Two-level clustering towards unsupervised discovery of acoustic classes

Ciro Gracia[1], Xavier Anguera[2], Xavier Binefa[1]
[1]Department of Information and Communications Technologies
Universitat Pompeu Fabra, Barcelona, Spain
[2]Telefonica Research, Edificio Telefonica-Diagonal 00, Barcelona, Spain
{ciro.gracia, xavier.binefa}@upf.edu, xanguera@tid.es

***Abstract*—In this paper we focus on unsupervised discovering of acoustic classes suitable for use in pattern recognition applications. Our approach is based on a two-level clustering of an initial acoustic segmentation of the audio data in order to allow for discovery and correct modeling of complex acoustic classes. Initially, in a first-level, the acoustic space is densely clustered in order to provide a first layer of acoustic variance reduction. In a second-level clustering we use the acoustic segmentation to infer a smaller number of super-clusters taking advantage of the intra-segment relationships between the first-level clusters. In this paper we compare three possible clustering methods to obtain super-clusters as sub-sets or linear combinations of first-level clusters. Results indicate that the proposed two-level approach improves the balance between Purity and inverse Purity evaluation measures while significantly improving the stability of the transcriptions obtained when using the resulting models to transcribe the same acoustic events in different spoken utterances.**

Index Terms: zero resources, clustering, query by example.

I. Introduction

In this paper we deal with the problem of automatic extraction of distinct acoustic classes from recorded audio. While supervised approaches solve the problem by using large transcribed corpora from which they learn both the language structure and its acoustics, unfortunately, the availability of such transcribed speech corpora is not universal, and their production is slow, costly and usually requires a deep knowledge of the structure of the language. For this reason, unsupervised approaches do not build classical phoneme based acoustic models but usually derive their own set of acoustic classes, that usually have a possible correspondence with the phonetic structure of the language. With the increasing popularity of unsupervised approaches to Keyword Spotting and Query-by-Example tasks [2], [6], [13], [10], the problem of how to automatically extract acoustic classes from raw speech data has grown in interest. Still, the question remains in how to automatically build a set of acoustic classes that can be most successfully used in such pattern-based speech recognition tasks.

In previous work, [4] defined a set of acoustic classes by using a segmental speech model. This approach first defines the acoustic data as a set of non overlapping segments obtained by an unsupervised segmentation algorithm that are then clustered into a predefined number of classes, used later to represent the data. To model the classes they use polynomial trajectories, together with a probabilistic model named Segmental Gaussian mixture model (SGMM). Its main drawbacks are the lack of a optimal decoding scheme as on novel data the system first requires an unsupervised segmentation of the acoustic data prior to labeling. In consequence, the complete decoding scheme propagates the segmentation algorithm errors.

Other methods like [3] use a similar approach composed on an unsupervised segmentation followed by a clustering step that uses Hidden Markov Models (HMM) as models. This approach allows for the use of an efficient HMM decoding scheme, avoiding unsupervised segmentation errors to propagate through the decoding step.

A slightly different approach [12] takes advantage of a Gaussian mixture tokenizer to cluster the segments. An initial segmentation provides a data breakdown into segments. Then instead of directly clustering the segments, a Gaussian mixture model(GMM) is trained by using the complete acoustic data and then each segment is labeled by using GMM as tokenizer. This avoids some of the problems coming from the segmental clustering step. As in many other approaches, the resulting decoded data in [12] is the basis for a posterior HMM modeling and embedded re-estimation.

Our objective is to find a set of acoustic classes to later be useful to transcribe acoustic data such that they provide us with a stable transcription for the same acoustic events in different spoken utterances. In order to evaluate the goodness of our models we measure how the resulting acoustic classes behave with respect to underlying phonemes by using the cluster purity and the inverse cluster purity measures. Our interest is to find a small set of classes that retain as much as language discrimination power between language acoustic patterns. In addition, we propose an alternative way to evaluate clustering by using the resulting transcriptions for a set of known utterances which contain the same acoustic events. For this we use Levenshtein distance (usually known as the Edit distance) in order to compare these transcriptions.

In this paper we present a two level clustering approach for building an set of acoustic classes in an unsupervised manner. We initially model the acoustic space with a first-level dense clustering in order to obtain simple classes while retaining acoustic resolution. In a subsequent step, we analyze relationships between initial classes in order to infer a smaller set of more complex, higher level classes. We investigate three different clustering approaches in order to obtain these higher

level classes as a subset or a linear combination of the first-level clusters. Resulting higher level acoustic classes improve the balance between both evaluation measures while obtaining significant lower transcription differences in all scenarios.

II. System Description

Our approach is based on a two-step clustering scheme. The objective of the first clustering step is to model the entire acoustic space. This task requires a big number of clusters to be able to obtain an accurate enough acoustic resolution with a high probability of not mixing different underlying phoneme sources. Despite of that, this first-level clustering alone is not suitable for the transcription of acoustic data. The resulting high number of classes in the transcription alphabet increases the variance of the transcriptions of acoustic data, where the same acoustic elements get modeled by slightly different clusters. For this reason we introduce next a second-level clustering in order to obtain a smaller number of final clusters, which we call super-clusters. Our objective is to reduce the variance in transcriptions obtained by using these clusters, while also maintaining a high purity and inverse purity values, all important for zero-resources query-by-example applications.

To illustrate the two-level clustering approach, in Figure 1 an example is given for two phoneme, 'ae' and 'ao', extracted from samples in the TIMIT database. The top-left figure in Figure 1 shows the two-dimensional projection of acoustic frames belonging to both phonemes where lines are drawn between temporally adjacent acoustic frames. The top-right figure in Figure 1 shows the results of the first-level clustering and, finally, the bottom figure in 1 shows a possible resulting second-level clustering where 3 clusters were defined. We can see that by using a two-level clustering strategy we can create more complex clusters that better represent the underlying acoustic data. Next we describe how we obtain each level of clusters.

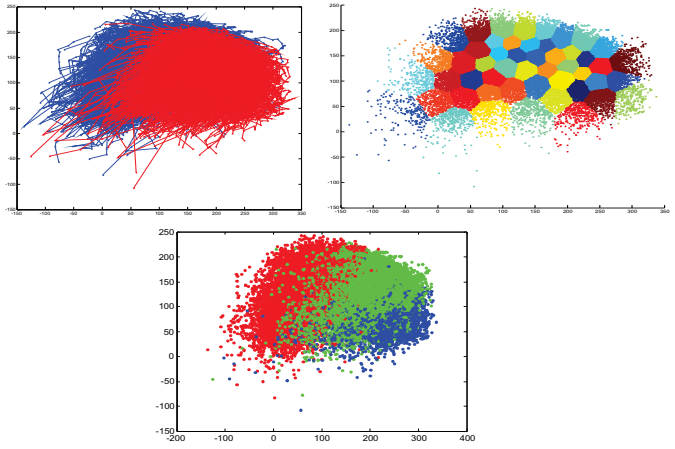

Fig. 1. Schematic view of the proposed approach for a toy ('ae','ao') phoneme scenario. From upper-left to bottom-right: acoustic segments, first-level clustering and finally the inferred super-clusters

A. First-Level Clustering

To obtain the first level clustering, we use a vector quantization (VQ) scheme in order to produce a compact representation of the acoustic space, applied to the pool of all acoustic frames available for training. This approach allows us to describe the acoustic space as a finite set of non overlapping classes. In applying VQ, we found the initialization of the VQ centroids to be a problem. As we wanted to compare different clustering parameters, variances produced by standard random initialization of VQ centroids must be avoided.

The approach we used to initialize the VQ centroids is based on a hierarchical division of the acoustic space. Initially, all the acoustic data is grouped into a single cluster which its VQ centroid set at the mean vector. Iteratively, we select and split the cluster with highest sum of square errors. The cluster splitting is performed by projecting the cluster points into the cluster principal axis. The resulting new clusters are formed as the points projected in positive and negative subspaces. We found it important to compute the VQ centroids from each cluster by using the medioid instead of mean function. This is because using the medioid function avoids the problem of empty clusters in posterior minimization. After initializing the clusters, we use a standard distortion minimization algorithm (k-means) until convergence [11].

B. Second Level Clustering

We propose a two level clustering algorithm as a tool for obtaining an accurate acoustic resolution and a low number of final clusters for data transcription. We can define super-clusters either as a hard set or as a soft mixture of first-level clusters. Initially we need to establish an heuristic in order to determine which clusters belong to the same underlying super-cluster. We do so by describing the acoustic data vectors x as a set of non overlapping segments and using the segments as prior. This assumes that the set of acoustic feature vectors inside each segment belongs to the same super-cluster. This segmental prior can be estimated by using an automatic acoustic segmentation algorithm, although in the current work, in order to independently evaluate the clustering approaches, we have used the ground truth phonetic transcription. Once the segmentation has been obtained, the resulting S_m segments are described by using the accumulated posterior probabilities to each of the C_n^1 first level clusters, as show in equation 1. The resulting matrix $DW \in R^{mxn}$, describes segments as pseudo occurrences between clusters and becomes the input for the second level clustering.

$$DW(S_j, C_k^1) = \sum^{x_i \in S_j} P(C_k^1 | x_i) \quad (1)$$

In this paper we explore three different approaches for second-level clustering, namely Hierarchical agglomerative clustering (AHC), Probabilistic Latent Semantic analysis (PLSA) and Non negative matrix factorization (NMF). The major differences between the approaches come both from the functional being optimized and from the representation of the super-clusters being used. Next we will describe how we apply each of these techniques for the creation of super-clusters.

1) Agglomerative Hierarchical Clustering: We use AHC to describe super-clusters as disjoint subsets of first-level clusters. By using matrix DW, AHC iteratively joins clusters which have a similar behavior with respect to the data until the desired

number of clusters is reached. In merging the closest clusters, AHC uses the Wards criterion [1] in order to obtain clusters by minimizing their data behavior variance.

2) Probabilistic Latent Semantic Analysis: Given that different acoustic units might share similar acoustic data (e.g. silence frames), we could assume that super-clusters should be a constructed as a mixture of simpler clusters. This leads us to explore approaches based on matrix factorization techniques, like Probabilistic Latent Semantic analysis and Non negative matrix factorization.

Probabilistic Latent Semantic analysis (PLSA) [8] is a method originally designed for text processing. In document clustering a set of documents D are composed from a set of vocabulary words W and the objective is to infer a set of topics T from the co-occurrences between words and documents. In fact, PLSA factorizes the frequency matrix $DW \in R^{mxn}$ into two matrices by means of the inferred topics $DW = DT*TW$. We use this approach for the creation of super-clusters by assuming that segments and clusters are equivalent to documents and words in the text scenario. Then, the objective is to infer a set of super-clusters as a mixture of simpler clusters by taking into account the relationship between clusters and segments.

3) Non-negative Matrix Factorization: Non-negative Matrix Factorization (NMF) [9] factorizes the relational matrix DW between segments and clusters by assuming that data can be decomposed into a sum of additive components. Given a non negative matrix DW^{mxn} and a positive integer k, the objective is to find two non negative matrixes $W \in R^{mxk}$, $H \in R^{kxn}$ such that they minimize the functional given by equation 2.

$$f(W,H) = \frac{1}{2}||DW - WH||^2 \tag{2}$$

Note that the main differences between PLSA and NMF come from the function being minimized. In fact, PLSA has been recognized as equivalent to an NMF model when a Kullback-Leibler divergence is used as cost function [5].

III. Experiments

We perform experiments using TIMIT corpus. In particular, we use the SA1 set from TIMIT TRAIN corpus to perform the experiments. This set consists on 420 utterances with the same text uttered by 420 different speakers. For each utterance we computed spectral features using a pre-emphasis filter (factor was set to 0.97) and a 20 millisecond Hanning window with a 5 millisecond shift. The Mel Filtered spectrogram is generated by a reference software [4] using 50 triangular Mel scale filters between 1Hz and 8000Hz. Instead of obtaining standard MFCC parameters through DCT, we decorrelate the feature vectors using PCA.

A. Evaluation measures

In order to evaluate clustering results we use the ground truth phonetic transcription as underlying segmentation of the data (we plan on using an automatic segmentation in the future). We first analyze how phoneme labels are mapped into clusters. To do this we use the cluster purity as defined in equation 3. Cluster purity evaluates how the resulting K discovered acoustic classes match with respect to the underlying phoneme classes. For each of the resulting acoustic classes $c_i \in C$, we compute the cluster purity as the weighted average of the maximum value among all phoneme probabilities ($ph_i \in PH$, being PH the phone set) in each cluster. Additionally, we also use the inverse cluster purity measure shown in equation 4 in order to evaluate the dispersion of the phoneme data into resulting classes.

$$Purity = \sum_{i=1}^{K} max(P(ph|c_i)) * P(c_i); \tag{3}$$

$$InvPurity = \sum_{i=1}^{PH} max(P(c|ph_i)) * P(ph_i); \tag{4}$$

In addition, we define a totally unsupervised evaluation measure based on the Levenshtein distance between data transcriptions. After mapping the acoustic data segments to the most likely class and generating the utterance transcription strings we align each possible pair of transcriptions using a dynamic programming algorithm and obtain their path length normalized Levenshtein distance [7]. We use the average distance among all possible utterance pairs as a measure of how well can our clustering model unseen data. Ideally, the Levenshtein distance between several instances of the same acoustic events should result in a low average value.

TABLE I. Results for 300 VQ and 47 super-cluster

Method	Purity	Inv. Purity	Levenshtein distance
VQ clustering	64.92%	10.35%	92.75%
Two-level AHC	50.88%	41.96%	63.79%
Two-level PLSA	49.11%	40.73%	68.03%
Two-level NMF	51.74%	38.1%	70.75%
Ground truth	100%	100%	21.9%

B. Result and discussion

Figure 2 shows results for the different proposed approaches w.r.t. cluster purity and inverse cluster purity as a function of the number of clusters used in the first-level clustering. As expected, increasing the number of first-level clusters also increases the cluster purity, but it produces a decreasing inverse cluster purity and also a higher average transcription distance. The reason is that by increasing the number of clusters we promote smaller clusters that relate to a very localized acoustic source, which does not map properly with the underlying phoneme classes.

For this reason, in this paper we introduce a second-level clustering in order to reduce the number of clusters. Figures 3, 2 and table I show the purity , inverse purity and average transcription distance when setting the number of super-clusters to 47 (similar to the number of phoneme classes expected in a standard language). In addition, the ground truth row exposes that despite utterances contain same text, the resulting pronunciations vary and in consequence their phonetic transcriptions have some differences. We can observe that despite loosing a 15% in Purity, we have a significant

[4]Dan Ellis. Lab Rosa Matlab Audio Processing Examples

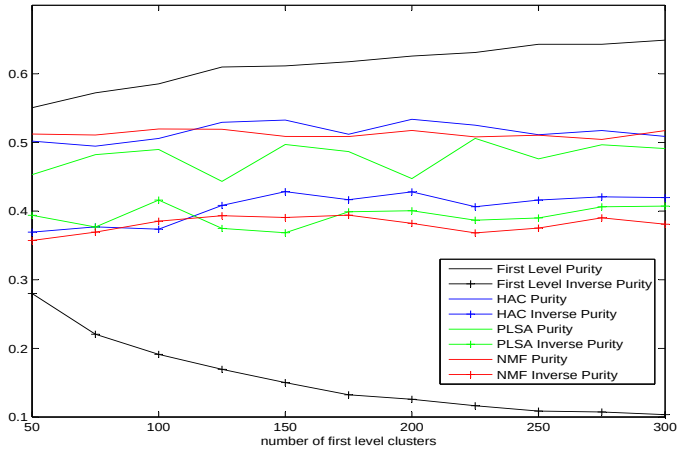

Fig. 2. Purity and inverse Purity evaluation for the different approaches

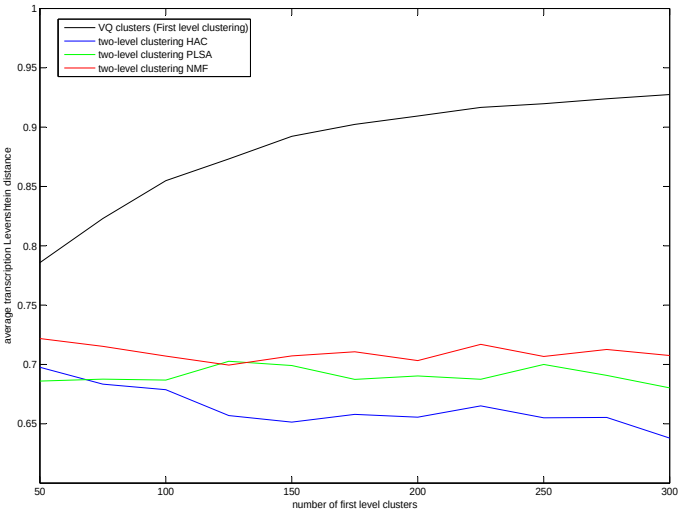

Fig. 3. Average edit distance between utterances for the different approaches

increase in Inverse Purity and lower average distance between transcriptions while decreasing the number of classes by a factor of 6. Despite that the three second-level approaches obtain similar Purity and inverse Purity measure, AHC outperform both NMF and PLSA. We believe that this is caused by their sensitivity to the cluster initialization, we leave for future work the task of defining a more deterministic initialization.

IV. Conclusions

In this paper we proposed a novel approach for the automatic discovery of acoustic classes from recoded data. Our approach focuses on obtaining classes that will later be useful for pattern recognition applications, therefore obtaining similar transcriptions to similar uttered patterns. Our approach presents a way to improve classical clustering by taking into account the segmentation of the acoustic data and introducing a second clustering level to define more complex clusters. The obtained results indicate that our approach provides lower differences between different instances of the same spoken utterances, while retaining clusters quality evaluations.

References

[1] V. Batagelj. Generalized ward and related clustering problems. *Classification and related methods of data analysis*, pages 67–74, 1988.

[2] C. Chan and L. Lee. Unsupervised spoken term detection with spoken query using segment-based dynamic time warping. 2011.

[3] G. Chollet, J. Cernocky, A. Constantinescu, S. Deligne, and F. Bimbot. Toward alisp: A proposal for automatic language independent speech processing. *NATO ASI series. Series F: computer and system sciences*, pages 375–388, 1999.

[4] A. Garcia and H. Gish. Keyword spotting of arbitrary words using minimal speech resources. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[5] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 601–602. ACM, 2005.

[6] T. Hazen, W. Shen, and C. White. Query-by-example spoken term detection using phonetic posteriorgram templates. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 421–426. IEEE, 2009.

[7] W. Heeringa. *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, University Library Groningen][Host], 2004.

[8] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[9] D. Lee, H. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[10] A. Muscariello, G. Gravier, F. Bimbot, et al. Zero-resource audio-only spoken term detection based on a combination of template matching techniques. In *INTERSPEECH 2011: 12th Annual Conference of the International Speech Communication Association*, 2011.

[11] T. Su and J. Dy. A deterministic method for initializing k-means clustering. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 784–786. IEEE, 2004.

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

[13] Y. Zhang and J. Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 398–403. IEEE, 2009.