Technical Report
The IRIT-UPS system @ ZeroSpeech 2017
Track1: unsupervised subword modeling

Thomas Pellegrini, Céline Manenti, Julien Pinquier
IRIT, Université de Toulouse, CNRS, Toulouse, France
{thomas.pellegrini,celine.manenti,julien.pinquier}@irit.fr

July 2017

Abstract

In this report, we describe the IRIT-UPS approach for the unsupervised discovery of sub-lexical units in speech in the framework of the *Zero Resource Speech Challenge* 2017 edition. We derive unsupervised representations that consist of the distances between the MFCC vectors extracted from the speech signal and a hundred cluster centroids estimated on millions of these vectors with an efficient scalable implementation of the k-means algorithm. We show that using a whitening transformation (ZCA) to pre-process the MFCCs is crucial to outperform the baseline MFCC representation. Furthermore, we obtain slight but consistent improvements by performing data selection before estimating the centroids. This simple approach yielded competitive results for the within-speaker condition but generalized less well in the between-speaker one.

1 Introduction

Manually annotated speech data abound for the most widely spoken languages but it is not the case for the vast majority of the other languages and dialects. In order to build speech automatic tools, in particular automatic speech recognition systems, one needs unsupervised approaches to cope with this lack for these so-called *under-resourced* languages [1]. For acoustic modeling, unsupervised discovery of sub-lexical units in continuous speech gained momentum in recent years, encouraged by initiatives such as the *Zero Resource Speech Challenges* (ZRSC) [2, 3]. In these challenges, two unit discovery tasks are proposed: the first one at sub-lexical level and the second one at word level. In this document, we report experiments related to the first task only, in which the participants are expected to construct representations of speech sounds robust to within- and between-talker variations, referred to as *within* and *across* in the remainder of this report.

In the 2015 challenge edition, a variety of models were used by the participants: Deep Neural Networks (DNNs) [4, 5, 6], Dirichlet Process Gaussian Mixture Models (DPGMMs) [7]. Methods based on supervised models learned on other well-resourced languages were also used, such as phone posteriorgrams [7] and articulatory information [8].

The DPGMM approach yielded the best results and in one case even outperformed the supervised topline. This approach uses speaker-normalized MFCCs as input to a DPGMM. The DPGMM posteriors were shown to capture phoneme discriminability. DPGMMs, also referred to as infinite

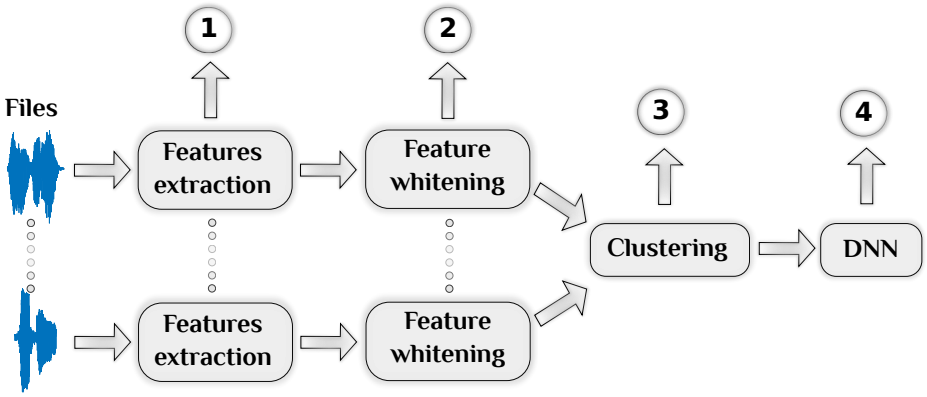

Figure 1: System overview. The numbers inside circles refer to our different feature representations reported in Table 2.

Gaussian Mixture Models, are Bayesian nonparametric models. They can automatically learn the number of components according to the observed data, and as such, are well suited to an unsupervised learning task where the characteristics of the target language are not known (*e.g.*, the number of phonemes).

Concerning neural networks, variants of Auto-Encoders (AEs) were tested: standard, correspondence, and denoising AEs [4]. With these models, the feature representations correspond to the activations of a hidden layer with markedly less neurons than the other layers, for this reason these features are called bottle-neck features. Binarized AE outputs with six features only were shown to perform better than standard MFCCs [6]. The information carried by the speech signal is reformulated in a condensed form and the AE is expected to capture the most salient features of the training data. Studies have shown that, in some cases, AE posteriorgrams outperform those of GMMs [9, 10]. Correspondence AEs (cAEs) no longer seek to reconstruct the input data but data previously mapped in a certain way. They require a first step of grouping speech segments into similar pairs (pseudo-words, etc.) found by Dynamic-Time Warping (DTW). In [5], Siamese networks, a method similar to cAEs that also needs speech segment pairs for training, achieved better results than AEs.

As an alternative to the latter approaches, we wanted to use the k-means algorithm to derive pseudo-classes instead of DPGMMs or DTW. K-means were already used for the same purpose in [11] to obtain phone posteriorgrams with standard AEs, and in [12] also with AEs and graph clustering.

This report is organized as follows. Section 2 describes our approach based on k-means applied to ZCA-whitened MFCCs, eventually followed by DNNs trained with the k-means pseudo-classes. In Section 3, after a brief overview of the speech material and the evaluation metric used in the challenge, we report and analyze the experiments conducted first on French, second on the five languages of the challenge.

2 System description

Our approaches are illustrated in Figure 1. The system can be evaluated at each step: ① feature extraction (baseline representations), ② feature whitening, ③ clustering and ④ DNN-based classifi-

cation.

2.1 Feature extraction: MFCC

As baseline feature representations (called *baseline* system, see section 3.3), the organizers of the challenge used 13-d MFCCs + Δ + $\Delta\Delta$ computed every 10 ms on 25 ms windows. In the starting code kit provided by them, the feature extraction module was made available. Figure 2 shows the 39-d MFCCs extracted from a 10-second speech file taken from the ZRSC challenge (top), and the same features after ZCA whitening (bottom). As one can see, after ZCA, all the 39 descriptors share the same dynamic ranges centered around zero.

2.2 Feature whitening with ZCA

Zero Components Analysis (ZCA) is a feature whitening technique often used in image processing [13]. Its purpose is to standardize and whiten features while retaining their original spatial orientation of the data points, as shown in Figure 3[1]. ZCA differs from Principal Component Analysis (PCA) by a rotation only. More precisely, consider the data arranged in a matrix X, with each row being a data point and each column a descriptor. First, X is zero-mean centered into $\overline{X}$. Let U, D be a singular decomposition of $\text{cov}(\overline{X})$, the covariance matrix of $\overline{X}$, with U the matrix of unit-length eigenvectors and D the diagonal matrix with the corresponding eigenvalues, such that $\text{cov}(\overline{X}) = UDU^t$. In practice, X is whitened with the following formula:

$$X_{\text{ZCA}} = U(D + \epsilon)^{-1/2}U^t\overline{X}$$

The ϵ value chosen can have a significant impact on the results. We studied its impact in section 3.4. Besides preventing from potential illegal divisions by null eigenvalues, this hyperparameter plays the role of a low-pass filtering that limits the impact of the eigenvector axes associated to small eigenvalues.

In speech processing, and speech recognition in particular, to the best of our knowledge, ZCA is not commonly used. We rather found speaker recognition studies making use of PCA whitening applied to i-vectors, a transform called Eigen-Factor Radial (EFR) [14].

We tested ZCA whitening in two ways: 1) a single transformation for a given language/pair dataset, 2) transformations on a *per-file* basis (one transformation per audio file). In the first case, we could have used PCA whitening since a rotation on the whole dataset would not change the results of a subsequent k-means. But, in the second case, only ZCA was feasible since we needed to retain the original data points' orientation to perform k-means afterwards. As we shall see here-after, the second strategy was the best option.

2.3 Clustering with k-means

The k-means algorithm, also called "vector quantization", is a well-known clustering technique for assigning classes to data samples in an unsupervised way. We used the classic version, in which cluster centroids are found by minimizing the L^2 distance between data points and the nearest centroid. We also tested the cosine distance, more precisely an inner-product distance, but the L^2 norm led to the best results.

In our work, we use k-means to learn feature representations in the form of the distances between data points and the cluster centroids as candidate features for the ZRSC task. In [13], the authors

[1] code available at `https://github.com/topel/demo-zca`

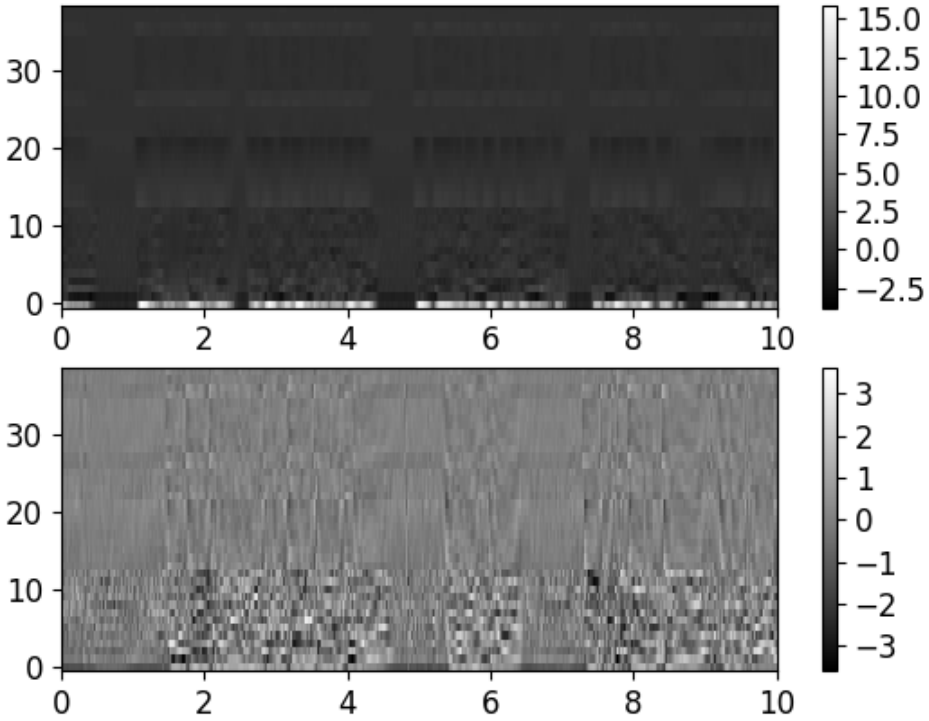

Figure 2: Top: 39-d raw MFCCs for a 10-s speech audio file, bottom: after ZCA whitening.

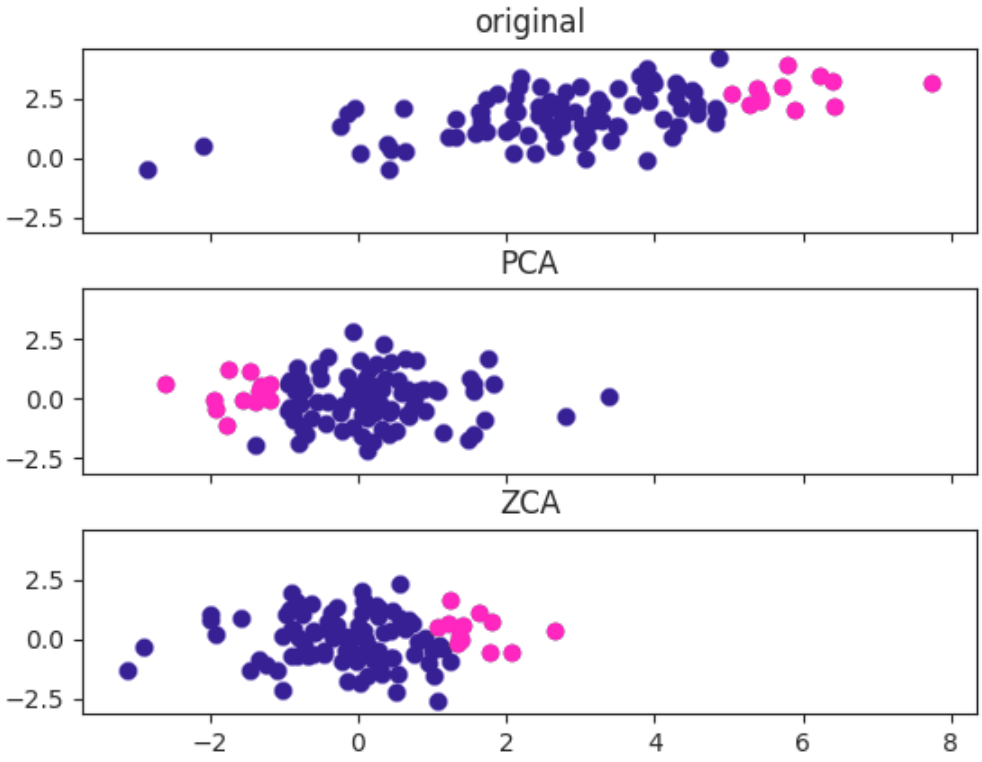

Figure 3: Comparison between PCA and ZCA on a dummy 2-D set of points.

give several recommendations and useful tricks to properly use k-means as an unsupervised feature representation learning tool, and they report excellent results achieved by k-means in learning large-scale representations of images. In particular, the feature whitening pre-processing step (ZCA) is said to be crucial and as we shall report hereafter it is also the case in our experiments. They also recommend to use the so-called spherical variant of k-means [15], in which the centroids are normalized to lie on the unit sphere, but this degraded performance in our case.

Besides using distances as feature representations, we also used the cluster assignments to perform data selection to refine the cluster centroid estimates, and also to train DNNs in a supervised fashion.

2.4 Neural Networks

Based on the pseudo-labels obtained with k-means, we explored the use of two types of DNNs trained in a supervised fashion to recognize the pseudo-labels: Fully-Connected feed-forward (FCNNs) and Convolutional Neural Networks (CNNs). For FCNNs, only comprised of dense layers, the ZCA-whitened 39-d MFCCs were used as an input, while log-filterbank coefficients of dimension 32×40 (one central 25-ms frame and its five left and right neighbor frames) fed the CNNs [16]. The FCNNs are comprised of five dense 256-unit layers with leaky-rectify non-linearities, trained for five training epochs with the categorical cross-entropy loss function (we consider the k-means assignments as hard decisions with a single assignment per data point), Adam [17] weight update rule, a $1e-3$ learning rate, and a batchsize of 200 samples.

The CNNs comprise 3 convolution layers followed by two dense layers. The convolution filters chosen are of size 5×4, with a 1×2 max-pooling before the last convolution layer. The two dense layers for prediction have respectively 200 and 100 neurons and are followed by a dropout layer with a 0.5 probability. We used batch-normalization for all the layers.

Three feature representations were tested: 1) the activations of a given layer, 2) the posteriorgrams, 3) the distances to new cluster centroid estimates.

The best results were not obtained by the posteriorgrams but by the activations of the last hidden layer with the model types. Nevertheless, these representations did not improve the results obtained with k-means.

3 Experiments and results

3.1 Speech material

Participants were provided with raw audio speech files and also with low-level parameters already extracted by the challenge organizers, namely 13 MFCCs and their first and second derivatives.

Several language/audio recording duration pairs were to be considered with the aim to assess the approach robustness against language dependency, speaker identity, and file duration. File duration is relevant if one uses some kind of speaker adaptation, for instance.

For development purposes, three large open-source datasets were released in French, English and Mandarin, together with manual or automatically aligned phonetic transcriptions at triphone level. Evaluation is performed on these three languages and also on two additional "surprise" datasets comprised of speech in two languages unknown to the participants: LANG1 and LANG2.

For each language, the datasets include a training set and a test set. The speakers of the test sets are different from those of the training sets. The participants are free to use the training subsets or not.

To set up the models and adjust the hyperparameter values, we only used test subsets, since no performance gain was observed by also considering the training subsets. Table 1 shows the statistics

Table 1: Corpus description.

Language	English	French	Mandarin
Duration (hours)	27	17	25
Language	LANG1	LANG2	
Duration (hours)	25	10	

of these languages. For the three "development" languages, we separately used 27 hours in English, 17 hours in French and 25 hours in Mandarin. The two "surprise" languages LANG1 and LANG2 comprise 25 hours and 10 hours of speech, respectively.

3.2 Evaluation metric: the ABX error rate

The feature representations generated are evaluated using the ABX error rate [18, 19]. The ABX-discriminability of category x from category y is defined as the probability that samples A and X are further apart than a sample B with X (A and X are from category x and B is from category y). This is based on a distance d over the (model-dependent) space of feature representations for these sounds. Given two sets of feature representations that we wish to evaluate, $S(x)$ and $S(y)$ from category x and y, respectively, the metric estimates this probability using the following formula:

$$\frac{1}{m(m-1)n} \sum_{A \in S(x)} \sum_{B \in S(y)} \sum_{X \in S(x) \setminus \{A\}} (\mathbb{1}_{d(A,X) > d(B,X)} + \frac{1}{2} \mathbb{1}_{d(A,X) == d(B,X)})$$

with $m = \#S(x)$ and $n = \#S(y)$.

Two ABX error rates are computed for each language/duration pair: one calculated for a same speaker (the *within* condition) and one between different speakers (the *across* condition). The smaller the ABX, the better the feature representations.

3.3 Baseline and topline

The baseline feature representations, provided by the challenge organizers, consist of the 39-d MFCCs extracted every 10 ms on 25-ms frames. For instance, for the 10-s duration condition, the baseline ABX rates for the three development languages for the *within*/*across* conditions were 12.1%/23.4% for English, 12.6%/25.5% for French and 11.5%/21.3% for Mandarin. Results for the two evaluation languages were 9.3%/23.2% (lowest error) and 14.3%/29.5% (highest error).

The topline results, achieved by the organizers with supervised phonetic automatic decoders or manual annotations, show much lower ABX rates, such as 9.1%/6.8% for French/10-s duration. All these results are given in Table 3.

3.4 Preliminary experiments

In this section, we report experiments and results for French, with the set of files of 10-s duration. ABX scores of a selection of experiments are given in Table 2.

As we saw earlier, ZCA makes it possible to whiten the data while retaining the data point orientation. It can therefore be applied globally, or separately for each speaker (file by file). Indeed, whatever the duration condition considered, a given file contains speech from a single speaker only. The second strategy gave us better results than the first one. The decrease in ABX with the per-file ZCA compared to the global ZCA is much greater for the *across* condition (-4.7%) than for the

Table 2: Selection of results (ABX error rate) for our approaches on French / 10 seconds. Our submission to the challenge is indicated by (∗).

Approach	within	across
① MFCCs, baseline	12.6	25.5
② ZCA (*per-file* basis)	10.2	19.8
③ k-means	9.7	17.6
+ Selection for centroid re-estimation (∗)	9.5	17.7
+ Small centroid removal	**9.3**	**17.3**
④ DNN, output layer (posteriorgrams)	17.3	27.0
DNN, centroid re-estimation	9.4	17.6
DNN, last hidden layer	10.4	20.1
CNN, output layer (posteriorgrams)	12.2	21.2
CNN, last hidden layer	9.8	19.1

Table 3: Results of baseline, topline and our submitted system, for *across* and *within* conditions.

language	English			French			Mandarin			LANG1			LANG2		
file duration (s)	1	10	120	1	10	120	1	10	120	1	10	120	1	10	120
Across															
baseline	23.4	23.4	23.4	25.2	25.5	25.2	21.3	21.3	21.3	23.6	23.2	23.0	30.0	29.5	29.5
topline	8.6	6.9	6.7	10.6	9.1	8.9	12.0	5.7	5.1	12.8	10.5	10.4	7.1	3.6	4.3
ZCA + k-means	17.6	16.2	16.3	20.1	17.7	17.3	14.7	13.5	13.4	19.2	16.3	16.0	23.3	23.3	23.1
Within															
baseline	12.0	12.1	12.1	12.5	12.6	12.6	11.5	11.5	11.5	10.3	9.3	9.4	14.1	14.3	14.1
topline	6.5	5.3	5.1	8.0	6.8	6.8	9.5	4.2	4.0	8.7	7.1	7.0	6.6	4.6	3.4
ZCA + k-means	9.8	8.1	8.2	11.6	9.5	9.3	10.9	8.4	8.1	8.8	6.6	6.3	13.1	11.7	11.7

within one (-0.5%). Applying the per-file ZCA thus is less speaker-dependent than the global ZCA. ZCA outperforms the baseline non-whitened MFCC features: -5.7% *across* and -2.4% *within*.

Then, we experimented clustering with k-means on the ZCA-whitened features (number ③ in Figure 1 and Table 2). Several million points per language need to be clustered. To do so, we used FAISS[2], an open-source library for efficient similarity search and clustering of dense vectors [20]. On an Intel Xeon CPU E5-2623 with 16 processors, clustering 3 Million points with 100 means takes less than 10 minutes.

Applying k-means directly on the baseline features (13 MFCCs + $\Delta + \Delta\Delta$) yields poor results (18% ABX compared to the 12% baseline). Applying ZCA was crucial before applying k-means. With $k = 100$ means after ZCA allowed us to further decrease the error rate by 0.5% for *within* and 2.2% for *across*, compared to ZCA whitened MFCCs.

The ϵ hyperparameter for ZCA has a strong impact on the results. Its optimal value differs according to the condition (language, *across* or *within* scores), as we can see in Figure 4. By comparing the results on the three development languages for *across* and *within*, we opted for an ϵ value of 0.01 as a compromise value since it is not the optimal value for all the languages and duration conditions. On average, there are about 2/3 of the eigenvalues that fall below this threshold.

The number of clusters is another hyperparameter that we set to $k = 100$ empirically after

[2] `https://github.com/facebookresearch/faiss`

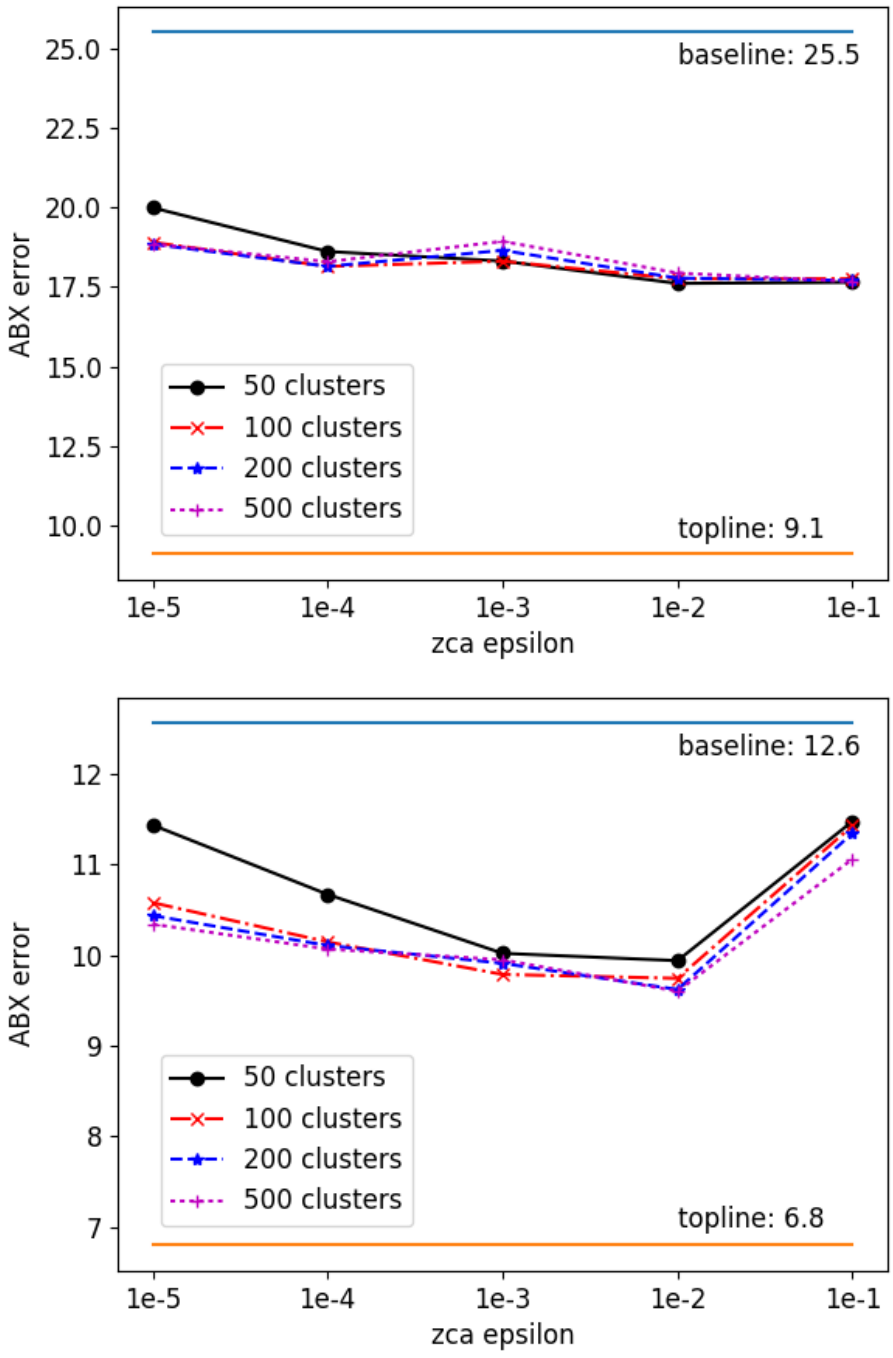

Figure 4: Influence of the ϵ ZCA hyperparameter on the results obtained with the distance-to-centroid representations with k-means for French/10s for *across* (top) and *within* (bottom). $\epsilon = 0.01$ is the best value in this case.

testing several values: 50, 100, 200 and 500 clusters as shown in Figure 4. Again, this number of pseudo-classes is not optimal for all the combinations language/duration from the development sets.

We considered the cluster occupation rates. The clusters are supposed to correspond to pseudo-phones, so an assigned cluster is expected to be stable for 40 to 100 ms of speech signal, which correspond to 4 to 10 consecutive frames with a 10 ms hop size. We assume that a given frame is more representative of its assigned cluster if its left and right nearest neighbor frames share the same cluster label. We have chosen to recalculate the centroids by retaining the frames that share their labels with at least its closest left and right neighbor frames. The effectiveness of this data selection did not revealed statistically significant. On the French corpus, this allowed us to gain 0.2% *within* but degraded the *across* score by 0.1%.

Then, by studying the distribution of the samples among the different pseudo-classes, we observed that after sample selection one of the clusters contained very few examples (less than 100 samples out of several Million points). Deleting this pseudo-class brought 0.2% and 0.4% absolute improvements in *within* and *across*, respectively. Nevertheless, this was not observed in English nor Mandarin so we decided not to use this filtering in our submission to the challenge.

We then tried to improve these results by further adding a neural network as described in Section 2 and depicted by a ④ symbol in Figure 1 and Table 2. As shown in this table, no significant gain was achieved no matter which the feature representation we tried: either posteriorgrams, activations of the last hidden layers or centroid re-estimation. Furthermore, these findings are similar when using DNNs or CNNs. In some cases (in English and Mandarin), though, small improvements were observed in the *within* condition but together with a score degradation in the *across* condition. This supports the claim that the pseudo-classes obtained with k-means capture information related not only to the underlying phonemes but also to the speakers. This would explain that these classes do not generalize well regarding the speaker identity. It is interesting to note that the predictions made by a DNN and the k-means hard assignments that were used to train the DNNs differ by less than 5% only. We observed that the DNN predictions seem to be more stable than the k-means assignments (less inconsistent variations in the pseudo-label assignation).

3.5 Submissions to the challenge

We made two submissions to the challenge official evaluation on the LANG1 and LANG2 surprise set. We only report the results obtained with the second system, the best one, denoted by $(*)$ in Table 2. These are given in Table 3 together with the baseline and topline. We outperform the baseline for the 5 languages tested: -4%, -3.1%, -3.1% for LANG2/1s-10s-120s *within*, respectively. Our representations are therefore robust to the target language. We also note that our results are strongly correlated to the results obtained by the baseline MFCCs: in a test case where the baseline is higher in ABX error rate than in another case, our system also provides results in the same order. For instance, in LANG1 and LANG2 10-s, the baseline is better for LANG1 with 9.3% than for LANG2 with a 14.3% value in *within*. Our system is also better for LANG1 (6.6%) than for LANG2 (11.7%). Our best results were obtained for Mandarin, for which we get the best results in proportion to the others: we have the second best *within* score among the participants.

Finally, the *across* condition is the weak point of our approach. The distance between the sub-lexical units is indeed much smaller for a same speaker than for two different speakers. They are therefore not that robust to speaker identity.

Figure 5 further illustrates our results obtained by our best submitted system (light blue). This figure allows us to visualize where we stand in relation to the baseline (dark blue), the topline (yellow) and the best scores published on the challenge results' webpage (green)[3]. This first ranked

[3] `http://sapience.dec.ens.fr/bootphon/2017/page_5.html`

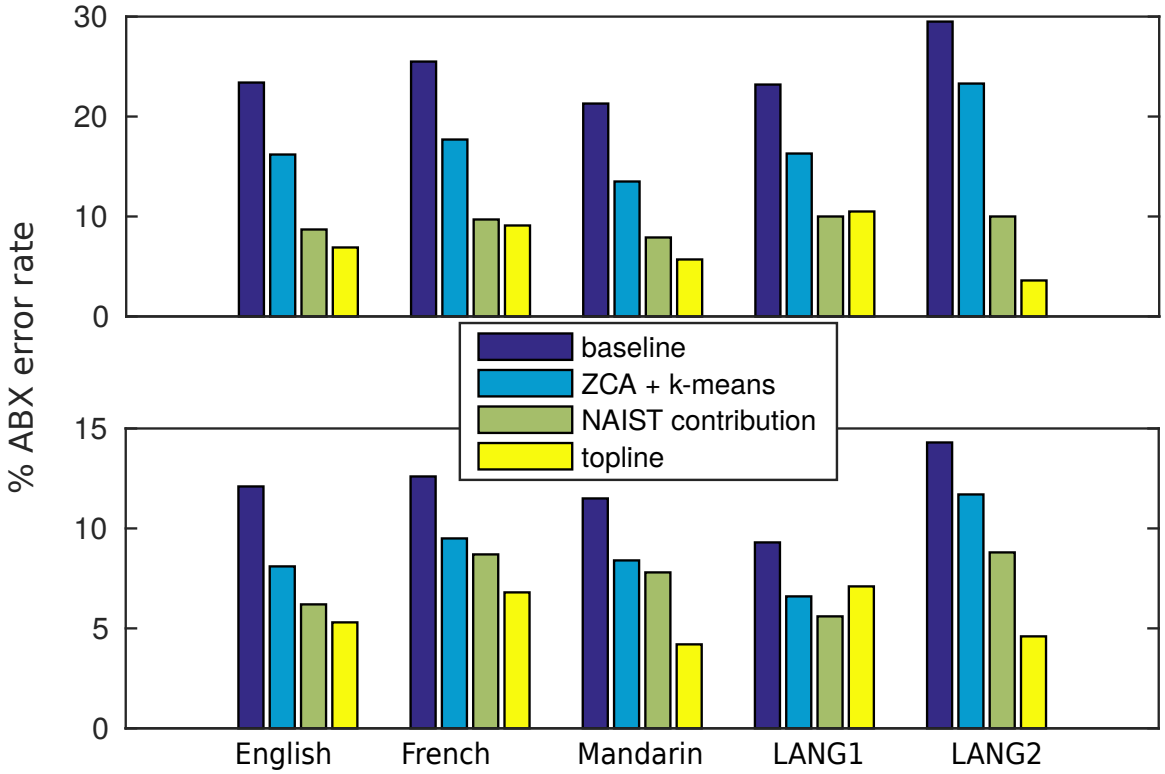

Figure 5: Results for French/10s: baseline, topline, our best submission and the best contribution of the challenge, for across (top) and within (down).

contribution is described by their authors as based on DPGMMs. We obtained competitive results in the *within* condition and even beat the topline for LANG1. The *across* results are relatively further apart from the best contribution. The code associated to our submissions is available on the ZeroSpeech2017 Zenodo webpage[4].

4 Conclusions

In this report, we described the IRIT-UPS approach for the unsupervised discovery of sub-lexical units in speech in the framework of the *Zero Resource Speech Challenge* 2017 edition. We proposed to use the distances between the data points (MFCCs and their first and second derivatives) and a hundred cluster centroids estimated on millions of these points.

We show that using a whitening transformation (ZCA) to pre-process the MFCCs is crucial to outperform the baseline MFCC representation, with ABX error rate reductions of 2.4% and 5.7% for the within-speaker and across-speaker conditions, respectively, in the case of the French language and the 10-s duration autio files. Furthermore, we obtained slight but consistent improvements by performing data selection before estimating the centroids (further decrease of about 0.5% and 2.2% for the same test case. This simple approach yielded competitive results for the within-speaker condition but generalized less well in the across-speaker one, probably due to the fact that the k-means clusters capture some information about the speaker identity, which prevents from yielding good generalization capabilities regarding the acoustic sub-lexical units.

In this work, we considered each language / duration test case as separate cases. As future work, we plan to conduct multilingual experiments, in which we use the development data from several languages. We could also mix the different duration test cases. Finally, the results with DNN-based posteriorgrams were deceiving. We plan to train the networks with a regression loss function on the the cluster assignment probabilities obtained with k-means, instead of the hard single pseudo-class assignments as we did so far. This is expected to use more information than a single pseudo-class per data point and also to prevent training on samples wrongly assigned by the k-means algorithm.

References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[2] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *INTERSPEECH*, pp. 3169–3173.

[3] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadyi, M. Bernard, L. Besacier, X. Anguerra, and E. Dupoux, "The zero resource speech challenge 2017," in *ASRU*.

[4] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *INTERSPEECH*, 2015, pp. 3199–3203.

[5] R. Thiollière, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *INTERSPEECH*, 2015, pp. 3179–3183.

[4] `https://zenodo.org/communities/zerospeech2017/`

[6] L. Badino, A. Mereta, and L. Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders," in *INTERSPEECH*, 2015, pp. 3174–3178.

[7] H. Chen, C. Leung, L. Xie, B. Ma, and H. Li, "Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling: a feasibility study," in *INTERSPEECH*, 2015, pp. 3189–3193.

[8] P. Baljekar, S. Sitaram, P. Kumar Muthukumar, and A. W. Black, "Using articulatory features and inferred phonological segments in zero resource speech processing," in *INTERSPEECH*, 2015, pp. 3194–3198.

[9] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *ICASSP*, 2014, pp. 7634–7638.

[10] V. Lyzinski, G. Sell, and A. Jansen, "An evaluation of graph clustering methods for unsupervised term discovery.," in *INTERSPEECH*. 2015, pp. 3209–3213, ISCA.

[11] H. Wang, T. Lee, and C.-C. Leung, "Unsupervised spoken term detection with acoustic segment model," in *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*. IEEE, 2011, pp. 106–111.

[12] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu, "Learning deep representation for graph clustering," *AAAI*, pp. 1293–1299, 2014.

[13] A. Coates and A. Y Ng, "Learning feature representations with k-means," in *Neural networks: Tricks of the trade*, pp. 561–580. Springer, 2012.

[14] P.-M. Bousquet, D. Matrouf, J.-F. Bonastre, et al., "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Interspeech*, 2011, pp. 485–488.

[15] I. S Dhillon and D. S Modha, "Concept decompositions for large sparse text data using clustering," *Machine learning*, vol. 42, no. 1, pp. 143–175, 2001.

[16] Thomas Pellegrini and Sandrine Mouysset, "Inferring phonemic classes from cnn activation maps using clustering techniques," 2016.

[17] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," in *INTERSPEECH*, 2013, pp. 1–5.

[19] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task (ii): Resistance to noise," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[20] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *arXiv:abs/1702.08734*, 2017.