Topic-based Speaker Recognition for German Parliamentary Speeches

Doris Baum

Fraunhofer IAIS
Department NetMedia
Sankt Augustin, Germany
`doris.baum@iais.fraunhofer.de`

***Abstract*—In the last decade, high-level features for speaker recognition have become a research focus, as they are believed to alleviate the weak point of the classical spectral/cepstral-feature-based approaches: mismatch in acoustic conditions or channel between training and test data. Identification cues such as prosody, pronunciation, and idiolect have been successfully investigated. Semantic speaker recognition, such as identifying people by the topics they frequently talk about, has not found an equal amount of attention. However, it is a promising approach, especially for broadcast data and multimedia archives, where prominent speakers can be expected to often talk about their specific subjects. This paper reports on our experiments with topic-based speaker recognition on German parliamentary speeches. Text transcripts of speeches of federal ministers were used to train speaker models based on word frequencies. For recognition, these models were applied to automatic speech recognition transcripts of parliamentary speeches and could identify the correct speaker surprisingly well, with an EER of 13.8%. Fusing this approach with a classical GMM-UBM system (with EER 14.3%) yields an improved EER of 8.6%.**

I. INTRODUCTION

With the ever growing amount of multimedia data available, for example in broadcast archives, automatic information extraction for efficient retrieval becomes an increasingly pressing issue. Automatic speaker recognition is a valuable tool for information retrieval in multimedia data, allowing for a search of audio recordings from a specific speaker. In combination with automatic speech recognition (ASR), this enables users to find utterances of a query phrase from a specific person, for example "What did Obama say about the financial crisis?"

However, due to the very heterogeneous acoustic conditions of the domain, including varying environments and emotional states of the speakers, reverberation, background talk, noise, or music, there is often considerable acoustic mismatch between training and test data. This makes classical voice-based speaker recognition with spectral / cepstral features [1], [2] difficult; thus it is desirable to use more cues than voice for recognition, just as humans do.

Especially in the last decade, so-called high-level features, such as prosody, speaker-specific pronunciation, idiolect, and dialogue structure, have been successfully investigated and found to carry valuable speaker information [3], [4], [5], [6]. Fusion of the new high-level systems with each other and with cepstral systems shows that these additional speaker information sources help improve the overall speaker recognition rate [7], [8].

For example, phone n-grams, recognised with an open-loop phone recogniser, can be used to capture speaker-specific pronunciation [5]. Speakers are modelled through their phone n-gram frequencies in the training data, and an equivalent background model is built from background training data, to represent unknown speakers. For recognition, speaker and background models are then compared with the n-gram frequencies in the test material and their scores are combined into a likelihood ratio for the speaker. A drawback of this method (as with other high-level approaches) is that it needs more training data than a cepstral system to properly train all relevant n-gram combinations, especially with larger n-gram sizes. The required training data can be reduced significantly by deriving the speaker models from the background model through *maximum a posteriori* (MAP) adaptation [9].

Similarly, word n-grams can be used to identify speakers by idiolect / word-usage, capturing stylistic phrases, such as "you bet", shown by Doddington in [6] with experiments on manual transcriptions and later verified with ASR transcriptions, for example in [8]. Again, a larger amount of training data is needed to reach satisfactory results. In the Switchboard and Mixer corpora used for these experiments, the conversation topic was predetermined by the recording system [6], [10] and not chosen by the speakers, so they were likely recognised by stylistic patterns and not so much by their favourite topics. However, Doddington mentions a residual content bias – which may not be desirable in an access control speaker verification system, but which may be exploited in a setting where speakers can be expected to have favourite topics and frequently talk about them. This is often the case with VIPs talking in the media, for example with well-known politicians shown in news broadcasts or giving interviews on TV.

This paper reports on experiments with topic-based speaker recognition, done on German parliamentary speeches of ministers. The test scenario was chosen because a) it is very similar to the challenging material in the broadcast scenario of our audio search application [11], and b) an evaluation corpus had to be built from scratch, and the parliamentary data was readily available and already structured by speaker.

Section II details the underlying method used for topic-based speaker recognition, while Section III describes the data used in more detail. Section IV gives the experimental results with the topic-based method, compares it with an idiolectal system based on [6] and the cepstral GMM-UBM approach from [1], and finally the topic-based and the cepstral approaches are fused to yield an improved speaker recognition system.

II. TOPIC-BASED SPEAKER RECOGNITION

A. Features

As, for example, Andrews [5] and Doddington [6] have shown, speaker-specific frequencies of (phone or word) n-grams, compared to frequencies in a background model, are a viable method to capture high-level speaker information such as pronunciation and idiolect. Thus, speakers' word frequencies in ASR transcriptions, compared with a background word frequency model, seem a natural choice for capturing their favourite topics through key term frequencies.

The necessary word frequencies are obtained in three steps: First, automatic speech recognition is done on the audio input files to get a word transcript. Second, the transcript is normalised through stemming and stop word removal to retain only the basic forms of semantically important words, as is common in text mining [12, Section 2.1.1]. Then, the occurrences of the remaining words are counted and used as features for speaker modelling.

The speech recognition system employed for the first step is based on the setup described in [13]. It uses the Julius[1] engine from ISTC [14] and Fraunhofer IAIS models for German with triphone acoustic models, a 3-gram language model, and 200k words in the dictionary[2]. For step two, the ASR output is processed with the Snowball stemmer and stop word list for German[3].

B. Training Phase

A common problem for high-level speaker recognition approaches is that the amount of audio training material required at high levels of abstraction is quite substantial and may not be present for the desired speakers / domain. There are two ways to tackle this problem: "Artificially" increasing the size of the available training set and reducing the amount of training data required.

To increase the size of the training set, text material by the speakers, taken from their web pages, was used to build the training set, thus considerably boosting the amount of available data. It can be assumed that this approach works better if the text training data closely matches the ASR output for the test material in style and diction, so for our current experiments, we retrieved text versions of speeches given by the speakers.

To reduce the necessary training set size, *maximum a posteriori* adaptation was chosen, as it was found to halve training data requirements for phone n-grams in [9]. Analogous to the approach taken in [1], speaker word counts H_i for speaker S_i are used to adapt background word counts H_{BG} to derive new speaker word counts $\hat{H}_i$. The influence of the speaker training data depends on the adaptation coefficient α, with high emphasis on the speaker data when $\alpha \rightarrow 1$ and high emphasis on the background data when $\alpha \rightarrow 0$. The MAP-adapted speaker count for word n is calculated as follows:

$$\hat{H}_i(n) = \alpha \cdot H_i(n) + (1 - \alpha) \cdot H_{BG}(n) \tag{1}$$

Using the adapted speaker word counts, the speaker information carried by word n can then be modelled by the log likelihood ratio $\lambda_i(n)$ between the speaker word frequency and the background word frequency,

$$\lambda_i(n) = log\left[\frac{\hat{H}_i(n)}{\hat{N}_i}\right] - log\left[\frac{H_{BG}(n)}{N_{BG}}\right], \tag{2}$$

where $\hat{N}_i$ and N_{BG} are the total word count for speaker S_i and the background data, respectively. The log likelihood ratios for all words in the training data form the speaker model.

C. Recognition Phase

Analogous to [5], for recognition, a speaker's score s_i is calculated by summing the speaker word scores $\lambda_i(n)$ for all words in the test data, weighted according to their number of occurrences:

$$s_i = \frac{\sum_n w(n)\lambda_i(n)}{\sum_n w(n)}, \tag{3}$$

where the weighting factor $w(n)$ is determined by the word count in the test data $c(n)$ and the discounting factor d:

$$w(n) = c(n)^{1-d} \tag{4}$$

The discounting factor determines the influence of the word count, with equal weights for all words regardless of their number of occurrences when $d = 1$ and weighting proportional to the number of occurrences when $d = 0$.

As the score is a weighted average of log likelihood ratios, a speaker is more likely than the background if $s_i > 0$ and vice versa.

D. Fusion with Voice-based Approach

A simple method for fusion between topic-based speaker recognition and a classical voice-based (cepstral) approach like [1] would be to weight the cepstral score v_i with the topic-based score s_i by simply multiplying the average likelihood ratios to yield a combined score η_i:

$$\eta_i = \exp(s_i) \cdot \exp(v_i) \tag{5}$$

However, experiments revealed that both approaches have different characteristics and may be combined in a more advantageous way: While the cepstral system can be better tuned to a high precision, rejecting impostors well (when allowing a high number of false rejections), the topic-based

---

[1] http://julius.sourceforge.jp/

[2] The large number of words is necessary because of the prevalence of compounding in the German language.

[3] http://snowball.tartarus.org/

system can be better adjusted to have a high recall, identifying true speakers well (when permitting a high number of false acceptances) – see Figure 1. This seems plausible: a favourite topic is not as specific to a speaker as cepstral characteristics, but it also is not as susceptible to the acoustic variation present in the real-life test material.

Employing these characteristics in fusion, topic-based speaker recognition can be used to narrow down the set of potential speakers by preselecting only the speakers known to talk about the subject of the test material at hand, that is where $s_i > 0$. The final hypothesis is then based on these speakers' cepstral similarity with the test material: Their voice-based average likelihood ratio $\exp(v_i)$ is weighted by a prior p_i, which depends on the topic-based likelihood ratio $\exp(s_i)$:

$$p_i = \begin{cases} 0 & s_i < 0 \\ \frac{\exp(s_i)}{\sum\limits_{s_j > 0} \exp(s_j)} & s_i > 0 \end{cases} \tag{6}$$

This yields the alternative fused score η'_i:

$$\eta'_i = p_i \cdot \exp(v_i) \tag{7}$$

The performance of both methods is detailed in Section IV.

III. DATA

To our knowledge, there is no readily available German speech corpus that allows us to test our hypothesis that topic-based speaker recognition is possible and adds helpful information when fused with a spectral/cepstral-based system[4]. Thus, an evaluation corpus fitting the application of recognising well-known persons in broadcast media had to be assembled, without recording and annotating from scratch a large amount of material.

The German government and parliament offer on their web pages suitable material which could be transformed into a corpus: video recordings of German parliamentary speeches given by politicians (and searchable by speaker) are available through the parliament's "Web-TV" service[5]. Also, the federal cabinet offers on its web page[6] textual versions of speeches from almost all federal ministers (sometimes from parliament but mostly from other occasions).

This led to the selection of 14 federal ministers (6 female and 8 male), for whom enough material was present from both sources, as the 7 test speakers and 7 impostors.

As training material for the topic-based and the idiolectal target speaker models, the textual speeches were used in order to have enough data – between 60 and 144 speeches per speaker. Audio from 5 video recordings of the target speakers' parliamentary speeches was extracted, segmented, and used to train the cepstral-based GMM-UBM speaker models, yielding at least 6 minutes of training data per speaker. The same recordings were also processed with ASR to be added to the idiolectal training material.

A subset of the politics news feeds from the German Press Agency (dpa) from 2006, amounting to approximately one million words, served as training material for the topic-based and the idiolectal background model. Broadcast and parliamentary recordings from other speakers, about 160 minutes, were employed as training data for the cepstral-based background model.

As test material, audio from 5 different videos of parliamentary speeches per speaker (or impostor) were used, each video between 10 minutes and 1 hour in length. Another 2 videos per speaker and impostor were set aside for development purposes in order to be able to optimise parameters. Great care was taken to ensure that the textual transcripts of the speeches used for training did not correspond to video recordings of the same speeches in the test or development set.

For the topic-based and the idiolectal approach, word transcriptions of the test material were produced by our ASR system. In order to get an impression of the quality of the speech recognition results, a small subset of the test set (2-3 minutes from 4 speakers, 10 minutes overall) was manually transcribed in order to be able to measure the word error rate (WER) with the NIST speech recognition scoring toolkit[7]. The WER varied between 25% for the best speaker and 32% for the worst, with the average WER at 28% — an estimate of the WER to be expected for the whole corpus. By manually checking the ASR output it became apparent that longer words were recognised far more reliably than short words — which is advantageous for the task at hand, because the important keywords characterising a topic will most likely be long words. This hypothesis is supported by the fact that application of stemming and stop-word-removal (to both manual and automatic transcripts) reduces the measured WER to 23%. Thus, all training and test material for topic-based speaker recognition was normalised with both techniques.

IV. EXPERIMENTS

For evaluation, all speaker and impostor test files were scored against all true speaker models, and the resulting scores were used to produce detection-error trade-off (DET) curves using the NIST DET-curve plotting software[7].

The MAP adaptation coefficient α and the discounting factor d for the topic-based speaker recognition (TBSR) system were empirically determined on the development set to be $\alpha = 0.98$ and $d = 0.45$.

Figure 1 shows the results for the proposed topic-based speaker recognition system with and without stemming and stop word removal, compared with an idiolectal setup similar to [6], and the classical cepstral GMM-UBM approach [1].

The idiolectal system used word bigrams and was trained on the ASR transcripts of the audio training files and the textual speeches. In contrast to [6], the speaker models were MAP

---

[4] The fact that speaker recognition corpora for English also seem to avoid topic-bias may be a reason why this approach has not drawn so much research attention so far.

[5] http://webtv.bundestag.de/iptv/player/macros/bttv/index.html

[6] http://www.bundesregierung.de/Webs/Breg/DE/Bundesregierung/Bundeskabinett/bundeskabinett.html

[7] http://www.itl.nist.gov/iad/mig/tools/

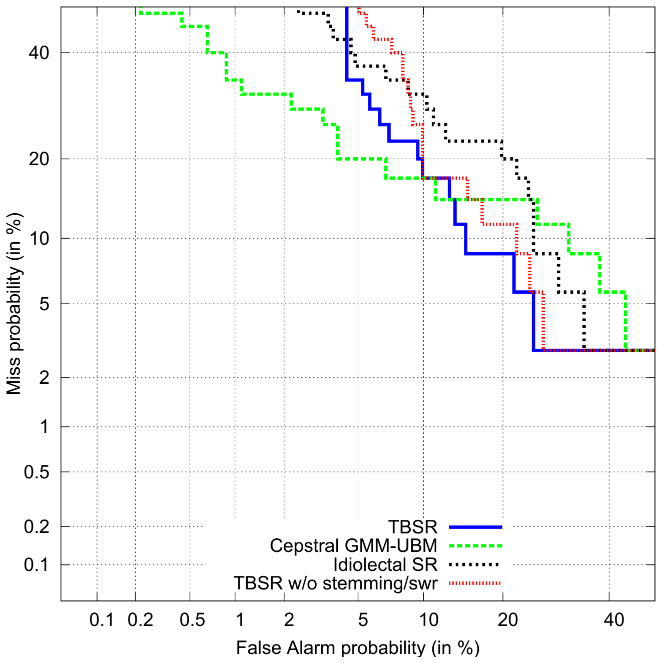

Figure 1. DET-curves for the topic-based speaker recognition system (TBSR) with and without stemming and stop word removal, for the idiolectal approach, and for the cepstral GMM-UBM system

adapted (with Equation 1, $\alpha = 0.98$) from a background model based on the news feed background training material in order to improve performance. As in Doddington's work, n-gram scores were weighted according to their number of occurrences in the test material (corresponding to $d = 0$ in Equation 4). The system had an equal error rate (EER) of 20%. Contrary to Doddington's experiments on the Switchboard corpus, with the given data there certainly is a content bias, raising the recognition performance above the level that could be expected from recognition of style only.

However, the topic based system ($\alpha = 0.98$, $d = 0.45$) gives better results: an EER of 14.6% and 13.8% without and with stemming and stop word removal, respectively. Using stemmed words instead of word bigrams and removing the stopwords – short, common words, which often form the characteristical stylistic phrases in Doddington's work – improves performance. Apparently, in this scenario, it is better to focus on topic instead of idiolectal mannerisms. This is plausible because the underlying ASR system recognises word stems of longer words more reliably than affixes and short words.

The GMM-UBM system had 512-mixture speaker models, which were MAP adapted from the background model, and used Mel-frequency cepstral coefficients, with 12 coefficients, energy, deltas, and deltadeltas, normalised with cepstral mean substraction.

When comparing the DET-curves, it becomes apparent that the cepstral system is better suited to reject impostors when a high number of misses is admissible, whereas the topic-based approach finds more true speakers when a high number of false alarms can be tolerated. This finding was used in the second fusion method (Equation 6 and 7).

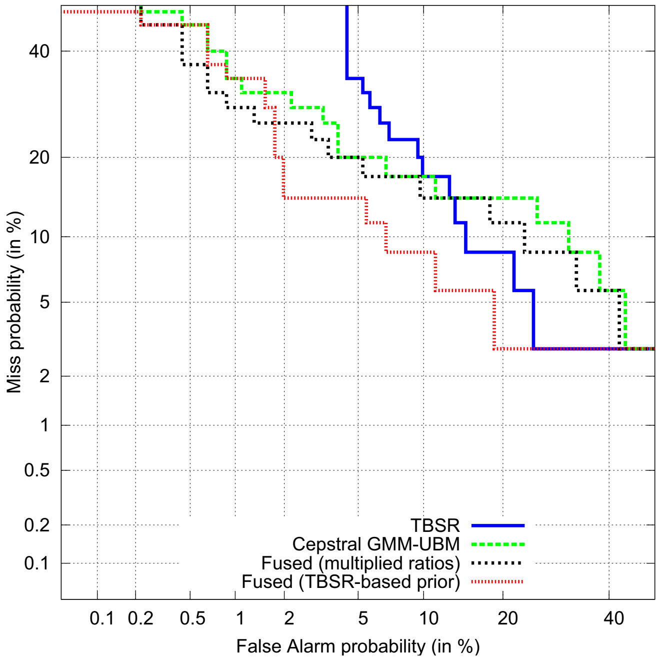

Figure 2. DET-curves for the fused systems, with two methods for fusion: multiplied likelihood ratios and TBSR-based prior

The results for the fusion methods in comparison to the original systems are shown in Figure 2: Fusion by multiplying likelihood ratios generally seems to improve on the original systems' performance, although it doesn't lower the EER.

However, the topic-based prior fusion method yields far better results, decreasing the EER to 8.6%.

V. CONCLUSION

It was confirmed that speakers' favourite topics can indeed be used to identify them in an appropriate scenario, where they choose the subject themselves or can be expected to be interviewed mostly about their areas of expertise, such as in political speeches or more generally in the broadcast domain.

Speaker specific information can be represented by speaker word frequencies in ASR transcripts, normalised with background word frequencies. The training data requirements can be reduced by MAP adapting the speaker models from the background model, and the training set can be successfully increased by using text sources. Focusing on topic instead of style yields improved results because of the ASR system's better performance on longer word stems, which carry semantic information. Fusion with the topic-based system improves the voice-based baseline, decreasing the Equal Error Rate from 14.3% to 8.6%; so apparently a new level of speaker information is added.

After these first experiments with the topic-based approach, we plan to further evaluate the proposed method on our German broadcast data corpus currently in development [15] in order to see how robustly the system copes with more difficult material and what the specific training and test data requirements are for that domain.

Also, a number of points for future research emerge, like boosting of the training material with text automatically retrieved from the web, augmentation of the speaker models with synonyms of their topics' keywords, and better speaker modelling through discriminative training. Possibly, the performance of a fused system can be further improved by more refined approaches to fusion, making use of information about weak and strong points of the individual systems.

ACKNOWLEDGEMENTS

Part of this study was funded by the German Federal Ministry of Economics and Technology through the CONTENTUS[8] scenario of the THESEUS project.

The author would like to thank Rolf Bardeli, Daniel Schneider, and Jochen Schwenninger for their helpful discussions.

REFERENCES

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, vol. 10, 2000, pp. 19–41.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[3] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03)*, vol. 4, Hong Kong, 2003, pp. 784–787.

[4] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS'02," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03)*, vol. 4, April 2003, pp. 792–795.

[5] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2002 (ICASSP '02)*, vol. 1, 2002, pp. 149–152.

[6] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, 2001, pp. 2521–2524.

[7] D. Reynolds, J. Campbell, W. Campbell, R. Dunn, T. Gleason, D. Jones, T. Quatieri, C. Quillen, D. Sturim, and P. Torres-Carrasquillo, "Beyond cepstra: Exploiting high-level information in speaker recognition," in *Workshop on Multimodal User Authentication*, Santa Barbara, California, December 2003, pp. 223–229.

[8] S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST speaker recognition evaluation system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 (ICASSP '05)*, vol. 1, Philadelphia, 2005, pp. 173–176.

[9] B. Baker, R. Vogt, M. Mason, and S. Sridharan, "Improved phonetic and lexical speaker recognition through MAP adaptation," in *ODYSSEY04 - The Speaker and Language Recognition Workshop*, Toledo, Spain, May 2004.

[10] C. Cieri, J. P. Campbell, H. Nakasone, D. Miller, and K. Walker, "The Mixer corpus of multilingual, multichannel speaker recognition data," in *4th International Conference on Language Resources and Evaluation (LREC '04)*, 2004, pp. 26–28.

[11] D. Schneider, J. Schon, and S. Eickeler, "Towards large scale vocabulary independent spoken term detection: Advances in the Fraunhofer IAIS Audiomining System," in *Proceedings of the ACM SIGIR Workshop "Searching Spontaneous Conversational Speech" held at SIGIR '08*, J. Köhler, M. Larson, F. Jong de, W. Kraaij, and R. Ordelman, Eds., Singapore, July 2008.

[12] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, vol. 20, no. 1, pp. 19–62, May 2005.

[13] T. Mertens and D. Schneider, "Efficient subword lattice retrieval for german spoken term detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2009 (ICASSP '09)*. IEEE, April 2009, pp. 4885–4888.

[14] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*. Aalborg, Denmark: ISCA, September 2001, pp. 1691–1694.

[15] D. Baum, B. Samlowski, T. Winkler, R. Bardeli, and D. Schneider, "DiSCo - a speaker and speech recognition evaluation corpus for challenging problems in the broadcast domain," in *Proceedings of the GSCL Symposium 'Sprachtechnologie und eHumanities' 2009*, W. Hoeppner, Ed., Duisburg, February 2009.

---

[8] http://theseus-programm.de/scenarios/en/contentus.html