OCR-aided person annotation and label propagation for speaker modeling in TV shows

Mateusz Budnik, Laurent Besacier, Ali Khodabakhsh, Cenk Demiroglu

OCR-AIDED PERSON ANNOTATION AND LABEL PROPAGATION FOR SPEAKER MODELING IN TV SHOWS

Mateusz Budnik, Laurent Besacier

Laboratoire d'Informatique de Grenoble
Univ. Grenoble-Alpes
Grenoble, France

Ali Khodabakhsh, Cenk Demiroglu

Electrical and Computer Engineering Department
Ozyegin University
Istanbul, Turkey

ABSTRACT

In this paper, we present an approach for minimizing human effort in manual speaker annotation. Label propagation is used at each iteration of an active learning cycle. More precisely, a selection strategy for choosing the most suitable speech track to be labeled is proposed. Four different selection strategies are evaluated and all the tracks in a corresponding cluster are gathered using agglomerative clustering in order to propagate human annotations. To further reduce the manual labor required, an optical character recognition system is used to bootstrap annotations. At each step of the cycle, annotations are used to build speaker models. The quality of the generated speaker models is evaluated at each step using an i-vector based speaker identification system. The presented approach shows promising results on the REPERE corpus with a minimum amount of human effort for annotation.

Index Terms— active learning, annotation propagation, clustering, speaker identification, OCR

1. INTRODUCTION

Thanks to the widespread availability of TV and the Internet, an immense quantity of videos are now available and can be used as a source for extracting a variety of useful information. However, due to complex nature of such data, annotation and indexing of these data can be costly and requires intensive human labor, which makes it impossible to be done at very large scale. To speed-up the process, several active learning methods are being developed to determine the most suitable instances/tracks for annotation, and by doing so, reducing the need for manual annotation [1].

Effective methods for finding the most informative samples for labeling have been an active field of research. An unsupervised algorithm for active learning is presented in [2] addressing some of the drawbacks of supervised active learning. An example of those is inability of selecting samples from new categories, which can be problematic in person identification tasks on broadcast data due to high number of distinctive classes with small number of instances [3]. A semantic approach of annotation propagation is proposed in [4], and a person name propagation algorithm is proposed in [5] for video data. It is observed that the final score for person identification can be higher than with a SVM trained on the same initial labeled data. The information of the cluster structure is used alongside the selection of the most representative samples in [6] to avoid sampling the same cluster.

In this paper, as an extension of the previous work done by the authors in [7], the effectiveness of our active learning and annotation propagation methods is evaluated for a true speaker identification task. The method provides a cheap solution for creating reference speaker models from raw video input. For this, we make use of a classifier's output on clustered speech data to determine the most informative samples for annotation [8]. This annotation is then propagated throughout its corresponding clusters. Reclustering is done afterwards and then a new sample is selected for annotation. To bootstrap the labeling process, Optical Character Recognition (OCR) can be used on overlaid text of broadcast data (e.g. when a person is shown on screen and his/her name is shown at the bottom of the screen).

Our goal here is to investigate how much annotation is needed to produce an accurate speaker model. Additionally, the comparison is made between models that were trained in an unsupervised manner (with the only source of labels being the overlaid names present in the video) and those that gradually use more human annotation, i.e. they become more and more supervised. The quality of the generated speaker models is evaluated through speaker identification tests on a separate set of unseen data.

A previously developed unsupervised system for label propagation [9] was used and it is shortly described here. In our work the active learning module is added on top of that unsupervised system. The results indicate the advantages of using both overlaid names for the cold start and propagation of annotation within clusters. Also, to obtain a model with good enough performance, it is shown that one does not necessarily have to label all the data.

The rest of the paper is organized as follows: Section 2 describes the overall design and proposes in-detail description

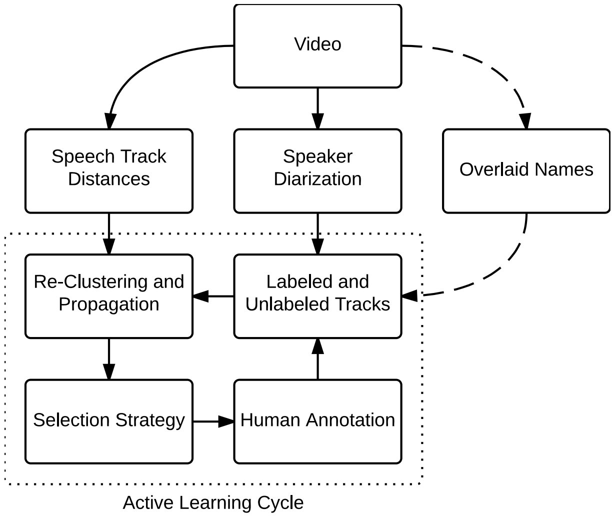

Fig. 1. Overview of the system structure with optional use of extracted overlaid names.

of the used systems. The experimental setup and results are presented in Section 3. Section 4 concludes the paper.

2. SYSTEM DESCRIPTION

An overview of the system used in this study is shown in Figure 1. After the extraction of speaker tracks (which are equivalent to segments and in this work both names are used interchangeably), the distances between tracks are calculated and agglomerative clustering is done. In the case of usage of OCR labels for cold start, OCR is applied on the videos and overlaid names are extracted. Then, the clusters are labeled with initial labels obtained accordingly. Next, the active learning (AL) cycle (pictured in Figure 1) is introduced by choosing a sample for labeling using different selection strategies (random, chronological, longest, and biggest cluster probability). Once the new label is obtained, we recalculate the clusters and propagate the annotation to the instances in the same cluster. During this step, some clusters may be combined and new clusters may be created. The new and modified cluster structure is then used as a starting point for the next cycle of active learning. Details about each component of this process are given in the next sub-sections.

2.1. Speaker Diarization and Speech Track Distance Calculation

After splitting the signal into acoustically homogeneous segments with the use of the conventional BIC-criterion [10], a similarity score matrix is calculated between each pair of speech tracks. Single full-covariance Gaussians are used and distances are normalized to have a maximum of 1. This is an automated process, which may lead to some of the speaker tracks containing a degree of background noise, music or silence.

2.2. Overlaid name extraction

In order to be able to automatically extract the overwritten names from video, an OCR system following the design proposed in [11] is used. Usually guests and speakers are introduced on a television show in their first appearance with an overlaid text containing their name. Using OCR, the texts visible on the screen are extracted, however, it is important to note that not all of the visible texts contain the name of the speaker.

This module is composed of a text detection step and a text recognition step. For text detection, the approach in [12] is adopted where coarse detection is obtained using Sobel filter and dilatation/erosion. To filter out the false positive text boxes, several temporally shifted images are extracted for the same text. This allows the system to overcome the shortcoming of binarization.

When text boxes are extracted, the publicly available `Tesseract`[1] OCR system developed by Google is applied for text recognition.

2.3. Selection Strategies and Re-clustering

During a single AL cycle, one track is selected for each show of the train collection. Four different selection strategies are applied in this study.

Random
: This strategy chooses a random track to be annotated for every show; this method is used as a baseline.

Chronological
: This strategy chooses the track to be annotated chronologically for a given show.

Longest
: For this strategy, the longest track is chosen for manual annotation. The tracks that are already annotated (or that benefit from label propagation) are left out.

Biggest Cluster Probability
: In this strategy, a track from a cluster with a low annotation ratio is chosen for annotation. For this, the score S_c is calculated for each cluster as $S_c = \frac{a_t}{n_t}$, where a_t is the number of annotated tracks in the cluster and n_t is the total number of tracks in the cluster. This score is used to bias a random variable for which tracks in clusters with low annotation ratio have higher probability to be selected.

2.4. Speaker Identification Systems

An automatic speaker identification system based on front-end factor analysis [13], also known as Total Variability Space (TVS), is chosen for this study.

Gaussian Mixture Models (GMM) are typically used as a generative model for representing the acoustic feature space

[1] http://code.google.com/p/tesseract-ocr/

in speaker recognition systems. A Universal Background Model (UBM) is first trained on speech data from multiple speakers, and then speaker-specific models are obtained using Maximum a Posteriori (MAP) adaptation.

Speaker models can be represented as high-dimensional supervectors of means of Gaussian distributions. Using probabilistic factor analysis on these supervectors, speaker models can be represented as a low-dimensional identity vector (i-vector). In this approach, the mean supervector of a speaker s is modelled with $M_s = M_0 + Tw_s$, where M_0 is the mean supervector of the UBM, w_s is the i-vector of speaker s, and T is the low-rank rectangular matrix representing the variability space of the i-vectors. In case of having multiple tracks for speaker modeling, i-vectors extracted from each speech track are averaged and the average i-vector is used as the speaker model. The averaging can be done in a weighted manner according to the logarithm of the duration of the tracks. Duration of the tracks in seconds is incremented by one prior to taking logarithm to avoid negative weights. Length normalization is done prior to generation of speaker models [14].

Identification is done by extraction of the i-vector of a target track and calculation of the cosine similarity score of the extracted i-vector with all the speaker models (each represented with an i-vector). The speaker corresponding to the model with the highest score is chosen as the identified speaker.

For feature extraction, Energy and Mel-Frequency Cepstral Coefficients (MFCCs) of 13 dimensions are extracted every 10 ms with a window length of 20 ms. These features along with their delta and delta-delta coefficients are concatenated. Feature warping [15] is done on these features and static energy feature is discarded resulting in a 41 dimensional feature vector per frame. Sound activity detection is done using bi-gaussian distribution on frame log-energies. The segmentation is done in a similar manner as described for speaker diarization. Segments with less than 150 ms of voice activity are ignored.

A UBM consisting of 1024 gaussians is trained on the training data for the GMM-UBM system. A UBM of the same size is used for training the T matrix. Total variability space is then learned using Expectation-Maximization (EM) algorithm on the segmented training data. The dimension of the output i-vectors is set to 400 and weighted averaging is done for generation of speaker model i-vectors. A modified version of MSR Identity Toolbox [16] is used for the experiments.

3. EXPERIMENTS

3.1. Database and Protocols

The REPERE corpus [17] was used for evaluation of the annotation propagation methods. This corpus consists of seven types of shows from the French TV channels BFM-TV and LCP, and is aimed for development of person identification methods on broadcast data. The train set consists of 58 videos and has a total duration of ˜28 hours. The development set and test set, with a total annotated duration of ˜8 hours and ˜13 hours, containing 57 and 90 videos respectively, are both used for evaluation. Only parts of the videos corresponding to specific shows are annotated. Length of these series differ, ranging from 3 minutes to half an hour, and therefore, the number of annotations per video vary widely from twenty to several hundreds.

3.2. Experimental Setup

In this study, simulated active learning is done on the training set, where all the manual annotations are available initially, but considered as unknown to the system. The labels are revealed to the system as the selection algorithm selects them for annotation. The simulation is done for the duration of 20 AL cycles and on each cycle only one annotation per show is selected. For each selection strategy, there are 10 replicas and randomly 20% of the data is completely left-out for each replica. Four different selection strategies are evaluated.

For every repetition of each AL cycle using each selection strategy, speaker models are created based on the output annotations of the active learning system (on the training data). The speaker identification performance is evaluated on the development and test sets separately using the standard F-measure. However, due to the fact that all the data is not annotated, the performance is only evaluated on the annotated tracks. Number of annotated tracks used in this study are 3490 and 4779 for development and test respectively.

Tests are done in an open-set manner, and results with correct labels corresponding to the maximum possible F-measure as well as results with a fully supervised system are reported for comparison. Tests with and without OCR cold start are also reported.

3.3. Results and discussion

Figure 2 and 3 give the F-measure results with the standard deviation starting with no available annotation. With a good enough selection strategy one can get close to the fully supervised performance in just a few steps. Overall, the best approach is to select the longest track first, which is not surprising given that a long uninterrupted speech segment from a single speaker would lead, in most cases, to a good speaker model.

The results with the overlaid names as a cold start can be seen in Figure 4 and 5 for the development and test set, respectively. This approach gives an initial boost in performance, which can be further increased with just a few additional annotations. Regardless of the use of the OCR, both approaches are able to arrive at the same level of performance (especially in the case of the longest strategy), even though the use of OCR makes it faster for every selection approach. This would indicate that the approach presented here could be successfully employed on different datasets where the OCR may not always be available. The weak performance of the

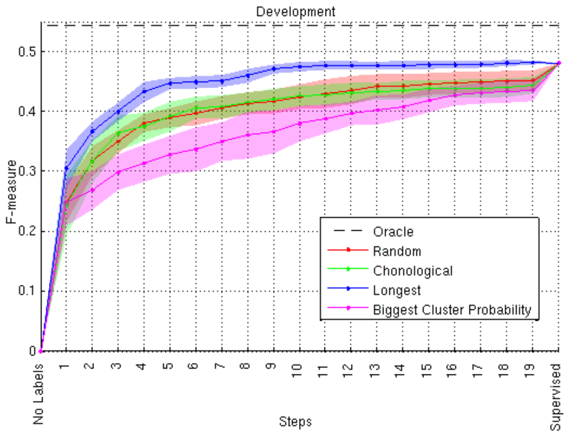

Fig. 2. Performance of the speaker identification system on the development data in terms of F-measure. The results with supervised speaker modeling as well as maximum possible F-measure in the open-set setup are also reported for comparison.

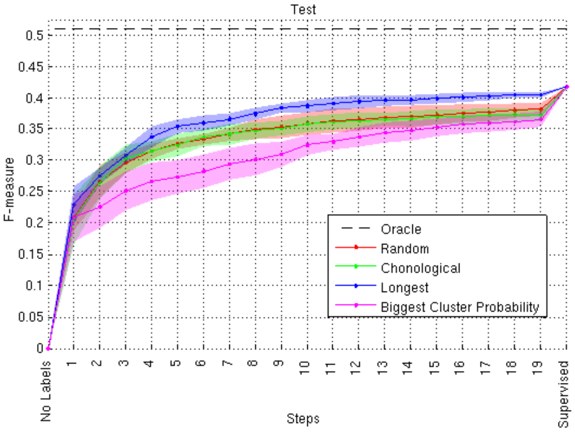

Fig. 3. Performance of the speaker identification system on the test data in terms of F-measure. The results with supervised speaker modeling as well as maximum possible F-measure in the open-set setup are also reported for comparison.

biggest cluster strategy can partially be attributed to the noisiness of the initial clusterings.

After around 6 steps (when using OCR) the performance increases only slightly. Without OCR it takes around 9 steps for the best strategy to get to a comparable level. It seems, therefore, that the use of such an active learning system can greatly reduce the number of annotation needed to produce competitive speaker models. Manual annotation is often expensive and time consuming. This approach can help to reduce this burden, especially when the final goal is to have reliable speaker models.

4. CONCLUSION

In this paper the evaluation of an active learning system is presented. The main goal was to train speaker models using

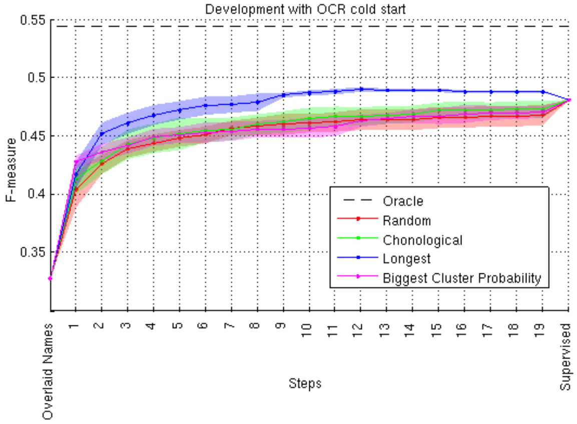

Fig. 4. Performance of the speaker identification system on the development data with use of overlaid names for initialization in terms of F-measure. The results with supervised speaker modeling as well as maximum possible F-measure in the open-set setup are also reported for comparison.

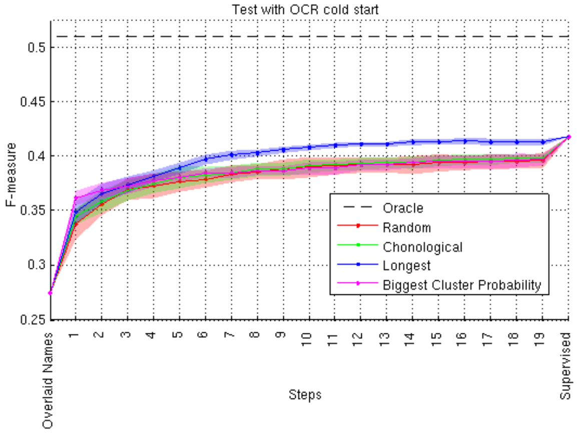

Fig. 5. Performance of the speaker identification system on the test data with use of overlaid names for initialization in terms of F-measure. The results with supervised speaker modeling as well as maximum possible F-measure in the open-set setup are also reported for comparison.

i-vectors and trying to determine if a similar performance can be achieved without fully annotating the dataset. The presented experiments indicate that when using a good selection strategy a comparable performance can be obtained with just a fraction of the labels.

Additionally, the use of automatic labels, extracted using an OCR system, is evaluated. As seen in the experiments, OCR names can further reduce the need for manual annotation.

Acknowledgements: This work is a part of a project aiming to create a collaborative annotation framework for 3M (multimodal, multimedia, and multilingual) data, called CAMOMILE[2] project.[3]

[2] http://camomile.limsi.fr

[3] This work was conducted as a part of the CHIST-ERA CAMOMILE

5. REFERENCES

[1] Stephane Ayache and Georges Quenot, "Evaluation of active learning strategies for video indexing," *Signal Processing: Image Communication*, vol. 22, no. 7, pp. 692–704, 2007.

[2] Weiming Hu, Wei Hu, Nianhua Xie, and Steve Maybank, "Unsupervised active learning based on hierarchical graph-theoretic clustering," *IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics*, vol. 39, no. 5, pp. 1147–1161, 2009.

[3] Tanzeem Choudhury, Brian Clarkson, Tony Jebara, and Alex Pentland, "Mulitmodal person recognition using unconstrained audio and video," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, 1999, pp. 176–181.

[4] Gilberto Zonta Pastorello, Jaudete Daltio, and Claudia Bauzer Medeiros, "Multimedia semantic annotation propagation," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, 2008, pp. 509–514.

[5] Phi The Pham, Tinne Tuytelaars, and Marie-Francine Moens, "Naming people in news videos with label propagation," *IEEE Multimedia*, vol. 18, no. 3, pp. 44–55, 2011.

[6] Hieu T Nguyen and Arnold Smeulders, "Active learning using pre-clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, pp. 79–86.

[7] Mateusz Budnik, Johann Poignant, Laurent Besacier, and Georges Quénot, "Active selection with label propagation for minimizing human effort in speaker annotation of tv shows," in *Workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2014, pp. 5–9.

[8] Bahjat Safadi and Georges Quenot, "Active learning with multiple classifier for multimedia indexing," *Multimedia Tools and Applications*, vol. 60, pp. 403–417, 2010.

[9] Johann Poignant, Hervé Bredin, Laurent Besacier, Georges Quénot, and Claude Barras, "Towards, a better integration of written names for unsupervised speakers identification in videos," in *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2013, pp. 84–89.

[10] Scott Chen and Ponani Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. Virginia, USA, 1998, p. 8.

[11] Johann Poignant, Laurent Besacier, Georges Quénot, and Franck Thollard, "From text detection in videos to person identification," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*. IEEE, 2012, pp. 854–859.

[12] Marios Anthimopoulos, Basilis Gatos, and Ioannis Pratikakis, "A two-stage scheme for text detection in video images," *Image and Vision Computing*, vol. 28, no. 9, pp. 1413–1426, 2010.

[13] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[14] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Interspeech*, 2011, pp. 249–252.

[15] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," 2001.

[16] Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker recognition research," *Speech and Language Processing Technical Committee Newsletter*, 2013.

[17] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The repere corpus: a multimodal corpus for person recognition.," in *LREC*, 2012, pp. 1102–1107.

project, which was funded by the ANR (Agence Nationale de la Recherche, France) and by the TÜBİTAK (TÜrkiye BİlimseI ve Teknolojik Araştırma Kurumu, Turkey) grant No 112E176.