UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

FACHGEBIET NACHRICHTENTECHNIK
PROF. DR.-ING. R. HÄB-UMBACH

Review

of the Dissertation by Lucas Ondel, entitled

Discovering Acoustic Units from Speech: A Bayesian Approach

Topic and Context

Acoustic units constitute the interface between the continuous-values speech signal and the discrete-valued inventory of linguistic symbols. The unit inventory is typically defined by expert phoneticians, and forms the basis of ``classical'' automatic speech recognition systems, which employ statistical models for each acoustic unit, typically phones in context, to map segments of speech to units.

This thesis is concerned with learning the size of the acoustic unit inventory and a hidden Markov Model for each of the units solely from spoken utterances, i.e., without any supervision information. This is an important research objective for at least three reasons. First, there is a large body of “low-resource'' languages, spoken by only a small population, for which the acoustic unit inventory is simply not yet known. A tool which could discover them from untranscribed audio recordings without the need of expert phoneticians would be highly appreciated. Second, the techniques can be employed to make use of untranscribed audio for the development of language processing tools. Finally, unsupervised acoustic unit discovery may serve as a computational model and shed some light into how infants acquire a language.

Lucas Ondel approaches unsupervised acoustic unit discovery with the toolset of non-parametric Bayesian statistics. This is a promising approach, as it allows to infer the number of units alongside the estimation of the unit models. The Bayesian framework, furthermore, allows to incorporate prior information, e.g., from related languages, and can lead to meaningful models even for small data sets.

Content

After a short introduction which motivates the thesis, gives a glimpse on related work and which summarizes the contributions, the author presents the non-parameric Bayesian Phone-Loop Model, which constitutes the core of his thesis. It consists of a set of hidden Markov Models (HMMs), one for each acoustic unit (AU), whose emission probabilities are modelled by Gaussian Mixture Models (GMMs). To account for an unknown and, in principle, infinite number of AUs, a Dirichlet Process prior is used, which allows to infer the size of the unit inventory from the data. Ondel proposes to reduce the hierarchical model, where an observation is explained as a sequence of HMMs, and each HMM as a sequence of HMM states, to a flat HMM, consisting of a sequence of HMM states, that encode both the HMM label and the state label within an individual HMM.

While this, at first sight, appears to be a minor change, it has considerable consequences: It avoids the introduction of a boundary variable as in [Lee and Glass, 2012] and reduces the complexity of the inference process. For this model the generative process, which generates the observed data, and an inference algorithm for the posteriors of the latent variables and parameters are derived. Rather than using the Chinese Restaurant formulation, the author prefers the stick breaking formulation, which allows to develop a computationally more efficient variational inference algorithm to estimate the posteriors. For the involved probabilities, Ondel chose their formulation as members of the exponential family, which enables a unified treatment of categorical and Gaussian random variables. The inference algorithm jointly estimates the posteriors of all variables, and thus also a segmentation.

This chapter is concluded with an experimental evaluation. Experiments are carried out on the well-known TIMIT database, and a truly low-resource language, MBOSHI, a Bantu language, spoken in Congo-Brazzaville. The resulting segmentation is compared with a reference segmentation in terms of precision and recall (with a collar of 20 ms), where the reference segmentation is from phonetic experts in the case of TIMIT, and from forced alignment by a HMM-GMM recognizer trained in supervised fashion in the case of MBOSHI.

AU discovery performance is measured by the “normalized mutual information’’ (NMI), which is the mutual information between the discovered and the ground truth acoustic units, normalized to the average of the entropies of the unit sets. This measure, however, should be taken with some care, because its value depends on the size of the acoustic unit inventory. Thus, the values for different sizes of discovered AU inventories are not directly comparable. To gain more insight in the performance, additional performance measures should have been used, such as, e.g., an equivalent phone error rate, where each discovered AU is mapped to the ground truth phone it overlaps with most. Also, the ABX score would have been insightful, which compares the same phones spoken by different speakers with different phones spoken by the same speaker, and thus is a measure of how speaker independent the discovered units are.

The following chapter “Generalized Subspace Model for Sound Representation’’, introduces a general framework for learning a subspace of probabilistic models. I liked the toy example which illustrates how the concept works. This framework is then used to derive the “Dirichlet Process Subspace Hidden Markov Model’’ to be used in the AU discovery task. The goal is to guide the AU discovery process to the relevant phonetic subspace. Moreover, it provides a principled way to include information from related languages, as their phonetic subspace can be used as an “educated prior’’ for the phonetic subspace discovery of the target language. Note that the subspace is learnt in a supervised fashion. This is o.k., since it is done on different languages and not on the target language for which the AU inventory has to be determined. Finally, the proximity of discovered phone models from known phones in the subspace allows an interpretation of the found models in terms of phonetic categories (e.g., nasals, etc.).

Using a subspace prior from other European languages led to significant improvements in NMI on TIMIT, while, not surprisingly, this was ineffective for MBOSHI, simply because MBOSHI is from a completely different language family than the languages from which the subspace was defined. However, the cheating experiment, where the subspace was learnt (in supervised mode) on the same language as the AU discovery, showed that even better priors do exist.

The final contribution of this dissertation is on improving the language modelling part of the model. While so far a unigram model has been used on the AU sequence, the chapter entitled “Phonotactic Language Model’’ illustrates how to employ higher-order language models. To this end, the Dirichlet Process is replaced by a hierarchical Dirichlet Process (HDP), more precisely, a HDP of order two (a Bigram). Thus, the probability of an Acoustic Unit becomes dependant on the preceding AU. Again, the author argues that inference with the (hierarchical) Chinese Restaurant process is computationally too

expensive and therefore replaces it by a hierarchical stick breaking construction according to Teh. Here, the parameters of the Beta distribution have to be chosen in a specific way, which, however, destroys the conjugacy of the distribution, making inference more difficult. Lucas Ondel derives the inference algorithm for this extended model. He further proposes a weighting of acoustic and language model contributions, thus accounting for their different modelling accuracy. Going from unigram to bigram gave an improvement of up to one percentage points in NMI, and the weighting improved it by another up to one percentage point.

The dissertation is complemented with a conclusion, the list of references and two appendices, which summarize variational Bayesian optimization and the theory of exponential family of distributions.

Evaluation

Acoustic Unit discovery using a non-parametric full Bayesian approach for a HMM-GMM model is an appropriate, although mathematically very demanding undertaking. Mr. Ondel masters this challenge with great success. The thesis clearly shows the author's proficiency in this field of advanced probability theory. The notation is concise, the text is well comprehensible and didactically well laid out. Although the thesis is full of formulas, it is nevertheless well readable. Mr. Ondel has done a great job in laying out a complex theory to the reader.

The topic is appropriate to this area of dissertation and is up-to-date from the viewpoint of the present level of knowledge. Nevertheless, two remarks are in order: One should mention, that in the context of language modeling, an extension of the Dirichlet Process, the Pitman-Yor process, is known to deliver language model probabilities that better account for the universal inverse power law, called Zipf's law. However, here, to the best of the reviewer's knowledge, no stick breaking construction is known, which gives good reasons to stick to the Dirichlet process. Nevertheless, a comment why the Pitman-Yor process was not considered would have been appreciated. Second, HMM-GMM methods are no longer considered state of the art in supervised automatic speech recognition, and also in the field of unsupervised unit discovery approaches, neural networks are making inroads. It would therefore be interesting to compare, at least on theoretical grounds, the approaches outlined in this thesis with recent developments based on variational approaches using neural networks.

The author presented original work. The main contributions to the state of the art, which also form the main chapters of this thesis, are as follows. There is, first, the unsupervised HMM-GMM phone loop model, which is a clear improvement in terms of computational efficiency and simplicity, over the original model by Lee and Glass. Second is the Dirichlet Process subspace HMM, which, besides being a theoretical very interesting concept, leads to significantly improved phone models. Finally, the bigram Dirichlet Process, which was previously known in the context of (pure) language modelling, brings a refinement with slightly improved NMI performance.

Lucas Ondel has conducted a series of experiments to corroborate his theoretical findings. Here, I would have liked to see a somewhat more extensive experimental evaluation. The quality of the discovered acoustic units is solely assessed by the normalized mutual information measure (NMI), although the literature on zero resource speech technologies offers more performance measures, that highlight different aspects of the models.

The core of the thesis has been published in the leading international conferences, mostly ICASSP and Interspeech. His participation in the 2016 and 2017 JSALT workshop furthermore **documents his excellent international standing**. I also appreciate that he has made some of the code open source.

In my opinion, the thesis clearly meets the requirements of the proceedings leading to PhD title conferement.

Paderborn, January 14, 2021

(Reinhold Häb-Umbach)