ITERATIVE TRAINING OF A DPGMM-HMM ACOUSTIC UNIT RECOGNIZER IN A ZERO RESOURCE SCENARIO

Michael Heck, Sakriani Sakti, Satoshi Nakamura

Augmented Human Communication Laboratory,
Graduate School of Information Science,
Nara Institute of Science and Technology,
Nara, Japan
{michael-h,ssakti,s-nakamura}@is.naist.jp

ABSTRACT

In this paper we propose a framework for building a full-fledged acoustic unit recognizer in a zero resource setting, i.e., without any provided labels. For that, we combine an iterative Dirichlet process Gaussian mixture model (DPGMM) clustering framework with a standard pipeline for supervised GMM-HMM acoustic model (AM) and n-gram language model (LM) training, enhanced by a scheme for iterative model re-training. We use the DPGMM to cluster feature vectors into a dynamically sized set of acoustic units. The frame based class labels serve as transcriptions of the audio data and are used as input to the AM and LM training pipeline. We show that iterative unsupervised model re-training of this DPGMM-HMM acoustic unit recognizer improves performance according to an ABX sound class discriminability task based evaluation. Our results show that the learned models generalize well and that sound class discriminability benefits from contextual information introduced by the language model. Our systems are competitive with supervisedly trained phone recognizers, and can beat the baseline set by DPGMM clustering.

Index Terms— acoustic unit discovery, Dirichlet process, unsupervised learning, unsupervised speech recognition, zero resource

1. INTRODUCTION

We speak of a *zero resource scenario* in the speech processing domain, when labeled training data and knowledge about the target language are not available. Current technology can not yet imitate capacities that are natural to humans to robustly learn acoustic and language models in an unsupervised way. Recently, evaluations such as the zero resource speech challenge [1] specialize in tackling this demanding task by asking the following question: Can we learn a whole language from scratch by deploying adaptive machine learning algorithms?

The absence of supervision makes it difficult to apply machine learning methods that are commonly used to build state-of-the-art HMM based speech processing systems. There has been work on estimating popular feature transformations without prior labels at hand [2, 3]. Much work has been done on training and adapting speech recognizers with little to no supervision [4, 5, 6]. Usually, automatic transcriptions for new adaptation data are produced using speech recognizers that were initialized on very small amounts of supervised data. Transcribing and re-training is normally repeated over multiple iterations to benefit from gradually improved models. Initial transcriptions can also be generated with only segmental tokenizers at hand [7]. Where the underlying sound inventory of the data is unknown, works such as [8, 9] perform acoustic unit discovery and modeling.

Works like these raise an interesting question: Can we go one step further and build a state-of-the-art acoustic unit recognizer for an unknown language given merely the audio data without any provided labels or other linguistic knowledge? Answering this question demands solving the tasks of (1) finding an inventory of the underlying sounds of the target language (2) modeling these sounds appropriately, preferably along with contextual linguistic information that supports a recognizer in handling ambiguity. Possible applications of such recognizers range from automatic transcription for language analysis and preservation, keyword discovery and topic classification to providing a basis for developing functional services for natural human-machine interaction.

Machine learning approaches to the first task are pattern matching [10, 11] on raw audio data and unsupervised sound unit detection [12]. These techniques have been successfully applied to solve tasks such as spoken term detection [13], topic segmentation [14] or document classification [15]. Bayesian models such as the Dirichlet process Gaussian mixture model (DPGMM) are a good choice for dealing with the problem of unknown model complexity. Chen et al. [16] cluster speech features by inferring a DPGMM and demonstrate its suitability for automatic detection of sound classes in untranscribed data. Their work is the best-performing contribution to the zero resource speech challenge 2015 [1]. We demonstrated in previous work [17, 18] that it is possible to

978-1-5090-4903-5/16/$31.00 ©2016 IEEE

unsupervisedly learn various feature transformations on automatically generated labels, and that these transformations can be used to produce feature vectors that considerably improve the DPGMM clustering performance.

In this work we further expand our unsupervised learning scheme in the zero resource scenario of our previous studies. We propose to build a full-fledged acoustic unit recognizer without prior labels. For that, we combine our iterative DPGMM clustering framework with a standard pipeline for supervised GMM-HMM acoustic model (AM) and n-gram language model (LM) training, along with a scheme for iterative model re-training. Specifically, we sample a DPGMM to find a dynamically sized set of acoustic units that are optimized with respect to sound class discriminability. These acoustic units are used to initialize a context dependent speaker adaptive AM and an acoustic unit based n-gram LM. Similar to [7, 19] we follow an iterative approach attempting to gradually improve the trained models by decoding and re-training, but we let the DPGMM sampler decide the amount and structure of the used sounds.

With our proposed framework it is possible to build a DPGMM-HMM acoustic unit recognizer that is competitive with supervisedly trained phone recognizers, according to the performance on the ABX sound class discriminability task [20]. The ABX test based evaluation measures class discriminability of posteriorgrams. This allows a direct comparison of the decoding quality with the clustering quality of the DPGMM. We show that our DPGMM-HMM recognizer can beat the baseline set by our previous studies on DPGMMs. We also show that the model re-training helps improve performance even over multiple iterations. Our results indicate that the contextual information encapsulated in the LM considerably helps the sound class discriminability. Useful models can be unsupervisedly learned even on minimal amounts of data. We argue that by utilizing the DPGMM-HMM framework it is possible to build a state-of-the-art acoustic unit recognizer without any prior supervision.

2. ACOUSTIC UNIT DISCOVERY

To solve the task of acoustic unit discovery, we utilize a DPGMM sampler to cluster extracted speech features into various sound classes. The set size is determined dynamically by the Bayesian approach. Our method is based on [16], but has been modified by us in previous work to incorporate automatically estimated linear feature transformations which proved to be very helpful in constructing good features for boosting the clustering quality [17, 18]. Because many useful feature transformations need labels for estimation, we use a multi-staged clustering framework that automatically finds frame-based class labels in a first run of clustering standard speech features, estimates feature transformations to transform these features and re-clusters the transformed input in a second run. The clustering scheme is depicted in Figure 1.

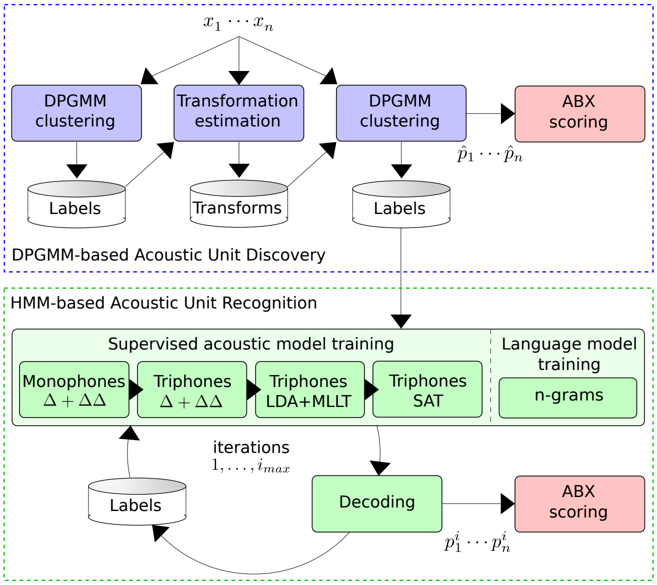
$x_1 \cdots x_n$
DPGMM clustering
Transformation estimation
DPGMM clustering
ABX scoring
$\hat{p}_1 \cdots \hat{p}_n$
Labels
Transforms
Labels
DPGMM-based Acoustic Unit Discovery
HMM-based Acoustic Unit Recognition
Supervised acoustic model training
Language model training
Monophones $\Delta + \Delta\Delta$
Triphones $\Delta + \Delta\Delta$
Triphones LDA+MLLT
Triphones SAT
n-grams
iterations $1, \ldots, i_{max}$
Labels
Decoding
ABX scoring
$p_1^i \cdots p_n^i$

Fig. 1. Scheme of the DPGMM-HMM acoustic unit recognition framework. $x_1 \cdots x_n$ denotes the input feature vectors. The model training for acoustic unit recognition is iterative, where the models of iteration $i = 1$ are trained on the initial labels from the acoustic unit discovery step, and the models of iteration $i \in \{2, \ldots, i_{max}\}$ are trained on the hypotheses of iteration $i - 1$. $\hat{p}_1 \cdots \hat{p}_n$ denotes the posteriorgrams after DPGMM sampling. $p_1^i \cdots p_n^i$ denotes the posteriorgrams after decoding in iteration i.

2.1. Dirichlet process Gaussian mixture model

DPGMMs (also known as infinite GMMs) extend finite mixture models by the aspect of automatic model selection: The model finds its complexity automatically given the data. Inference is typically sample based using a Markov chain Monte Carlo (MCMC) scheme such as Gibbs sampling. The sampler used here alternates between a non-ergodic restricted Gibbs sampler and a split/merge sampler to form an ergodic MCMC sampler. A super-cluster sampler groups similar clusters into super-cluster groups g, given a cluster similarity measure. The merge step of the split/merge sampler can be conditioned on g to only consider merge candidates within the same super-cluster that the current sample belongs to. For more in-depth informations regarding the used DPGMM sampler, please refer to [16, 21].

2.2. Unsupervised speech feature transformation

Speech feature transformations used by our framework help to project feature vectors into a more suitable sub-space for sound class discrimination by feature de-correlation and speaker adaptation. To estimate various transformations we train an AM by exploiting a standard pipeline for supervised training. During the course of the training we learn transformations via linear discriminant analysis (LDA), estimating maximum likelihood linear transforms (MLLT) and using feature-space maximum likelihood linear regression (fMLLR). LDA helps to minimize intra-class discriminability and maximize inter-class discriminability of the speech features and to enable dimensional reduction of high-dimensional stacked feature vectors. The state-dependent MLLTs maximize the likelihood of the target data. fMLLR helps to capture inter-speaker variability in speaker dependent transforms and to generate speaker independent state distributions instead.

2.3. Two-stage clustering

We produce automatic labels by sampling an initial DPGMM given standard feature vectors with their derivatives. The output consists of generic class labels and the hypothesized class membership of every speech frame. Each class is simply named with the numeric ID of the Gaussian that most likely produces the respective feature vector.

The frame-wise labels serve as basis for the subsequent transformation estimation. We collapse the labels for each utterance to emulate a more natural textual reference by compressing all subsequent tokens of the same type to a single token. We initialize an AM by context-independent monophone training. Then we subsequently train context dependent tri-phones on untransformed standard features. During this model training we automatically learn LDA transformations using the acoustic states as classes. The MLLTs are learned given the initialized HMMs, and fMLLR is based on alignments with speaker-independent features.

3. ACOUSTIC UNIT RECOGNITION

The automatic labels generated with the method described above can be used to train acoustic and language models fit for decoding. This step uses the same standard pipeline for supervised training as above, now with the objective to decode the target data with the resulting model in combination with the LM. The acoustic unit recognition scheme is depicted in Figure 1. The data sets we use in this zero resource setting are the only resources we have for training and testing, thus the entire training and decoding pipeline is designed for x-fold cross-validation.

3.1. Training

The *acoustic model training* makes use of automatic transcriptions that are produced by collapsing the class label output from the multi-stage DPGMM clustering. The transcriptions are used to initialize context and speaker independent GMM-HMM monophone models, see Figure 2. Multiple iterations of increasingly complex training followed by label writing result in speaker adaptively trained context-dependent tri-phones. The pre-processing produces LDA+MLLT+fMLLR transformed feature vectors.

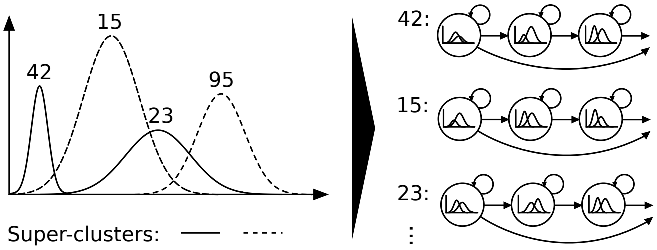

Fig. 2. *Left:* Scheme of a sampled DPGMM. Super-clusters are visualized with different line styles. Each Gaussian represents one sound class, denoted by a generic ID. *Right:* DGPMM-HMMs trained on the DPGMM label output.

A commonly used topology for acoustic modeling is left-to-right 3-state HMMs with or without skip states because of its suitability to model phone inventories crafted by linguists. It is not guaranteed, however, that automatically discovered acoustic units share the temporal properties of phones in the linguistic sense. Thus, our setup is designed to also operate with alternative HMM topologies.

The *language model training* produces an n-gram LM on the same automatic transcriptions, where the transcriptions are used as-is, i.e., no additional filtering or cleaning is performed prior to training. The LM is based on the class labels, thus captures the phonotactics of the data, given the generic acoustic units.

The DPGMM sampler used to generate the automatic labels can sample labels that group several clusters according to some cluster similarity measure. These super-cluster labels can be used as an alternative to the normal cluster labels, thus effectively reducing the amount of potential acoustic units to be trained. Making use of this reduced set of classes makes sense when the amount of clusters found during DPGMM sampling is considerably higher than the size of commonly used phone or sound inventories.

3.2. Decoding

The decoding is performed with the generic acoustic unit based AM and LM, and in turn produces acoustic unit based hypotheses, i.e., essentially resembling a “phone” recognizer. Because naturally we do not have a development data set at hand, we use default values for all parameters that might be subject to tuning, such as beam sizes and model weights.

3.3. Iterative re-training

A first system sys_1 is initialized with the help of the transcriptions that were produced by formatting the DPGMM output. By default, we iteratively re-train AM and LM simultaneously by using the hypotheses produced with system sys_{i-1} to build system sys_i in iteration $i \in \{2, \ldots, i_{max}\}$. The iterations after building system sys_1 can alternatively be restricted to one model type, i.e., either the AM or the LM is the sole subject of iterative re-training.

It is straightforward to replace the transcriptions of the previous training step with the hypotheses. Afer each iteration, we evaluate the system performance by extracting frame-wise acoustic unit posteriorgrams and measuring their ABX sound class discriminability.

4. EXPERIMENTS

4.1. Data

We use the official data set of the Interspeech zero resource speech challenge [1] for all our experiments, which contains two separate data sets of pure speech for American English (4h 59min) and Xitsonga (2h 29min), a southern African Bantu language. Each segment contains non-overlapping speech of exactly one speaker and is without noise or pauses. The English data is extracted from the Buckeye corpus and consists of conversational speech. The Xitsonga data is an excerpt of the NCHLT corpus and is comprised of read speech.

4.2. Evaluation

The evaluation metric we use to measure the cluster quality and the decoding quality is the ABX phone discriminability between phonemic minimal pairs [20], a method which is related to the ABX task used in psycho-physics [22]. The provided toolkit allows the easy evaluation of posteriorgrams which we can extract after DPGMM clustering as well as after decoding.

Each acoustic unit being found via DPGMM clustering (and used for acoustic modeling for the decoding approach) is considered a phone in the context of the evaluation. We compute GMM posteriorgrams for each speech frame after clustering as described in Section 2 and acoustic unit posteriorgrams after decoding as described in Section 3, and score them in the same manner. Both types of posteriorgrams share the same structure due to the fact that the sound units of the AM are identical with the DPGMM classes.

Let A and B be speech representations of sound categories a and b. The ABX phone discrimination accuracy is

$$c(a,b) = \frac{1}{|a| \cdot |b| \cdot (|a|-1)} \sum_{A \in a} \sum_{B \in b} \sum_{X \in a \setminus \{A\}} \left(\delta_{d(A,X)<d(B,X)} + \frac{1}{2}\delta_{d(A,X)=d(B,X)}\right) \quad (1)$$

where δ is an indicator function and $d(\cdot,\cdot)$ is the dynamic time warping (DTW) distance defined over vectors of frame-based features (in this case posteriors). As in Schatz et al. [20], we use the Kullback-Leibler divergence to compute the DTW distances. Our scores are the error rates within and across speakers. The rates are averaged over all contexts for a given pair of central phonemes and then over all pairs of central phonemes.

	English		Xitsonga	
Features	within	across	within	across
DPGMM ([16])	10.8	16.3	9.6	17.2
DPGMM ([18])	10.6	15.7	8.4	12.2
DPGMM-HMM (sup.)	12.5	16.6	6.7	11.2
DPGMM-HMM	**11.1**	**15.1**	8.2	11.6

Table 1. The baseline results provided by the DPGMM clustering (*DPGMM*), the top-line result provided by the supervisedly trained phone recognizer, and the optimal results for each condition given our proposed setup (both *DPGMM-HMM*).

4.3. Setup

For the feature vector clustering via DPGMM sampling, we use the same initialization and parameters than in [17, 18].

We use the Kaldi speech recognition toolkit [23] to extract PLP speech feature vectors for a frame length of 25 milliseconds and frame shift of 10 milliseconds. Mean variance normalization (MVN) and vocal tract length normalization (VTLN) is applied. All AMs used in our framework are likewise trained with Kaldi, following a standard scheme for speaker adaptive training (Kaldi recipe s5). All parameters that can be tuned are set to default values. To form the input for LDA estimation, we stack the standard PLP features with a context of 4, meaning that the 4 left and 4 right feature vectors are stacked on top of the current vector, which is the center vector. The LDA output dimensionality is 20 for feature transformation prior to DPGMM clustering, and set to the default value 40 for the decoding. We use a either a modified 3-state HMM topology with a skip from the first state to the next HMM, or a 1-state HMM topology.

To train the n-gram LMs for our experiments, we use the SRILM toolkit [24] with Witten-Bell discounting [25] and no pruning. We set $n = 4$ for all decoding experiments.

4.4. Clustering transformed speech features

The baseline for the DPGMM based feature vector clustering performance was set by Chen et al. [16], which won track one of the zero resource speech challenge 2015 [1]. This system has been outperformed by our clustering setup using feature transformations as described in our previous work [17, 18]. We found that PLP feature vectors are consistently leading to a higher clustering quality than MFCC feature vectors. We also found that the stacking context parameter $c = 4$ prior to LDA transformation and LDA output dimensionality $d = 20$ are good values to work with. With the application of LDA

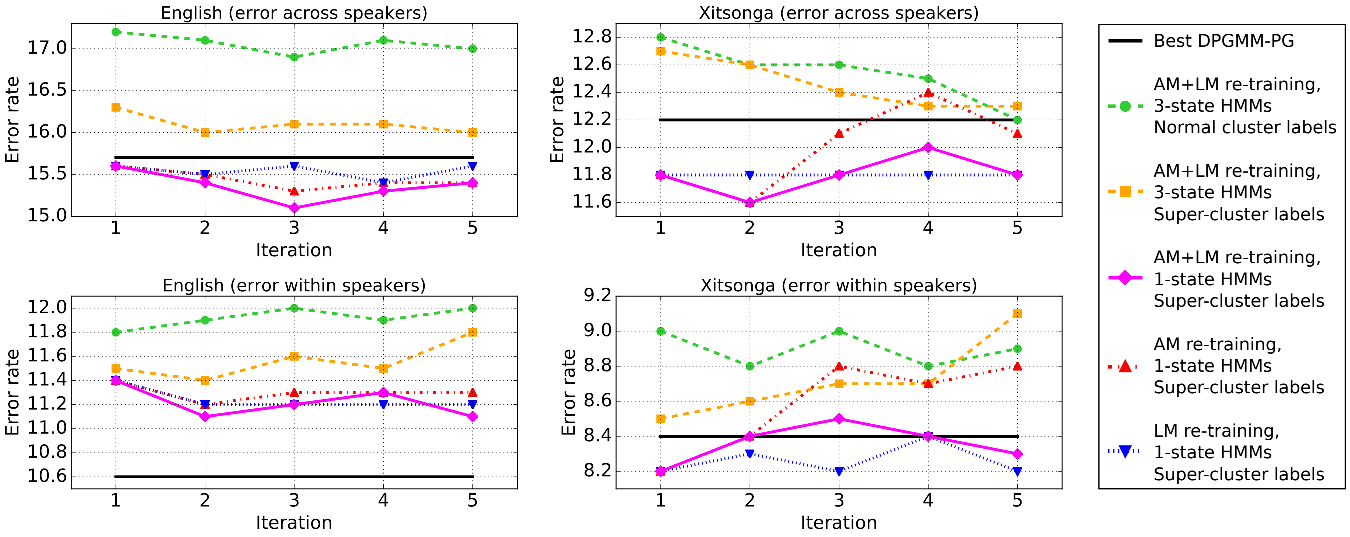

Fig. 3. Error rates within and across speakers for both languages in dependency of the model training iteration. The black horizontal line marks the baseline set by the best DPGMM clustering. *AM re-training* and *LM re-training* denote systems with exclusively re-trained AM or LM, respectively. *1-state HMMs* denotes systems that use the single state topology instead of the default. Systems have been trained either on the normal DPGMM label output or on the super-cluster labels.

we were able to produce feature vectors that considerably helped the DPGMM clustering process to find better clusters. Further, the transformations learned with fMLLR during the speaker adaptive training helped boost the discrimination capabilities across speakers. The details of these findings can be found in [17, 18]. The performance of Chen et al.'s and our setup is listed in Table 1.

4.5. Decoding with acoustic units

We trained an AM and a 4-gram LM given the classes discovered during the DPGMM clustering. Because training and test data are identical in our scenario, we use 12-fold cross-validation for training the models for decoding. The cross-validation ensures that the measured performance is an indicator of how well the learned models generalize, besides showing that they are generally capable of representing the training data. The models are used to decode the cross-validation left-out portion of the data. The decoding hypotheses were subsequently used to re-train the models for another iteration of decoding. This was done multiple times to measure a potentially positive effect of iterative unsupervised re-training on the decoder performance.

To get a top-line performance for the decoding with acoustic units, i.e., the kind of performance we can expect if we had an optimal set of acoustic units and (near) perfect transcriptions to learn models, we also trained a normal AM and phone-based 4-gram LM with the same setup given the original references and decoded the target data with 12-fold cross-validation. All results are listed in Table 1.

The performance of the DPGMM-HMM acoustic unit recognizers is depicted in Figure 3. Even though a general tendency to convergence is not observable, one can see that multiple iterations of model re-training tend to have a positive effect. The error across speakers drops for the recognizers for both languages even after 3 or more iterations, whereas the positive effect diminishes more rapidly within speakers.

The acoustic unit recognizers are competitive when compared to the supervisedly trained phone recognizers. For English, our proposed setup can even beat the supervised system according to ABX discriminability within speakers.

The posteriorgrams after decoding start off with a higher discriminability error than the posteriorgrams after DPGMM sampling, which were used to generate the labels for the decoder training in the first place. In other words, a performance loss is observable by attempting to train more complex models. However, a steady performance improvement is observable for the discriminability across speakers, while the error rate within speakers remains relatively stable. We take this as an indicator that the models do have the capacities of still learning more from the data.

4.6. Using super-cluster labels

The DPGMM sampler can sample labels that group several clusters according to some cluster similarity measure, in this case the J-Divergence [26]. We used the super-cluster labels as an alternative to the normal cluster labels to effectively reduce the amount of potential acoustic units. The number of clusters found during DPGMM sampling usually is in the hundreds, whereas the sampled super-clusters are in the range of tens, raising the hope that they resemble more phone-like units. As can be seen in Figure 3 we indeed observed a performance gain when training the models on super-cluster labels, supporting our assumption that the super-clusters might be more suitable to describe the target data.

4.7. Modeling sounds with single states

The fact that we were not able to beat the DPGMM clustering in the ABX task lets us assume that the acoustic units we found might not quite resemble phones as defined by linguistics. Thus we also conducted decoding experiments with 1-state HMMs instead of 3-state HMMs. By simplifying the models in this way we observed a considerable performance gain. Apparently, the data can be represented more accurately with chained single state HMMs. We take this as a sign that the found units are potentially too short to be modeled accurately with 3 states.

The posteriorgrams after decoding with 1-state HMMs outperform the DPGMM posteriorgrams in all but the within speaker discriminability test for English. It is also noteworthy that the model training seems to saturate after fewer iterations than before, possibly due to the reduced complexity of the AM. We now see optimal performance after the third iteration at the latest. The proposed system also clearly outperforms the supervisedly trained phone recognizer for English by showing a relative improvement of 9% to 11% in sound class discriminability performance. For Xitsonga, the performance of the automatic sound units is fairly close to the performance of the supervisedly trained recognizer.

4.8. Selective re-training

We conducted experiments to analyze the isolated effects of AM and LM re-training. In two lines of experiments we only re-trained one of the two model types each. The results that are depicted in Figure 3 allow conclusions regarding the importance of the amount of available data: For English we see an improvement when simultaneously re-training AM and LM. If both model types are re-trained exclusively, with the other model kept fix after iteration 1, the performance remains suboptimal. If the same test is done for Xitsonga however, one can see that the AM tends to deteriorate very quickly with new iterations of re-training. This is a strong indicator that the amount of training data is insufficient to reliably estimate models with multiple iterations. The LM re-training seems more robust but also suffers from multiple iterations. The combined re-training of both model types yields suboptimal performance compared to re-training the LM exclusively. The deteriorating AM is simply overpowering the benefits of an LM.

5. CONCLUSION

We proposed to build an acoustic unit recognizer without any provided labels by utilizing a Bayesian DPGMM sampler to unsupervisedly discover acoustic units in the target data for subsequent acoustic and language model training on automatically generated labels. The resulting DPGMM-HMM acoustic unit recognizer was used to solve the ABX sound class discriminability task. Multiple iterations of decoding and model re-training proved to be suitable to boost performance. We showed that the automatically discovered acoustic units seem to differ from phones in the sense that they seem generally shorter. We demonstrated that the contextual informations modeled by the LM considerably help discriminating sounds and that the sound class discriminability after DGPMM clustering can be outperformed by introducing such contextual knowledge. With our proposed framework it is possible to build a DPGMM-HMM acoustic unit recognizer that is competitive with supervisedly trained phone recognizers. Useful models can be unsupervisedly learned even on minimal amounts of data. A recognizer build in this way without any prior supervision can serve as basis for further and more sophisticated system development. In future work we plan to utilize such initialized systems to also infer lexical knowledge from the data to boost recognition performance and to enable automatic generation of lexica for new languages.

6. ACKNOWLEDGEMENTS

Part of this research was supported by JSPS KAKENHI Grant Number 24240032 and 26870371.

7. REFERENCES

[1] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux, “The zero resource speech challenge 2015,” in *Proceedings of Interspeech*, 2015.

[2] Chris Ding and Tao Li, “Adaptive dimension reduction using discriminant analysis and k-means clustering,” in *Proceedings of the International Conference on Machine learning*. ACM, 2007, pp. 521–528.

[3] Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu, “Discriminant analysis for unsupervised feature selection,” in *SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2014, pp. 938–946.

[4] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, “Unsupervised acoustic model training,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002, pp. 877–880.

[5] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, “Lightly supervised and unsupervised acoustic model training,” *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[6] Frank Wessel and Hermann Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[7] Man-hung Siu, Herbert Gish, Arthur Chan, William Belfield, and Steve Lowe, "Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. 28, no. 1, pp. 210–223, 2014.

[8] Aren Jansen and Kenneth Church, "Towards unsupervised training of speaker independent acoustic models," in *Proceedings of Interspeech*, 2011, pp. 1693–1696.

[9] Chia-ying Lee and James Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 40–49.

[10] Alex Park and James Glass, "Towards unsupervised pattern discovery in speech," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2005, pp. 53–58.

[11] Alex Park and James Glass, "Unsupervised pattern discovery in speech," *Transactions on Audio, Speech, and Language Processing*, pp. 186–197, 2008.

[12] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008, pp. 165–168.

[13] Yaodong Zhang and James Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2009, pp. 398–403.

[14] Igor Malioutov, Alex Park, Regina Barzilay, and James Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2007, pp. 504–511.

[15] Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church, "NLP on spoken documents without ASR," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 460–470.

[16] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *Proceedings of Interspeech*, 2015.

[17] Michael Heck, Sakriani Sakti, and Satoshi Nakamura, "Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario," in *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced Languages*, 2016.

[18] Michael Heck, Sakriani Sakti, and Satoshi Nakamura, "Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering," in *Proceedings of Interspeech*, 2016.

[19] Cheng-Tao Chung, Cheng-Yu Tsai, Hsiang-Hung Lu, Chia-Hsiang Liu, Hung-yi Lee, and Lin-shan Lee, "An iterative deep learning framework for unsupervised discovery of speech features and linguistic units with applications on spoken term detection," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2015, pp. 245–251.

[20] Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proceedings of Interspeech*, 2013.

[21] Jason Chang and John Fisher III, "Parallel sampling of DP mixture models using sub-cluster splits," in *Advances in Neural Information Processing Systems*, 2013, pp. 620–628.

[22] Neil Macmillan and Douglas Creelman, *Detection theory: A user's guide*, chapter 9, Psychology press, 2004.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE.

[24] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[25] Ian Witten and Timothy Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1094, 1991.

[26] Jianhua Lin, "Divergence measures based on the Shannon entropy," *Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.