Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 73 – 79

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Unsupervised Linear Discriminant Analysis for Supporting DPGMM Clustering in the Zero Resource Scenario

Michael Heck*, Sakriani Sakti, Satoshi Nakamura

Augmented Human Communication Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

Abstract

In this work we make use of unsupervised linear discriminant analysis (LDA) to support acoustic unit discovery in a zero resource scenario. The idea is to automatically find a mapping of feature vectors into a subspace that is more suitable for Dirichlet process Gaussian mixture model (DPGMM) based clustering, without the need of supervision. Supervised acoustic modeling typically makes use of feature transformations such as LDA to minimize intra-class discriminability, to maximize inter-class discriminability and to extract relevant informations from high-dimensional features spanning larger contexts. The need of class labels makes it difficult to use this technique in a zero resource setting where the classes and even their amount are unknown. To overcome this issue we use a first iteration of DPGMM clustering on standard features to generate labels for the data, that serve as basis for learning a proper transformation. A second clustering operates on the transformed features. The application of unsupervised LDA demonstrably leads to better clustering results given the unsupervised data. We show that the improved input features consistently outperform our baseline input features.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: acoustic unit discovery; Bayesian nonparametrics; Dirichlet process; feature transformation; Gibbs sampling; unsupervised linear discriminant analysis; zero resource

1. Introduction

In a zero resource scenario, large amounts of labeled training data, parallel data, and knowledge about the target language are unavailable for developing speech processing systems with supervised techniques. Where infants are capable of robustly modeling acoustic and language models in an unsupervised way, current speech technology is not yet capable to imitate these capacities.

Confronted with an unknown language, human experts usually attempt to define a set of acoustic units that fully covers the underlying sound repertoire. Core techniques of machine learning approaches to this task are pattern

* Corresponding author. Tel.: +81-743-72-5265 ; fax: +81-743-72-5269.
E-mail address: michael-h@is.naist.jp

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.032

matching [1,2] on raw audio data and unsupervised sound unit detection [3] and have already been successfully applied to solve tasks such as spoken term detection [4], topic segmentation [5] or document classification [6].

In non-clinical situations where development data is usually unavailable, model complexity is not known a priori. Bayesian models such as the Dirichlet process Gaussian mixture model (DPGMM) automatically adjust the model complexity given the data. DPGMMs have been successfully applied to speech processing tasks such as unsupervised lexical clustering [7]. Previous work [8] clustered standard MFCC speech features by inferring a DPGMM. Each Gaussian was interpreted as modeling a specific sound class. The posteriorgrams were evaluated to show that DPGMM is a suitable technique to automatically detect sound classes in untranscribed data.

It is straightforward to assume that more advanced feature representations may lead to a better classifier performance. For instance, context information is an important factor to correctly classify speech features in common speech processing systems. Chen et al. [8] use MFCC features with first and second derivatives for clustering. The derivatives help cover a small context but triple the dimensionality. Feature stacking can cover a much larger context, but at significantly higher expenses in terms of dimensionality. A feasible processing of high-dimensional feature vectors makes dimension reducing feature transformations mandatory.

Traditional supervised acoustic modeling typically makes use of feature transformations such as linear discriminant analysis (LDA) [9] to minimize intra-class discriminability, to maximize inter-class discriminability and to extract relevant informations from high-dimensional features spanning larger contexts. Class discriminating properties of feature vectors are critical for clustering. However, LDA needs class labels to estimate the feature transformations, making it difficult to use in a zero resource setting where the classes and even their amount are unknown.

In this work we attempt to improve the DPGMM clustering by introducing unsupervised LDA to the sampling pipeline. There has been work that utilize k-means clustering to automatically obtain pseudo labels for LDA estimation [10,11]. We similarly attempt to automatically produce class labels, but we want to overcome the limitation of having to predefine the size of the label set. For that, we use a non-parametric DPGMM sampler to generate labels for our untranscribed data. Our contribution is an easy to understand two-staged clustering framework that automatically finds a dynamically sized set of framewise class labels for unsupervised LDA transformation to project high-dimensional large context covering feature vectors into a more suitable subspace for DPGMM clustering.

2. 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 it's complexity automatically given the training data. Inference is typically sample based using a Markov chain Monte Carlo (MCMC) scheme such as Gibbs sampling. The sampler used here is combining a restricted Gibbs sampler with a split/merge sampler. For more in-depth informations, please refer to [12,8].

2.1. Generative process

Let $X = x_1, \cdots, x_n$ be a set of observations. The generative process of X given a DPGMM is as follows:

- Mixing weights $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$ are generated according to a stick-breaking process
- GMM parameters $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_k\}$ are generated according to a prior distribution $\mathrm{NIW}(m_k, S_k, \kappa_k, \nu_k)$
- A label z_i is assigned to every x_i, according to $\boldsymbol{\pi}$
- A data point x_i is generated according to the z_i-th Gaussian component

$\theta_k = \{\mu_k, \Sigma_k\}$ are Gaussian parameters, and the parameter set of the prior Normal-inverse-Wishart (NIW) distribution consists of a prior m_0 for μ_k, a prior S_0 for Σ_k, the belief-strength κ_0 in m_0 and the belief-strength ν_0 in S_0.

2.2. Inference

The parallelizable sampler used here alternates between a non-ergodic restricted Gibbs sampler and a split/merge sampler to form an ergodic MCMC sampler.

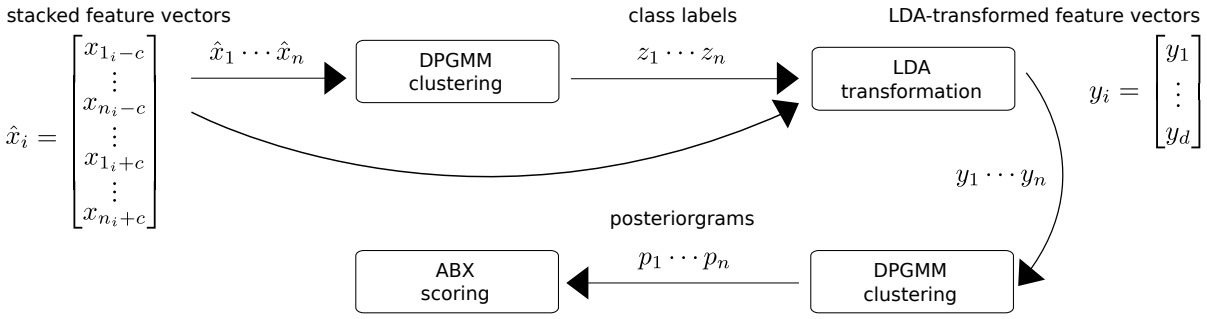

Fig. 1. Scheme of the sampling process. The dimensionality of the stacked feature vectors $\hat{x}_i$ is $13(2c + 1)$, where c is the stacking context. The dimensionality of the LDA-transformed feature vectors y_i is $d \ll 13(2c + 1)$.

Restricted Gibbs sampling allows labels z_i to be sampled from a finite set Z. By definition of the DPGMM, the distribution of the mixture weights follows a Dirichlet distribution.

Split/merge sampling performs operations on the existing components. To provide good split candidates, each component is augmented with two sub-clusters with mixing weights $\pi_{k,l}, \pi_{k,r}$ and parameter sets $\theta_{k,l}, \theta_{k,r}$, and each observation of a component is augmented with a sub-cluster label $z_{sub_i} \in l, r$.

The Split/merge sampler proposes split and merge moves in a Metropolis-Hastings fashion. A Hastings ratio H is computed according to the momentary assignment of observations of a component to it's sub-clusters, and a move is accepted with a probability $\min(1, H)$. For the merge step, merges of randomly picked components are proposed.

2.3. Posteriorgram generation

The posterior probability of cluster c_k, given observation x_i is

$$p(c_k|x_i) = \frac{\pi_k N(x|\theta_k)}{\sum N(x|\theta_j)} \tag{1}$$

and $P_i = (p(c_1|x_i), \cdots, p(c_K|x_i))$ forms a posteriorgram for observation x_i.

3. Unsupervised Linear Discriminant Analysis

To improve the quality of DPGMM based speech feature vector clustering, we propose to utilize LDA in an unsupervised fashion. The idea is to automatically find a projection that maps high-dimensional feature vectors into a more suitable space for DPGMM clustering. Using LDA for feature transformation is motivated by several reasons: First of all, LDA is a simple linear transformation and a standard technique that attempts to minimize intra-class discriminability and to maximize inter-class discriminability of speech features. As such, LDA is widely used in automatic speech processing systems. Naturally, class discriminating properties would greatly benefit clustering approaches for acoustic unit discovery. Secondly, LDA allows for dimensional reduction of high-dimensional stacked feature vectors that span a larger context by omitting lower-ranked coefficients. Thirdly, LDA transformations are fast and easy to compute and only require the feature vectors and respective class labels.

The need of class labels makes it difficult to use LDA in a zero resource setting out of the box, since the class identities and even the amount of classes are unknown, let alone the class memberships of our feature vectors. To overcome this issue, we propose to use a two-staged clustering framework that automatically finds frame-based class labels in a first clustering run on the untranscribed data, performs LDA estimation and feature transformation and runs a second run of clustering on the transformed vectors. Figure 1 is a graphical overview of the proposed framework.

The DPGMM is a Bayesian non-parametric model that automatically detects the optimal number of classes given a set of data. We make use of this property and run an initial clustering on standard feature vectors to get a set of class labels and the hypothesized class membership of every speech frame. These classes are generic and simply named

with the numeric ID of the Gaussian that most likely produced the respective feature vector. Having the labels for every frame at hand, we can simply estimate an LDA transformation.

Context is an important source of information to correctly classify speech features. Feature stacking can cover a much larger context than appending the first and second derivatives. Thus, we compute the LDA for stacked MFCC feature vectors, where we use a context of c, meaning that we stack the c left and c right feature vectors on top of the current vector, which is the center vector. To keep the dimensionality low for any feasible clustering, we reduce the output dimension d of the LDA transformation to a significantly lower value than the input dimension.

Once we produced new feature vectors by using the LDA transformation, we perform another run of DPGMM based clustering on these. The class discriminating properties of the LDA are assumed to support the clustering and lead to a better cluster quality. We demonstrate in Section 4 that the LDA in fact significantly boosts the clustering quality.

4. Experiments

4.1. Data

All our experiments were conducted on the official data set of the Interspeech zero resource speech challenge[13], which contains two separate data sets of pure speech for American English (4h 59min) and Xitsonga (2h 29min), a southern African Bantu language. Segments contain no overlapping speech, noise or pauses, and speech of exactly one speaker. 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 used to measure the cluster quality is based on the minimal pair ABX phone discriminability task[14]. After clustering, GMM posteriorgrams can be computed for each speech frame, as described in Section 2.3. Evaluation is performed on these posteriorgrams, that serve as new speech representations for each frame. Let A and B be stimuli belonging to sound categories a and b. Then the ABX phone discrimination accuracy is

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

where $d(a,b)$ is the DTW divergence and δ is an indicator function. We followed Schatz et al.[14] and used the cosine distance for standard features and KL-divergence for posteriorgrams to compute the DTW divergences. Our scores are the error rates within and across speakers, where the rates are averaged over all found contexts for a given pair of central phonemes and then over all pairs of central phonemes.

4.3. Setup

To get as close as possible to the setup of Chen et al.[8] we use the same parameters. Thus, DPGMM sampling is done for 1500 iterations, and the priors are set so that m_0 is the global mean, S_0 is the global covariance, $\kappa_0 = 1$ and $\alpha = 1$. The value of ν_0 slightly varies and is set to the toolkit's default of $\nu_0 = D + 3$, where D is the dimension of the input feature vectors. All feature vector types are extracted for a frame length of 25 milliseconds and frame shift of 10 milliseconds and make use of mean variance normalization (MVN) and vocal tract length normalization (VTLN).

4.4. Baseline

As a baseline we extract 39 dimensional MFCC+Δ+$\Delta\Delta$ as input to the DPGMM sampler. We test both the ABX discrimination error of the raw features as well as of the posteriorgrams as result of the DPGMM clustering. We also compare to the official results of Chen et al.[8], the details are listed in Table 1.

Despite using the same sampling setup and input feature types, there is a mismatch between the results of Chen et al.[8] and our baselines. We believe this mismatch is caused by the fact they use a custom voice activity detection

Table 1. The baselines for this work are the results of Chen et al.[8] and our own numbers produced with the official data segmentation. DPGPG stands for Dirichlet process Gaussian posteriorgrams.

System	English		Xitsonga	
	within	across	within	across
MFCC+Δ+ΔΔ[8]	17.2	26.8	19.6	30.8
DPGPG[8]	10.8	16.3	9.6	17.2
MFCC+Δ+ΔΔ	15.7	25.5	19.7	30.0
DPGPG	12.2	19.5	8.9	14.2

for segmenting the full 10 hours of English data and do not mention any segmentation attempts for the 5 hours of Xitsonga data, where we use the officially provided segmentation that limits both data sets to about half the amount. Due to the differences we start with a higher error rate on English, but a lower error rate on Xitsonga.

4.5. PCA vs. LDA

We started our experiments by doing a comparison of LDA and the closely related principal component analysis (PCA)[15,16] in order to find out whether simpler and entirely unsupervised techniques could serve us better than the proposed approach. PCA is a simple orthogonal linear transformation to de-correlate variables. The difference to LDA is that the class memberships of the features subject to transformation are not taken into consideration. The data is simply mapped to a new coordinate system so that the first principal component has the largest variance, the second component the second largest, and so on. Dimensional reduction is performed in a similar way to LDA. The results are displayed in figure 2.

The transformations take stacked standard feature vectors without their derivatives as input. With the stacking context parameter set to $c = 4$, and the output dimensionality set to $d = 20$, we found that the application of PCA led to opposite results for the two target languages. Where the transformation helped improve the clustering process of English speech feature vectors according to both error rates, within and across speakers, using PCA did not help beat the baseline for the Xitsonga data.

The LDA transformation on the other hand was able to produce feature vectors that significantly helped the DPGMM clustering process to find better clusters. The error rates for both languages dropped consistently, and especially across speakers a clear performance boost is observable. By using the LDA transformed features, it was possible to outperform our own baseline, and the error rates for English also undercut the numbers of Chen et al.[8], thus compensating for the deficit in the baseline numbers that we had to begin with.

4.6. Input features for LDA

One of the advantages of LDA is that it can be applied to a multitude of input features. To find out whether there is a better choice than MFCCs, we conducted experiments on perceptual linear prediction (PLP) feature vectors for comparison. The PLP features produced evaluation errors comparable to and slightly lower than the baseline set by MFCC: By using PLP+Δ+ΔΔ as input to the sampler we observed improvements for both languages given the error rate within speakers. We have seen an error rate decrease across speakers for Xitsonga and a stable error rate across speakers for English. When using PCA and LDA transformed PLP features, the latter proved to generate the lower error rates in all comparisons. No matter whether MFCC or PLP features served as input, PCA was inferior to LDA, leading us to the joint conclusion that unsupervised LDA is the better choice for mapping features into a lower space to perform clustering, and that PLP feature vectors seem to carry some information that benefits class discriminability. Again, the results are plotted in figure 2.

4.7. Input and output dimensions for LDA

The stacking context parameter c directly regulates the input dimensionality of the LDA transformation. The output dimensionality of LDA can be adjusted by setting the parameter d. Since the right choice of input and output dimensionality of the LDA transformation might be critical to the clustering performance, we conducted a grid search

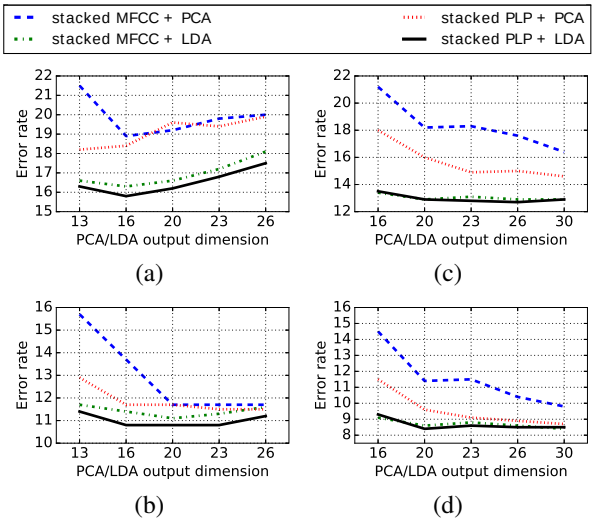

Fig. 2. Error rates across and within speakers in dependency of feature type and transformation, computed for several output dimensionalities. (a)-(b) Error rates for English, (c)-(d) Error rates for Xitsonga.

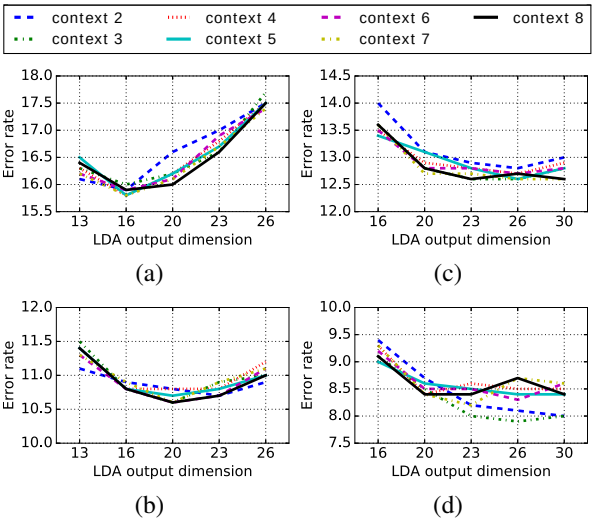

Fig. 3. Error rates across and within speakers in dependency of stacking context size and the LDA output dimensionality. (a)-(b) Error rates for English, (c)-(d) Error rates for Xitsonga

on these two parameters. We restricted these tests to PLP features only, given our previous findings. The results of our tests are visualized in figure 3.

It is noteworthy that the usefulness of the LDA transformation does not seem to depend on the stacking context size. Our results demonstrate that any context larger than 2 is a suitable choice for the transformation estimation. The output dimensionality of the transformation however does have an impact on performance. Increasing dimensionality has an opposite effect on the two target languages. Where Xitsonga benefits from a higher output dimensionality, best results for English are produced using a lower dimensionality. This effect might be attributable to two circumstances: Firstly, it is to point out that the data sets differ in nature. The English data is comprised of conversational speech, thus a transformation that maps into a significantly lower dimensional space might produce more stable features suitable for the DPGMM clustering. The Xitsonga data on the other hand has the characteristics of read speech, thus the higher dimensions of the transformed features may still be able to carry useful information for sound class discrimination. Secondly, having a look at the manually crafted phoneme sets of the original corpora (Buckeye for English, NCHLT for Xitsonga), one finds that the optimal output dimensionalities for the respective data sets roughly correlate with the set sizes. Where the English data is originally labeled with a 39-elemental phoneme set, and the optimal output dimension in our experiments is located around 20, the Xitsonga data is labeled with 53 distinct tokens, and the optimal output dimension seems to lie around 26. The results further allow to assume that class discrimination across speakers benefits from a slightly higher dimensionality than the within speaker class discrimination. We believe this is because more dimensions might be necessary to discriminate classes including speaker variations.

For in-the-field experiments we would naturally either have to apply parameters tuned on an out-of-language development set or parameters that proved to work well for solving similar tasks. The results displayed in figure 3 show that it is possible to find good parameters for a known language that work considerably well for a new language: By using the parameters we tuned on English, we achieve a sound discrimination quality that is only slightly lower than the one that could be achieved with an optimal parameter set. Table 2 lists the best results of our tests.

5. Conclusion

We were able to demonstrate that dimension reducing unsupervisedly estimated LDA is an efficient way to map speech feature vectors into a subspace that is more suitable for DPGMM based clustering. The Gaussian posteriorgrams that can be extracted from a DPGMM sampled on transformed vectors carry better sound class discriminating characteristics than the ones sampled on untransformed standard features. We showed that unsupervised LDA esti-

Table 2. The optimal results for each input feature type.

Features	English		Xitsonga	
	within	across	within	across
MFCC	12.2	19.5	8.9	14.2
MFCC+PCA	11.7	19.2	9.8	16.4
MFCC+LDA	11.1	16.6	8.4	12.9
PLP	11.8	19.6	8.5	13.9
PLP+PCA	11.7	18.4	8.7	14.6
PLP+LDA	10.6	16.0	8.0	12.6

mation based on automatically generated labels works reliably across languages and clustering performance is fairly independent of the stacking context prior to LDA computation. We further showed that the output dimensionality of the transformation does influence clustering quality, and even with a value that has been tuned on an out-of-language test data set, a near optimal clustering quality can be achieved.

The results demonstrate that unsupervised LDA transformation supported DPGMM clustering is particularly suitable for low-resource languages and the zero-resource scenario. It is easy to conclude that this approach might be of help for more general purposes beyond low-resource languages. In future work, we will continue our research on unsupervised feature transformations by exploring the usefulness of the proposed approach for tackling other tasks such as unsupervised acoustic model training.

Acknowledgements

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

References

1. Park, A., Glass, J.. Towards unsupervised pattern discovery in speech. In: *Workshop on Automatic Speech Recognition and Understanding*. IEEE; 2005, p. 53–58.
2. Park, A., Glass, J.. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing* 2008; **16**(1):186–197.
3. Varadarajan, B., Khudanpur, S., Dupoux, E.. Unsupervised learning of acoustic sub-word units. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics; 2008, p. 165–168.
4. Zhang, Y., Glass, J.. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams. In: *Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE; 2009, p. 398–403.
5. Malioutov, I., Park, A., Barzilay, R., Glass, J.. Making sense of sound: Unsupervised topic segmentation over acoustic input. In: *Association for Computational Linguistics Annual Meeting*; vol. 45. Citeseer; 2007, p. 504.
6. Dredze, M., Jansen, A., Coppersmith, G., Church, K.. NLP on spoken documents without ASR. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2010, p. 460–470.
7. Kamper, H., Jansen, A., King, S., Goldwater, S.. Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings. In: *Spoken Language Technology Workshop (SLT)*. IEEE; 2014, p. 100–105.
8. Chen, H., Leung, C.C., Xie, L., Ma, B., Li, H.. Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In: *Proceedings of Interspeech*. 2015.
9. Fisher, R.A.. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 1936;**7**(2):179–188.
10. Ding, C., Li, T.. Adaptive dimension reduction using discriminant analysis and k-means clustering. In: *Proceedings of the 24th international conference on Machine learning*. ACM; 2007, p. 521–528.
11. Tang, J., Hu, X., Gao, H., Liu, H.. Discriminant analysis for unsupervised feature selection. In: *SDM*. SIAM; 2014, p. 938–946.
12. Chang, J., Fisher III, J.W.. Parallel sampling of dp mixture models using sub-cluster splits. In: *Advances in Neural Information Processing Systems*. 2013, p. 620–628.
13. Versteegh, M., Thiolliere, R., Schatz, T., Cao, X.N., Anguera, X., Jansen, A., et al. The zero resource speech challenge 2015. In: *Proceedings of Interspeech*. 2015.
14. Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., Dupoux, E.. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In: *Proceedings of Interspeech*. 2013.
15. Pearson, K.. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 1901;**2**(11):559–572.
16. Hotelling, H.. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 1933;**24**(6):417.