Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services

Michael Feld[1], *Felix Burkhardt*[2], *Christian Müller*[1]

[1] German Research Center for Artificial Intelligence
[2] Deutsche Telekom

{mfeld, cmueller}@dfki.de
Felix.Burkhardt@telekom.de

Abstract

Car manufacturers are faced with a new challenge. While a new generation of "digital natives" becomes a new customer group, the problem of aging society is still increasing. This emphasizes the need of providing flexible in-car dialog that take into account the specific needs and preferences of the respective user (group). Along the lines of this year's Interspeech motto "Spoken Language Processing for All", we address the question how we find out which group the current user belongs to. We present a GMM/SVM-supervector system (Gaussian Mixture Model combined with Support Vector Machine) for speaker age and gender recognition, a technique that is adopted from state-of-the-art speaker recognition research. We furthermore describe an experimental study with the aim to evaluate the performance of the system as well as to explore the selection of parameters.

Index Terms:age recognition, automotive, personalization

1. Introduction

In recent years, the car has begun a rapid transition from a transport vehicle to a full-fledged personal information and entertainment center. Among the generation of people who grow up "always connected", the so-called "digital natives", there is a strong demand for a seamless transition in their digital environment from and into the car. However, the way that people actually use technology and interact with it depends very much on the individual. While it would be very difficult and never perfect trying to provide an optimal interface to everyone, there are certain large groups of people with similar requirements. Along the lines of this years Interspeech motto "Spoken Language Processing for All", we consider here the phenomenon of the aging society, which is enlarging the gap between e-included ("digital natives") and e-excluded people (inter alia elderly users). Supporting them in accessing the systems is a first and also much needed step towards modern user-centric car HMI (Human-Machine Interface) design, as they differ from younger people among other things in interaction errors, perception and indices of workload [4]. In order to provide systems tailored to the specific needs and preferences of younger and elderly users, a flexible service provisioning and HMI concept needs to be established. For elderly people, additional safety features like warnings or sustained in-car display of road traffic signs can be enabled by default, in order to compensate decreased vision [9] or reduced cognitive capacity [3]. For "digital natives", an exhaustive choice of services and connectivity to communication media and social platforms could be provided[1]. A key factor in the design of user-adaptive systems is that the information upon which decisions are based should be acquired transparently and in a non-intrusive way. Incorporating a "senior mode" respectively "digital native mode" that has to be enabled manually would obviously be a rather poor choice from a psychological point of view, but even simply asking for age and gender explicitly is often considered intrusive. As such data would be used by external services as well, privacy concerns are justified. An additional disadvantage of such an explicit specification is that it would require a persistent user profile to be managed, with the additional overhead of selection and log-in procedures when a car is used by more than one person.

[1]Another line of work of our research group is concerned with investigating novel interaction paradigms for communication and social platforms in the car with a special focus to minimize driver distraction

2. Knowledge Sources for User-Adaptive HMI and Belief States

Sensors can provide valuable cues about user-related properties, and information derived from speech belongs into this category as well. When abstracting from the individual knowledge sources, we can talk about *belief states* of the system for a user. It describes the "level" of user knowledge that is present. The estimated age and gender of a user is an example of a such a belief state. In this case, the system has a stereotypic belief about the user. This is more than just knowing the user's existence but less than knowing his or her identity. The classified speech signal is the *evidence* which causes the user to shift between belief states. The actual analysis of evidence has to be done by an evidence fusion component.

Compared to explicit information, sensor-based knowledge is subject to uncertainty to a much higher extent. With traditional numeric hardware sensors like a weight sensor, there is usually a fixed estimation error that can be anticipated. A speaker classification task introduces an additional challenge when mapping audio data to speaker classes. In pattern recognition, this phenomenon is referred to as *indistinguishability error* or *Bayes error*: a residuum of misclassifications that is inherent to the problem, characterized by a large overlap in the feature space between classes. It is its particularly high Bayes error that makes speaker age recognition a challenging task. There are several measures that can be taken to mitigate the implications of this. One is to use a cost function when adapting an application based on uncertain knowledge. If the cost of a misclassification is too high, i.e. an adapted feature may have an undesired side-effect for the user, it should be evaded. In the case of speaker classification, the likelihood ratio of a classification result depends on the actual input sample and hence changes dynamically. Information from several sources with different degrees of uncertainty involved can also be combined (knowledge fusion). In practice, speech-based cues are not the

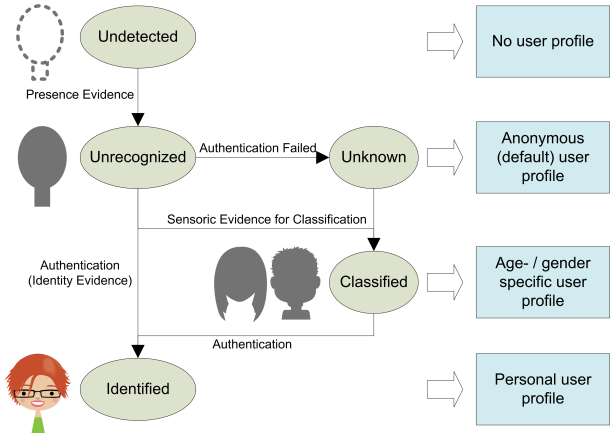

Figure 1: Schematic diagram of how evidence from different types of knowledge sources changes the system's perception of the user and activates specific or generic profiles. The circles represent belief states, the arrows evidence flow and the boxes user profiles.

Table 1: Age and gender class scheme applied in this experiment.

Class		*Ages*
1: children (C)		< 15
2: young females	3: young males	15 – 24
4: adult females	5: adult males	25 – 54
6: senior females	7: senior males	> 54

only cues that contain hints of the user's age. Video data and interaction behavior provide valuable data as well. As a consequence, the performance of a final application can even surpass the numbers presented in this study. Our primary goal however should be to improve the classification algorithms in order to reach a higher overall classification accuracy.

3. The Speaker Classification Problem

The idea of a speaker classification system is to map a given audio sample containing speech to a category in a pre-defined set of speaker classes. In order to function properly, having sufficient data to train the algorithms on is a crucial factor. Equally important is the selection and filtering of features derived from the speech signal, which make up the decision criteria. In this case, we seek functional measures derived from the raw signal representing the effects of vocal aging as described above. Finally, the selection of target classes also needs to be done carefully. In case of the automatic recognition of speaker age and gender, also referred to as the AGENDER approach in previous work, literature studies and corpus analyses have been conducted to find suitable classes. Particularly for age, they revealed characteristic class boundaries reflecting changes and degradations in the vocal tract. However, age classes are not only motivated by the classification task itself. Application-specific constraints also play a role as a reasonable mapping between speaker classes and adaptation strategies needs to be possible. Therefore, we used a modified version for speaker classes in this paper which is detailed in Table 1. Particularly, we present some of the recent efforts to improve the performance of speaker classification. In a substantial experimental study, a GMM/SVM-supervector system (Gaussian Mixture Model combined with Support Vector Machine) trained on Mel-Frequency Cepstral Coefficients (MFCCs) is evaluated and a selection of parameters is explored. The approach is adopted from state-of-the-art speaker recognition.

4. Experimental Setup

Although not part of a regular, formal, and open evaluation, we established for this paper an evaluation procedure that resembles the ones performed by the National Institute of Standards and Technology (NIST) as much as possible. We did this by clearly distinguishing between an evaluation site and a research site. The evaluation site was responsible to collect development and evaluation data as representative as possible for the application. Prior to the evaluation, the research site was provided with the development data only. At a pre-determined date, the blind evaluation data was provided to the research site for processing. The system's output was submitted to the supplier in NIST format. This procedure guarantees a fair, "one-shot" evaluation without any parameter tuning on the evaluation set. The task itself was adapted to the NIST evaluations as well. Instead of the identification task, we adopted the *detection task*: Given a speech segment (s) and an age class to be detected (target age class, A_T), the task was to decide whether A_T was present in s (yes or no), based on an automated analysis of the data contained in s. The system performance was evaluated by presenting it with a set of trials. Each test segment was used for multiple trials, with one trial for each of the age classes. Besides the decision as to whether the age class of interest was actually present in s, the output of the system contained a score indicating the system's confidence in its decision. More positive scores indicated greater confidence that the segment contained the target age class. The performance of the system was calculated in terms of detection *miss* and *false alarm* probabilities. Miss probability was computed separately for each target age class and false alarm probability was computed separately for each target/non-target age class pair. In addition, these probabilities were combined into a single number representing the cost performance of a system, according to the following cost model:

$$C(A_T, A_N) = C_{Miss} * P_{Target} * P_{Miss}(A_T) + C_{FA} * (1 - P_{Target}) * P_{FA}(A_T, A_N),$$

where A_T and A_N are the target and non-target age classes, C_{Miss}, C_{FA} are the costs for a miss and a false alarm, respectively, and P_{Target} is the prior probability of the target age class. Costs and priors are application model parameters. Here, we used two sets of application parameters. The first one is $C_{Miss} = C_{FA} = 1$ and $P_{Target} = 0.5$. It is similar to the Equal Error Rate (EER), a very common error measure in speech classification, i.e. favors an operating point (decision threshold) where the difference between the miss rate and false alarm rate is minimal.

The corpus that was used for this experiment contains the voices of 954 German speakers, each taking part in 18 turns of up to six sessions. The audio was recorded over cell phones and landline connections in 8000 Hz, 16 bit mono format. The selection of speakers is approximately evenly distributed over the seven target classes, with class 1 also being balanced for gender. The data consists of an altered version of the SpeechDat text material, containing short fixed and free text typical for automated call centers. A typical utterance is about 2 seconds

in length, but there are also some utterances between 3 and 6 seconds. In total, the corpus consists of 47 hours of speech. Of all speakers in the corpus, 90% were labeled and used for the experiment. Three working sets were created on that data: A training set (40%), a development test set (30%) and an evaluation set (30%), each with non-overlapping speakers. The test and eval sets were balanced based on the amount of available material.

5. The GMM-SVM Supervector Approach for Speaker Age Recognition

The GMM-SVM supervector approach was adopted from speaker recognition and first applied to the problem of speaker age recognition by [1]. It combines the strengths of the generative Gaussian Mixture Model with Universal Background Model (GMM-UBM) approach and the discriminative power of a large-marginal discriminative method like the Support Vector Machine (SVM). Our version of the GMM-SVM supervector system for speaker age recognition is described in the following section. We use the abbreviation IE for "investigated experimentally" to point out parameters that are tuned on the development test set. From training and background data, mel-frequency cepstral coefficients (MFCCs) are extracted (IE: step width, window size, number of coefficients and deltas). Frames with a low overall intensity (amplitude) are filtered out to reduce noise (IE: ratio of silence-filtered frames). For each sample respectively all data stemming from one speaker, a GMM is trained. A single, large GMM is trained for all background data (IE: number of mixtures – same for all GMMs). The background GMM is initialized either from a random selection of instances or using the k-means algorithm (IE: GMM initialization method). The speaker/utterance specific GMMs are not trained on the relatively short training data alone but deviated from the background model using the Maximum A Posteriori (MAP) method (IE: Parameter map_{rel} that weighs the importance of the new material relative to the background model).

The resulting GMMs are never applied (tested) in the conventional sense. Instead, the stacked means are extracted and used as input features for the backend SVM. In this way, for every speaker/utterance, one feature vector of dimensionality *number of mixtures* times *number of coefficients* is obtained. To compensate inter-speaker and inter-session variability, "nuisance" dimensions of the resulting supervector space are projected out. This is done on the basis of the ratio between the within-class and between-class variance of all dimensions (IE: the threshold of that ratio for keeping or projecting out dimensions). Features are afterwards normalized by mean/variance. One SVM is trained for every age class in a one-against-all fashion i.e. training vectors from one age class (e.g. CHILDREN) are used as positive examples and the training vectors of all other age classes are used as negative examples. The bias resulting from a larger negative training set is compensated by weighing the training errors on the positive cases higher (IE: SVM kernel function). The development test set one and two as well as later the blind evaluation data is processed analogously from the feature extraction step until the normalization step. Scores from all SVMs are obtained. The highest score is taken to determine the "winner"-class for the objective function *accuracy* (ACC) and decision thresholds for every model are calibrated for the NIST *decision cost function* (DCF). The latter is done using development test set two (IE: devtest one and two are identical to obtain a larger test set). After all parameter tuning was done, blind evaluation data was processed according to the "pseudo NIST evaluation procedure". Two score sets were delivered corresponding to the respective objective function.

6. Evaluation Results

The parameters indicated as "IE" were tuned on the development test set (not evaluation set) in a series of experiments. Table 2 provides an overview over the results. We found 5ms to be the best step width in conjunction with a relatively large window size. Despite the fact that higher MFCC coefficients are known to convey more speaker-specific characteristics whereas lower ones convey more phone-related information, adding the higher coefficients or classifying only on basis of the higher coefficients degraded the results. Also, using the deltas did not help. Filtering out training material based on an intensity filter lead to a lower performance as well. Obviously, the models picked up useful information from frames with lower intensity. The best performance with respect to the number of mixtures has been found at 128, which is a reasonable number if we consider that with a rather general age model (compared to speaker models) a higher number of Gaussians can be expected to lead to overfitting. A small map_{rel} factor turned out to be the best choice. This makes sense since the length of the training utterances is rather short. An elaborate initialization of the GMMs did not help in this case – more precisely it even degraded the performance compared to initializing randomly. What is really surprising is that removing dimensions in the supervector space based on the ratio of inter vs. intra class variability did not help. This is clearly counter intuitive if we consider that many dimensions should rather reflect nuisance attributes. In future studies, this step should be replaced by factor analysis. The fact that the linear kernel outperformed the other three confirms the results by [1].

The final system was applied to the evaluation set using the Pseudo-NIST evaluation paradigm described earlier. In a "one shot" evaluation, the accuracy reached 38% (versus 43% on the test set), which can partly be attributed to the fact that utterances in the eval set were on average shorter.

7. Using the Classification System in the Car for Tailoring Dialog and Mobile Services

In order to get to the point where we can use age-adapative functionality from our services, a couple of more steps are necessary. The experiments described earlier have been conducted using a *Speech-Based Classification Framework*. In practice, the step from such experimental results to the integration of algorithms into other systems can be quite a challenge, as now speed may start to become an issue, and the system may have to be adapted to work on multiple processors or external interfaces have to be written. If the development prototype was heavily script-based, serious performance and reliability issues may prevent a smooth integration in other components. As in-car systems are often embedded systems, this is an area where this matters particularly. A special feature of the aforementioned framework is the generation of fast run-time modules also called Embedded Modules. An Embedded Module consists of components, interfaces and pins, which can be composed freely to realize many different classification scenarios. The module's external interfaces allow it to be embedded into C++ code or to be called from Java applications.

Once we have the classification module available, we still

Table 2: Parameter tuning results on the development test set. *Italic*: initial settings (as in previous experiments); **bold face**: best result (kept for successive experiments). See explanation and interpretation in the text.

position in system	(experiment #) parameter	value	accuracy	DCF
front end	(1) step width	**5 ms**	40.3	26.4
		10 ms	40.5	27.6
		25 ms	40.4	28.4
		50 ms	37.4	29.0
	(2) window size	*1x step*	36.7	30.0
		3x step	40.3	26.4
		6x step	40.2	26.1
	(3) MFCC coefficients	***12 (1-12)***	40.2	26.1
		8 (12-19)	37.1	26.6
		19 (1-19)	39.0	28.1
	(4) delta coefficients	***excluded***	40.2	26.1
		included	39.7	27.7
frame filter	(5) intensity-based removal	***0%***	40.2	26.1
		20%	39.9	27.1
		40%	37.7	27.1
		60%	38.4	26.7
GMM training	(6) number of mixtures	64	40.8	26.6
		128	40.2	26.1
		256	40.1	27.2
		512	41.8	26.6
	(7) map_{rel}	0.01	43.5	24.9
		0.1	43.5	24.9
		2	42.7	25.2
		16	40.2	26.1
		100	38.4	28.0
	(8) initialization	***random***	43.5	24.9
		k-means	42.4	26.5
nuisance variability compensation	(9) features removed	***0%***	43.5	24.9
		10%	43.1	25.1
		30%	42.9	25.2
		65%	42.3	25.6
SVM training	(10) kernel type	***linear***	**43.5**	**24.9**
		polyn. (2)	26.0	42.4
		polyn. (3)	25.1	36.8
		radial basis	13.4	50.0

need to make the information it generates available to dialogs and services which use it. Instead of leaving it up to each service to collect the individual pieces of information from all different sources, it is more consistent and efficient to have a centralized, structured storage for user knowledge. In fact it is furthermore advantageous to store vehicle and external context information (e.g. about traffic jams and road obstacles) side-by-side with user data because of the numerous interdependencies. This also leads to the question of how user knowledge should be represented. The approach we are pursuing has a knowledge component at its core, which exposes a heterogenous (because of the variety of sources) collection of knowledge in the form of an ontology, which allows applications to query it using common RDF-based methods in addition to more specialized interfaces. A city guide service that wishes to emphasize different localities or events based on the target group (e.g. male/female or young/old) can then simply read the corresponding information from this knowledge base. Similarly, the data is used by a dialog system that wants to increase font size for elderly people or a warning function trying to decide how much in advance to present its message.

The previous two examples also illustrate another part of our architecture, which focuses on the adaptation task: early observations indicate that many adaptation strategies, such as adjusting display times, changing size and color, or reducing information, are not necessarily application-specific, but are rather generic strategies applied to application-defined content. Hence, we also provide an adaptation component logically placed between the knowledge component and the service. This means that we can more easily make existing services to some extent age and gender specific without defining new adaptation logics.

8. Conclusion

In order to create a flexible service provisioning and HMI that takes into account the specific need of a certain group of users, three top-level problems have to be solved: 1. How do we find out which group the current user belongs to?; 2. How should the knowledge be represented and linked to knowledge on the system / service in order to support adaptation?; 3. What would be the best adaptation strategy? In this paper, we addressed the first question. We presented a GMM/SVM-supervector system for speaker age and gender recognition, a technique that is adopted from state-of-the-art speaker recognition research. We furthermore described an experimental study with the aim to evaluate the performance of the system as well as to explore the selection of parameters. Additional contributions of this paper are: a structured itemization of experimental results, which shed light on the effect of various design decisions, as well as a concrete conceptual outline with respect to problem two (knowledge representation) and three (adaptation strategies).

9. References

[1] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, 'Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines', in *Proceedings of ICASSP 2008*, Las Vegas, NV, (2008).

[2] L. Cerrato, M. Falcone, and A. Paoloni, 'Subjective age estimation of telephonic voice', *Speech Communication*, **31**(2-3), 107–112, (2000).

[3] R. J. Davidse, M. P. Hagenzieker, P. C. van Wolffelaar, and W. H. Brouwer, 'Effects of in-car support on mental workload and driving performance of older drivers', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **51**(4), 463–476, (2009).

[4] A. Gruenstein, J. Orszulak, S. Liu, S. Roberts, J. Zabel, B. Reimer, B. Mehler, S. Seneff, J. Glass, and J. Coughlin, 'City browser: developing a conversational automotive hmi', in *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pp. 4291–4296, Boston, MA, USA, (2009). ACM New York, NY, USA.

[5] N. Minematsu, K. Yamauchi, and K. Hirose, 'Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques', in *Proceedings of Eurospeech 2003)*, pp. 3005 – 3008, Geneva, Switzerland, (2003).

[6] P.H. Ptacek and E.K. Sander, 'Age recognition from voice', *Journal of Speech and Hearing Research*, **9**, 273–277, (1966).

[7] S. Schötz, *Perception, Analysis and Synthesis of Speaker Age*, Ph.D. dissertation, University of Lund, Sweden, 2006.

[8] S. Schötz, 'Acoustic Analysis of Adult Speaker Age', in *Speaker Classification*, ed., Christian Müller, volume 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*, Springer, Heidelberg - Berlin - New York, (2007). this issue.

[9] J. M. Wood, 'Aging, driving and vision', *Clinical and Experimental Optometry*, **85**(4), 214–220, (2002).