A Visual Analytics Approach to Finding Factors Improving Automatic Speaker Identifications

Pierrick Bruneau
LIST
L-4362 Esch-sur-Alzette
pierrick.bruneau@list.lu

Mickaël Stefas
LIST
L-4362 Esch-sur-Alzette
mickael.stefas@list.lu

Hervé Bredin
LIMSI - CNRS
F-91405 Orsay
herve.bredin@limsi.fr

Johann Poignant
LIMSI - CNRS
F-91405 Orsay
johann.poignant@limsi.fr

Thomas Tamisier
LIST
L-4362 Esch-sur-Alzette
thomas.tamisier@list.lu

Claude Barras
LIMSI - CNRS
F-91405 Orsay
claude.barras@limsi.fr

ABSTRACT

Classification quality criteria such as precision, recall, and F-measure are generally the basis for evaluating contributions in automatic speaker recognition. Specifically, comparisons are carried out mostly via mean values estimated on a set of media. Whilst this approach is relevant to assess improvement w.r.t. the state-of-the-art, or ranking participants in the context of an automatic annotation challenge, it gives little insight to system designers in terms of cues for improving algorithms, hypothesis formulation, and evidence display. This paper presents a design study of a visual and interactive approach to analyze errors made by automatic annotation algorithms. A timeline-based tool emerged from prior steps of this study. A critical review, driven by user interviews, exposes caveats and refines user objectives. The next step of the study is then initiated by sketching designs combining elements of the current prototype to principles newly identified as relevant.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Speaker identification; Visual Analytics

1. INTRODUCTION

Identifying speakers in audio signal, e.g., phone records, radio and TV broadcasts is an active field of research [4]. Speaker recognition systems generally proceed in a succession of steps: the audio signal is first segmented in speech turns, then feeding a clustering algorithm (i.e., diarization step). These clusters are then labelled using supervised speaker models. Obtained clusters are finally relabelled by matching them to a bank of speaker models trained supervisedly (a short review of such systems is given in [4]).

Annotations associate meta-data to segments in media. Such meta-data can be of any type, potentially complex [11]. The scope of this paper is limited to speaker annotations, i.e., meta-data restricted to speaker names, following the pattern *Firstname_LASTNAME*. Each name embodies a class (or equivalently category). Annotations are gathered in tracks characterized by their nature, e.g., reference manually annotated ground truth, or a hypothesis resulting from a given speaker identification algorithm. The difference between a hypothesis and the respective reference track can then be seen as an additional annotation track taking values in 0 and 1, respectively for error and success.

Established evaluation campaigns [6, 8] have supported notable improvements in algorithmic performance. Whilst ranks in evaluations and challenges rely on global error rates, researchers have also devoted efforts to more thorough performance analysis, with a view to understand factors leading to effective recognition. For example, the influence of speaker-specific properties (e.g., speech style) and acoustic conditions has been investigated [7].

Supportive evidence is then generally hard to build, and exploring hypotheses is costly. Interactive tools are expected to alleviate this bottleneck. The purpose of this paper is to consider the performance analysis task with a visual analytics approach. This process aims at harnessing recent progress in performance analysis, with visual and interactive tools enabling fast hypothesis formulation and verification. We address this problem within a design study framework, that iterates capturing requirements, agile implementations and evaluation steps [12].

Initial requirements led to proposing a tool to facilitate the differential analysis of algorithmic results with respect to a reference ground truth [3]. This early contribution is summarized in Section 2 among other related work. The first contribution of this work is to put the prior steps of our design study into perspective. The critical view in Section 3 relies on user interviews and a recent study in performance analysis [4]. From this critical view, sketches for the next step of the design study are proposed in Section 4. A summary of the contribution, along with perspectives for future work, conclude the paper in Section 5.

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.

DOI: http://dx.doi.org/10.1145/2818346.2820769.

2. RELATED WORK

The most natural way to visualize annotations for researchers in multimedia processing is to display them as glyphs in tracks of a timeline, marking punctual or time-spanning events. For instance, this view is prevalent in professional video editing software such as Adobe Premiere [1], and in existing tools for scholar media annotations edition and visualization. For example, some tools use this view to facilitate input and inspection of complex multimodal annotations [2, 11]. However these approaches do not concern themselves with difference between annotation tracks, and analyzing their distribution. StoViz [5] uses a timeline in the context of multimodal algorithmic errors for story arcs identification. Segmentation errors were emphasized by a dual use of animation and color-coding.

To our knowledge, the analysis of scholar annotations has not much been considered by the information visualization community. Yet annotations can be seen as categorical data, and models for handling and processing such data are classically based on tables of frequencies, also known as contingency tables [14]. Actually, confusion tables (see Section 4) are specific instances of contingency tables in the context of classifier output. Log-linear models (among which are logistic regression and ANOVA models) are then central tools for analyzing patterns of dependence or independence between variables. Contributions in visual and interactive support to these models [14] strengthen their relevance as building blocks for this paper.

ROC curves are a classical way to visualize classifier quality, but are restricted to binary classification, and lack interaction. Alternatively, Kapoor et al. bring interaction by allowing interactive adaptation of a classifier [10]. Often classifiers have probabilistic outputs translated to classes using thresholds, intuitively interpreted as costs of misclassification. The approach uses a confusion matrix that can be directly modified. Alpha-blended colors indicate the positive or negative influence associated to potential modifications. This principle inspired the interactive equation presented in Section 4. However in this paper misclassification costs are uniform across speakers and the objective is not so much on directly improving a classifier, than on locating speakers and media subject to error, and help hypothesis formulation in an outer step.

Following initial discussions with two users expert in speaker annotation algorithms, a video playback interface has been synchronized with a timeline view [3]. For supporting monomodal speech processing a waveform view could also have been considered. But the additional information regarding speaker identities may be useful for users, and open the way for handling multimodal data and algorithms.

Glyph widths in Figure 1 encode respective time segments, and their colors encode the associated meta-data (i.e., either speaker name or recognition success). For speakers, an automatic categorical color palette is used (see [3] for details), and difference between hypothesis and reference is mapped to a polar color scale to efficiently denote success and error. Figure 1 illustrates these features using data from the REPERE challenge [6].

Total durations for each category can be inspected in summary charts (i.e., pie chart, bar chart or treemap, see Figure 1b). A Focus+Context navigation bar supports rectangular selection, enabling zooming in the medium for local inspection in the displayed focus, while recalling the general distribution of annotations with grey-shaded glyphs (see Figure 1a), and instantly adapting displayed summary views to the visible annotations. Hovering over glyphs in the timeline reveals the meta-data and respective timings. Clicks on timeline glyphs command the playback position, enabling inspections of the associated medium content.

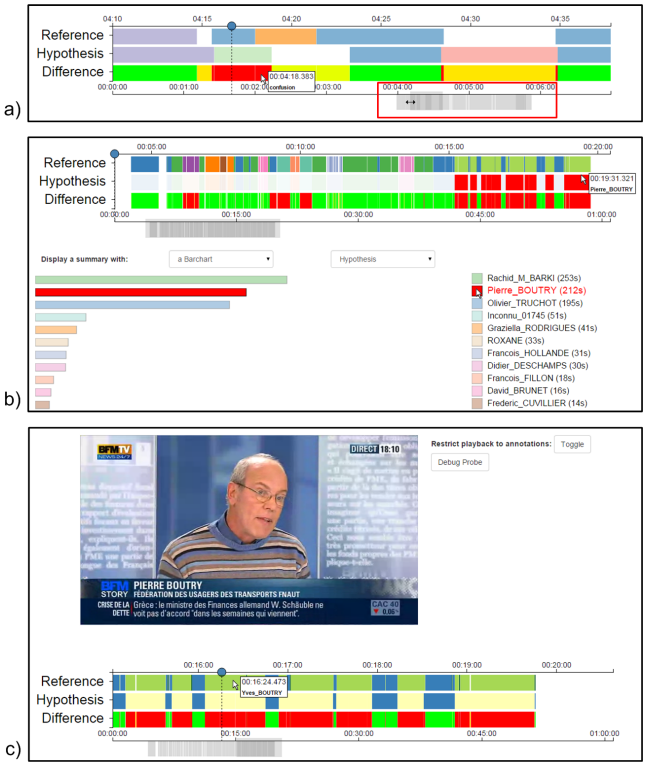

Figure 1: a) Segmentation problems in the hypothesis track are revealed. The red rectangle highlights the context bar at the bottom. The rectangular selection can be moved and its lateral bounds modified interactively, updating the focus accordingly. b) The interactive summary view (here bar chart) highlights associated annotations in the timeline. c) Joint use of the video playback and the timeline helps checking the validity of errors.

Summary views can be used to navigate in the set of speakers, e.g., identifying speakers strongly associated to errors. In the case shown in Figure 1b, it is clear that inferring speaker *Pierre_BOUTRY* is always leading to errors. Using the navigation bar to narrow down the view and seeking in the medium reveals this inference is actually correct, the homonym speaker *Yves_BOUTRY* being mistakenly recorded in the ground truth (see Figure 1c).

3. CRITICAL ANALYSIS

In the context of challenges such as REPERE [6], prior to being ranked, participants have the opportunity to flag mistakes in the reference track: this is the adjudication phase. Confronting users to the software proposed in [3] qualitatively revealed its ability to spot small-scale patterns, and more generally its fit as support to this adjudication activity.

Though the zooming feature may for example emphasize speech turn segmentation problems (see Figure 1a), that were shown as positively correlated with error in speaker diarization and identification [4], discussions established the timeline-based view as rather ineffective when it comes to hypothesizing factors influencing speaker recognition.

The adequate tool must allow to take a wider scope, as actually does a recent batch analysis of annotation errors [4]. For example, the error distribution in supervised speaker identification systems is there shown as highly speaker-specific, i.e., most often a speaker is either always or never correctly recognized [4]. The authors also found this distribution to be even more bimodal when using a manually curated segmentation in speech turns.

Several speaker characteristics have been hypothesized as influential to unsuccessful recognition, either *internal* (i.e., computed from data available in the corpus, e.g., total and average duration of speech turns, pause between speech turns) or *external* (e.g., amount of supplemental data for training respective dictionary speaker models). A post-hoc analysis was then performed by learning a decision tree classifier evaluating the influence of these characteristics on identification performance for the respective speakers. This process defines a hierarchy of importance among explanatory variables and led to conclude on the prevalence of the total duration of speech turns for effective recognition.

This study helped the users identify the true objective that an interactive tool must pursue, i.e., finding characteristics in media and speakers that influence, positively or negatively, their effective recognition. Moreover, considering a collection of annotated media, to date this analysis has been performed on each medium separately. The advantage of using a whole corpus of data holistically has to be explored.

Specifically, media, tracks and speakers would then be seen as independent dimensions. Analyses would then proceed by selections and projections on these axes, with adequate interactive filtering tools. For example, within this framework, the timeline view discussed in Section 2 can be seen as combined projections and selections on the media and tracks axes. Building upon the observations above, the next section describes simulated scenarios of use that serve to introduce new adequate statistical and interactive tools in context.

4. DESIGN SKETCHING

In this section we assume only one hypothesis track is available, originating from the algorithm the user wants to investigate and improve. Reference annotations are then background information that serves to compute quality metrics. We follow definitions in [4] and define $T_i^{\text{reference}}$ and $T_i^{\text{hypothesis}}$ as the total speech duration of speaker i over the corpus, respectively in the reference and the hypothesis tracks. Likewise, T_i^{correct} is the duration of correct identification of speaker i. Precision, recall and F1-measure defined with respect to these quantities are used as metrics of recognition quality in the remainder of the paper.

Instead of imposing the selection of a specific medium before actually visualizing annotations, we propose to initially aggregate over all media, leading to an *overview* of the set of speakers. Speakers can be sorted according to a metric among those mentioned above. The list, potentially large, can be filtered on demand. In Figure 2a, a bar chart is used for this display. The bar length encodes the metric. Behind the scenes category colors are mapped to speakers, but as each bar refers to a distinct speaker, it is not necessary to show them at this stage.

One or more speakers can then be selected before proceeding. When only one speaker is selected, a *unidimensional* analysis is triggered. The distribution of the speaker over media is then displayed, along with a filtered confusion matrix. The respective reference speech durations are mapped to glyph sizes, and the metric is mapped to glyph colors (see Figure 2b). Here the user can estimate the distribution of the speaker across the corpus and see whether the associated error is uniform, and immediately spot dubious media. Clicking on a medium glyph reveals a video playback with an associated timeline. The timeline presented in Section 2 is adapted by filtering displayed annotations to the selected speaker, along with any other mistaken with him.

Confusion durations are normalized and mapped to colors in the respective table. Clicking on a cell here reveals the distribution of media relating the two selected speakers, with similar media glyph mapping as shown in Figure 2b. The process steps are recalled on top of the view. The filters applied at the *overview* level remain valid (Figure 2a). Prior steps of the process can thus be recalled and updated, dynamically affecting subsequent steps.

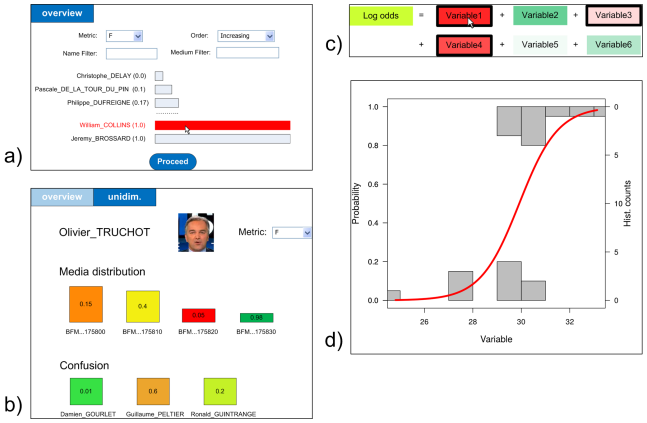

Figure 2: a) Initial overview of the speaker set. b) Unidimensional, single-speaker analysis. Related speakers and media are indicated and mapped with relevant metrics. c) Interactive logit model relating the performance metric with the set of explaining factors. d) Visualization of the logit model. The red curve shows the predicted probabilities, and observed responses are reported in bar charts [13]

From the *overview* step, multiple (potentially all) speakers can be selected. Proceeding from such a state launches a *multidimensional* analysis akin to the post-hoc classification presented in [4]. Figure 2c shows an interactive fitting process, that tentatively predicts the performance metric for speakers using a set of pre-defined explanatory variables (see Section 3) with a logit (i.e logistic regression) model. The latter is generally used for binary classification, whereas performance metrics defined in this paper lie in $[0, 1]$. As in [4], this situation could be handled by discretizing measures to binary values using a threshold. Yet the logit model grounds upon a probabilistic formalism, the loss function of which

can be easily adapted to fuzzy rather than discrete values. Moreover, logit models seamlessly handle categorical and numerical variables.

In this model, no hierarchy of variables is defined, and alternatively finding relevant explaining factors is seen as a model selection procedure. Exhaustive search in the space of models (i.e., subsets of retained explanatory variables) is computationally prohibitive, especially if interactivity is expected. Combining statistics and interactive features has already been proposed as a way to alleviate this problem [9].

We take an interactive top-down approach, by initially including all explanatory variables. The model is represented by an interactive equation (see Figure 2c). The overall goodness-of-fit of the model is mapped to the box on the l.h.s. of the equation. The contribution of each explanatory variable can be tested by comparing outcomes of the models including or omitting it [14]. Stroked bounding boxes indicate selected variables. The strength of the negative impact of removing (respectively the positive impact of adding) variables is mapped to the alpha channel of boxes on the r.h.s.

The user can interactively add or remove explanatory variables. Color mappings are updated according to renewed test results. Assuming N explanatory variables are available, we note that each addition or removal requires the estimation of $N + 1$ models and fitting tests. This amount remains small compared to the complete set of $N!$ models: yet in case computations are too long for interactive expectations, a progress bar can be included.

Logit model plots usually link the response probability to a single explanatory variable (see Figure 2d). In the context of the *multidimensional* analysis, this view can be enabled for a selected variable by collapsing all remaining variables to the intercept. The specific contribution of the variable to the current model is thus illustrated. Additional features, such as adding confidence intervals, and empirical sample estimates have also been studied [14].

5. CONCLUSION

The critical analysis in this paper helped identify the relevance of the timeline-based tool for user adjudication, but highlighted its inadequacy for performance analysis as generally coined in the literature. Directions drawn in Section 3, supported by our literature review, led to corpus-level, speaker-centric, novel designs.

Two complementary analysis procedures were described. When focused on a single speaker, it exploits the speaker-specific error distributions, with a view to emphasize dubious media or confusion patterns with other speakers. Alternatively, multiple speaker selections from the overview lead to analyze factors influencing the associated recognition performance. An interactive logistic regression model then supports user investigations.

Perspectives of course include implementing the sketches in Section 4, and carry out thorough qualitative and quantitative evaluations as soon as preliminary user interviews are satisfactory.

For the sake of clarity, segmentation-specific error was not explicitly accounted for in the designs in Section 4. Segmentation-oriented analyses could be derived by designing metrics aggregating those mentioned in Section 4 and alignment measures between reference and hypothesis.

6. ACKNOWLEDGMENTS

This work was done in the context of the CHIST-ERA CAMOMILE project funded by the ANR (Agence Nationale de la Recherche, France) and the FNR (Fonds National de la Recherche, Luxembourg).

7. REFERENCES

[1] Adobe Premiere Pro CC, 2015.

[2] O. Aubert and Y. Prié. Advene: active reading through hypervideo. In *ACM conference on Hypertext and hypermedia*, pages 235–244, 2005.

[3] P. Bruneau, M. Stefas, H. Bredin, A.-P. Ta, T. Tamisier, and C. Barras. A Web-Based Tool for the Visual Analysis of Media Annotations. In *IV 2014*, pages 145–150, 2014.

[4] D. Charlet, J. Poignant, H. Bredin, C. Fredouille, and S. Meignier. What Makes a Speaker Recognizable in TV Broadcast? Going Beyond Speaker Identification Error Rate. In *ERRARE Workshop*, 2015.

[5] P. Ercolessi, H. Bredin, and C. Sénac. StoViz: story visualization of TV series. In *ACM Multimedia*, pages 1329–1330, 2012.

[6] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The REPERE Corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102–1107, 2012.

[7] C. S. Greenberg, A. Martin, and M. Przybocki. The 2011 best speaker recognition interim assessment. In *Odyssey*, pages 275–282, 2012.

[8] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, and J. Hernandez-Cordero. The 2012 NIST speaker recognition evaluation. In *INTERSPEECH*, pages 1971–1975, 2013.

[9] S. Johansson Fernstad, J. Shaw, and J. Johansson. Quality-based guidance for exploratory dimensionality reduction. *Information Visualization*, 12(1):44–64, 2013.

[10] A. Kapoor, B. Lee, D. Tan, and E. Horvitz. Interactive optimization for steering machine classification. In *CHI*, pages 1343–1352, 2010.

[11] M. Kipp. *Anvil: The video annotation research tool*. Oxford University Press, 2011.

[12] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE TVCG*, 18(12):2431–2440, 2012.

[13] C. J. Stubben and B. G. Milligan. Estimating and analyzing demographic models using the popbio package in R. *J. Stat. Softw.*, 22(11), 2007.

[14] F. W. Young, P. M. Valero-Mora, and M. Friendly. *Visual statistics: seeing data with dynamic interactive graphics*. John Wiley & Sons, 2011.