Automatic Recognition of Speakers' Age and Gender on the Basis of Empirical Studies

Christian Müller

German Research Center for Artificial Intelligence
Germany
Christian.Mueller@dfki.de

Abstract

This paper describes a system that exploits the paralinguistic information in the speech to estimate the speakers' age and gender. Compared with previously published work, the so called AGENDER approach involves finer grained speaker classes and achieves a significantly higher classification accuracy. The introduction encompasses various application examples representing the actual AGENDER project context. Then hypotheses, method and a representative selection of results from extensive corpus analyses are presented, that build the empirical basis for the machine learning. Finally, the AGENDER approach on speaker classification is outlined, involving the comparison of different classification methods as well as evaluation results. The paper finishes with an outlook on extensions that are scheduled for the next project phase.

Index Terms: speaker classification, age and gender recognition, machine learning.

1. Introduction

Over the recent decades, the significance of speech technology, and in particular the automatic processing of spoken natural language, has been continously increasing. State-of-the-art research systems are able to process speech on a deep linguistic level. *Verbmobil* for example translates spontaneous speech between three languages (German, English, and Japanese) in the domain of appointment-scheduling and travel arrangement [1]. In human communication however, speech not only transports the semantics of an utterance, but also *paralinguistic* information, which allows among other things to infer some of the characteristics of the speaker. In everyday life we characterize people that we are talking with on the telephone solely on the basis of their voices and adapt our communicational behavior accordingly.

Developing systems that adapt their (dialog) behavior according to the needs of the user is subject to the discipline of *user modeling*. This research field is gaining an increasing importance as computer applications that are detached from the desk are emerging and now impact many areas of our lifes. The special requirements of such systems are as variant as the situations they are used in: A mobile pedestrian navigation system should for example take into account the fact, that the user is standing on a noisy street crossing where the environment consumes most of his attention, whereas at other times the user may be sitting on a bench in a quiet park and is able to focus completely on the interaction with the system [2].

Since ASRs for mobile devices are now available [3], it stands to reason to investigate, whether speech can be used as an information source to acquire a user model. This is the major objective of the work presented here. According to the example speaker characteristics *age* and *gender*, the approach was developed under name AGENDER. However, its domain-independent aspects can also be applied to other speaker characteristics like cognitive load, emotions or accent. Compared with previously published work (see for example [4]), this approach involves finer grained speaker classes and achieves a significantly higher classification accuracy. AGENDER was developed within the project *m3i* (Mobile Multi-Modal Interaction), which is a part of the project COLLATE (Computational Linguistics and Language Technology for Real Life Applications). The project is sponsored by the German Federal Ministry for Research and Education (BmBf) and is being carried out by the German Research Center for Artificial Intelligence (DFKI).

The target applications of AGENDER involve mobile shopping as well as pedestrian navigation systems. A central issue of these applications is the interaction style, which is multi-modal i.e. consists of gestures, speech, handwriting, and a combination of these. On the basis of the user models that are provided by AGENDER, the shopping assistant is able to make a specific selection of products (e.g. digital cameras): When the speaker is recognized as being a female person, the system can e.g. choose a camera that is especially designed for women. Analogously, the navigation system can adapt the selection of alternative routes: When the user is a child, a tourist guide could choose sights especially interesting for kids.

The vital interest in the AGENDER technology from telecommunications industry yielded another application domain, namely telephone-based spoken dialog systems (SDSs). The ambitions for improving SDSs through the AGENDER technology concern the increase of costumer satisfaction: On the basis of the speaker model, the selection of products shall be tailored to meet the requirements of the respective customer group. At the same time, the dialog shall be adapted accordingly as illustrated in table 1.

The AGENDER approach on speaker classification represents a combination of data-driven and knowledge-based aspects. The models are built on the basis of data stemming from extensive empirical analyses, which are presented in section 2. Section 3 then describes the characteristics of the approach including an evaluation in terms of classification accuracy. Section 4 provides an outlook on extensions that are scheduled for the next project phase.

2. Empirical Studies

The age classes that are investigated here are defined as follows: The class CHILDREN represents speakers up to and including an

caller 1:	‘What kind of mobile phone contracts do you offer?’
AGENDER:	Recognizes a young, male speaker and provides this information to the application.
system:	‘The contract xy is exactly right for you. You can send 150 free SMSes per month.’
caller 2:	‘What kind of mobile phone contracts do you offer?’
AGENDER:	Recognizes an elderly, male speaker and provides this information to the application.
system:	‘We would recommend you the contract abc. Besides a low base fee it has the advantage of a complete cost control, even abroad.’

Table 1: Example dialogs taken from the application scenario ’telecommunication’.

class	vs. class	feature	tendency
Cw	Cm	no or minor differences	
CwCm	YwYm	F0	–
		articulation rate	+
		voice quality	–
Yw	Ym	F0	–
YwYm	AwAm	F0	–
		voice quality	+
Aw	Am	F0	–
		voice quality	○
Aw	Sw	F0	–
Am	Sm	F0	+
AwAm	SwSm	pitch range	–
		voice quality	–
		articulation rate	–
		speech pauses	+

Table 2: Summarization of the hypotheses as described in [5, chap. 3]; ○ = tendency unclear.

age of 12 years. The class TEENAGER encompasses speakers between 13 and 19 years. Speakers between 20 and 64 years belong to the class (younger) ADULTS. The class of SENIORS begins with 65 years. Hence, in conjunction with the gender, the classification task consists of a total of eight classes. The speaker classes are denominated with one of the capital letters C,Y,A or S representing the age class followed by one of the lower case letters m or w representing the gender.

The hypotheses concerning the acoustic manifestations of the speakers’ age and gender were assembled from literature studies and are summarized in table 2. Despite of the simplifications – the interested reader is referred to [5, chap. 3] for a detailed discussion – they build a basis for an informed assembly of features to be investigated. In addition it is interesting to mention, that there’s a consensus about the fact of aging effects being more severe with men than with women. Also note the gender-specific development of the speaking fundamental frequency between younger adults and seniors.

The overall corpus used in this study consists of three parts: the German corpus BAS [6], the English corpus Timit [7] and an English corpus that was provided by Nuance[1] for this purpose. It comprises spontaneous speech as well as read words and sentences. The number of utterances was balanced on the basis of the speaker class for which the least data was available, namely Sm with 2037 utterances. With the other classes, the samples where randomly selected to avoid effects of utterance length and dialog type. The data was converted to telephone quality (8 KHz) beforehand.

The analyses were conducted on a ten-node cluster running the operating system LINUX. For the distribution of the workload we used m3i CAT, a corpus analyzing toolkit that was developed within the project m3i. It is characterized by the facility of incorporating a heterogenous set of analyzing scripts and storing the results in a homologously structured database. The majority of the speech features that are investigated here where extracted using analyzing scripts based on the program PRAAT [8]. Besides this, the tools SRSAD [9] and MRATE [10] were used. The former system detects, which parts of the sample contain speech and hence can be used as a basis for the calculation of speech pauses. Previous tests have shown, that SRSAD is very stable with respect to background noise. The latter system, MRATE, estimates the articulation rate.

The set of features was assembled according to the above described hypotheses: (1) **Pitch**, the speaking fundamental frequency, expressed in Hz with the statistical derivates mean, min, max and standard deviation. The difference between min and max is considered to be correlated with the pitch range and the standard deviation with the global variations of fundamental frequency (tremor). (2) **Jitter and shimmer**, i.e. microvariations of the F0-frequency and amplitude. Both features were measured with multiple algorithms including RAP and PPQ for jitter and AQP3 and APQ11 for shimmer [11]. With the variants discussed here, consecutive periods are measured and compared to an average value, which is calculated over an interval of various sizes, depending on the algorithm. In this way, the information relating to the actual pertubations is obtained while the global (linguistic) pitch-dynamics are largely ignored. Jitter and shimmer values are expressed in percent. (3) The **harmonics-to-noise-ratio** which quantifies the relative amount of additive noise in the voice signal. It thus reflects the dominance of harmonic (periodic) over noise (aperiodic) levels in the voice and is quantified in terms of dB. (4) The **articulation rate** expressed in syllables per second. (5) The **number of speech pauses** relative to the utterance length expressed in number per second. (6) The **duration of speech pauses** relative to their number expressed in seconds per pauses.

Table 3 provides a selection of the results of the corpus analyses: Each of the categories fundamental frequency (pitch), voice quality, speech rate, and pauses is represented by one prototype feature (see [5, chap. 5] for a comprehensive list). Due to the mutual dependency of age class and gender, it is reasonable to draw the tendencies for all eight classes instead of interpreting the results separately. Hence, the x-axes of the graphs represent the speaker classes (capital letter = age class, lower case letter = gender) while the y-axes represent the normalized values of the respective features. By normalizing, the domain of the values is mapped to a range from -1 to +1, allowing a direct comparison of the features. The tendencies between the classes are labeled on top of the graph. They are all statistically significant (t-test, $p \leq 0.01$) except the bracketed cases.

The tendencies of the mean fundamental frequency form a characteristic stair-shaped curve; the steps occur where the difference between succeeding age classes is larger than the gender-specific difference within one age class. The global negative tendency from Cw to Sm confirms the hypotheses. However, the dif-

[1] http://www.nuance.com (2006/02/10).

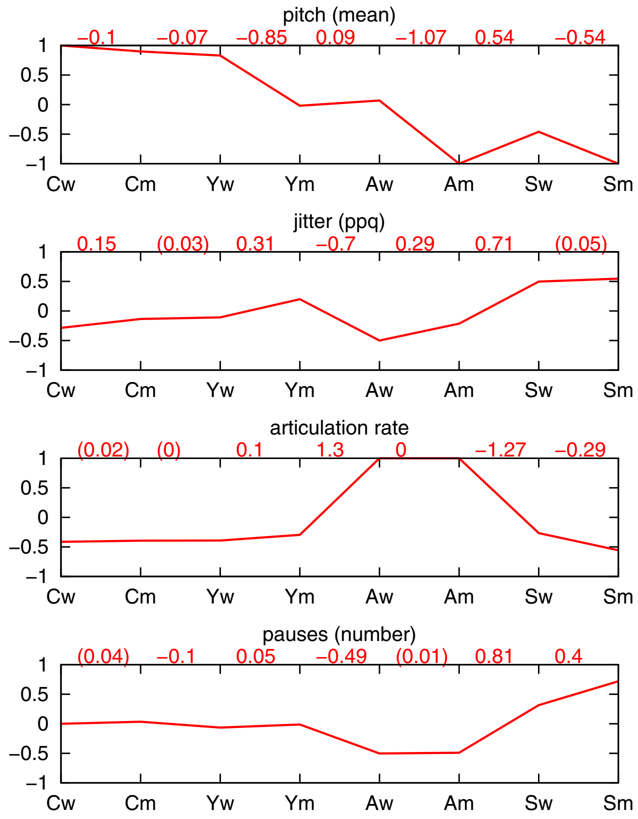

Table 3: A representative selection of the results of the corpus analyses.

ference between Am and Sm is marginal.

With respect to voicy quality – here on the example of jitter (ppq) – the hypotheses are likewise confirmed (note that higher values correspond to lower voice qualities). The positive tendency between children and youngsters can attributed to the loss of voice control caused by the rapid anatomical changes during puberty. This explanation is supported by the facts, that a) the effects are larger with boys and b) there's a large negative tendency between Yw/Ym and Aw/Am. As expected, seniors show the highest jitter values.

The curve of the articulation rate is dominated by the significantly high values of Aw/Am while children, youngsters and seniors lie on a similar low level. The results concerning the speech pauses conform with this picture: younger adults make less pauses than any other age class. Here however, the values of the seniors are significantly higher the those of children and youngsters.

3. Automatic Age and Gender Recognition

The above presented empirical results allow theoretical considerations on the effects of vocal aging and their gender-specific occurrences. In this paper however, they serve as a basis for the training of models that are able to automatically recognize the speakers' age class and gender – a task which can be seen as a pattern recognition problem. In AGENDER, the phases of pattern recognition, which concern the feature extraction and classification, are called the *first layer*. With respect to the classification, the following well known machine learning methods have been investigated: 1. Naive Bayes (NB), 2. k-Nearest-Neighbor (KNN, k=5), 3. C 4.5 Decision Trees (C45), 4. Support-Vector-Machines (SVM) and 5. Artificial Neural Networks (ANN (Multilayer Perceptron).

8-class-problem total accuracy 63.50 %

	Cw	Cm	Yw	Ym	Aw	Am	Sw	Sm
Cw	**76.09**	4.07	13.6	5.06	0.54	0.05	0.44	0.15
Cm	54.25	**12.37**	12.52	15.51	1.13	0.25	3.78	0.2
Yw	54.15	2.41	**27.44**	13.16	1.28	0.1	1.37	0.1
Ym	20.08	3.98	6.33	**59.25**	1.03	1.13	4.96	3.24
Aw	0.25	0	0.2	0.54	**84.73**	3.44	6.92	3.93
Am	0	0	0	0.74	3.53	**87.87**	1.57	6.28
Sw	0.59	1.13	0.15	2.5	3.78	0.93	**77.07**	13.84
Sm	0	0.05	0	1.67	1.18	1.47	12.47	**83.16**

Table 4: Confusion matrix for the 8-class-problem with an ANN. The total accuracy is 65.50 % with a chance level of 12.5 %.

	Aw	Am
Aw	**90.63**	9.37
Am	4.36	**95.64**

Table 5: Confusion matrix for the gender recognition problem with an ANN. The total accuracy is 93.14 %.

The results are very promising: The classification accuracy of all methods in the test were significantly higher than the chance level. Table 4 shows a confusion matrix of the best-performing method ANN. The columns represent the actual speaker class and the rows the results of the classifier. Hence, the diagonal (bold numbers) contains the correctly classified cases, the so called true positive rates (TPRs). The values are percentages that were calculated by a ten-fold cross validation.

The overall accuracy for the eight-class problem that was obtained with the method ANN is 64.5 % which is five times better than the chance level (12.5 %). With TPRs between 77.07 and 87.87 %, the accuracies for adults and seniors are very satisfying, while – on the first look – those for the remaining speaker classes (except Cw) are not. The confusion matrix however shall not only be interpreted in terms of TPRs; it is likewise important to consider the distribution of the misclassified cases. The majority of misclassified Cm for example has been categorized as Cw, a fact that absolutely conforms with our hypotheses. In general, most of the confusion occurred within consecutive cells whereas "long distance" confusions (indicating noisy classifiers) occurred rather seldom. This interpretation is supported by the high accuracy of 94.61 % (1.89 times chance level) provided in table 6 where the age classes are grouped in a way that seniors are discriminated from all other classes. Likewise, with respect to a pure gender estimation an accuracy of 93.14 % (1.86 times chance level) was achieved (see table 5).

Figure 1 compares the performances of the various classification methods. The x-axis represents the total accuracy (average TPR) and the y-axis the balance (standard deviation of the TPRs). As already mentioned, the neural network (ANN) performed best, followed the k-nearest-neighbor model (KNN). The rather

	CwCmYwYmAwAm	SwSm
CwCmYwYmAwAm	**92.24**	7.76
SwSm	3.02	**96.98**

Table 6: Confusion matrix for the discrimination of seniors from all other age classes with an ANN. The total accuracy is 94.61 %.

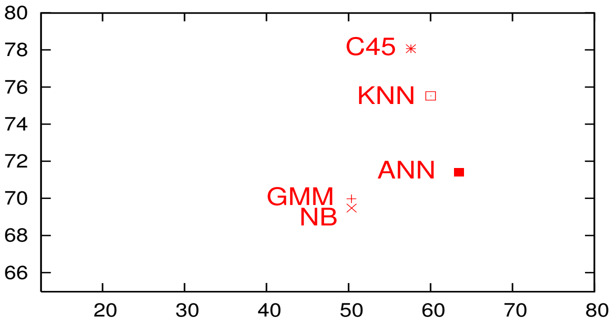

Figure 1: Comparison of different classification methods.

simple decision tree method (C45) also performed surprisingly well, especially with respect to the balance of the TPRs. The parametric methods Gaussian mixture models (GMM) and naive bayes (NB) however fell short of our expectations. Note, that the GMM implementation used in this test didn't learn the weight vector using the EM-algorithm but estimated it on the basis of an upstream evaluation (see [5, chap. 8] for a detailed discussion).

Despite of these positive results the AGENDER speaker classification approach distinguishes itself by means of a special post processing technique, the so called *second layer*: Multiple post processing problems are solved with one single mechanism, namely dynamic baysian networks (DBNs). [5] provides examples on how DBNs can be used for: 1. explicitely modeling the classification inherent uncertainty; 2. incorporating top down knowledge into the decision making process, like e.g. the fact that depending on the context, certain classifiers are more reliable than others; 3. fusing the results of multiple classifiers with respect to one utterance (static fusion) as well as several consecutive utterances (dynamic fusion).

4. Outlook

Due to the vital interest from the telecommunication industry, the AGENDER approach is mainly extended to meet the requirements of telephone-based applications which involve the optimization of the classification time as well as the optimization of the accuracy. Although this primarily concerns the AGENDER implementation, it entails the need for explicit benchmarking of the various methods.

The immediate goal of future work will be a finer distinction of age classes. Therefore it is necessary to extend the set of features e.g. by adding cepstral features, that are used in the related field of speaker identification. In addition, the use of the formant frequencies f1 and f2 are considered to be beneficial. With respect to the pattern recognition, additional methological alterations must be taken into account. The true positive rate as a measure of the performance may no longer be suitable and should be replaced by a measure that incorporates the distance between the real class and the estimated class. Other classification methods might also come to the forefront, which are based on numerical regression rather than an assignment of discrete classes. The advantage would be inter alia that an immediate comparison with the human ability of estimating the speaker's age would be possible. The infrastructure that was built up for the studies of the automatic retrieval of speaker characteristics will be applied to other classification problems: the determination of the (auditory) context as well as the recognition of the language.

5. References

[1] Wolfgang Wahlster, *Verbmobil: Foundations of Speech-To-Speech Translation*, Ellis Horwood Series in Artificial Intelligence. Springer, Berlin - Heidelberg - New York, 2000.

[2] Christian Müller, "Multimodal Dialog in a Pedestrian Navigation System," in *Proceedings of ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany, 2002, pp. 42 – 44.

[3] Rainer Wasinger, Christoph Stahl, and Antonio Krüger, "M3I in a Pedestrian Navigation & Exploration System," in *Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices*, Pisa, Italy, 2003, pp. 481–485.

[4] Christian Müller, F. Wittig, and J. Baus, "Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs," in *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)*, Geneva, Switzerland, 2003, pp. 1305 – 1308.

[5] Christian Müller, *Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender]*, Ph.D. thesis, Computer Science Institute, University of the Saarland, Germany, 2005.

[6] Florian Schiel, "Speech and Speech-Related Resources at BAS," in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998, pp. 343–349.

[7] J. et Al. Garofolo, *DARPA TIMIT CD-ROM: An Acoustic Phonetic Continous Speech Database*, National Institute of Standards and Technology, Gaithersburg, MD, USA, 1998.

[8] Paul Boersma, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 9, no. 5, pp. 341–345, 2001.

[9] D.C. Smith, J.L. Townsend, D. Nelson, and D. Richman, "A Multivariate Speech Activity Detector Based on the Syllable Rate," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)*, Phoenix, USA, March 1999, pp. 73–76.

[10] Nelson Morgan and Eric Fosler, "Combining Multiple Estimators of Speaking Rate," in *Proceedings of the 23rd International Conference on Acoustics, Speech, and Signal Processing (ICASSPP'98)*, may 1998, pp. 729 – 732.

[11] R.J. Baken and R.F. Orlikoff, *Clinical Measurement of Speech and Voice*, Singular Publishing Group, San Diego, Ca, USA, 2. edition, 2000.