UNSUPERVISED VALIDITY MEASURES FOR VOCALIZATION CLUSTERING

Kuntoro Adi[1], *Kristine E. Sonstrom*[2], *Peter M. Scheifele*[3], *Michael T. Johnson*[1]

[1]Department of Electrical and Computer Engineering, Marquette University, Milwaukee, USA
[2]Department of Animal Science, University of Connecticut, Storrs, USA
[3]Communication Science and Disorders, University of Cincinnati, Cincinnati, OH
{kuntoro.adi, mike.johnson}@marquette.edu, kristine.sonstrom@uconn.edu, scheifpr@ucmail.uc.edu

ABSTRACT

This paper describes unsupervised speech/speaker cluster validity measures based on a dissimilarity metric, for the purpose of estimating the number of clusters in a speech data set as well as assessing the consistency of the clustering procedure. The number of clusters is estimated by minimizing the cross-data dissimilarity values, while algorithm consistency is evaluated by calculating the dissimilarity values across multiple experimental runs. The method is demonstrated on the task of Beluga whale vocalization clustering.

Index Terms— speech/speaker clustering, unsupervised validity, dissimilarity value, validation of classifiers.

1. INTRODUCTION

Unsupervised clustering of acoustic waveforms based on waveform similarity has important application to both human speech tasks and animal vocalization analysis. The underlying goal of this clustering is to identify vocalizations with similar patterns for tasks such as repertoire analysis, or in the case of human speech, for tasks such as document indexing and speaker clustering. Recent work in this area has included efforts in speaker indexing [14], labeling speaker turns [11], utterance similarity using speaker clustering [19], speaker diarization [15], and speaker segmentation for meetings [10].

Many different clustering algorithms have been developed [8]. In the clustering process, there are neither pre-specified classes nor observations that would show what kind of desirable relationship is valid among the observations. One of the most difficult parts of clustering is validation of the results, since by definition the task is unsupervised and it is not possible to quantify whether the results are "correct". Nonetheless, it is still possible to make measurements regarding cluster similarity and algorithm consistency as indicators of confidence in the clustering results.

In general, there are three methods to investigate the validity of a cluster, based on assessment using external criteria, internal criteria, and relative criteria [7, 8]. An external assessment criterion evaluates the clustering result using an *a priori* known structure, i.e. it is a supervised approach to validation. Internal criteria evaluation determines if the structure is intrinsically appropriate for the data by using only comparisons of data, while a relative criteria assessment compares two structures resulting from the same algorithm to find out which one is more stable or more appropriate for the data. Thus internal criteria methods are data-based and relative criteria methods are algorithm-based. Both internal and relative validation criteria are unsupervised approaches.

Whenever possible, researchers utilize supervised validity (external criteria) to evaluate clustering results. The degree of similarity between the resulting structure and known partition of the data is calculated using similarity measures such as cluster purity [1, 17], partition misclassification count (PMC)[13]; or some statistical measures such as the Rand statistic, the Jackard coefficient, or the Fowlkes and Mallows index [8].

The research presented here addresses unsupervised cluster validity using a cluster dissimilarity method to estimate the number of clusters in a data set, and to improve and assess the accuracy of a given clustering procedure. The method is developed based on Lange et al. [12], which itself is built upon the early work of Brekenridge [2] as later generalized and extended by Fridlyand and Dudoit [6] in their Clest algorithm.

This research is organized as follows: after the introduction in section 1, section 2 describes the notion of dissimilarity and its usage to estimate the number of clusters in the data and to assess the accuracy of the clustering procedure. Section 3 provides experimental validation through the HMM-based *k*-model clustering of Beluga whale repertoire data and discussion of the results; followed by conclusions in section 4.

2. METHODS

This section introduces the clustering distance method of Lange et al. [12], leading to a metric of dissimilarity. The number of clusters is then estimated from the cross-data dissimilarity analysis and the clustering results are evaluated for consistency using their multi-run dissimilarity.

2.1. Dissimilarity metric

Let a data set **D** consist of N observations $\mathbf{D} = \{ X_1, \ldots, X_N \}$. $X\mathrm{i} = \{\mathbf{x}_{i1}, \ldots \mathbf{x}_{it}\}$ is an observation of length t composed of potentially multivariate feature vectors **x**. The problem of clustering is to find a partition of the data set into k disjoint

clusters. A clustering algorithm A_k builds a solution $\mathbf{L}= A_k(\mathbf{D})$, where $\mathbf{L} = \{L_1, \ldots L_n\}$ is a vector of labels, and $L_i \in \{1, \ldots k\}$ denotes the cluster label. Note that the algorithm A_k is not a classifier itself, but rather a software tool to establish a matching between a specific finite data set and associated cluster labels.

Consider a comparison of solutions computed on two different data sets. Let $\mathbf{L}_1= A_k(\mathbf{D}_1)$ be defined with regard to a data set $\mathbf{D}_1$ and $\mathbf{L}_2= A_k(\mathbf{D}_2)$ for data set $\mathbf{D}_2$. The goal would be to compare two solutions $\mathbf{L}_1$ and $\mathbf{L}_2$ and to assess their similarity or dissimilarity. They are, however, not directly comparable since they come from different data sets. To assess the distance of clustering solutions, Lange et al. devise a predicted label or classifier that renders the solutions comparable.

In general, supervised classification generates a classifier function C that assigns an arbitrary observation from a designated feature space to one of k classes based on a labeled input data. A dataset $\mathbf{D}_1$ together with clustering solution $\mathbf{L}_1$ can be considered as a training data set used to construct a generalized classifier function. This classifier C trained from $(\mathbf{D}_1, \mathbf{L}_1)$ predicts a new label $\mathbf{L}_3= C(\mathbf{D}_2)$ for data set $\mathbf{D}_2$. These labels $\mathbf{L}_3= C(\mathbf{D}_2)$, then, can be compared to those generated by the clustering algorithms, that is, with $\mathbf{L}_2= A_k(\mathbf{D}_2)$.

Lange et al. define a measure of the distance of $\mathbf{L}_3$ and $\mathbf{L}_2$ using a normalized Hamming distance measure as follows:

$$d(\mathrm{L}_3, \mathrm{L}_2) := \frac{1}{n}\sum_{i=1}^{n} 1\{L_{3i} \neq L_{2i}\} \qquad (1)$$

where $\mathbf{1}\{L_{3i}\neq L_{2i}\}=1$, if $L_{3i}\neq L_{2i}$ and zero otherwise. Equation (1), compares two sets of labels that are not necessarily in natural correspondence. This measure quantifies the average distance of two clustering solutions. It can be seen as a misclassification risk with respect to class labels produced by a clustering algorithm.

One significant problem with this approach is the non-uniqueness of label representations. Two partitionings of a data set $\mathbf{D}_2$ might be structurally equivalent although the labelings $\mathbf{L}_3$ and $\mathbf{L}_2$ are differently represented. For instance, a cluster labeled 2 in the first solution might correspond to the one labeled 1 in the second solution, and vice versa. This ambiguity poses a problem for validation.

To overcome the non-uniqueness of representation, the label indices in one solution need to be optimally permuted as to maximize the agreement between the two solutions under comparison. The distance value, then, is modified as follows:

$$d_{Pk}(\mathrm{L}_3, \mathrm{L}_2) := \min_{\pi \in P_k} \frac{1}{n}\sum_{i=1}^{n} 1\{\pi(L_{3i} \neq L_{2i})\} \qquad (2)$$

where P_k is the set of all permutations of the label elements. Equation (2) quantifies the fraction of differently labeled points and can be regarded as the empirical misclassification risk of classifier C with respect to the clustering algorithm A_k.

To use this concept of solution distance in a way that indicates overall dissimilarity value, we denote $Dis(A_k)$ as the average of $d_{Pk}(\mathrm{L}_3, \mathrm{L}_2)$ obtained for r times of split over the data $\mathbf{D}$:

$$Dis(A_k) = \frac{1}{r}\sum_{i=1}^{r} d_{Pki}(\mathrm{L}_3, \mathrm{L}_2) \qquad (3)$$

Lange et. al. refer to this metric as "stability"; however, since its value increases rather than decreases with distance between solutions, we here refer to it as a "dissimilarity" index of clustering solutions with regard to the distribution of the data.

2.2. Number of cluster estimation in a data set

Applying this dissimilarity value to estimate the number of clusters in the data, equation (3) is normalized using the misclassification rate of a random labeling $Dis(R_k)$ that assigns an observation to cluster v with probability $1/k$:

$$\overline{Dis}(A_k) := \frac{Dis(A_k)}{Dis(R_k)} \qquad (4)$$

The smaller the value of $\overline{Dis}(A_k) \in [0,1]$, the more similar are the solutions being compared.

Using this approach to estimate the cluster number we have the following algorithm:

For each number of clusters $k \in \{k_{\min}, \ldots, k_{\max}\}$ perform the following steps

1. Estimate $\hat{Dis}(A_k)$ by averaging r splits of the data:
   a. Split the given data set into two halves $\mathbf{D}_1$, $\mathbf{D}_2$ and apply a clustering algorithm A_k to both
   b. Construct classifier C using $\mathbf{D}_1$ and its cluster labels $\mathbf{L}_1 =A_k(\mathbf{D}_1)$; then compute $\mathbf{L}_3= C(\mathbf{D}_2)$
   c. Use equation (2) to calculate the distance of the two solutions $\mathbf{L}_3= C(\mathbf{D}_2)$ and $\mathbf{L}_2= A_k(\mathbf{D}_2)$
2. Sample s random k-labels, compare pairs of these, and compute the empirical average of the dissimilarities to estimate $\hat{Dis}(R_k)$
3. Normalize each $\hat{Dis}(A_k)$ with $\hat{Dis}(R_k)$ to get an estimate for $\overline{Dis}(A_k)$ using equation (4)

Return the estimated number of clusters
$\hat{k} = \arg\min_k \overline{Dis}(A_k)$.

2.3. Cluster validity measure

The notion of dissimilarity can be further employed to assess the consistency of clustering results. If all partitions into k clusters obtained from running algorithm A_k on data $\mathbf{D}$ are close in structure to partition $\mathbf{L}$, then $\mathbf{L}$ can be considered to be consistent. The dissimilarity value used for this is a generalization of the cross-data cluster dissimilarity computation mentioned in the previous section. To implement this, the clustering algorithm is run t times on the same data set using different initial conditions (or different parameter settings if parameter variation is of interest) and computes the average dissimilarity value of the labeling results as follows:

$$d_{Pk}(\mathrm{L}_i, \mathrm{L}_j) := \min_{\pi \in P_k} \frac{1}{n}\sum_{m=1}^{n} 1\{\pi(L_i \neq L_j\}; \; i, j = 1, \ldots, t \qquad (5)$$

where P_k is the permutation of all label elements, $\mathbf{1}\{L_i\neq L_j\}=1$, if $L_i\neq L_j$ and zero otherwise; and $\mathbf{L} = A_k(\mathbf{D})$; with $\mathbf{L}$ = labeling result, A_k = a clustering algorithm, $\mathbf{D}$ = data to cluster. The smaller the multi-run dissimilarity value $\in [0,1]$, the more

consistent is the clustering algorithm across this dataset. To incorporate the impact of data inclusion as well as initial conditions, this idea can easily be extended to use random subsets for each run in a resampling-with-replacement fashion.

3. EXPERIMENTS

An illustration of this method is given using a Beluga whale repertoire clustering experiment. Figure 1 shows the basic approach:

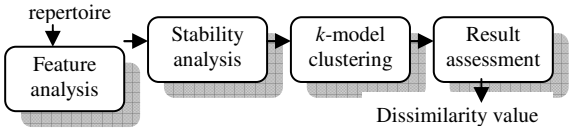

Figure 1. Block diagram to cluster Beluga repertoire data

3.1. Beluga data set

The vocalizations were made by a population of beluga whales residing in the Saint Lawrence River Estuary. Belugas use echolocation to locate their prey, to find breathing holes in the Arctic ice sheet, and to navigate in the waters. They produce many different calls, including clicks, squeals, and whistles. The Beluga population has been extensively studied regarding their seasonal distribution, size, age structure, and pathology. They are susceptible to the effect of human-made noise due to the large amount of shipping in the Saint Lawrence River, and the focus of this study has been on the analysis of the impact of this noise [18] and on evaluating similarity between established social groups. Even though the repertoire of the species is not well known, the vocalizations needed to be accurately categorized for the study, so vocalization clustering was implemented.

The vocalizations were selected from five study sites. They were recorded in July and August over the course of six years by Scheifele [18]. The acoustic data was collected with an omni-directional hydrophone and recorded on a Sony TCD-D8 DAT tape with 16-bit quantization.

3.2. Hidden Markov Model-based *k*-model clustering

A hidden Markov model (HMM)-based *k*-model clustering [3] is used for vocalization clustering. On each iteration every repertoire data is assigned to a single cluster represented by an HMM. The HMM parameter updates are influenced only by data items currently in the associated clusters.

The HMM-based *k*-models algorithm is a generalization of the standard *k*-means approach, with the cluster centroid vectors being replaced by the probabilistic models (in this case, HMMs). The criterion to re-assign data to clusters is maximization of the likelihood of the data points. The re-assignment of the data employs the Viterbi algorithm. The computation of clusters is done by re-estimating the model parameters using a Baum-Welch re-estimation algorithm.

Features for this experiment consist of Greenwood function cepstral coefficients (GFCC) [4] which are generalizations of well-known Mel-frequency cepstral coefficients (MFCCs) [5]. Thirty-six element feature vectors are extracted. They consist of cepstral coefficients along with delta and acceleration coefficients. The repertoires are Hamming windowed with frame and step sizes of 3 ms and 1.5 ms. Cluster models for the experiment are 15-state left-to-right HMMs with each state contains a single diagonal-covariance Gaussian.

3.3. Experimental results - discussion

Initially the dissimilarity metric is used for feature selection to find out the best features for Beluga repertoire clustering in terms of overall dissimilarity. Figure 2 shows the dissimilarity results on three different features, the cepstral coefficients (GFCC) and their delta (D) and acceleration (A) coefficients, the mean-normalization cepstral coefficients (GFCCDA-MN), and the variance-normalization cepstral-coefficients (GFCCDA-VN) [9] across different number of clusters for beluga data06 data set.

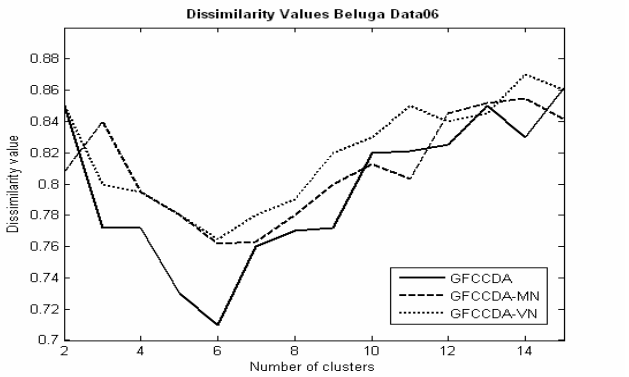

Figure 2. The dissimilarity index values of the beluga data06 from three different cepstral coefficient features

Of the three features, GFCCDA leads to the best performance. This feature is then selected for the rest of the clustering procedure.

Figure 3 shows the use of the cross-data dissimilarity method to estimate number of clusters k from four different beluga data sets. Estimates are $k = 3$ (data96), $k = 5$ (data03, data07) and $k = 6$ (data06).

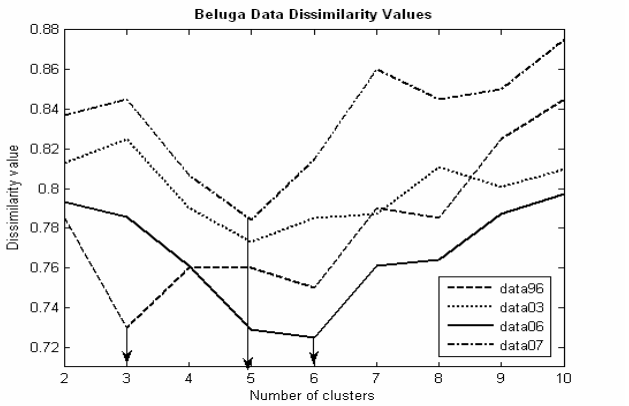

Figure 3. Cluster estimates for four different Beluga data sets.

To illustrate the consistency of the results, Table 1 presents the dissimilarity values and their respected standard deviations from 10 separate runs of the clustering from 5 different beluga data sets.

	Data	Number of clusters	Dissimilarity value
1	data96a	2	0.026±0.003
2	data96	3	0.130±0.012
3	data03	5	0.473±0.046
4	data05	6	0.359±0.049
5	data06	6	0.211±0.019

Table 1. The estimated number of clusters and dissimilarity values of 5 different beluga data sets

Results indicate that the dissimilarity value has a significant range for the different datasets. It gives an almost perfect match (0.026±0.003 dissimilarity value) for data96a. Groups of repertoires in this data are assigned consistently to the same clusters in each experimental run. In comparison, the clustering result presents a relatively high dissimilarity value (0.473±0.046) for data03. There are several possible hypotheses for this variation. Inconsistent clustering runs as shown by a high dissimilarity value may indicate that there are a relatively large range of vocalization types in the data set with only a few examples of each, so that data limitation prevents accurate grouping. Another possibility is that the vocalizations are relatively similar with a more continuous variation.

4. CONCLUSION

The idea of using dissimilarity to assess the quality of clustering has been introduced, as a way of evaluating the distance between clustering solutions. The dissimilarity metric is implemented both to estimate clustering parameters such as number of clusters, as well as for overall assessment of consistency in the final clustering results. The approach is able to infer the natural partitions of complex acoustic data such as Beluga repertoire data, and, more importantly, is able to provide a confidence measure regarding consistency of these clustering results without known *a priori* vocalization labels.

5. REFERENCES

[1] Ajmera, J., Bourlard, H., Lapidot, I., McCowan, I., (2002), "Unknown-multiple speaker clustering using HMM," *International Conference on Spoken Language Processing*.

[2] Breckenridge, J. N., (1989), "Replicating cluster analysis: method, consistency and validity," *Multivariate Behavioral Research*, 24, 147-161.

[3] Clemins, P. J., (2005), *Automatic classification of animal vocalizations, Ph. D. Dissertation*, Marquette University, WI.

[4] Clemins, P.J., Trawicki, M.B., Kuntoro Adi., Jidong Tao, Johnson, M.T., (2006), "Generalized perceptual features for vocalization analysis across multiple species," *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP Proceedings, vol. 1: 14-19.

[5] Davis, S. B., Mermelstein, P., (1980), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28:357-366.

[6] Fridlyand, J., Dudoit, S., (2001), *Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method*, Technical Report 600, Department of Statistics, University of California, Berkeley, CA.

[7] Halkidi Maria, Batistakis Yannis, Vazirgiannis Michalis, (2001), "On clustering validation techniques," *Journal of Intelligent Information Systems*, The Netherlands, Kluwer Academic Publishers, 17:2/3, pp. 107-145.

[8] Jain, A. K., Dubes, R. C., (1988), *Algorithms for clustering data*, Prentice Hall

[9] Kuntoro Adi, Johnson Michael T., (2006), "Feature normalization for robust individual identification of the Ortolan bunting (*emberiza hortulana L*)," *The Journal of the Acoustical Society of America*, vol. 119, no. 5.

[10] Jin Qin, Laskowski Kornel, Schultz Tanja, Waibel Alex, (2004), "Speaker segmentation and clustering in meetings," *Proceeding of the International Conference of Spoken Language Processing*, South Korea.

[11] Johnson, S. E., (1999), "Who spoke when? – Automatic segmentation and clustering for determining speaker turns," *Proc. Eurospeech*.

[12] Lange Tilman, Roth Volker, Braun Mikio L., Buhmann Joachim M., (2004), "Stability-based validation of clustering solutions," *Neural Computation* 16:1299-1323.

[13] Li Cen, Biswas Gautam, (2000), "A Bayesian approach to temporal data clustering with hidden Markov model representation," *Proceeding of the Seventeenth International conference on Machine Learning*.

[14] Masafumi Nishida, Tatsuya Kawahara, (2003), "Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion," *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[15] NIST, Rich transcription 2004 Spring meeting recognition evaluation, http://www.itl.nist.gov/iad/894.01/test/rt/rt2004/spring

[16] Qin Jin, Kornel Laskowski, Tanja Schultz, Alex Waibel, "Speaker segmentation and clustering in meetings," *Proc. ICSLP'04.*

[17] Solomonoff, A., Mielke, A., Schmidt, M., and Gish, H., (1998), "Clustering speakers by their voices," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 757–760, 1998.

[18] Scheifele, P. M., (2003), *Investigation into the response of the auditory and acoustic communication systems in the beluga whale (Delphinapterus leucas) of the St. Lawrence River estuary to noise, using vocal classification*, Ph. D. Dissertation, University of Connecticut, Hartford, CT.

[19] Wei-Ho Tsai, Shih-Sian Cheng and Hsin-Min Wang, (2004), "Speaker clustering of speech utterances using a voice characteristic reference space," *Proc. ICSLP'04*