

TOWARDS DOMAIN INDEPENDENT SPEAKER CLUSTERING

Yvonne Moh¹, Patrick Nguyen², and Jean-Claude Junqua²

¹ Lehrstuhl für Informatik VI
Computer Science Department
RWTH-Aachen – University of Technology
52056 Aachen, Germany
yvonne@halifax.rwth-aachen.de

² Panasonic Speech Technology Laboratory (PSTL)
Panasonic Technologies Company
3888 State Street, Suite 202,
Santa Barbara, CA 93105, U.S.A.
{nguyen, jcj}@research.panasonic.com

ABSTRACT

Speaker clustering is a key component in many speech processing applications. We focus on Broadcast News meta data annotation and speaker adaptation. In this setting, speaker clustering consists of identifying who spoke, and when they spoke in a long news broadcast. Speaker clustering is given a set of short audio segments. Ideally, it will discover how many people are speaking in the broadcast, and when they are speaking. The same problem can be transposed to a different domain.

In this paper, we present two techniques that do not require *a priori* training. The speaker clustering is based on information collected solely on encountered test data. They aim at being portable across domains.

The first method is based on a Bayesian Information Criterion (BIC), with single full-covariance Gaussians. It is fairly primitive but effective. The second method, called speaker triangulation, constructs a coordinate system based on conditional likelihoods of the audio segments. Clusters are located in this coordinate system. We are able to achieve state-of-the-art performance on NIST evaluations across different data sets.

1. INTRODUCTION

Audio indexing has become more popular and usable in the last years. This is reflected by a need for offering more than just speech-to-text (STT) transcriptions. This year, NIST presented a ground breaking evaluation paradigm, called Rich Transcription, that seeks to enrich STT transcriptions with meta data. In parallel, the speaker recognition benchmark introduced a new data set, with unknown conditions. The goal was to study portability to a new domain.

Meta data are additional information that can be displayed for improved readability or downstream processing. For instance, speaker clustering is providing meta data that can be consumed in key frame detection or speaker adaptation.

The paper has four remaining sections. Firstly, we introduce speaker clustering and its basic theory. Secondly, we motivate our research and present our clustering methods. Thirdly, algorithms are validated on NIST evaluation sets. Finally, we conclude and present some directions for future research.

This work was done while Yvonne Moh was an intern in PSTL. We would also like to thank the NIST for providing the data and for their work in defining the evaluation framework. We also acknowledge Jean-François Bonastre and Sylvain Meignier for many helpful discussions.

2. SPEAKER CLUSTERING

Speaker clustering can be applied in a number of speech processing applications. We will focus on speech recognition and meta data generation. In a typical speech recognition system, the audio is first partitioned into small segments, with gender / bandwidth classification. This is called segmentation. Then, speaker clustering groups those audio segments into larger clusters. Ideally, each cluster will correspond to a unique speaker, and vice-versa. There are conflicting goals in meta data generation and speech recognition:

- Meta data generation is concerned about *classification* of speakers. It needs to separate sound-alike speakers. Performance measures include frame error, BBN index, and Rand Index.
- Speaker adaptation is concerned about *regression* of speakers. If two speakers are reasonably indistinguishable, then they should be considered equal. Performance is measured in improvements over baseline recognition.

We will study speaker clustering under these two performance goals.

2.1. Clustering theory

Speaker clustering has been an active research field for many years. We can account for roughly four fundamental problems and solutions:

- agglomeration: how to form the clusters? We can either use divisive or agglomerative techniques.
- stopping criteria: how many speakers are in the stream? In other words, we decide to stop the merging/splitting process.
- distance measures: how close are two audio segments considered? For instance, a Mahalanobis distance between the means could be envisioned.
- set distances: how close are two *clusters* of segments? The maximum / minimum linkage are two paramount examples.

In this paper, we attack the problem from the point of view of distance measures.

2.2. Definitions

Let $X = \{x_1, x_2, \dots, x_n\}$ be an audio stream. x_i represents an observation vector of X , and n is the total number of observation vectors (frames) in audio stream X . The dimension of each vector is D .

Since we are involved in first segmenting this audio stream X , and then clustering it, further refinements have to be specified.

2.2.1. Segments:

X can be broken into segments. These (non-overlapping) segments may eventually constitute a subset of X , for instance, in the case where silences are omitted. We represent the segments as X_1, X_2, \dots, X_S . X_i is one of S segments in X . Note that $X_i \subset \bigcup_{i=0}^S X_i \subset X$. We write $X_i = \{x_{i,1}^s, x_{i,2}^s, \dots, x_{i,n_i}^s\}$, such that $x_{i,j}^s$ represents the j -th observation vector in segment X_i . X_i has a total of n_i frames. $N = \sum_{i=0}^S n_i$ indicates the number of non-discarded frames.

In subsequent chapters, we will be using the means and the covariances to describe the segments. For segment X_i , we denote the sample mean μ_i^s and covariance Σ_i^s as follows:

$$\mu_i^s = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}^s \quad (1)$$

$$\Sigma_i^s = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{i,j}^s - \mu_i^s)(x_{i,j}^s - \mu_i^s)' \quad (2)$$

When two segments X_i and X_j are merged, we refer to the resulting mean and covariance as:

$$\mu_{i \cup j}^s = \frac{1}{n_i + n_j} (n_i \mu_i^s + n_j \mu_j^s) \quad (3)$$

$$\Sigma_{i \cup j}^s = \frac{1}{n_i + n_j} \left(\sum_{k=1}^{n_i} (x_{i,k}^s - \mu_{i \cup j}^s)(x_{i,k}^s - \mu_{i \cup j}^s)' + \sum_{k=1}^{n_j} (x_{j,k}^s - \mu_{i \cup j}^s)(x_{j,k}^s - \mu_{i \cup j}^s)' \right) \quad (4)$$

2.2.2. Clusters:

As for clusters, we let $C(K)$ represent a clustering of K clusters C_1, C_2, \dots, C_K . Each cluster C_i has segments $X_{i,1}^c, \dots, X_{i,S_i}^c$, we denote the number of frames in segment $X_{i,j}^c$ as n_{j,S_i}^c . The clusters are all disjoint. We can concatenate the frames from the S_i segments to represent the set of observation frames in C_i . These will be referred to as $\{x_{i,1}^c, x_{i,2}^c, \dots, x_{i,N_i}^c\}$ where N_i indicates the number of frames in cluster C_i . We have $N_i = \sum_{X_j \in C_i} n_j = \sum_{j=1}^{S_i} n_{j,S_i}^c$. Note that $N = \sum_{i=0}^K N_i$.

$$\mu_i^c = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j}^c$$

$$\Sigma_i^c = \frac{1}{N_i} \sum_{j=1}^{N_i} (x_{i,j}^c - \mu_i^c)(x_{i,j}^c - \mu_i^c)'$$

As with segments, when merging two clusters C_i and C_j , we get the following:

$$\mu_{i \cup j}^c = \frac{1}{N_i + N_j} (N_i \mu_i^c + N_j \mu_j^c) \quad (5)$$

$$\Sigma_{i \cup j}^c = \frac{1}{N_i + N_j} \left(\sum_{k=1}^{N_i} (x_{i,k}^c - \mu_{i \cup j}^c)(x_{i,k}^c - \mu_{i \cup j}^c)' + \sum_{k=1}^{N_j} (x_{j,k}^c - \mu_{i \cup j}^c)(x_{j,k}^c - \mu_{i \cup j}^c)' \right) \quad (6)$$

3. PORTABLE CLUSTERING METHODS

3.1. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) was introduced for speaker clustering in [1]. Let X_i and X_j be two segments. We model the observations from each segment as a single Gaussian, i.e. $X_i = \mathcal{N}(\mu_i^s, \Sigma_i^s)$ and $X_j = \mathcal{N}(\mu_j^s, \Sigma_j^s)$.

BIC is given by:

$$\text{BIC} = (n_i + n_j) \log |\Sigma_{i \cup j}^s| - n_i \log |\Sigma_i^s| - n_j \log |\Sigma_j^s| - \lambda P,$$

with

$$P = \frac{1}{2} (D + \frac{1}{2} D(D+1)) \log N.$$

Excessive splitting is prevented by the penalty P .

Merits of this method were proven for Broadcast News. However, on NIST evaluations carried out on Switchboard data, it is not deemed a viable alternative. In our experiments, we show that with full covariance matrices with static coefficients, state-of-the-art performance can be achieved. This approach has several properties:

- No training is required: all data is contained within the data test set.
- Tuning: only one parameter, λ needs to be estimated. We found that the value of the parameter was independent of the database.
- The system looks at the global system before making a local merging decision. The system is changed at every step.
- The selection is based on covariances: the most consistent solution is found. In other words, clusters that are internally homogeneous are good.
- Modeling with one full covariance with static coefficients was instrumental in our success.

Contrarily to speech recognition, correlation between cepstral features conveys useful information.

3.2. Triangulating speakers

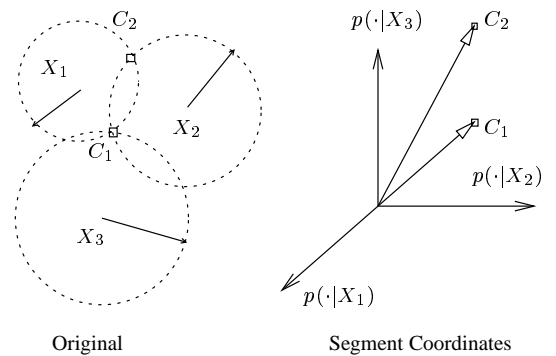


Fig. 1. Clusters $C_{1,2}$ are identified by the distance of their centroid to segments $X_{1,2,3}$. On the left, we see how the distance is measured. On the right, we see the coordinate system.

The use of background speakers, and universal background models is very popular in many approaches. Background speaker modeling provides natural means for normalization. However, background speakers must be trained on a database that is reasonably close to the

target environment. After training, one assumes that enough data was collected for reliable estimation of the background speakers.

We decide to use audio segments as reference models. Since they are trained on disparate amounts of data, further processing is required. A set of segments is used to generate a “referential”. This is shown on Fig. 1. All clusters are located in this coordinate system.

Classical segment-to-segment define a distance in a segment’s reference system. In the case of Gaussians, it is a Euclidean distance centered around the segment’s mean weighted its precision. To combine multiple segments reference systems, we use a triangulation method: a point in space is uniquely represented by the distance to reference points. Triangulation is popular in constructing maps, when an absolute coordinate system is not available. When there are K true speakers, they form at most a K -dimensional space. Each point in this space is described by its relative distance to at least $(K - 1)$ points. An over-complete system with more reference then required $S > (K - 1)$, should produce more robust estimation.

Let C_k be a putative speaker cluster. In the speaker triangulation method, we define a S -dimensional vector $p(k)$ for each of these clusters, which represents the conditional probability of C_k given all segments. If $p_s(k)$ is the s -th component of $p(k)$, we define:

$$p_s(k) = p(C_k | X_s), \quad s = 1, \dots, S. \quad (7)$$

A single, full covariance Gaussian emission probability served as $p(\cdot | X_s)$. Now suppose that we present another cluster C_j as a candidate for clustering with C_k . They will be considered equivalent if the correlation $p(k)^T p(j)$ is large:

$$D(k, j) = \sum_s p_s(C_k | X_s) p_s(C_j | X_s) \approx p(C_k, C_j). \quad (8)$$

Informally, it is the probability of two events C_k and C_j of the same speaker occurring simultaneously. It can be thought of as a vectorized GLR [2]. On Fig. 1, the correlation is a representation of having many identical segments indicating that C_k and C_j are near or far to them.

Fig. 2 shows the relative likelihood of each segment relative to each individual segment in a show. For better visualization, the segments have been sorted according to the speakers. As expected, each speaker creates a box of high relative probabilities, and seen by the dark boxes along the diagonal. Two rows that are highly correlated are believed to belong to the same speaker.

We observe that this method can be characterized by several properties:

- No training is required: the method can be ported from Broadcast News to Switchboard without modification. There is no training of cohort or universal background models.
- Condition normalization: a stream with segments in mixed conditions may be processed. For instance, wide bandwidth segments are less confusable intrinsically. For narrow bandwidth, one has to account small changes in the feature space. Classical systems weigh wide bandwidth inordinately.
- Self-reference: the reference system is based on the audio stream itself. Therefore, it will naturally cover all the space. There is no need for careful selection of “background” speaker models.
- Likelihood correlations, instead of pairwise Kullback-Leibler distances, do not suffer from bias in the number of frames.
- Dimensionality: during successive merges, the dimensionality S of the coordinate system remains constant. Merging errors do not propagate through the variance.
- Localization: triangulation is very sensitive around densely segment populations. The density can be due to either intrinsic confusability or many events of the same speaker.

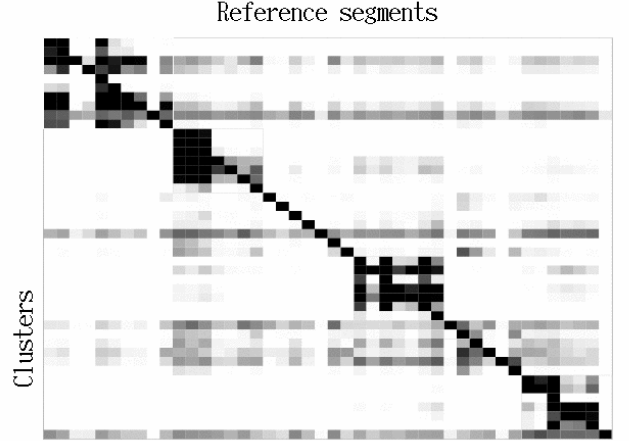


Fig. 2. Likelihood Matrix for Broadcast News set BN / RT-02, show 1. Dark dots mean higher likelihood. A dark dot off diagonal indicates the pairwise distance cluster-reference is small. A row with dark dots on the same columns means that they originate from the same speaker.

- Coherence: clusters with very different centroids are considered different. Incoherent merges, or merges with centroids that are not the same, are discarded. There is no notion of consistence, where one considers the homogeneity of a candidate merge.

4. EXPERIMENTS

To assess the performance of the speaker clustering scheme, we present two common embodiments: automatic speech recognition of Broadcast News, and meta data annotation for Broadcast News and Switchboard.

4.1. Experimental framework

The Broadcast News automatic speech recognition system [3] employs MFCC features, with delta and acceleration coefficients, and normalized by the cepstral mean on a causal sliding window of 2 seconds. A total of 192k Gaussians per gender were trained for about 2000 context-dependent tied states. The language model contains over 67M trigrams and 17M bigrams, for a lexicon of 57k words. The audio was pre-segmented using condition and gender dependent GMMs, plus silence. The first and the second pass are identical in nature: the second pass uses speaker-cluster adapted models. MLLR was applied in block-diagonal mode with 7 regression classes.

The meta data system used the same MFCC features for BN (16 kHz), and PLP features replaced during SWB experiments (8 kHz). We made no effort to optimize the front-end processing. We present results with the NIST Speech Activity Detection (SAD) segments. Best results were obtained with a BIC stopping criterion, and nearest neighbor clustering. In BN systems, the gender is determined automatically. Clusters may not cross gender boundaries. However, since the same speaker can appear in narrow bandwidth and wide bandwidth, we allow cross bandwidth clusters. The system was scored using official tools provided by NIST.

4.2. Results

On Table 1, we show results on meta data. The high performance of both approaches is a testimony of the success of the portability. In

System	Test Set	Frame Err
BIC	BN - SID-02	21.6%
Triangulation	BN - SID-02	21.0%
BIC	SWB - SID-02	8%
Triangulation	SWB - SID-02	13.3%
BIC	BN - RT-02	15.0%
Triangulation	BN - RT-02	3.6%

Table 1. Frame error rate for meta data on different sets.

RT-02, there were 6 10-min excerpts from an hour-long show. In this case, the triangulation method can perform significantly better than our standard baseline. In SID-02, those excerpts were concatenated.

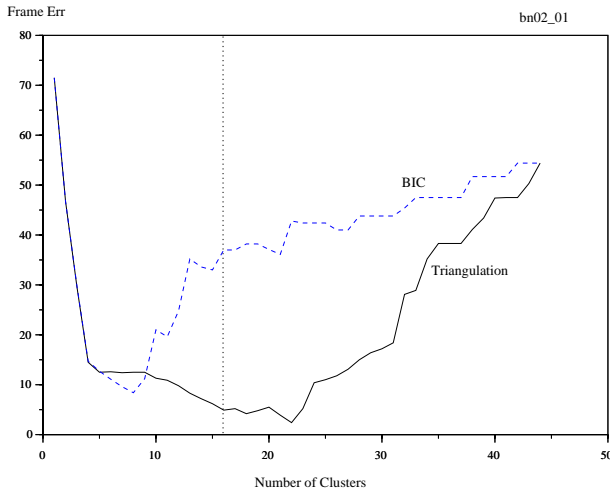


Fig. 3. BIC vs Correlation: NIST Frame Error on segments from test BN/RT-02, show 1. There are 16 speakers.

On Fig. 3, we show how the NIST frame error varies with the stopping threshold. The baseline can reduce its average frame error rate by merging many small duration segments. This is also reflected in Table 1. We see that the properties of the approaches are independent of the domain, but related with the total number of speakers and segments. Triangulation is shown to work well when there are relatively fewer segments to merge. On the other hand, BIC works better with a few number of speakers. Our explanation is based on the properties of the methods. They can be characterized by the use of the clusters' covariance. Triangulation is effective in situations in the initial phases of merging when there are many clusters. This is due to the over-determination of the coordinate system: when there are too many segments, fine-grained differences due to *intra* speaker variability are taken into account. On the other hand, BIC will blur difference with the covariance collected during merges. BIC learns covariance from the data obtained by successful (correct) merges. As we go towards a system where only a few merges are necessary, BIC does not have enough data to build a correct estimate for the variance, and does not bode well with a disparate amount of frames in the segments. Additionally, the global merging rule has a tendency of quickly merging narrow-band speakers, because the ratio between variance (consistency) and squared mean difference (coherence) is low. On the other hand, triangulation does not rely on the variance, but rather on a relative position of the centroid. It knows where centroids are, regardless of their intrinsic variability.

System	Test Set	WER
BIC	BN - RT-02	19.5%
Triangulation	BN - RT-02	19.5%
BIC	BN - H498	20.3%
Triangulation	BN - H498	20.3%

Table 2. Word error rates (WER) for speech recognition

On Table 2, we see results on speech recognition. As we can see, in our range of meta data accuracy, there is no difference for speech recognition. This is readily explained by the fact that the error rate is small, and by construction, confusable segments come from sound-alike speakers. Therefore, as far as regression of speakers is the goal, there does not seem to be an advantage using either approach.

5. CONCLUSION AND FURTHER WORK

We have presented two approaches that are portable across domains. The first approach (BIC) employs a blind clustering that is distinguished by its simplicity, specifically in the lack of *a priori* parameters that it requires. To our surprise, it performed very well: we attribute its success to modeling via full covariances matrices of static coefficients. The second approach, called speaker triangulation, builds a coordinate system based on the segments presented to clustering. It is simple and computationally attractive.

Experiments on Broadcast speech and Switchboard show that we can achieve state-of-the-art clustering on recent NIST evaluation test sets with both Broadcast News and conversational telephone speech data. Experiments on speech recognition show that precise meta data may not be crucial for speaker adaptation.

Further work will concentrate on improving clustering specifically for adaptation. Also, the gap in error rate between large and small sets should be reduced. Both methods seem to have their strengths that should be combined.

6. REFERENCES

- [1] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Proc. ICASSP*, 1998, vol. 2, pp. 645–648.
- [2] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc. ICASSP*, 1991, vol. 2, pp. 873–876.
- [3] P. Nguyen, L. Rigazio, Y. Moh, and J.C. Junqua, "Rich Transcription 2002 Site Report, Panasonic Speech Technology Laboratory (PSTL)," 2002.
- [4] R. O. Duda and P. B. Hart, *Pattern Classification and Scene Analysis*, Wiley, 1973.
- [5] S. E. Johnson, "Who spoke when? - automatic segmentation and clustering for determining speaker turns," in *Proc. Eurospeech*, Budapest, Hungary, 1999, vol. 5, pp. 2211–2214.
- [6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," October 2000.
- [7] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S.J. Young, "Segment generation and clustering in the htk broadcast news transcription system," in *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998, pp. 133–137.