_{rL,(r'-1)L+1}\}$, where $1 \leq r, r' \leq R$ and $1 \leq l \leq L-1$ and $1 \leq l' \leq L$.
- An emission model $a_k(\mathbf{O}_t)$ to measure the similarity between the hidden state S_k and the observation $\mathbf{O}_t$.

To summarize the above description, the HMM is described by a group of parameters $\mathbf{\Lambda} = \{a_k(.), \mathbf{T}\}$. According to the configuration of the emission probability $a_k(\mathbf{O}_t)$, we consider two types of HMMs as follows.

2.1 Semi-Continuous Density HMM

In semi-continuous density HMM (SCDHMM), a Gaussian mixture model (GMM) is utilized to compute the likelihood of observation $\mathbf{O}_t$ given state S_k,

$$a_k(\mathbf{O}_t) = \sum_m A_{m,k} \mathcal{G}(\mathbf{O}_t; \boldsymbol{\mu}_m, \mathbf{\Sigma}_m) \tag{1}$$

where $\mathbf{A}$ is the Gaussian weight matrix of the states, $\boldsymbol{\mu}_m$ is the mean of the m-th Gaussian and $\mathbf{\Sigma}_m$ is the covariance matrix of the m-th Gaussian which is diagonal here. The Gaussians are shared across all the hidden states in SCDHMM.

2.2 Continuous Density HMM

In a continuous density HMM (CDHMM), every hidden state S_k has its own Gaussian mixture model (GMM) to compute the likelihood of observation $\mathbf{O}_t$,

$$a_k(\mathbf{O}_t) = \sum_m A_m^{(k)} \mathcal{G}(\mathbf{O}_t; \boldsymbol{\mu}_{m.k}, \mathbf{\Sigma}_{m,k}) \tag{2}$$

where $\mathbf{A}_m^{(k)}$ is the Gaussian weight vector of the state S_k, $\boldsymbol{\mu}_{m,k}$ and $\mathbf{\Sigma}_{m,k}$ are the mean and covariance matrix of the m-th Gaussian of the k-th state, respectively. The Gaussians are tied to a specific hidden state.

3. Nonnegative Matrix Factorization for (S)CDHMM Learning

Before introducing the nonnegative matrix factorization (NMF) approach to SCDHMM learning, we first represent each utterance as its Gaussian posteriorgram,

$$\mathbf{X}_1, \ldots, \mathbf{X}_{T_n} \tag{3}$$

where $\mathbf{X}_t$ contains the posterior probabilities of frame $\mathbf{O}_t$ on the Gaussians. By computing and accumulating the Gaussian posterior probabilities,

$$\begin{aligned} &\sum_{t=1}^{T_n-1} \Pr(X_t^{(n)} = \mathcal{G}_m, X_{t+1}^{(n)} = \mathcal{G}_{m'}) \\ &= \sum_{t=1}^{T_n-1} \Pr(X_t^{(n)} = \mathcal{G}_m) \Pr(X_{t+1}^{(n)} = \mathcal{G}_{m'}), \end{aligned} \tag{4}$$

we obtain the co-occurrence matrix of Gaussians of the n-th utterance. By vectorizing, i.e. stacking the columns of the co-occurrence matrix, one utterance is represented by a $M \times M$-dimensional vector where M is the number of Gaussians. For all the training utterances, we obtain a matrix $\mathbf{V}_{M\times M,N}$.

3.1 NMF on Gaussian Co-Occurrences to Extract Repeating Spoken Units

NMF has good performance on finding repeating patterns. We thus utilize it to find repeating spoken units. The process is described by the following optimization problem where the columns of $\mathbf{W}$ represent the discovered spoken units in their Gaussian co-occurrence form and $\mathbf{H}$ stores the presence probability of the units in the utterances.

$$\{\mathbf{W}, \mathbf{H}\} = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} \operatorname{KLD}(\mathbf{V} || \mathbf{WH}), \tag{5}$$

where KLD refers to the Kullback-Leibler divergence which fits well to the count data. This model has been successfully discovered word-like units as reported in [11].

However, for short unit discovery, e.g. phones, one should better first cut the whole utterance into short segments and treat each segment as an artificial utterance to conduct the above process. Once the segments are short, the activation matrix $\mathbf{H}$ should be constrained to be sparse since only a few of the spoken units appear in each *short* segment, which is helpful to alleviate the confusion between spoken units in a long utterance. Thanks to the zero-locking property of NMF, the segments do not have to be accurately located in the non-silence region, which is an advantage compared to HMM where strict beginning/ending frames are required.

3.2 NMTF Links NMF Outputs to Initialize the SCDHMM

Subsequently, each column of $\mathbf{W}$ is matricized (inverse of vectorization) to a square co-occurrence matrix $\mathbf{C}$. If the data is generated by a HMM, this matrix is structured $\mathbf{C} \approx \mathbf{ATA}^T$. Hence, the mixture weight matrix $\mathbf{A}$ and transition matrix $\mathbf{T}$ of each sub-HMM can be estimated from a nonnegative matrix tri-factorization (NMTF) [11].

$$\{\mathbf{A}, \mathbf{T}\} = \underset{\mathbf{A},\mathbf{T}}{\operatorname{argmin}} \operatorname{KLD}(\mathbf{C} \| \mathbf{ATA}^T), \tag{6}$$

In a DDHMM, the discrete emission probabilities are estimated as $\mathbf{A}$ in this process. Here, only the Gaussian weight matrix $\mathbf{A}$ and the transition matrix $\mathbf{T}$ of the sub-HMM is obtained, while the Gaussian means and co-variance matrices are kept fixed.

Subsequently, the sub-HMMs are initialized by the Gaussians from the input (e.g. from unsupervised k-means clustering or GMM training), the Gaussian mixture weights and the transition probabilities of the hidden states both estimated from the NMTF. Baum-Welch training is then conducted to update all the parameters involved.

3.3 Updated the Gaussian Posteriors in NMF by the Outputs of SCDHMM

In Baum-Welch training, Gaussian means and co-variances are updated. So in the next cycle we have to update the Gaussian posteriorgram in (3) by computing the posterior probabilities of the frames on the updated Gaussians. By sequentially executing the steps in Sect. 3.1, Sect. 3.2 and Sect. 3.3, one cycle of SCDHMM learning is completed. Since NMF and NMTF can be interpreted as nonnegative Tucker decomposition (NTD) and to be consistent with the terminology in [11], we name this algorithm as the alternative training of NTD and SCDHMM (NTD.Alt.SCD.BW in short in Fig. 1 where "SCD.BW" refers to the BW training of the SCDHMM).

3.4 CDHMM Refinement of SCDHMMs

In state-of-the-art ASR, CDHMM with tied Gaussians for each state has shown to have strong ability to represent the probabilistic distribution of a steady part of the speech signals in a speaker independent way. However, the training of CDHMM directly from input observations is severely affected by the initialization. Our experiments showed that random initialization of CDHMM shrank to one sub-HMM while leaving the remaining sub-HMMs unused. We thus utilize the above SCDHMM to initialize the CDHMM.

For each hidden state in the yielded SCDHMM, take the top-N Gaussians with the highest weights and these Gaussians are subsequently assigned to this state. In this process, the same Gaussian tied to different states will be regarded as different Gaussians in the following CDHMM training. Baum-Welch learning is again adopted to update the involved parameters.

4. Experiments and Results

4.1 Word Discovery on TIDIGITS

The proposed recipe was evaluated on word discovery from TIDIGITS to see if it outperforms the discrete density version of [11]. We utilized an analysis window with 25ms long and 10ms shift. Mean purity of the discovered units is again taken as the evaluation metric. The purity is computed from the confusion matrix of the units and the ground truth words. An utterance is first decoded into a unit sequence using the Viterbi algorithm. The confusion matrix is then constructed by counting the overlapping frames between the decoded unit sequence and the ground truth word sequence from a supervised speech recognizer. The confusion matrices from unsupervised training of SCDHMM with two different random initializations (attempts) are illustrated in Fig. 1. Since a continuous density is used to describe the emission density of a hidden state, a relatively small number $M = 500$ of Gaussians are used in the above experiments instead of 1000 Gaussians in our previous work [11]. R=25 sub-HMMs each of which has $L = 10$ hidden states are created and initialized randomly. The BW training uses 25 EM passes, while 5 NMF-NMTF passes and 5 EM passes are applied in the proposed recipe. Compared to the results reported in [11], sub-HMMs discovered by the BW training of CDHMM have higher purity than the ones from the BW training of discrete density HMM. The proposed recipe yields the best results regarding the pure units corresponding to digits. Our experiments also showed that the improvement from SCDHMM to CDHMM (5 Gaussians per state, thus 1250 ones in total) was marginal, so the performance of CDHMM was not included here.

4.2 Phoneme Discovery on TIMIT

To extract phones (not words) from the TIMIT dataset we utilized a finer analysis window with 20ms long and 5ms shift. For the NMF learning stage, every utterance was cut into small segments by using a 100-frames window and with 50-frames shift. A segment was represented in its Gaussian posteriorgram and its co-occurrence vector forms a column of the NMF data matrix. The number of sub-HMMs was set as R=125 to give sufficient complexity to model male and female versions of the 61 phonemes. To treat the large number of segments (around 60k) on a personal computer, online NMF learning was adopted instead of the batch learning [13], where the epoch size was 4000 (segments). The number of NMF training pass was 2, i.e. the training data was covered twice. Subsequently, NMTF was computed to

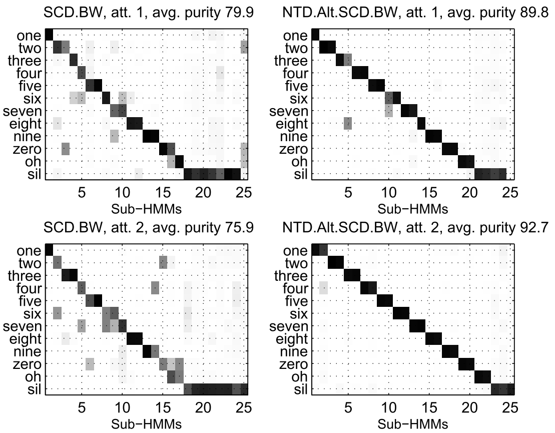

Fig. 1 Confusion matrices on TIDIGITS from the BW training and the proposed recipe. For DDHMM results, the reader is referred to Fig. 5 of [11] for comparison, where the highest purity from DDHMM is 88.6.

Table 1 Results on TIMIT dataset with all the speakers.

	Method	Purity	Purity (no sil.)	NMI	NMI (no sil.)
Phn. Cls	BW	44.1	40.1	35.4	35.5
	NTD.Alt.SCD.BW	44.3	41.0	37.2	35.0
	NTD.Alt.CD.BW	44.0	42.2	36.0	36.4
Phones	BW	33.2	42.4	38.2	34.4
	NTD.Alt.SCD.BW	34.2	43.1	39.0	34.5
	NTD.Alt.CD.BW	33.4	44.5	38.6	35.7

Table 2 Results on the utterances of female speakers of TIMIT only

	Method	Purity	Purity (no sil.)	NMI	NMI (no sil.)
Phn. Cls	BW	46.7	42.9	38.5	39.9
	NTD.Alt.SCD.BW	47.5	43.7	39.8	38.8
	NTD.Alt.CD.BW	47.1	44.8	39.0	40.4
Phones	BW	36.3	44.1	40.8	38.1
	NTD.Alt.SCD.BW	37.6	45.4	42.0	37.8
	NTD.Alt.CD.BW	37.7	46.0	41.3	40.0

yield emission matrices and transition matrices of 125 sub-HMMs. Every sub-HMM has 3 states.

The purity and normalized mutual information (NMI) [6] of the confusion matrix between the discovered units and the phoneme (or phoneme classes) were reported in Table 1 and 2 where *Purity (no sil.)* and *NMI (no sil.)* refer to the corresponding metrics after removing the units strongly related to silence. *Phoneme classes* refers to the re-calculated results by collapsing the 61 phones into 39 phoneme classes [6]. From the tables, we see an improvement of the model's performance from the BW/NTD.Alt.SCD.BW to NTD.Alt.CD.BW by approximating the state-of-the-art results obtained by using a pre-segmentation [6].

5. Conclusion

In this paper, we extended our recently proposed unsupervised learning algorithm of discrete density HMM to continuous HMM. The algorithm was able to discover units of different length by adjusting the number of hidden unit and the discovered units had strong correspondence to the ground truth units. The algorithm performed pretty well on the small vocabulary dataset TIDIGITS by reaching high purity of the spoken units. On the performance on the phoneme dataset TIMIT, the proposed model approximated the state-of-the-art results which used a pre-segmentation and spectral clustering of Gaussians. Additional constraints from a few human labels are likely to improve the purity.

Acknowledgments

M. Sun, Y. Wang and X. Zhang were funded by NSFC (61471394, 61402519, 11404407) and the Natural Science Foundation of Jiangsu Province (BK20140071, BK20140072). H. Van hamme was funded by the KULeuven research grant GOA/14/005 (CAMETRON).

References

[1] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, C.-Y. Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.8111–8115, 2013.

[2] L. Boves, L.T. Bosch, and R. Moore, "ACORNS - towards computational modeling of communication and recognition skills," 6th IEEE Int. Conf. Cognitive Informatics, pp.349–356, 2007.

[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] Y. Zhang and J.R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," ICASSP, pp.4366–4369, 2010.

[5] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Acoustic segment modeling with spectral clustering methods," IEEE/ACM Trans. Audio, Speech and Lang.Process., vol.23, no.2, pp.264–277, Feb. 2015.

[6] H. Wang, T. Lee, C.C. Leung, B. Ma, and H. Li, "A graph-based gaussian component clustering approach to unsupervised acoustic modeling," INTERSPEECH, pp.875–889, 2014.

[7] M.H. Siu, H. Gish, S. Lowe, and A. Chan, "Unsupervised audio patterns discovery using hmm-based self-organized units," INTERSPEECH, pp.2333–2336, 2011.

[8] O. Walter, T. Korthals, R. Haeb-Umbach, and B. Raj, "A hierarchical system for word discovery exploiting dtw-based initialization," Automatic Speech Recognition and Understanding Workshop (ASRU), pp.386–391, 2013.

[9] C. Lee and J.R. Glass, "A nonparametric bayesian approach to acoustic model discovery," Proc. 50th Annual Meeting of the Association for Computational Linguistics, pp.40–49, 2012.

[10] A.H.H.N. Torbati, J. Picone, and M. Sobel, "Speech acoustic unit segmentation using hierarchical dirichlet processes," INTERSPEECH, pp.637–641, 2013.

[11] M. Sun and H. Van hamme, "Joint training of non-negative tucker decomposition and discrete density hidden markov models," Computer Speech and Language, vol.27, no.4, pp.969–988, 2013.

[12] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–286, 1989.

[13] J. Driesen and H. Van hamme, "Modelling vocabulary acquisition, adaptation and generalization in infants using adaptive Bayesian PLSA," Neurocomputing, vol.74, no.11, pp.1874–1882, 2011.