Unsupervised Word Discovery from Speech using Bayesian Hierarchical Models

Oliver Walter and Reinhold Häb-Umbach

Paderborn University

Abstract. In this paper we demonstrate an algorithm to learn words from speech using non-parametric Bayesian hierarchical models in an unsupervised setting. We exploit the assumption of a hierarchical structure of speech, namely the formation of spoken words as a sequence of phonemes. We employ the Nested Hierarchical Pitman-Yor Language Model, which allows an a priori unknown and possibly unlimited number of words. We assume the n-gram probabilities of words, the m-gram probabilities of phoneme sequences in words and the phoneme sequences of the words themselves as latent variables to be learned. We evaluate the algorithm on a cross language task using an existing speech recognizer trained on English speech to decode speech in the Xitsonga language supplied for the 2015 ZeroSpeech challenge. We apply the learning algorithm on the resulting phoneme graphs and achieve the highest token precision and F score compared to present systems.

1 Introduction

Automatic speech recognition (ASR) systems mostly rely on supervised learning, with an acoustic model and a language model, trained from transcribed speech and text data. Both, the inventory of words and phonemes are known, and a lexicon with word pronunciations in terms of phoneme sequences is given.

Here we consider a setting, where neither the pronunciation lexicon nor the vocabulary are known in advance, since the acoustic training data come without labels. In general, the phoneme inventory is not know either, however here we use the acoustic models of another language to decode the acoustic data, demonstrating the effectiveness of cross language transfer.

As depicted in Figure 1 an audio recording is typically represented as a time series of feature vectors. A symbolic representation can be learned by discovering repeated sequences of vectors and assigning the same labels to similar sequences, corresponding to phone-like units [1, 19, 17, 13]. On this label sequence again similar sequences are discovered and given labels from another label set, thus arriving at a segmentation into words [18, 8, 4, 5, 7]. An n-gram language model is learned simultaneously and used to calculate the probabilities of words, depending on their $n-1$ preceding words.

Figure 2 depicts the generative model: a language Model $\mathbf{G}$ and the lexicon are generated from a prior process, the Nested Hierarchical Pitman-Yor process. Within the nested process, a word language model is drawn from a Hierarchical

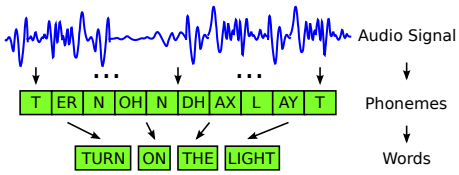

Fig. 1. Hierarchical model of Speech, phonemes and words.

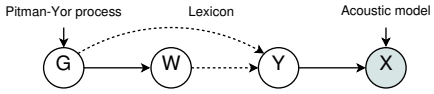

Fig. 2. Language Model **G**, words **W**, phonemes **Y** and feature vectors **X**.

Pitman-Yor process, whose base distribution is a distribution over all possible phoneme sequences, calculated by a phoneme language model. Phoneme Sequences not corresponding to a word in the lexicon, and therefore new words, are obtained as draws from the same phoneme language model whose prior is again a Hierarchical Pitman-Yor process with a uniform base distribution over phonemes. The words **W** are generated (drawn) using the language model and mapped to phoneme sequences **Y** using the lexicon. Acoustic feature vectors **X** are finally generated employing an acoustic model.

Here we will focus on the discovery of words from phoneme sequences, where the phoneme sequences have been generated by a phoneme recognizer, trained with another language, assuming a phoneme set and acoustic models for each of the phonemes to be known.

2 Unsupervised Word Segmentation

If neither the pronunciation lexicon nor the language model are known, and we are left with the task to segment a phoneme string into the most probable word sequence, we have to learn the language model together with the words. We use the Nested Hierarchical Pitman-Yor Language Model (NHPYLM), denoted by **G**, which is a Bayesian language model and allows new, previously unseen words, to evolve and assign probabilities to them. It is based on the Pitman-Yor process prior, which produces power-law distributions that resemble the statistics found in natural languages [14, 15, 7].

An n-gram language model $G_{\mathbf{u}}$ is a categorical distribution over the N words of the vocabulary, conditioned on the $n-1$ preceding words $\mathbf{u} = w_{l-1}, \dots w_{l-n+1}$: $G_{\mathbf{u}} = \{P(w^1|\mathbf{u}), \dots, P(w^N|\mathbf{u})\}$. In a Hierarchical Pitman-Yor process, $G_{\mathbf{u}}$ is modeled as a draw

$$G_{\mathbf{u}} \sim PY(d_{|\mathbf{u}|}, \theta_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}) \tag{1}$$

from a Pitman-Yor process with base measure $G_{\pi(\mathbf{u})}$, strength parameter $d_{|\mathbf{u}|}$ and discount parameter $\Theta_{|\mathbf{u}|}$ [15]. The base measure corresponds to the expected probability distribution of the draws and is set to the language model $G_{\pi(\mathbf{u})}$ of the parent $(n-1)$-gram. This process is repeated until the parent language model is a zerogram, which in the supervised case means that all words have the same probability, given by one over the number of words. Since in the unsupervised setting the vocabulary size is not known in advance, the zerogram cannot be

specified. It is therefore replaced by the likelihood for the word being a phoneme sequence, calculated by a Hierarchical Pitman-Yor Language Model (HPYLM) of phonemes $\mathbf{H}'$ where a hierarchy of phoneme language models is built up to some order m, similar to (1). The phoneme zerogram is finally set to a uniform distribution over the phoneme set. The resulting structure is the NHPYLM, which consists of a HPYLM for words and a HPYLM for phonemes.

Since we now have to learn the NHPYLM along with the words and the phoneme sequence, the maximization problem becomes:

$$\begin{aligned}(\hat{\mathbf{W}}, \hat{\mathbf{G}}, \hat{\mathbf{Y}}) &= \arg\max_{\mathbf{W},\mathbf{G},\mathbf{Y}} P(\mathbf{W}, \mathbf{G}, \mathbf{Y}|\mathbf{X}) \\ &= \arg\max_{\mathbf{W},\mathbf{G},\mathbf{Y}} P(\mathbf{W}, \mathbf{Y}|\mathbf{X}, \mathbf{G}) P(\mathbf{G}) \end{aligned} \tag{2}$$

The Nested Hierarchical Pitman-Yor process prior $P(\mathbf{G})$ over the language model is introduced. Instead of having one particular language model, we have to find that pair of language model, word sequence and phoneme sequence which maximizes the joint probability.

The maximization is carried out by Gibbs sampling, first jointly sampling a word and phoneme sequence from $P(\mathbf{W}, \mathbf{Y}|\mathbf{X}, \mathbf{G})$ [8], by keeping $\mathbf{G}$ constant in (2) and then sampling the NHPYLM from $P(\mathbf{G}|\mathbf{W})$ [7] in an alternating and iterative fashion for each utterance. To avoid the recomputation of the acoustic model scores with every iteration, we use a speech recognizer to produce a phoneme lattice, containing the most likely phoneme sequences.

Joint sampling of the word and phoneme sequence can be very costly. To reduce the computational demand, the phoneme sequence is first sampled from the speech input according to $P(\mathbf{Y}|\mathbf{X}, \mathbf{H})$ and then a word sequence from that phoneme sequence according to $P(\mathbf{W}|\mathbf{Y}, \mathbf{G})$ [5, 4]. For the sampling of the phoneme sequence, an additional phoneme HPYLM $\mathbf{H}$, which includes the word end symbol, is employed. To incorporate knowledge of the learned words, the phoneme HPYLM is sampled from $P(\mathbf{H}|\mathbf{W})$ using the sampled word sequence and their corresponding word sequence.

3 Experiments

We evaluate the segmentation algorithm on datasets provided for the 2015 ZeroSpeech challenge [16]. The datasets consist of an English dataset containing conversational speech from the Buckeye corpus [10] and a second dataset containing prompted speech in Xitsonga, a south African Bantu language, from the NCHLT Xitsonga corpus [2]. Our goal is to demonstrate the possibility of using existing acoustic models from another language to perform the word segmentation, we use acoustic models trained on prompted English speech for both datasets. The English dataset is used to demonstrate the segmentation performance when using acoustic models of the same language. The Xitsonga corpus serves as the low resource language for which we assume to only have audio data available but no transcriptions.

We use the tools provided for the 2015 ZeroSpeech challenge for the evaluation and to be able to compare our results to previous publications. We focus on the type and token scores. The type scores are a measure for the quality of the discovered lexicon and therefore the set of discovered words. The token scores are a measure for the quality of the discovered word tokens and therefore the transcription of the speech, also called parsing quality. A detailed description of the evaluation framework and evaluation measures can be found in [16].

3.1 Setup

For the acoustic model we use a p-norm DNN-HMM triphone speech recognizer [20] trained on English speech from the WSJ0+1 corpus [9]. We build the recognizer using the nnet2 p-norm recipe for WSJ provided with the Kaldi [11] speech recognition toolkit. The recipe was modified to enable phoneme recognition without a word lexicon by building a simple lexicon, mapping each triphone to its middle phoneme.

The recognizer uses LDA transformed 13 dimensional MFCC feature vectors extracted with a frame rate of 10 ms and a context of ± 3 frames at a target dimensionality of 40. FMLLR speaker adaptation of the LDA transformation is performed by a two pass decoding scheme where we assume the speaker ID to be known.

The recognizer is used to create phoneme lattices for both datasets which are processed by the segmentation algorithm. We varied the word- and character language model order in the segmentation algorithm from 1 to 2 (WLM) and 1 to 8 (CLM) to evaluate the performance with different model complexities. Gibbs sampling is performed until iteration 150 to generate the segmentation of a sentence and to update the language model. From iteration 151 Viterbi decoding is performed to generate a segmentation. From iteration 176 the fallback probability to the character model is set to zero to disable the discovery of new words and clean up the language model by removing infrequently, especially uniquely, discovered words. The thresholds were chosen so that in each step the algorithm converged.

3.2 Results

Evaluating the performance of the segmentation algorithm on the Xitsonga dataset delivers insight into its usefulness for low resource language processing. We treat the Xitsonga language as a low resource language by assuming that only audio data is available but no transcriptions. We also assume that no acoustic model is available and instead use the English acoustic model to create phoneme graphs for the segmentation algorithm. This concept is also called cross language transfer, where knowledge from one language is transfered to another.

Figure 3 shows the type F scores for different language model orders and decoding settings. It can be seen that the performance increases with increasing character language model order. The overall scores are fairly low though. This is

mainly due to the mismatch in acoustic models and the resulting errors and noisiness of the phoneme latices. Viterbi decoding delivers a little lower performance although for higher character language model orders it matches the performance with Gibbs sampling. This might partly be due to the noisy characteristics of the input phoneme Lattices. Viterbi decoding is supposed to find the result with the highest probability. Due to the noise this might not be the optimal result. While Gibbs sampling delivers samples from the distribution of segmentations and language models seems to result in better performance. Deactivating the character language model deteriorates the results. Most likely the input data is too noisy resulting in many infrequent words which are being removed in this case. Increasing the word language model order from one to two also does not change the results significantly. The scores are a little higher for the lower order character language models but almost the same for the higher order language models. It seems that word context improves the performance for lower order character language models and noisy input but not for more complex models, contrary to previous results on less noisy data [5].

Figure 4 shows the token F scores. The behavior is similar to the type F score. The result deteriorates with Viterbi decoding and deactivating the character language model. Increasing the word language model order from one to two results in marginally better results. The biggest issue in this low resource setup seems to be the noisy input data making it difficult to learn appropriate models at higher word language model orders.

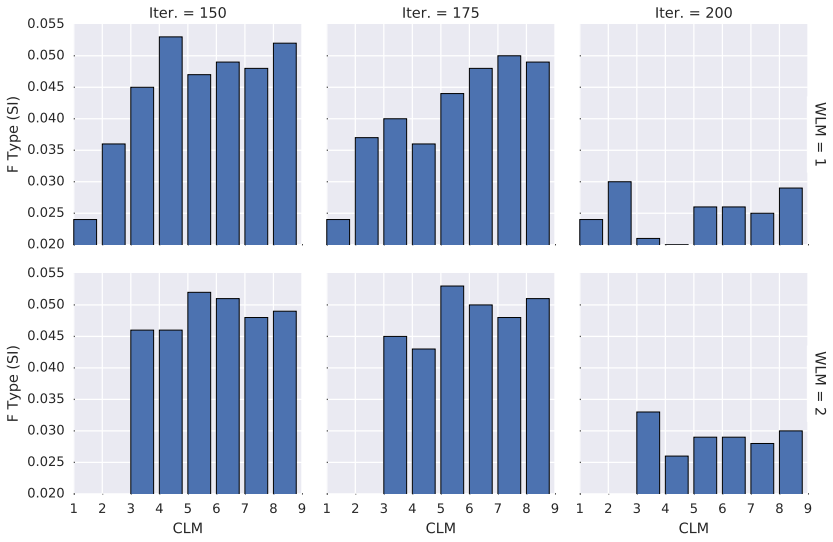

Fig. 3. Type F-score with varying word- and character language model order for Xitsonga dataset. Iter.: 150 (Gibbs), 175 (Viterbi), 200 (No character model fallback)

Fig. 4. Token F-score with varying word- and character language model order for Xitsonga dataset. Iter.: 150 (Gibbs), 175 (Viterbi), 200 (No character model fallback)

3.3 Comparison with previous results

In the 2015 ZeroSpeech challenge two types of systems participated. The two systems can be classified into segmentation systems that segment, cluster and label the complete utterance. Our system also falls into this category. On the other hand Spoken Term Discovery (STD) based systems discover similar segments and only clusters and labels those, leaving segments not discovered as similar to others unlabeled. In Table 1 we compare our results to the two types of systems. For the challenge only two systems were submitted [3] and we compare to the best setups of each.

Osc. Seg. is based on a simple segmentation algorithm finding minima in a particular oscillation frequency of the speech similar to the theta-rhythm brain oscillations and segment according to those. Fixed length representations of the Discovered segments are then clustered, labeled and n-grams of those clusters, sorted in ascending order from longest to shortest, labeled as words [12].

STD is a system based on finding similar segments, building a graph with edges connecting those similar segments with weights proportional to their similarity and clustering them using graph clustering algorithms [6].

For our system we compare the best setups with word language model order one and highest type F score (NHPYLM 1) and word language model oder two and highest token F score (NHPYLM 2) to the other systems. The system performs best in both settings on the English dataset, since we are using English acoustic models. On the Xitsonga dataset our system performs best on the token precision and F score and second best in all three token performance measures.

It also performs better then the Osc. Seg. system. For the type performance our system performs second best in all measures after the STD system. Since our system is a segmentation system it performs better on the token measures while the STD system is able to discover a better lexicon but not label all segments, resulting in higher type measures on the Xitsonga dataset.

Since we are using English acoustic models, the comparison on the English is to be understood as a baseline in case of known and partly matching models.

Table 1. Precision (P), Recall (R), F-score (F) for Type and Token on English and Xitsonga dataset with different algorithms. Red: best score, blue: second best score.

	English						Xitsonga					
	Type			Token			Type			Token		
System	P	R	F	P	R	F	P	R	F	P	R	F
Osc. Seg.	14.1	12.9	13.5	22.6	6.1	9.6	2.2	6.2	3.3	2.3	3.4	2.7
STD	3.1	9.2	4.6	2.4	3.5	2.8	4.9	18.8	7.8	2.2	12.6	0.8
NHPYLM 1	18.1	38.7	24.6	28.8	19.0	22.9	3.9	8.2	5.3	4	2.7	3.2
NHPYLM 2	17.8	36.7	24.0	24.5	25.5	25.0	3.7	8.5	5.1	4.1	3.4	3.7

4 Conclusion

Our system demonstrated a higher performance over a comparable segmentation system while still suffering from noisy input data. Although we achieved better performance than the STD system on the tokes, type quality is still behind STD systems. It is still an open question how to deal with noisy input data. Future research will investigate the integration of the acoustic model into the learning process and how to extend the system to deal with errors in the phoneme lattices and pronunciation variants.

References

1. Chaudhuri, S., Harvilla, M., Raj, B.: Unsupervised learning of acoustic unit descriptors for audio content representation and classification. In: Proc. of Interspeech (2011)
2. De Vries, N.J., Davel, M.H., Badenhorst, J., Basson, W.D., De Wet, F., Barnard, E., De Waal, A.: A smartphone-based asr data collection tool for under-resourced languages. Speech communication 56, 119–131 (2014)
3. Dupoux, E.: The Zero Resource Speech 2015 Challenge Results (2015 (accessed July 14, 2016)), http://www.lscp.net/persons/dupoux/bootphon/zerospeech2014/website/page_5.html
4. Heymann, J., Walter, O., Haeb-Umbach, R., Raj, B.: Unsupervised Word Segmentation from Noisy Input. In: Automatic Speech Recognition and Understanding Workshop (ASRU) (Dec 2013)

5. Heymann, J., Walter, O., Haeb-Umbach, R., Raj, B.: Iterative bayesian word segmentation for unspuervised vocabulary discovery from phoneme lattices. In: 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014) (may 2014)
6. Lyzinski, V., Sell, G., Jansen, A.: An evaluation of graph clustering methods for unsupervised term discovery. In: Proceedings of Interspeech (2015)
7. Mochihashi, D., Yamada, T., Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (2009)
8. Neubig, G., Mimura, M., Kawaharak, T.: Bayesian learning of a language model from continuous speech. IEICE TRANSACTIONS on Information and Systems 95(2) (2012)
9. Paul, D., Baker, J.: The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the workshop on Speech and Natural Language (1992)
10. Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E.: Buckeye corpus of conversational speech (2nd release). Columbus, OH: Department of Psychology, Ohio State University (2007)
11. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. No. EPFL-CONF-192584, IEEE Signal Processing Society (2011)
12. Räsänen, O., Doyle, G., Frank, M.C.: Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In: Proc. Interspeech (2015)
13. Siu, M.h., Gish, H., Chan, A., Belfield, W., Lowe, S.: Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery. Computer Speech & Language 28(1), 210–223 (2014)
14. Teh, Y.W.: A Bayesian interpretation of interpolated Kneser-Ney (2006)
15. Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2006)
16. Versteegh, M., Thiolliere, R., Schatz, T., Cao, X.N., Anguera, X., Jansen, A., Dupoux, E.: The zero resource speech challenge 2015. In: Proceedings of Interspeech (2015)
17. Walter, O., Despotovic, V., Haeb-Umbach, R., Gemmeke, J., Ons, B., Van hamme, H.: An evaluation of unsupervised acoustic model training for a dysarthric speech interface. In: INTERSPEECH 2014 (2014)
18. Walter, O., Haeb-Umbach, R., Chaudhuri, S., Raj, B.: Unsupervised Word Discovery from Phonetic Input Using Nested Pitman-Yor Language Modeling. ICRA Workshop on Autonomous Learning (2013)
19. Walter, O., Korthals, T., Haeb-Umbach, R., Raj, B.: A Hierarchical System For Word Discovery Exploiting DTW-Based Initialization. In: Automatic Speech Recognition and Understanding Workshop (ASRU) (Dec 2013)
20. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 215–219 (May 2014)