TOPIC IDENTIFICATION OF SPOKEN DOCUMENTS USING UNSUPERVISED ACOUSTIC UNIT DISCOVERY

Santosh Kesiraju[1,3]*, Raghavendra Pappagari*[2]*, Lucas Ondel*[1]*, Lukáš Burget*[1]*,*
Najim Dehak[2]*, Sanjeev Khudanpur*[2]*, Jan "Honza" Černocký*[1]*, Suryakanth V Gangashetty*[3]

[1] Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic
[2] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, U. S. A
[3] International Institute of Information Technology, Hyderabad, India

ABSTRACT

This paper investigates the application of unsupervised acoustic unit discovery for topic identification (topic ID) of spoken audio documents. The acoustic unit discovery method is based on a non-parametric Bayesian phone-loop model that segments a speech utterance into phone-like categories. The discovered phone-like (acoustic) units are further fed into the conventional topic ID framework. Using multilingual bottleneck features for the acoustic unit discovery, we show that the proposed method outperforms other systems that are based on cross-lingual phoneme recognizer.

Index Terms: topic identification, acoustic unit discovery, unsupervised learning, non-parametric Bayesian models

1. INTRODUCTION

Recent advances in machine learning and spoken language technologies have given rise to many daily life applications. This progress is mainly coming from the so called "deep learning" methods, that require large amounts of labelled data for training. Unfortunately, for many languages the lack of labelled data precludes the direct application of state-of-art spoken language technologies.

The need for automatic analysis of spoken documents is important, since the amount and the ability to store multimedia data is increasing day-by-day. The technologies developed in this regard are primarily useful for tasks such as query based document retrieval, topic identification (topic ID), key-word spotting, etc. Most of these information retrieval tasks rely on the semantics in a document, where the notion of topics play an important role. One particular task of interest is topic ID, where the goal of a system is to identify the topics of the spoken documents in a given collection. This can also be seen as a supervised task, where, a given document has to be classified into one of the pre-defined topics.

A majority of the systems for topic ID of spoken documents use word or phoneme based automatic speech recognition (ASR) as the pre-processing step, followed by the application of techniques developed by the text retrieval community [1, 2, 3]. It is possible to train ASR systems for English on large amounts (1000 hours) of publicly available data [4] and software [5]. But, not every language is rich in resources for building ASR systems, hence there is a need for developing techniques that are useful for languages with low or zero resources. Earlier works on the analysis of spoken documents in zero resource scenarios were based on identifying recurrent patterns of speech (spoken words), where dynamic time warping (DTW) based algorithms were used [6, 7]. However, these are not scalable to large amounts of data. An alternative is to use phone recognizers from other languages. This idea was explored for the task of topic ID in [8]. Under limited resource conditions, i. e., with limited vocabulary for training an ASR, topic ID of spoken documents was explored in [9]. In this paper, we propose a topic ID system that relies on the unsupervised discovery of acoustic (phone-like) units using a non-parametric Bayesian model.

We have recently proposed an infinite phone-loop model [10], similar to [11], to automatically segment unlabelled speech into phone-like categories. By using variational Bayes rather than Gibbs sampling, we have shown that this model can be trained efficiently on large speech corpora with greater accuracy [10]. We use this model as a front-end to a topic ID system. A similar idea was proposed in [12, 13], where the authors used "self-organizing-units" to represent speech into meaningful tokens. In our work, we jointly learn the speech segmentation and the parameters of the acoustic model in a completely unsupervised fashion, whereas the earlier approaches [12, 11], learn the segmentation independent of the acoustic model. In [12], the acoustic model is learnt together with the language model, whereas we limit ourselves to model the acoustic data.

The infinite phone-loop model is described in Section 2, and our topic ID framework is explained in Section 3. Section 4 includes the details of the data set, description of the baseline and the proposed systems. We provide the results of topic ID systems in Section 5, followed by conclusions in Section 6.

2. THE INFINITE PHONE-LOOP MODEL

2.1. Model

The model aims at segmenting and clustering unlabelled speech data into phone-like categories. It is similar to a phone-loop model in which each phone-like unit is modelled by an HMM, and each HMM state distribution is represented by a GMM. This phone-loop model is fully Bayesian in the sense that:

- it incorporates prior distributions over HMM state transition probabilities, and parameters of state emission GMM distributions,
- it has a prior distribution over the units modelled by a Dirichlet process [14].

Informally, the Dirichlet process prior can be seen as a standard Dirichlet distribution prior for a Bayesian mixture with an infinite number of components. However, we assume that our N data samples have been generated with only M components ($M \leq N$) from the infinite mixture. Hence, the model is no longer restricted to have a fixed number of components but instead can learn its complexity (i. e. number of components used, M) according to the training data. The generation of a data set with M speech units can be summarized as follows:

1. sample the vector $\mathbf{v} = v_1, ..., v_M$ with
$$v_i \sim \mathrm{Beta}(1, \gamma) \tag{1}$$
where γ is the concentration parameter of the Dirichlet process
2. sample parameters of M HMMs, θ_1,...,θ_M from the prior (base) distribution of the Dirichlet process.
3. sample each segment as follows:
   (a) choose a HMM parameters with probability $\pi_i(\mathbf{v})$ (using *stick breaking process* [15]) defined as:
$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j) \tag{2}$$
   (b) sample a path $\mathbf{s} = s_1, ..., s_n$ from the HMM transition probability distribution
   (c) for each s_i in $\mathbf{s}$:
      i. choose a Gaussian components from the mixture model
      ii. sample a data point from the Gaussian density function

2.2. Model parameters

In the absence of information about the prior distribution of the parameters of the model, it is convenient to use conjugate prior (distribution), which greatly simplifies the inversion of the model: indeed, due to the conjugacy, the posterior distribution of each parameter of the model will have the same parametric form of the prior. The distribution of the mean $\boldsymbol{\mu}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}$ with diagonal $\boldsymbol{\lambda}$ is modelled by a Normal-Gamma density: $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, (\kappa_0\boldsymbol{\lambda})^{-1})$ $\mathrm{Gamma}(\boldsymbol{\lambda}|\alpha_0, \boldsymbol{\beta_0})$ where $\boldsymbol{\beta}_0$ is the rate parameter of the Gamma distribution. The prior of the weights $\boldsymbol{\pi}$ of a GMM and the row r of the transition matrix of an HMM are modelled by Dirichlet distributions parametrized by the vectors $\boldsymbol{\eta}_0^{(gmm)}$ and $\boldsymbol{\eta}_0^{(hmm,r)}$ respectively. Finally, the prior distribution over the proportions v_i is the $\mathrm{Beta}(1, \gamma)$ distribution. The model also has 3 set of hidden variables:

- $\mathbf{c}$, where c_i the index of the HMM for the i^{th} segment in the data set
- $\mathbf{S}$, where s_{ij} the HMM state of the j^{th} frame in the i^{th} segment
- $\mathbf{M}$, where m_{ij} the GMM component of the j^{th} frame in the i^{th} segment.

2.3. Inference

We would like to invert the model previously defined to obtain the probability of the parameters, and the hidden variables which define the segmentation, given the data. Following variational Bayes (VB) framework, it can be achieved by optimizing a lower-bound on the log-evidence of the data with respect to the distribution over the parameters q:

$$\begin{aligned} \log p(X) \geq & E_q[\log p(\mathbf{X}, \mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}|\boldsymbol{\Phi}_0))] \\ & - E_q[\log q(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta})] \end{aligned} \tag{3}$$

where $\mathbf{X}$ is the entire set of features of the N segments, $\mathbf{c} = c_1, ..., c_N$, $\mathbf{S} = s_{11}, ..., s_{NL_N}$, $\mathbf{M} = m_{11}, ..., m_{NL_N}$, $\boldsymbol{\Theta}$ is the set of all the parameters and $\boldsymbol{\Phi}_0$ is the set of the hyper-parameters of the prior distribution over the parameters. The equality is achieved if and only if $q(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}) = p(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta} \mid \mathbf{X})$. Because of the conjugate prior distribution described in Section 2.2, we have a closed form solution [15] for a coordinate ascent algorithm, when considering the mean-field approximation:

$$q(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}) = q(\mathbf{c}, \mathbf{S}, \mathbf{M})q(\boldsymbol{\Theta}), \tag{4}$$

where we have assumed the statistical independence between the parameters and the hidden variables of the model. Following [15], another approximation is done to cope with the infinite number of components in the mixture; we set $v_T = 1$ to force the weight of any component greater than T to zero. By using the factorization in (4) and variational calculus, one can show that the (log) distributions that maximizes the bound (3) are:

$$\begin{aligned} \log q^*(\mathbf{c}, \mathbf{S}, \mathbf{M}) = E_{q(\boldsymbol{\Theta})}[\log p(\mathbf{X}, \mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}|\boldsymbol{\Phi}_0)] + \text{const} \\ \log q^*(\boldsymbol{\Theta}) = E_{q(\mathbf{c}, \mathbf{S}, \mathbf{M})}[\log p(\mathbf{X}, \mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}|\boldsymbol{\Phi}_0)] + \text{const} \end{aligned} \tag{5}$$

Maximizing the bound (3) minimizes the KL divergence between (4) and the true posterior distribution of model parameters. Therefore (4) can be taken as the approximate posterior, which is found by evaluating each factor in turn using (5) until convergence. Details about the update equations can be found in [10].

The mixture of HMMs can be interpreted as a single compound HMM, which allows us to easily evaluate the approximate posterior distribution $q(\mathbf{c}, \mathbf{S}, \mathbf{M})$ using the standard forward-backward (Baum-Welch) algorithm. Similarly, Viterbi algorithm can be used for decoding the sequences of the discovered acoustic units. This subtlety simplifies the inference algorithm as we do not need any pre-segmentation of the speech data.

3. TOPIC ID FRAMEWORK

3.1. Topic ID in low resource scenarios

Let D be the collection of documents comprising a vocabulary V, and let each document belong to one and only one topic from a set of T topics. Let d, w and t be the variables for denoting documents, tokens in the vocabulary and topics respectively. Assuming the *bag-of-words* approach, each spoken document d is represented in the form of a vector, whose dimension is equal to the size of the vocabulary V. In the conventional topic ID framework, the vocabulary V is simply the set of words as seen in the document collection. In low resource scenarios, when a reliable word based ASR is not available, the vocabulary could be made from phoneme n-grams (usually $n = 3, 4$). It was observed that the topic ID based on phoneme trigrams is a robust alternative to a word based topic ID system [2]. Since the infinite phone-loop model discovers phone-like units, we experimented with 3-grams and 4-grams as the terms (word-types) in the vocabulary.

3.2. Vocabulary selection

In a supervised setting, vocabulary selection plays an important role as it can drastically reduce the dimension of the document vectors and significantly improve the performance of the classifier. The n-grams for vocabulary are chosen based on conditional probabilities as used in [2]. The conditional probability of topic t given a n-gram w is estimated as follows:

$$P(t \mid w) = \frac{f_{wt} + |T|\, P(t)}{f_w + |T|}, \tag{6}$$

where f_{wt} is the number of times the n-gram w appeared in documents related to topic t, f_w is the total number of times n-gram w appeared in all the documents from the training set. $P(t)$ is the probability of topic t as estimated from training corpus. The conditional probability in (6) is computed for every topic and the vocabulary is formed by considering top N_t n-grams per topic with the highest probabilities (6).

3.3. Document representation

If f_{wd} represents the frequency of token w in document d, then the smoothed TF-IDF (term frequency - inverse document frequency) representation (v_{wd}) is given by,

$$v_{wd} = f_{wd} \, . \log\left(\frac{|D|}{1 + N_{dw}}\right) + 1, \tag{7}$$

where N_{dw} represents the number of documents in which the term w appears. The resulting document vectors are further ℓ_2 normalized, such that the sum of the squares of elements equals to 1.

3.4. Document classification

For classifying the documents, we have used linear support vector machines, trained using stochastic gradient descent [16, 17]. The SVMs are used in a one-versus-all strategy for multi-class classification. On the training data, we used 5-fold cross validation and performed grid search over the choice of hyper-parameters (i. e., choice of ℓ_1, ℓ_2, elastic net regularization and the regularization coefficient) of the classifier. Using the best of hyper-parameters, the classifier is trained again using all the training data to predict the topic labels of the test documents.

4. EXPERIMENTAL SETUP

4.1. Data set

Our experiments on topic ID are conducted on the Fisher phase 1 English corpus, which is a collection of recordings from conversational telephone speech. Each document represents one telephone conversation that includes both sides of the call, and is associated to one and only one topic. We chose a *subset* that consists of the same 6 topics as in [7], but relatively more number of documents per topic. The details of this subset of data used in our experiments is given in Table 1. This subset was chosen to study the acoustic unit discovery (AUD) model. We have also experimented on a larger set of 40 topics with the same data splits as used in [2, 3, 8].

4.2. Oracle system

The oracle system is based on the English phoneme recognizer trained on Fisher corpus with large amounts ($\sim$ 500 hrs.) of data. The motivation for using such a setup is to show the performance of a topic ID system in scenarios where the target language is known and considerably large amounts of training data is also available. We used DNN based phoneme recognizer built with the Kaldi toolkit following the recipe described in [18].

Table 1. *The statistics show number of recordings per topic from a subset of Fisher corpus used in the preliminary experiments.*

Topic Name	**# docs.**	
	Training set	**Test set**
Anonymous Benefactor	20	56
Corporate Conduct in the US	20	38
Education	20	57
Holidays	20	58
Illness	20	71
Minimum Wage	20	144
Total duration (hrs).	21.67	77.28

4.3. Baseline systems

The baseline systems are based on phoneme recognizers from various languages: Czech, Hungarian, Russian, which were trained with split temporal context features [19]; and Turkish, from the Babel program, which was trained in a similar framework as described in [20]. The Hungarian phone recognizer was used as a baseline comparison for the task of topic ID in [2, 8, 12, 13].

4.4. Proposed system

The proposed system is based on the discovered acoustic units from the infinite phone-loop model. We explored the following set of input speech features for training the model:

1. 13 dimensional MFCCs + Δ + $\Delta\Delta$
2. Multilingual bottleneck features (Babel-MBN) [21].
3. Multilingual bottleneck features (global phone dataset, GP-MBN) [22].

The Babel-MBN features are extracted using bottleneck neural network trained on data comprising of Cantonese, Pashto, Tagalog, Turkish and Vietnamese and GP-MBN are trained on data comprising of Czech, German, Portuguese, Russian, Spanish, Turkish and Vietnamese languages. Both the neural networks are trained in the same fashion as described in [21].

The hyper-parameters of the infinite phone-loop model play a significant role in quality and quantity of the discovered acoustic units. We primarily experimented with the concentration (γ) of the Dirichlet process prior and the truncation (M). The effect of these hyper-parameters is explained in the following section along with the results. The rest of the hyper-parameters i. e., states per HMM ($\mathcal{S} = 3$) and Gaussian components per state ($\mathcal{C} = 2$) are fixed. We also investigate the importance on the amount of data used to train the infinite phone-loop models.

5. RESULTS

In the first section of the results, we give the comparison of topic ID systems across various baselines and AUD systems. All the systems are based on 1-best sequence from the recognizers. These experiments are performed on a *subset* of 6 topics from the corpus as detailed in Table 1. In the later section, we show the topic ID results on a *larger* set of 40 topics from the same corpus.

5.1. Topic ID on the subset

The AUD model was trained on the 21 hr. training set as presented in Table 1, and the trained model was used to automatically transcribe

Table 2. *Comparison of Topic ID accuracy (in %) on the subset of 6 topics across various systems for the best set of 3-gram vocabulary.*

Recognizer	Acc. (%)	Vocabulary size ($\|V\|$)
Hungarian (HU)	70.19	2428
Czech (CZ)	67.36	5856
Russian (RU)	60.90	3027
Turkish (TU)	55.04	12041
Proposed (AUD)	**76.48**	3029
Oracle (EN)	98.96	9516

both the training and test data in terms of the discovered acoustic units. The resulting automatic transcription was fed into the topic ID framework that was described in Section 3. Here, both the AUD and topic ID models are trained on the same 21 hr. training set (Table 1).

The classification accuracy (in %) of the topic ID systems based on various phoneme recognizers (baseline and oracle) and the discovered acoustic units (AUD) are presented in Table 2. The infinite phone-loop model outperforms all other phone recognizers except for the one trained on the English (target language). This shows that systems trained on another phone set than the target one are far from being optimal, and it is preferable to use unsupervised methods instead. The vocabulary size (set of all unique trigrams) of the proposed system is however much bigger than baseline systems, as the number of discovered acoustic units is 100 (which is larger than the number phoneme set of the other phone recognizers). In Table 2, the results are reported only for the vocabulary size for which the classification accuracy is observed to be highest.

5.1.1. *Topic ID across various AUD systems*

This section presents the comparison of several AUD systems that were explained in Section 4.4. We primarily experimented with various types of input speech features and concentration (γ) parameter of the Dirichlet process. Higher concentration ($\gamma > 1$) encourages more number of clusters (i. e., in the stick-breaking process, higher concentration results in more number of smaller chunks of the stick). From Table 3, we can observe that multilingual bottleneck features are a better representation of speech for unsupervised learning of acoustic units, and therefore results in better topic ID accuracy.

Table 3. *Comparison of Topic ID accuracy (in %) on the subset of 6 topics across various AUD systems.*

Feature type	Accuracy	
	$\gamma = 1.0$	$\gamma = 10.0$
MFCC	36.33	39.27
Babel-MBN	63.41	75.47
GP-MBN	72.74	**76.48**

5.2. Topic ID on the large set

The details of the topic ID training and test splits on a large set of 40 topics from Fisher corpus are presented in Table 4. These are the same splits as used in [2, 3, 8]. For these experiments, we have trained two AUD models, one with 26 hrs. (AUD-26) and the other with 52 hrs. (AUD-52), and neither of them overlap with any of the topic ID training or test data from Table 4. These two AUD models are trained with concentration, $\gamma = 10$ and GP-MBN input speech feature representation, as this combination was observed to be giving the best topic ID performance earlier (Table 3). After the AUD models are trained, they are used to transcribe the topic ID training and test data (Table 4) in terms of the discovered acoustic units, followed by the topic ID framework described earlier in Section 3.

Table 4. *Statistics of the data splits from large set of Fisher phase 1 corpus used in the experiments.*

Set	# docs.	Duration (hrs.)	# topics
Topic ID training	1374	244	40
Topic ID test	1372	226	40

We chose the best baseline system (i. e., Hungarian, HU) from Table 2 and perform the topic ID experiments on this large set of 40 topics in the same framework. All these results are presented in Table 5, and we can observe that the proposed AUD systems are better than the baseline, but still far from the oracle system (DNN based English phoneme recognizer). This is partly because we have a more difficult task of classifying 40 topics.

Table 5. *Comparison of Topic ID accuracy (in %) on the large set of 40 topics for the best set of n-grams from the vocabulary.*

Recognizer	Acc. (%)	$\|V\|$	n-gram	AUD params.
AUD-26	**53.84**	6061	3	$M = 200, \gamma = 10$
AUD-52	**55.54**	2140	4	$M = 100, \gamma = 10$
HU	47.92	25351	3	-
EN	91.41	11236	3	-

From these experiments, we observe that in an unknown scenario and/or language, it is better to borrow knowledge from the other languages at a lower (feature) level (multi-lingual bottleneck features) than at a much higher level (phone recognizer) and rely on the unsupervised (data driven) methods to discover the acoustic units from the data and use them for further tasks.

6. CONCLUSIONS

This work focuses on the importance and application of unsupervised acoustic unit discovery for the task of topic identification. We showed that using multilingual bottleneck features for learning the acoustic units, the performance of the topic ID system could be improved significantly. Our experiments on a corpus of conversational telephone speech showed that the proposed system performs better than the other systems which rely on the cross-lingual phoneme recognizers. Although the results are encouraging, there is still a significant space for improvement to reach the performance of supervised speech recognition systems. One step towards achieving this would be to jointly learn the language model and the infinite phone-loop parameters in an unsupervised fashion.

7. ACKNOWLEDGEMENTS

The work reported here was carried out during the 2016 Jelinek Memorial Summer Workshop on Speech and Language Technologies, which was supported by Johns Hopkins University via DARPA LORELEI Contract No HR0011-15-2-0027, and gifts from Microsoft, Amazon, Google, Facebook. It was also supported by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project ”IT4Innovations excellence in science - LQ1602”. Santosh Kesiraju is partly funded by TCS Research Fellowship program, India.

8. REFERENCES

[1] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek, “Approaches to topic identification on the switchboard corpus,” in *IEEE ICASSP*, Apr 1994, vol. i, pp. I/385–I/388 vol.1.

[2] Timothy J. Hazen, Fred Richardson, and Anna Margolis, “Topic Identification from Audio Recordings using Word and Phone Recognition Lattices,” in *IEEE Workshop on ASRU*, December 2007, pp. 659–664.

[3] Timothy J. Hazen, “MCE Training Techniques for Topic Identification of Spoken Audio Documents,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2451–2460, Nov 2011.

[4] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An ASR Corpus Based on Public Domain Audio Books,” in *IEEE ICASSP*, April 2015, pp. 5206–5210.

[5] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The Kaldi Speech Recognition Toolkit,” in *IEEE Workshop on ASRU*. Dec 2011, IEEE Signal Processing Society.

[6] Rémi Flamary, Xavier Anguera, and Nuria Oliver, “Spoken WordCloud: Clustering Recurrent Patterns in Speech,” in *International Workshop on Content-Based Multimedia Indexing*, June 2011, pp. 133–138.

[7] David F. Harwath, Timothy J. Hazen, and James R. Glass, “Zero Resource Spoken Audio Corpus Analysis,” in *IEEE ICASSP*, May 2013, pp. 8555–8559.

[8] Timothy J. Hazen, Man-Hung Siu, Herbert Gish, Steve Lowe, and Arthur Chan, “Topic Modeling for Spoken Documents using only Phonetic Information,” in *IEEE Workshop on ASRU*, December 2011, pp. 395–400.

[9] J. Wintrode and S. Khudanpur, “Limited resource term detection for effective topic identification of speech,” in *IEEE ICASSP*, May 2014, pp. 7118–7122.

[10] Lucas Ondel, Lukás Burget, and Jan Cernocký, “Variational Inference for Acoustic Unit Discovery,” in *SLTU-2016, 5th Workshop on Spoken Language Technologies for Under-resourced languages*, May 2016, pp. 80–86.

[11] Chia-ying Lee and James Glass, “A Nonparametric Bayesian Approach to Acoustic Model Discovery,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2012, vol. 1 of *ACL ’12*, pp. 40–49.

[12] Herbert Gish, Man-Hung Siu, Arthur Chan, and William Belfield, “Unsupervised training of an hmm-based speech recognizer for topic classification,” in *INTERSPEECH*, September 2009, pp. 1935–1938.

[13] 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.

[14] Charles E. Antoniak, “Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems,” *Annals of Statistics*, vol. 2, no. 6, November 1974.

[15] David M. Blei and Michael I. Jordan, “Variational Inference for Dirichlet Process Mixtures,” *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[16] Léon Bottou and Olivier Bousquet, “The tradeoffs of large scale learning,” in *Advances in Neural Information Processing Systems*, J.C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20, pp. 161–168. NIPS Foundation (http://books.nips.cc), 2008.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey, “Sequence-discriminative training of deep neural networks,” in *INTERSPEECH*, August 2013, pp. 2345–2349.

[19] Petr Schwarz, *Phoneme Recognition based on Long Temporal Context*, Ph.D. thesis, Brno University of Technology, 2009.

[20] Martin Karafiát, Frantisek Grézl, Mirko Hannemann, Karel Veselý, and Jan Cernocký, “BUT BABEL system for spontaneous Cantonese,” in *INTERSPEECH*, August 2013, pp. 2589–2593.

[21] František Grézl and Martin Karafiát, “Adapting Multilingual Neural Network Hierarchy to a New Language,” in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under- resourced Languages SLTU*, 2014, pp. 39–45.

[22] Tanja Schultz, “Globalphone: a multilingual speech and text database developed at Karlsruhe university,” in *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH*, September 2002.