VISUAL SPEAKER MODEL EXPLORATION

Christian Beecks, Thilo Stadelmann*†*, Bernd Freisleben*†*, Thomas Seidl**

*Data Management and Data Exploration Group, RWTH Aachen University, Germany
{beecks, seidl}@cs.rwth-aachen.de

†Dept. of Mathematics & Computer Science, University of Marburg, Germany
{stadelmann, freisleb}@informatik.uni-marburg.de

ABSTRACT

We present an interactive visualization system for the analysis of Gaussian mixture speaker models. The system exhibits the inner workings of the model intuitively by visualizing graphical representations of its parameters and of the underlying acoustical data at the same time. This enables the exploration of new modeling possibilities in the context of speaker clustering tasks.

Keywords— speaker model analysis, Gaussian mixture model, Signature Quadratic Form distance, visualization

1. INTRODUCTION

The most prominent speaker model is the Gaussian mixture model (GMM). It is predominantly used to capture the speaker specific characteristics in sets of mel-frequency cepstral coefficient (MFCC) feature vectors [1, 2]. Recently, this design has been challenged by showing its lack of expressibility in the light of more complex tasks like speaker clustering [3]. On the other hand, strong empirical evidence speaks for the GMM as a speaker model, and what has been learned in the almost 20 years of its engineering and application can be used in guiding the search for a better model formulation.

In this paper, an eidetic design approach [4] is taken to impart the inner workings of the GMM (or any other speaker model with a signature representation [5, 6]) in the context of a speaker clustering task. A visualization is presented that arranges multimodal representations of the voices in a distance-preserving way on the two-dimensional plane of an interactive graphical user interface. The voices are thereby represented by (a) plots of the GMMs' marginal distributions *and* (b) spectrograms of the underlying utterances. Clicking these representations shows the respective model's parameters and starts a playback of the utterance. This way, similarity relationships as revealed by the speaker model distances (thus being immanent in the models) can be analyzed in conjunction with the perceived distances of the listening experience. The (dis-)agreement of both modalities reveals valuable aspects of what a GMM is trying to and able to accomplish.

This paper is organized as follows: Section 2 briefly introduces the way GMM visualizations and spectrograms are created. Section 3 enlarges on the exploration system and on how model distances are computed and represented. Section 4 draws conclusions and briefly outlines areas of future work.

2. REPRESENTATIONS OF A VOICE

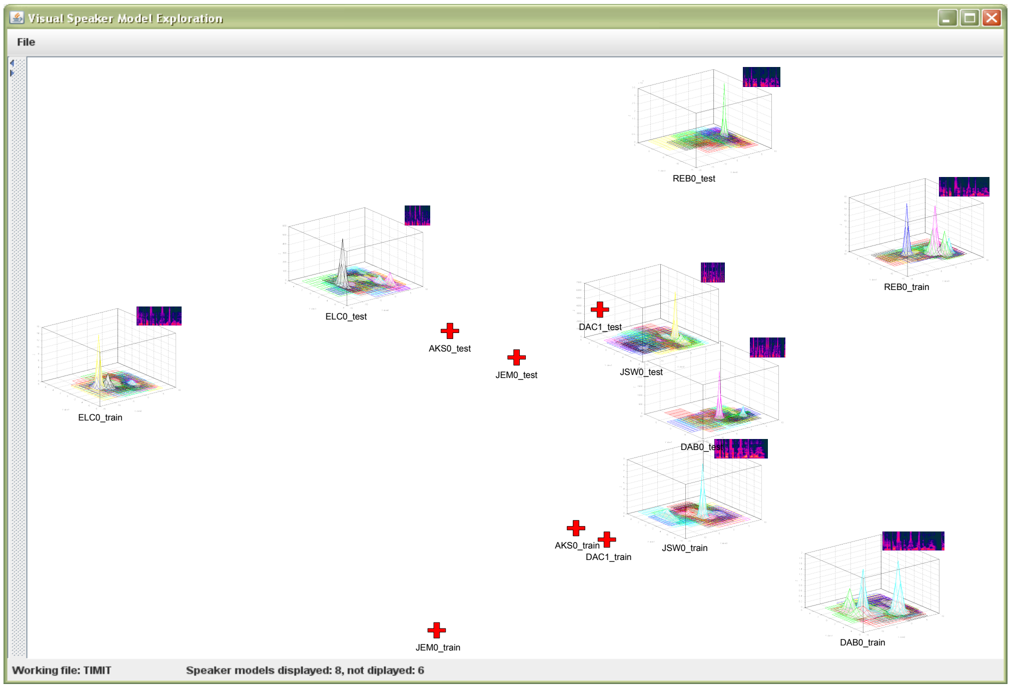
A voice is characterized, at least for a human listener, by samples of its speech. One transformation of the speech sound that reveals several of its aspects graphically is the spectrogram: a time-frequency-energy plot that shows for very short spans of time (frames) which frequency components are present at which intensity. A trained user is able to deduce e.g. gender, dialect, or what has been said from this plot.

For a technical system, a voice is characterized by the probability density function of the corresponding feature vectors' distribution. A GMM is a renowned way to model the density due to its ability to approximate arbitrary distributions. Individual Gaussians adapt to broad phonetic classes, and by averaging a number of sounds falling into this class they capture the inherent speaker characteristics. One possibility to visualize a GMM is by plotting the individual mixtures' marginal and joint densities per dimension based on their parameters. This predominantly shows the model's spread and fit instead of visualizing the underlying data.

3. VISUAL SPEAKER MODEL EXPLORATION

The proposed visual speaker model exploration system reflects the relationships among complex speaker models in a two-dimensional space. These relationships are well captured by the distances as given by the Signature Quadratic Form distance [5, 6], which showed good performance in speaker model comparison as opposed to competitors [7] in pretests. Multidimensional scaling is used to map each voice onto a two-dimensional point in the visualization, based on the GMM distances. At this point, the system displays the aforementioned graphical representations of a voice, namely the GMM visualization and the spectrogram. The user can either show or hide the model's properties by clicking the red pluses or visualizations, respectively. Additionally, clicking a point depicting no model will extend the visualization by searching the best-fitting GMM in the underlying GMM database.

978-1-4244-7493-6/10/$26.00 ©2010 IEEE



Working file: TIMIT   Speaker models displayed: 8, not diplayed: 6

Fig. 1. The Visual Speaker Model Exploration System. Spectrograms show the first 3 seconds of each utterance; model plots display MFCC dimension one vs. two in all 32 mixtures; labels reflect speaker names and membership to train/test corpus.

The visual speaker model exploration system takes into account both the visual and auditory modality in order to analyze speaker models: while the plotted GMM allows for a visual analysis of what the *model* has learned, the perception of the spectrogram and corresponding sound give an experience of what the underlying *data* represents. Relating *both* impressions to each other, and comparing the depicted technical distance of the models to the perceived “distances” of the sounds (spectrograms) helps in exploring the nonlinear interrelation of perceived and measured voice similarity. As the system enables an interactive and intuitive interface to examine the GMM relationships and their inherent properties, we believe that this novel form of speaker model relationship visualization will give new insights into the adaptation of GMM parameters to specific databases, tasks, and beyond. In particular for the task of speaker clustering, our system can be used in guiding the search for a better model formulation as it can visualize the relationships among hundreds of speaker models effectively and efficiently.

4. CONCLUSIONS AND FUTURE WORK

We presented a system for interactive visual speaker model analysis and exploration. It arranges the relationships among Gaussian mixture speaker models according to their distances, thereby allowing for intuition into their inherent properties. Taking into account both modalities, visual and auditory, this system can be used to guide the search for a better speaker model formulation and its parameter adaptation.

In the future, we plan to extend the audio-visual exploration to a wider range of models that have a valid signature representation.

5. REFERENCES

[1] D. A. Reynolds and R. C. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models,” *IEEE Trans. Speech and Audio Proc.*, vol. 3, pp. 72–83, 1995.

[2] T. Kinnunen and H. Li, “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors,” *Speech Comm.*, vol. 52, pp. 12–40, 2010.

[3] T. Stadelmann and B. Freisleben, “Unfolding Speaker Clustering Potential: A Biomimetic Approach,” in *Proc. of ACM Multimedia*, 2009, pp. 185–194.

[4] T. Stadelmann, Y. Wang, M. Smith, R. Ewerth, and B. Freisleben, “Rethinking Algorithm Development and Design in Speech Processing,” in *Proc. of ICPR*, 2010.

[5] C. Beecks, M. S. Uysal, and T. Seidl, “Signature Quadratic Form Distances for Content-Based Similarity,” in *Proc. of ACM Multimedia*, 2009, pp. 697–700.

[6] C. Beecks, M. S. Uysal, and T. Seidl, “Signature Quadratic Form Distance,” in *Proc. of CIVR*, 2010.

[7] T. Stadelmann and B. Freisleben, “Fast and Robust Speaker Clustering Using the Earth Mover’s Distance and MixMax Models,” in *Proc. of ICASSP*, 2006, vol. 1, pp. 989–992.