WHO SPOKE WHEN? - AUTOMATIC SEGMENTATION AND CLUSTERING FOR DETERMINING SPEAKER TURNS

S.E. Johnson

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
sej28@eng.cam.ac.uk

ABSTRACT

The problem of labelling speaker turns by automatically segmenting and clustering a continuous audio stream is addressed. A new clustering scheme is presented and evaluated using a clustering efficiency score which treats both agglomerative and divisive clustering strategies equally. Results show an efficiency of 70% can be obtained on both manually and automatically derived segments on the 1996 Hub4 development data.

For the task of identifying potentially unknown anchor speakers within broadcast news shows, the frame classification error rate is very important. To reflect this, a frame-based cluster efficiency is defined and the results show a 90% frame-based efficiency can be achieved. Finally a frame-based comparison between the manually and automatically derived segment/cluster sets shows that approximately one third of the errors are introduced during segmentation and two-thirds during clustering.

1. INTRODUCTION

In recent work we have described automatic methods for both segmenting and clustering a continuous audio stream input [2, 4] . These methods were shown to be an important part of our overall recognition system for broadcast news. The segmenter is designed to produce segments of of between 1 and 30 seconds duration which are acoustically homogeneous (i.e. they contain only one speaker and noise/channel condition). The clusterer is designed to place acoustically similar segments into groups (clusters) of a certain minimum occupancy (generally 30 seconds). This allows Maximum Likelihood Linear Regression (MLLR) to be applied, thus improving the overall performance of the recognition system.

In this work, attention is switched to the problem of determining speaker turns when the speakers (and number of speakers) are unknown. The aim therefore is to produce just 1 pure cluster for every speaker, independent of the amount of time the speaker is talking. The previous speaker-adaptation clustering strategy is modified and a new recombination procedure introduced, to reflect this new aim of speaker-identification.

Initially the clustering performance for both these systems is evaluated on an utterance basis. This reflects the task when the user has a database of recorded utterances and wishes to retrieve one from a given speaker as quickly as possible. By presenting perfect speaker clusters the number of utterances the user has to listen to in order to find the appropriate message is dramatically reduced. The clustering efficiency from [5] is used to present the results and it is shown that setting the free parameter can produce results which equate the no-clustering case in both divisive and agglomerative clustering.

The performance measure is then moved from an utterance-level to a frame-level basis. This allows greater emphasis to be placed on longer segments and models the task of tracking (unknown) speakers through a broadcast news show when the user may not be interested in very short utterances. The frame-based approach also allows the separate errors introduced by the segmentation and clustering stages to be quantified.

This paper describes briefly the segmenter and clusterer in section 2, introduces the clustering performance measures and derives formulae for the critical case in section 3, gives experimental details in section 4, presents utterance-based and frame-based results in sections 5 and 6 and offers conclusions in section 7.

2. SEGMENTER AND CLUSTERER

The automatic segmenter splits the input audio on silences and then uses a GMM classifier to label the data as pure (wideband) speech, telephone speech, pure music, or music with speech. The pure music is discarded and the music with speech is combined with the pure speech and labelled wideband. A phone recognition pass, clustering and the application of some heuristic rules then follows to produce segments between 1 and 30 seconds duration labelled by bandwidth. Gender labels are then added based on the likelihood of the data using gender-specific models in a word-level recognition pass. Further details of the segmentation process are given in [2].

The clusterer (described in [4]) represents each segment by a single correlation matrix. The arithmetic harmonic sphericity [1] is used as the distance measure. A top-down split-and-merge algorithm is used for the clustering. Each node is split into 4 child nodes and the new correlation matrices for the child nodes are calculated by concatenating the data within them. The segments are then assigned to the closest child node, the statistics recalculated and the process repeated until no more segments move. This is repeated until all the nodes have been split completely.

It is necessary to define when a split is allowable to prevent the data being split back into its constituent segments. The speaker-adaptation scheme sets a minimum occupancy requirement of 3000 frames (30 seconds) on the final clusters to ensure robust speaker adaptation can follow. For the speaker-identification scheme no such restriction is necessary and alternative stopping criterion must be found. New parameters are introduced which model the minimum gain required from splitting and the maximum level of overlap between child nodes to allow the split to go ahead. Another parameter is added to deal with the special case of singleton clusters where the intra-node distance is zero. By changing these parameters whilst keeping the minimum occupancy required to zero, different levels of recombination for the speaker-identification scheme can be achieved.

3. EVALUATING CLUSTERING PERFORMANCE

In order to judge different clustering algorithms a quantitative measure of performance is required. Large single-speaker clusters should be rewarded whilst multi-speaker clusters and incomplete clustering should be penalised. This discourages both grouping utterances from different speakers together and not grouping utterances from the same speaker. For the metrics used in this paper [1] the following terms are defined:

N_s	Total number of speakers
N_c	Total number of clusters
N_u	Total number of utterances
n_{ij}	# utterances in cluster i from speaker j
$n_j = \sum_i n_{ij}$	# utterances said by speaker j
$n_i = \sum_j n_{ij}$	# utterances in cluster i
$p_i = \sum_j \frac{n_{ij}^2}{n_i^2}$	purity of cluster i

The Rand Index

The first metric used in this paper is the Rand Index [3].

$$I_{RAND} = \frac{1}{2}\left(\sum_i n_i^2 + \sum_j n_j^2\right) - \sum_i \sum_j n_{ij}^2$$

This gives the number of utterance pairs that are from the same speaker and are not in the same cluster or that are from different speakers but are in the same cluster. Smaller I_{RAND} therefore represents a better speaker split, with perfect speaker split having an I_{RAND} of zero.

Clustering "Efficiency"

The second metric used is the clustering efficiency from [5]. This is based on the BBN metric[6]:

$$I_{BBN}(C) = \sum_i n_i p_i - QN_c$$

where Q is a user-defined parameter which represents the trade off between producing a few large clusters which may contain multiple speakers and incomplete clustering where certain speakers may have more than one cluster associated with them.

Clustering Efficiency is then defined in terms of perfect clustering, I(P), and the singleton cluster set, I(S), which represents the case of no clustering for an agglomerative scheme. Note that this value is not a true efficiency as it is possible to obtain a negative value for η.

$$\eta = \frac{I_{BBN}(C) - I_{BBN}(S)}{I_{BBN}(P) - I_{BBN}(S)}$$

For the singleton clusters (each utterance is a cluster):
$p_i = n_i = 1 \;\; \forall i$ and $N_c = N_u$ so $I(S) = N_u(1 - Q)$.
For perfect clustering:
$p_i = 1 \;\; \forall i$ and $N_c = N_s$ so $I(P) = N_u - QN_s$.
With this metric perfect clustering produces a score of 1.0 whilst the singleton cluster set scores 0.0. Note however, that another limit on performance exists, namely grouping all the utterances into 1 large cluster. This may produce a negative efficiency score, depending on the choice of Q.

Choosing Q

Experiments with Q set to 0.5 are reported in this paper to allow comparisons with previous work in this area [6, 5]. However, this gives an efficiency of around -1 for the case of a single cluster for the data used in this paper. It would be nice to have a baseline score of zero for the case of no clustering irrespective of whether the clustering is implemented in a divisive of agglomerative scheme. To achieve this, the value of Q is set to a critical value such that the one-cluster case also has a cluster efficiency, $I(1)$, of zero:
For 1 cluster: $N_c = 1$; $n_i = N_u$ for $i = 1$ and 0 otherwise
hence: $I(1) = \left(\frac{1}{N_u}\sum_j n_j^2\right) - Q$
hence setting $I(1) = I(S)$ so that $\eta(1) = 0$ gives:

$$Q_{crit} = \frac{N_u^2 - \sum_j n_j^2}{N_u(N_u - 1)}$$

Note that $0 \le Q_{crit} \le 1$ since:

$$N_u = \sum_j n_j \le \sum_j n_j^2 \le \left(\sum_j n_j\right)^2 = N_u^2$$

For the experiments reported in this paper, $Q_{crit} \approx 0.95$.

4. EXPERIMENTS

The 1996 Hub-4 Broadcast News Transcription development data was used for all the experiments reported in this paper. An initial set of segments was generated and labelled by bandwidth and gender (either manually or automatically). The clusterer was then run on a gender-dependent, bandwidth-dependent basis and the final cluster set was formed by concatenating the sets from the 4 bandwidth/gender conditions.

Results are presented for the cases of one overall cluster, (one_c), singleton clustering, (singleton_c), perfect clustering, (perfect_c), speaker-adaptation clustering, (adapt_c), and two speaker-identification systems (speak_1_c, speak_2_c). The speaker-adaptation scheme is that used in our overall recognition system before speaker

[1] Further theoretical justification for using these two metrics can be found in [6].

adaptation [4], whilst the speaker-identification systems use the scheme described in section 2 with different levels of recombination. The automatic segmentation is done using our 1997 segmenter [2].

Section 5 reports the results using the utterance-based metrics described in section 3 for the cases of Q=0.5 and $Q = Q_{crit}$. Section 6 uses the same cluster sets but gives the results on frame-based metrics.

5. UTTERANCE CLUSTERING

5.1. Perfect Segmentation

This experiment uses the manually generated segments. In this case each of the 488 segments contains only 1 speaker and has no length restrictions. [2]

Condition	N_c	I_{RAND}	η_{BBN}	η_{BBN}
			Q=0.5	Q=0.949
one_c	1	112807	-1.065	0.000
singleton_c	488	6021	0.000	0.000
adapt_c	81	5376	0.336	0.646
speak_1_c	92	4286	0.476	0.707
speak_2_c	165	4937	0.464	0.616
perfect_c	77	0	1.000	1.000

Table 1: Utterance-based clustering performance for perfect segments

The results given in Table 1 show that when considering Q=0.5, the basic speaker-adaptation scheme used in the speech recognition process offers reasonable speaker groupings from these manually derived segments. However, as expected, performance can be increased by switching to a speaker-identification scheme by removing the occupancy constraint and increasing recombination to compensate for the increased number of clusters. Note that the speak_1_c scheme, which has more recombination than speak_2_c, performs better.

It is interesting to note that the results for the critical value of Q show a slightly different pattern, namely that the speaker-adaptation scheme scores higher than the speak_2_c scheme, due to the smaller number of clusters.

5.2. Automatic Segmentation

This experiment was repeated using the automatically generated segments. This still represents the same task of utterance clustering, but assumes no boundary information is given and thus the utterances have to be generated from automatically segmenting the audio stream. The assumption of there being only one speaker in any given utterance is now, in general, false but the clustering performance can still be measured by assuming the dominant speaker is the only speaker of interest in an utterance. Also note that since the clustering is done on a bandwidth and gender dependent basis, the starting groups for clustering may have changed due to classification errors in the segmenter. The results are given in Table 2

Condition	N_c	I_{RAND}	η_{BBN}	η_{BBN}
			Q=0.5	Q=0.956
one_c	1	145850	-1.037	0.000
singleton_c	553	6778	0.000	0.000
adapt_c	106	6309	0.380	0.638
speak_1_c	119	4999	0.506	0.691
speak_2_c	151	5144	0.485	0.649
perfect_c	68	0	1.000	1.000

Table 2: Utterance-based clustering performance for automatically generated segments

These results are very similar to the manually-segmented case and show the same trends, namely that speaker-adaptation clustering gives a reasonable performance in the utterance-clustering task, but the performance can be increased further by switching to the speaker-identification scheme. The approximation that each segment only contains the dominant speaker does not seem to affect the results unduly.

6. FRAME-BASED SCORING

In order to be able to look at the relative effects of automating both the segmentation and the clustering on the overall performance, the definition of the scoring metric must be redefined to work on a frame basis. [3] This also reflects the true performance on certain tasks more accurately than utterance-based metrics. For example, for the identification of a (potentially unknown) anchor speaker in a broadcast news show, an error with a long utterance may be more significant than an error with a shorter utterance. A new frame-based efficiency is therefore defined. The previous formulae remain the same but the definitions are altered to:

N_f	Total number of FRAMES
n_{ij}	# FRAMES in cluster i from speaker j
$n_j = \sum_i n_{ij}$	# FRAMES said by speaker j
$n_i = \sum_j n_{ij}$	# FRAMES in cluster i

and the resulting baseline cases become:

perfect clustering:	$I^F(P) = N_f - QN_s$
singletons (each frame separate):	$I^F(S) = N_f(1-Q)$

$$\text{with } \; Q^F_{crit} = \frac{N_f^2 - \sum_j n_j^2}{N_f(N_f-1)}$$

The results for the cluster sets given in section 5 recalculated using this frame-based score are given in Tables 3 and 4 for the manual and automatically derived segments respectively. The frame rate was 100Hz and the number of frames after segmentation was approximately 600,000.

These results show that for manual segmentation a very high frame-based clustering efficiency of 81% (90% for Q_{crit}) can be obtained from automatic clustering. For the automatic segmentation compared to the automatic baseline, the results are almost identical to the manual case, with a frame-based efficiency of 79% (89% for Q_{crit}).

[2] This represents the scenario of retrieving a spoken message from a database when the speaker is known[6].

[3] As the initial segments are not the same for the manual and automatic case.

Condition	N_c	I^F_{RAND}	η^F_{BBN}	η^F_{BBN}
			Q=0.5	Q=0.951
one_c	1	1.664e+11	-0.902	0.000
singleton_c	591639	8.611e+09	0.000	0.000
adapt_c	81	6.382e+09	0.585	0.782
speak_1_c	92	4.355e+09	0.673	0.828
speak_2_c	165	4.648e+09	$\boxed{0.811}$	0.900
perfect_c	77	0	$\boxed{1.000}$	1.000

Table 3: Frame-based clustering performance for perfect segments

Condition	N_c	I^F_{RAND}	η^F_{BBN}	η^F_{BBN}
			Q=0.5	Q=0.951
one_c	1	1.795e+11	-0.902	0.000
singleton_c	614510	9.288e+09	0.000	0.000
adapt_c	106	7.962e+09	0.672	0.827
speak_1_c	119	5.785e+09	0.762	0.875
speak_2_c	151	5.943e+09	0.789	0.889
perfect_c	68	0	1.000	1.000

Table 4: Frame-based clustering performance for automatically generated segments (compared to the automatic segments)

6.1. Overall Error Analysis

Using the new frame-based efficiency the errors introduced by automating both the segmentation and clustering stages can be quantified. Segmentation can introduce errors by getting the boundaries wrong for speaker changes in the audio, wrongly detecting and discarding pure music/non-speech events or misclassifying the gender/bandwidth of the data. Since the clustering is done separately for the 4 possible permutations of bandwidth and gender, it is impossible to recover from any errors in the classification part of the segmenter. Errors in clustering are due to combining segments which are not from the same speaker or not combining segments which are from the same speaker.

The score from comparing the frame labels from the automatically segmented/clustered set with the perfectly segmented/clustered set are given in Table 5. Note that the Rand Index for the perfect case and the cluster efficiency at the previous value of Q_{crit} are no longer zero, due to errors introduced in the segmenter when non-speech events are removed.

Condition	N_c	I^F_{RAND}	η^F_{BBN}
one_c	1	1.641e+11	-0.902
singleton_c	614510	8.551e+09	0.000
adapt_c	106	7.394e+09	0.620
speak_1_c	119	5.525e+09	0.699
speak_2_c	151	5.649e+09	$\boxed{0.723}$
perfect_c	68	5.859e+08	$\boxed{0.892}$

Table 5: Frame-based clustering performance for automatically generated segments compared to the perfect segmentation (Q=0.5 or 0.95)

These results for cluster efficiency (Q = 0.5) are summarised in Table 6. After the automatic segmentation has occurred, the perfect speaker clustering results in 89.2% efficiency (as compared to the manually generated speaker baseline). Automatic clustering of these segments results in this number falling to 72.3%. Note the drop due to automatic clustering is around 18% for both the manual and automatic segmentation confirming that the errors introduced in segmentation are largely independent of those made in clustering.

Segmentation:	Manual	Automatic
Manual Clustering	1.000	0.892
Automatic Clustering	0.811	0.723

Table 6: Summary of Results for Q=0.5

7. CONCLUSIONS

These results show our methods of automatic segmentation and clustering produce a frame-based efficiency of 72.3% on the 1996 Hub4 development data. The loss of 27.7% from the perfect case is 39% due to errors in the automatic segmentation process, and 61% from the clustering procedure.

The concept of clustering efficiency has been extended to score divisive and agglomerative clustering schemes evenly, and a new frame-based scheme has been introduced. It is interesting to note that the relative performance of both the speaker-identification clustering schemes and the speaker-adaptation scheme depends on which definition of efficiency is used. In tasks such as following (potentially unknown) speakers through broadcast news shows, where frame error is more important, the speaker-identification system with moderate recombination performs the best. It is clear that to get optimal performance in segregating speakers, the task must be clearly defined before deciding how to run the clusterer.

Acknowledgements

This work is in part supported by an EPSRC grant on “Multimedia Document Retrieval” reference GR/L49611.

8. REFERENCES

1. F Bimbot & L Mathan. *Text-Free Speaker Recognition using an Arithmetic Harmonic Sphericity Measure.* Proc. Eurospeech, 1993, Vol. 1. pp. 169-172
2. T Hain, S E Johnson, A Tuerk, P C Woodland & S J Young. *Segment Generating and Clustering in the HTK Broadcast News Transcription System.* Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, pp. 133-137
3. L Hubert & P Arabie. *Comparing Partitions* Journal of Classification, 1985, Vol. 2 pp. 193-218
4. S E Johnson & P C Woodland. *Speaker Clustering Using Direct Maximisation of the MLLR-Adapted Likelihood.* Proc. ICSLP’98 Vol. 5 pp. 1775-1779
5. D A Reynolds, E Singer, B A Carlson, G C O’Leary, J J McLaughlin & M A Zissman. *Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics* Proc. ICSLP’98 Vol. 7 pp. 3193-3196
6. A Solomonoff, A Mielke, M Schmidt & H Gish. *Clustering Speakers by their Voices* Proc. ICASSP’98 Vol. 2 pp. 757-760