LETTER

Robust Speaker Clustering Using Affinity Propagation*

Xiang ZHANG[†a)], **Ping LU**[†], ***Nonmembers***, **Hongbin SUO**[†], ***Student Member***, **Qingwei ZHAO**[†], ***and*** **Yonghong YAN**[†], ***Nonmembers***

SUMMARY In this letter, a recently proposed clustering algorithm named affinity propagation is introduced for the task of speaker clustering. This novel algorithm exhibits fast execution speed and finds clusters with low error. However, experiments show that the speaker purity of affinity propagation is not satisfying. Thus, we propose a hybrid approach that combines affinity propagation with agglomerative hierarchical clustering to improve the clustering performance. Experiments show that compared with traditional agglomerative hierarchical clustering, the hybrid method achieves better performance on the test corpora.

key words: *speaker clustering, agglomerative hierarchical clustering, affinity propagation, generalized likelihood ratio*

1. Introduction

Speaker clustering refers to the task of grouping speech utterances into clusters such that each cluster contains speech from one speaker and also speech from the same speaker is grouped into the same cluster [1]–[3]. Currently, most speaker clustering approaches follow an agglomerative hierarchical clustering (AHC) framework, which is comprised of three major components: computation of pair-wise distances, generation of a cluster tree, and determination of the number of clusters [1]. The stopping criterion is critical to good clustering performance and depends on how the output is to be used. In our study, the agglomerative hierarchical clustering works with the commonly used *BIC* stopping criterion.

Recently, a new clustering algorithm named affinity propagation has been proposed, and it is being used to cluster images of face, identify representative sentences, detect genes, and perform other tasks [4], [5]. Affinity propagation exhibits fast execution speed and finds clusters with low error. In this letter, we introduce it to cluster speech segments in telephone conversations and broadcast news audio with an unknown number of speakers. Although adopting affinity propagation for speaker clustering can produce high cluster purity, the experiment results show that it may generate a far larger number of clusters which deteriorates the speaker purity dramatically. AHC performs well in determining the number of clusters. Thus an improved clustering method named APAHC is proposed, which is a hybrid approach combining affinity propagation with AHC. This approach uses AHC procedure to re-cluster the results of affinity propagation to achieve better clustering performance, especially the speaker purity. Our experiments show that the APAHC approach is superior to the traditional agglomerative hierarchical clustering approach and affinity propagation.

Manuscript received March 31, 2008.

Manuscript revised May 26, 2008.

[†]The authors are with the ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, China.

[*]This work is partially supported by MOST (973 program · 2004CB318106), National Natural Science Foundation of China (10574140, 60535030), The National High Technology Research and Development Program of China (863 program, 2006AA010102 · 2006AA01Z195).

a) E-mail: zhangxiang@hccl.ioa.ac.cn

DOI: 10.1093/ietisy/e91–d.11.2739

2. Speaker Clustering via Affinity Propagation

Affinity propagation assumes that all the speech segments are potential centers. By viewing each segment as a node in a network, affinity propagation recursively transmits real-valued messages along edges of the network until a good set of centers and corresponding clusters emerges. As described later, messages are updated on the basis of simple formulas during the procedure with pre-computed similarities. By simultaneously considering all the data points as candidate centers and gradually identifying clusters [4], affinity propagation can avoid many of the poor solutions caused by unlucky initializations and hard decisions. Thus, we introduce affinity propagation for the task of speaker clustering.

The *similarity* $s(i,k)$, where $i \neq k$, *preference* $s(k,k)$, *responsibility* $r(i,k)$ and *availability* $a(i,k)$ are the four main elements in affinity propagation. Affinity propagation takes a collection of real-valued similarities between speech segments as input, where the similarity $s(i,k)$ indicates how well the segment k is suited to be the center for the segment i. The preference $s(k,k)$ is a real number for each segment k. The segments with larger values of $s(k,k)$ are more likely to be chosen as centers. If a priori, all the segments are equally suitable as centers, the preferences should be set to a common value. The responsibility $r(i,k)$ reflects the accumulated evidence for how well-suited segment k is to serve as the center for segment i, taking into account other potential centers for segment i. The availability $a(i,k)$ reflects the accumulated evidence for how appropriate it would be for segment i to choose segment k as its center, taking into account the support from other segments that segment k should be a center.

In our affinity propagation speaker clustering, each similarity is set to the negative generalized likelihood ratio (GLR):

$$s(i,k) = -d_{GLR}(x_i, x_k),\ i \neq k \tag{1}$$

where, x_i and x_k are the speech feature vectors of the two segments, which can be modeled with two Gaussian models $N(\mu_{x_i}, \sum_{x_i})$ and $N(\mu_{x_k}, \sum_{x_k})$. $d_{GLR}(x_i, x_k)$ is the GLR distance between x_i and x_k, which is defined as follows [6]:

$$d_{GLR}(x_i, x_k) = \frac{L(x_i; \mu_{x_i}, \sum_{x_i}) \cdot L(x_k; \mu_{x_k}, \sum_{x_k})}{L(y; \mu_y, \sum_y)} \tag{2}$$

Where, $L(*)$ is the likelihood function and y is the union of x_i and x_k which indicates the concatenation of both feature vectors.

The suitable input preferences are very important to influence the final number of clusters. The larger the values of preferences, the more clusters will be produced. We set all the preferences to the same value – the median of total input similarities as mentioned in [4]. Responsibility and availability are two kinds of message exchanged between speech segments, which are iteratively updated by the formulas (3), (4), (5), reflecting the affinity of segments. They are computed as follows:

$$r(i,k) = s(i,k) - \max_{j:j\neq k}[s(i,j) + a(i,j)] \tag{3}$$

For $a(i,k)$, if $k = i$,

$$a(i,k) = \sum_{i':i'\neq k} \max[0, r(i',k)] \tag{4}$$

If $k \neq i$,

$$a(i,k) = \min\left\{0, r(k,k) + \sum_{i':i'\notin\{i,k\}} \max[0, r(i',k)]\right\} \tag{5}$$

For the first iteration, the availabilities are initialized to zero.

Affinity propagation combines the responsibilities and availabilities to control the center decisions. For segment i, the segment k which maximizes $r(i,k) + a(i,k)$ either identifies the segment i as a center if $k = i$, or identifies the segment that is the center for segment i if $k \neq i$. The whole affinity propagation procedure terminates after a fixed number of iterations or after the center decisions stay unchanged.

3. Speaker Clustering via Proposed APAHC

Our experiment results show that affinity propagation can achieve high performance of cluster purity, but it usually produces extra number of clusters, which makes the speaker purity quite low. This motivates us present an improved unsupervised speaker clustering approach named APAHC, which is a hybrid algorithm combining affinity propagation with AHC. We hope APAHC could take advantage of the satisfying cluster-purity performance of affinity propagation and the good performance of determining the number of clusters of AHC to generate better overall clustering performance than that of affinity propagation method, especially the performance of speaker purity. Experiments show that the proposed APAHC method also produces comparable results to conventional AHC.

APAHC firstly runs affinity propagation to under-cluster (where the number of clusters is believed to be far greater than the number of speakers) the speech segments using (6). This reduces the probability that the speech from different speakers will be classified into one cluster. This step is useful for the final clustering of APAHC and can produce clusters with quite high cluster purity.

$$s(k,k) = \underset{i=1:N, j=1:N, i\neq j}{\text{median}}\{s(i,j)\} - P, k = 1:N \tag{6}$$

Where, N is the number of total speech segments, *median* denotes the median of total input similarities, P is an empirical value used to modify the preference, which needs to be tuned for changes in audio type and features.

After the process of affinity propagation, clusters with a reasonable amount of speech can be produced. Each cluster's data is used to train a Gaussian model, and the GLR distance between any two clusters is calculated. These distances are then used to drive an agglomerative hierarchical speaker clustering based on the *BIC* stopping criterion to reduce the number of clusters by merging. The expression of ΔBIC is given by [7]:

$$\Delta BIC = \frac{1}{2}\{n \log|\Sigma_y| - n_1 \log|\Sigma_{x_i}| - n_2 \log|\Sigma x_k|\} - \lambda P \tag{7}$$

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log n \tag{8}$$

Where, n_1 and n_2 are the numbers of frames of x_i and x_k respectively, $n = n_1 + n_2$, Σ the covariance matrix and d the dimension of the feature vector. If the pair of clusters is best represented by a single Gaussian model, the ΔBIC will be low, whereas if there are two separate distributions, implying two speakers, the ΔBIC will be high. For each merging a new Gaussian speaker model can be trained with the combined data and distances of remaining clusters to the new cluster are updated. The process of AHC is generally stopped when the ΔBIC of the nearest pair of clusters is greater than a specified threshold, usually 0.

4. Experiment

4.1 Speech Corpora and Evaluation Metrics

Experiments for the proposed APAHC approach are carried out on 4-hour hand-labeled telephone conversations in the NIST 2004 Speaker Recognition Evaluation and all the broadcast news audio in MDE RT-04, respectively. The telephone data consists of 48 conversations, each of which is about 5-minute long. The duration of each broadcast news audio ranges from 30 minutes to 1 hour, and the number of speakers from 20 to 40. The number of speech segments ranges from 16 to 150 for each telephone conversation, and 76 to 320 for each broadcast news audio. Speech features of 14 line spectrum pair (LSP) are extracted from these data for

Table 1 Performance on the telephone conversations.

	AHC	Affinity Propagation	Proposed APAHC
cp (%)	98.1	94.5	95.8
sp (%)	83.9	84.4	90.1
K (%)	90.7	89.3	92.9

Table 2 Performance on the broadcast news.

	AHC	Affinity Propagation	Proposed APAHC
cp (%)	93.5	98.8	93.8
sp (%)	83.2	53.6	83.5
K (%)	88.2	72.8	88.5

every 20-ms Hamming-windowed frame with 10-ms frame shifts. For comparing, affinity propagation and AHC are also evaluated on the test sets. We evaluate our experiments with commonly used criteria [8], [9]: cluster purity (cp) and speaker purity (sp).

$$cp = \sum_{i=1}^{c} \max_{j\in[1:k]}(n_{ij}) \Big/ \sum_{i=1}^{c}\sum_{j=1}^{k} n_{ij} \tag{9}$$

$$sp = \sum_{j=1}^{k} \max_{i\in[1:c]}(n_{ij}) \Big/ \sum_{i=1}^{c}\sum_{j=1}^{k} n_{ij} \tag{10}$$

Where, k is the total number of speakers, c is the final number of clusters, and n_{ij} denotes the number of speech frames in cluster i spoken by speaker j.

In order to facilitate comparison between approaches, we also use an overall evaluation criterion [9]:

$$K = \sqrt{cp * sp} \tag{11}$$

4.2 Experimental Results

The traditional AHC, affinity propagation, and the proposed APAHC are implemented respectively.

Table 1 lists the results of the three approaches on telephone conversations. We can see that APAHC achieves both higher cluster purity and speaker purity than affinity propagation, and APAHC generates significant improvement in speaker purity compared with AHC with slightly decrease in cluster purity. The value of the overall evaluation criterion K also shows that APAHC is superior to affinity propagation and AHC on the telephone test data.

Table 2 displays the performance of the algorithms on broadcast news audio test set. We can observe that APAHC produces the best overall clustering performance. APAHC leads to about 0.3% improvement in both cluster and speaker purity compared with AHC, and achieves about 30% improvement in speaker purity compared with affinity propagation with about 5% decrease in cluster purity.

All the results in the two tables show that the proposed APAHC has better overall clustering performance than affinity propagation and AHC, and it can be well applied for speaker clustering in both telephone conversations and broadcast news data.

5. Conclusions and Future Work

In this letter, we introduce affinity propagation for speaker clustering. In order to overcome the limitation of the algorithm, an improved speaker clustering approach named APAHC is proposed, which aims at processing real-world media with unknown number of speakers. APAHC combines affinity propagation with AHC to cluster speech segments, and experiments show that it can achieve better overall clustering performance than affinity propagation and AHC, especially the performance of speaker purity. APAHC firstly over-clusters the speech segments and produces clusters with reasonable amount of speech. The future of the work is to use the Gaussian Mixture Models (GMMs) adapted from the universal background model (UBM) to represent the clusters in the AHC step, and use the cross likelihood ratio (CLR) to calculate pair-wise distances.

References

[1] D. Liu and F. Kubala, "Online speaker clustering," Proc. ICASSP'03, vol.1, pp.333–336, 2003.
[2] A. Iyer, U. Ofoegbu, R. Yantorno, and B. Smolenski, "Blind speaker clustering," Proc. ISPACS'06, pp.343–346, 2006.
[3] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," IEEE ASRU Workshop, pp.411–416, 2003.
[4] B. Frey and D. Dueck, "Clustering by passing messages between data points," Science, vol.315, pp.972–976, 2007.
[5] B. Frey and D. Dueck, "Supporting online material for clustering by passing messages between data points," www.sciencemag.org/cgi/data/1136800/DC1/1, 2007.
[6] H. Suo, M. Li, P. Lu, and Y. Yan, "Automatic language identification with discriminative language characterization based on svm," IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.567–575, March 2008.
[7] M. Cettolo and M. Vescovi, "Efficient audio segmentation algorithms based on the bic," ICASSP 2003, pp.537–540, 2003.
[8] W. Wang, P. Lv, Q. Zhao, and Y. Yan, "A decision-tree-based online speaker clustering," Lecture Notes in Computer Science, vol.4477, pp.555–562, 2007.
[9] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using hmm," Proc. ICSLP'02, pp.573–576, 2002.