INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Inferring phonemic classes from CNN activation maps using clustering techniques

Thomas Pellegrini, Sandrine Mouysset

[1]Université de Toulouse, UPS, IRIT, Toulouse, France
{thomas.pellegrini,sandrine.mouysset}@irit.fr

Abstract

Today's state-of-art in speech recognition involves deep neural networks (DNN). These last years, a certain research effort has been invested in characterizing the feature representations learned by DNNs. In this paper, we focus on convolutional neural networks (CNN) trained for phoneme recognition in French. We report clustering experiments performed on activation maps extracted from the different layers of a CNN comprised of two convolution and sub-sampling layers followed by three dense layers. Our goal was to get insights into phone separability and phonemic categories inferred by the network, and how they vary according to the successive layers. Two directions were explored with both linear and non-linear clustering techniques. First, we imposed a number of 33 classes equal to the number of context-independent phone models for French, in order to assess the phoneme separability power of the different layers. As expected, we observed that this power increases with the layer depth in the network: from 34% to 74% in F-measure from the first convolution to the last dense layers, when using spectral clustering. Second, optimal numbers of classes were automatically inferred through inter- and intra-cluster measure criteria. We analyze these classes in terms of standard French phonological features.
Index Terms: Convolutional Neural Network, phonemic categories, clustering

1. Introduction

Through advances in machine learning training algorithms, hardware computing capabilities, and the availability of very large data sets, deep neural networks (DNNs) have become the state-of-the-art technique in acoustic modeling [1, 2] and end-to-end large vocabulary automatic speech recognition (ASR) [3, 4]. To get insights on the feature representations learned by DNNs, several recent studies characterized how the high intrinsic variability in speech is reduced by the successive layers in deep networks [5, 6]. In [6], for example, 2-d representations of the same speech segments spoken by different speakers align better when using projections of unit activations of the deepest layers in a network. The 2-d visualizations were extracted using the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimension reduction technique that we also use in this study [7]. More recently, Nagamine *et al.* [8] studied how invariant phonemic categories are formed within a DNN trained for phone recognition, elucidating patterns of feature organization similar to the ones observed in the human auditory cortex. They found that some nodes are selective to broad phonetic features such as voicing, manner and place of articulation, by plotting average node responses to phonemes and by clustering them according to a phoneme selectivity index. In the present work, we report similar experiments but carried out on a convolutional neural network (CNN) rather than a DNN comprised of dense hidden layers only. An important difference between these types of models lies in the fact that DNNs ignore input topology (input feature order is of no importance for DNNs), whereas convolution filters try to model local correlations that do exist and that are strong in spectral representations of speech [9]. Hence, similarly to the findings of [8], we can expect activation maps of convolution layers to exhibit patterns specific to certain phonetic or phonological properties. To explore this direction, we carried out two sets of clustering experiments on activation maps of a CNN trained for phone recognition for French. First, we imposed a number of 33 classes equal to the number of context-independent phone models for French, in order to assess the phoneme separability power of the different layers. Second, an optimal number of classes was automatically inferred through inter- and intra-cluster measure criteria. We analyze the resulting classes in terms of standard French phonological features.

The paper is organized as follows. First, we describe the CNN model used in this work. Then, we give a brief overview of the two dimension reduction techniques: principal component analysis (PCA) and t-SNE, the linear and non-linear clustering approaches (k-means and spectral clustering) at the core of our study, and their evaluation metrics. We report and analyze the clustering experiments in Section 4 before concluding.

2. CNN model

The model used in this work comprises an input layer composed of 11 frames of 40 log filter bank static, delta, and delta-delta coefficients for each frame, extracted from a 20 ms frame to which have been added the five previous and following neighboring frames for a total of 11 consecutive frames. Two adjacent frames are separated by 10 ms. Two convolution layers with 3×5 filters, followed by respectively 1×3 and 1×2 downsampling (*max-pooling*) layers, produce respectively 32 and 64 activation maps that serve as input parameters for three 1024 unit dense hidden layers with a rectified linear unit (ReLu) activation function. Pooling is applied on frequency only and not time, as it was shown optimal for ASR [10]. Finally, the output dense layer comprises 33 units and uses a sigmoid activation function. The network weights were initialized using the "Xavier" method [11], and trained with gradient descent with Nesterov momentum, with a categorical cross-entropy cost function. The regularization *dropout* method ($p = 0.5$) was used with the dense hidden layers only. This model is not very deep but appears to be sufficient to get insights on its phonetic feature representation capabilities. To carry out our work, we used the

http://dx.doi.org/10.21437/Interspeech.2016-1299

Theano [12] and Lasagne toolkits[1].

To train the model, we used the BREF corpus, comprised of 100 hours of read speech collected from 120 native French speakers, who read texts from the French newspaper *Le Monde* [13]. We divided the corpus into a training and a development (*Val*) sub-corpora, in the 90%/10% proportions equivalent to 1M/150K examples, respectively. The development part was used to tune the learning hyper-parameters of the CNN (learning rate, stopping criterion). Training stopped if the cost computed on *Val* did not decrease more than 1e-3 for at least one over three consecutive epochs. The final model achieved a 26.7% phone error rate (PER) on *Val*. For information, a 20.1% PER on the TIMIT corpus was reported in [10] with a CNN trained for U.S. English phone recognition. We were not looking for the best performance possible but rather for a network which size allows the study of its parameters.

3. Methods

In order to characterize the CNN internal layers in terms of phone/phonetics modeling, we feed-forwarded 100 input samples per phone through the network. These samples were chosen within *Val* samples that were correctly classified by the model. Phones typically have a duration larger than 20 ms, hence, several successive samples separated by 10 ms (our hop size for feature extraction) share a same phone label. We systematically chose the third samples with the idea that these central samples are good representatives of the phones of interest.

The outputs of each layer (activations) are extracted and clustering is performed on them after reshaping them in 2-d matrices, if needed. For instance, the second max-pooling layer gives a tensor of dimension $3300 \times 64 \times 7 \times 4$, where 3300, 64 and 7×4 correspond respectively to the number of samples (100 per phone), the number of maps and the filter size of this layer. This tensor is reshaped into a 3300×1792 matrix before clustering.

3.1. Dimension reduction

Due to the fact that we attempt to identify phonemes and phonetic features with high-dimensional activation maps, embedding techniques may help in extracting pertinent piece of information on these data distributions. We consider two preprocessing steps: Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) [7].

PCA is processed on the whole activation maps. Since the contrast could vary from one layer to another, we set for each layer the number of principal components that keeps at least 90% of the covariance matrix spectrum.

The t-SNE method relies on random walks on neighborhood graphs to extract the local structure of the data and also reveal important global structure. Based on the pair-wise similarities in both the original data space and in the embedding space defined by conditional probabilities, the divergence between the two distributions is then minimized. For the initialization of the algorithm, a PCA is first performed to reduce the dimensionality of the initial data. We used the same numbers of principal components as the experiment with PCA only.

3.2. Clustering techniques

To study the impact of CNN for determining single phoneme or clusters of phonemes, we consider the two most popular clustering techniques based on either linear separation or non-linear separation: K-means and spectral clustering. K-means is computed with Manhattan distances [14] to take into account the main directions of the data distribution. As the dataset contains 100 samples per phoneme, the initialization of the centroïds is based on preliminary clustering 10% of the data.

Spectral Clustering (SC), which has a theoretical connection with weighted kernel k-means and normalized graph cut [15], aims at selecting dominant eigenvectors of a parametrized Gaussian affinity matrix in order to build an embedding space in which the clustering is made [16]. This method can detect arbitrarily shaped clusters with an appropriate choice of the Gaussian kernel similarity function. The inherent parameter of the Gaussian affinity measure which role is to threshold the affinity measure is based on both the distribution and the dimension of the dataset [17].

In Section 4.1, we report clustering experiments with a fixed number of clusters equal to 33, the number of distinct phones we use for French. In Section 4.2, we derived optimal numbers of clusters automatically, in order to analyze which phones the network groups together in terms of standard phonetic characteristics. To do so, two different clustering quality assessment criteria for K-means and SC have been used. These criteria rely on the same principle: the comparison between intra- and inter-cluster affinity. A good quality clustering is performed, in terms of affinity, when the affinity values between and within the clusters are low and high, respectively [18]. In terms of distances, the distance between the data points that belong to different clusters must be larger than the distance between points within a same cluster.

For K-means, the within- and between-cluster sums (WCS and BCS) of point-to-centroïd distances are computed for each possible number of clusters, up to 33 clusters. Mean ratios between WCS and BCS are computed and the optimal number of clusters corresponds to the one that gives the smallest ratio.

For SC, the Gaussian affinity matrix is exploited after indexing the data points with their assigned cluster label [19]. Thus, the off-diagonal blocks will represent the affinity between clusters and the diagonal ones the affinity within clusters. The mean ratios between the Frobenius norm of the off-diagonal blocks and that of the diagonal ones is then computed for each candidate number of clusters. The optimal number of clusters is again defined so as to minimize this ratio.

3.3. Evaluation

To evaluate the resulting clusters with a fixed number of 33 clusters, we use Precision, Recall and F-measure. Precision, denoted P, is the fraction of retrieved phonemes in the clusters. Recall, denoted R, is the fraction of the phonemes that are relevant in the clusters and that are successfully retrieved. F-Measure, denoted F, combines precision and recall as the harmonic mean of precision and recall. These measures are summarized by the following equations:

$$P = \frac{tp}{tp + fp},\ R = \frac{tp}{tp + fn},\ F = 2\frac{P.R}{P + R}$$

where tp, fp and fn respectively represent the number of true positives, false positives and false negatives. To compute these metrics, the phoneme assigned to a given cluster is the one that is the most represented in that cluster. The best possible cluster for a phone would be a cluster regrouping all the 100 samples of that phones, with no samples of any other phone.

[1] https://github.com/Lasagne/Lasagne

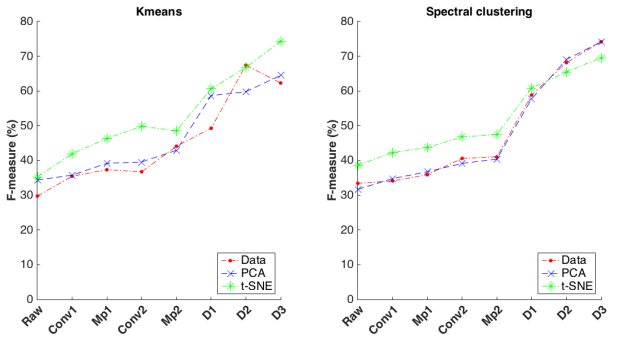

Figure 1: Clustering performance in F-measure with K-means (left) and SC (right), both without dimension reduction, with PCA and with t-SNE. Here, a number of 33 clusters equal to the number of target phones was imposed (*cf.* Section 4.1).

4. Results

4.1. Phone-specific clusters become more explicit with layer depth

Figure 1 shows clustering F-measure values obtained with K-means (left hand-side) or with SC (right-hand side), both without dimension reduction, with PCA or t-SNE. Each curve is made of 7 points: the leftmost point denoted "Raw" corresponds to the F-measure obtained when clustering the raw acoustic input features, and from left to right the clustering results with the activation maps outputted from the successive network layers: 'conv1', 'mp1', 'conv2', 'mp2' for the convolution and max-pooling layers, followed by 'D1', 'D2' and 'D3' the three dense hidden layers. We did not perform clustering on the output layer, which provides the phone probabilities for the 33 units, since we were interested in characterizing the inner layers' behavior. Clustering with the raw input features gives F-measure values between 0.3 and 0.4, with the best one obtained with t-SNE and SC. As expected, clustering the activations of the subsequent layers progressively improves the results. It can happen that multiple clusters are assigned to a same phone, it happened indeed with the shallower layers but not with the deepest ones. For both K-means and SC, t-SNE (green curves) outperforms significantly the two other approaches (no dimension reduction and PCA) when using the convolution/sub-sampling layer activations as input. It suggests that t-SNE indeed captures the global structure of the complex speech signal more efficiently than PCA does. Then, with the dense layer activations used as input (D1 to D3), the advantage of using t-SNE vanishes and similar performance is obtained in all the three conditions. Furthermore, K-means and SC achieve similar performance values. This indicates that a linear clustering algorithm such as K-means is as efficient as a more sophisticated non-linear approach such as SC. We can eventually relate these results of a performance increase when going deeper into the network to the idea that successive layers reduce data variability and progressively linearize the complex latent structure of speech [20]. This idea that DNNs, and CNNs in particular, are markedly successful thank to their structural linearizing capabilities has been illustrated with nice images in [21], in which simple arithmetic operations such as addition, subtraction and interpolation, are applied on the linearized feature representations of images learned by a network.

4.2. Broad phonetic classes are learned by the network

Our phone set for standard French is comprised of 33 context-independent units, modeling 17 consonants, 3 semi-vowels, 10 oral and 3 nasal vowels. There are three more vowels in French (two oral and one nasal) that we did not consider because of their lower number of occurrences in the speech corpus. In [8], the authors observed that some neuron units activate for phones that share phonetic features, for instance, plosives. In the present study, our findings with convolution activation maps are similar. In this paper, we use symbols from the SAMPA French Phonetic Alphabet.

By applying automatic Spectral Clustering, an optimal number of clusters of 7 was identified in order to minimize the ratio of inter- and intra-cluster affinity. Distinct single clusters regroup 82.7% of the closed vowels /y/, /i/, /e/, also grouped with the semi-vowel /H/, 81.3% of the /S/, /s/, /f/, /Z/ fricatives, 93% of medium to open vowels /a/, /E/, /9/, 92% of the nasal consonants /n/, /m/ and /J/, 60% of the nasal vowels /a~/, /o~/, /U~/, 68% of the plosives consonants /p/, /t/, /k/, /b/, /d/, /g/ and 76% with the rounded vowels /o/,/u/, /O/ and the /w/ semi-vowel. These clusters are completely coherent in terms of the phonological characteristics typical from the French sound system. But some specific phoneme like /R/ or /l/ are not well detected and are spread across several clusters. Very similar clusters were obtained with K-means but due to the fact that the optimized number of clusters is 17 for K-means, some phonological clusters are subdivided. For example, the plosive consonants /p/, /t/, /k/, /b/, /d/, /g/ are subdivided in three clusters: one with 71.3% of /b/,/d/, /g/, another one with 87% of the phoneme /t/ and a last one with 92% of /p/ and /k/.

In the remaining of the paper, we further illustrate these results. Figure 2 is a 3-d PCA projection of the first convolution layer activation maps obtained for the 33 phones. The maps correspond to averaged activation maps obtained when feed-forwarding 100 input samples per phone through the first layer. One can see clusters of similar phones, which have a phonetic interpretation. On the right-hand side, a cluster regroups the /p/, /t/ and /k/ voiceless plosives, close to another cluster with the /b/, /d/, /g/ voiced counterparts. Similarly, the /f/, /s/ and /S/ voiceless and /v/, /z/, /Z/ voiced fricatives also form two clusters. In the central part of the figure lies an axis with the three semi-vowels /H/, /J/ and /w/, and, behind, almost all the vowels are grouped together. The /m/ and /n/ nasal consonants are closed to each other, so as the /a~/ and /o~/ nasal vowels. We made available online interactive versions of this figure[2] and also of the figure obtained with t-SNE[3].

Figure 3 illustrates an example activation map for the three non-rounded front vowels /i/, /e/, /E/, the three rounded back vowels /u/, /o/, /O/, and the /a/ back vowel. They correspond to one map among the 32 maps of the 'conv1' layer, and they are the maps averaged on the 100 samples per phone. The seven maps are represented within the F1-F2 formant plane, with F1 being the vertical axis, and F2 the horizontal one. The map positions in the plane are just indicative, based on standard formant values for French vowels [22]. One can see a pattern in this figure according to the F1 axis: the activation maps of two vowels with close F1 values are very similar, and the strong activation values depicted in red become more central as F1 increases. It indicates that this map is sensitive to F1 variations,

[2] https://www.irit.fr/~Thomas.Pellegrini/research/pca3d.html

[3] https://www.irit.fr/~Thomas.Pellegrini/research/tSNE3d.html

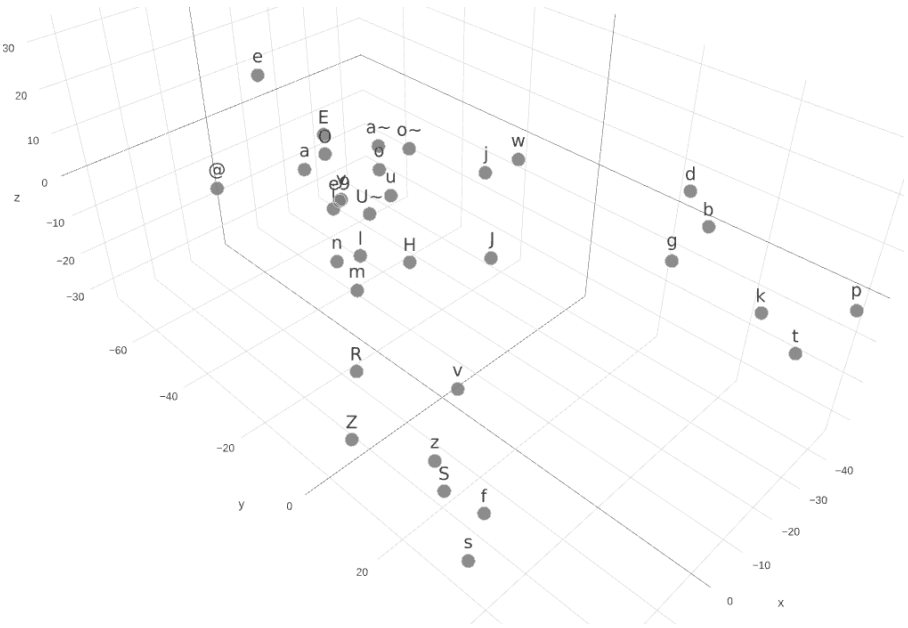

Figure 2: 3-d PCA projections of the first convolution layer activations averaged over 100 examples per phone. An interactive figure is available online.

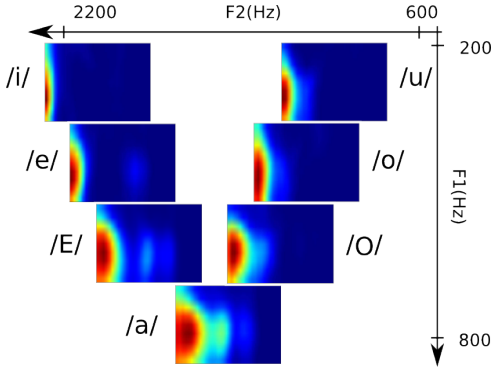

Figure 3: Example activation map of layer 'conv1' for closed to open vowels. F1 and F2 axes stand for first and second formant values. Vowel positions are indicative.

F1 being related to mouth aperture in the literature, and relatively insensitive to the F2 axis, usually associated to vowel anteriorityposteriority, which qualifies the place of articulation of the oral vowels. Figure 4 shows an example mean map of the first convolution layer for the voiced and voiceless plosives, which are very similar to each other and very different from the vowels activation maps of Figure 3.

5. Conclusions

In this paper, we reported our first attempts in elucidating the capacity of a CNN model trained for phoneme recognition in French to learn broad phonemic and phone specific features. For this purpose, clustering experiments were performed on activation maps extracted from the different layers of a CNN comprised of two convolution and sub-sampling layers followed by three dense layers. Two directions were explored with both linear and non-linear clustering techniques. First, we imposed a number of 33 classes equal to the number of

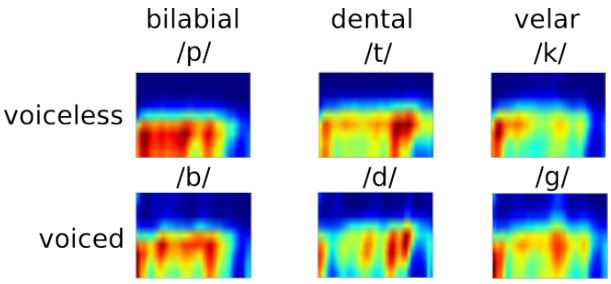

Figure 4: Example activation map of layer 'conv1' for plosives.

context-independent phone models for French, in order to assess the phoneme separability power of the different layers. As expected, we observed that this power increases with the layer depth in the network: from 34% to 74% in F-measure from the first convolution to the last dense layers, when using t-SNE as a dimension reduction technique followed by spectral clustering. Second, optimal numbers of classes were automatically inferred through inter- and intra-cluster measure criteria. Standard French phonological features such as place of articulation of French oral vowels was illustrated by representing activation maps as heat maps on a F1-F2 formant plane. We used 100 samples for each phone of interest and these samples were extracted 20 ms after the first occurring of the phones. We plan to repeat our experiments with more samples and we will try to use samples averaged on the whole speech segments labeled with a same phone label. We expect to increase the clustering performance by this way. Another question that we would like to tackle would be to characterize the phonetic modeling capabilities of the convolution layers at filter-level by considering their weights instead of the activation maps.

6. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

7. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1764–1772.

[4] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[5] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[6] A.-r. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4273–4276.

[7] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[8] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[10] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.

[11] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10)*, 2010.

[12] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. of the Python for Scientific Computing Conference (SciPy)*, 2010.

[13] J.-L. Gauvain, L. Lamel, and M. Eskenazi, "Design considerations and text selection for BREF, a large french read-speech corpus," in *Proc. ICSLP-90*, 1990, pp. 1097–2000.

[14] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[15] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 551–556.

[16] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[17] S. Mouysset, J. Noailles, and D. Ruiz, "Using a global parameter for gaussian affinity matrices in spectral clustering," in *High Performance Computing for Computational Science-VECPAR 2008*. Springer, 2008, pp. 378–390.

[18] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *Journal of the ACM (JACM)*, vol. 51, no. 3, pp. 497–515, 2004.

[19] S. Mouysset, J. Noailles, D. Ruiz, and R. Guivarch, "On a strategy for spectral clustering with parallel computation," in *High Performance Computing for Computational Science–VECPAR 2010*. Springer, 2010, pp. 408–420.

[20] S. Mallat, "Understanding deep convolutional networks," *arXiv preprint arXiv:1601.04920*, 2016.

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[22] C. Gendrot and M. Adda-Decker, "Impact of duration on f1/f2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in french and german." *Variations*, vol. 2, no. 22.5, pp. 2–4, 2005.