Combining Short-term Cepstral and Long-term Pitch Features for Automatic Recognition of Speaker Age

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

[1]International Computer Science Institute, Berkeley, USA
[2]T-Systems, Germany
cmueller@icsi.berkeley.edu, felix.burkhardt@t-systems.com

Abstract

The most successful systems in previous comparative studies on speaker age recognition used short-term cepstral features modeled with Gaussian Mixture Models (GMMs) or applied multiple phone recognizers trained with the data of speakers of the respective class. Acoustic analyses, however, indicate that certain features such as pitch extracted from a longer span of speech correlate clearly with the speaker age although the systems based on those features have been inferior to the before mentioned approaches. In this paper, three novel systems combining short-term cepstral features and long-term features for speaker age recognition are compared to each other. A system combining GMMs using frame-based MFCCs and Support-Vector-Machines using long-term pitch performs best. The results indicate that the combination of the two feature types is a promising approach, which corresponds to findings in related fields like speaker recognition. **Index Terms**: speaker classification, age recognition, GMM, SVM

1. Introduction

In human communication, 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. One of these characteristics is speaker age. In the field of phonetics, several aspects of speaker age have been addressed. From speech technology research, however, only few systems automatically recognizing the age of the speaker emerged so far (see e.g. [1], [2]), even though there is no lack of applications. Besides the difficulty of the task this might be due to the fact, that – in opposition to related areas like for example speaker identification – neither internationally approved and commonly used databases exist for age recognition nor officially organized evaluation series.

However, the Deutsche Telekom recently invited four German sites to participate in a benchmark workshop. This is to our knowledge the first attempt to compare multiple age recognition systems directly on a common dataset. The results of that workshop are summarized in [3]. The most successful systems used short-term cepstral features modeled with a GMM or applied multiple phone recognizers trained with the data of speakers of the respective class and inferred back to that class on the basis of their confidence measures. Acoustic analyses, however, indicate that long-term prosodic features correlate clearly with the speaker age [4] although the systems based on those features have been inferior to the before mentioned approaches. In this paper, three novel systems, combining short-term cepstral features and long-term prosodic features for speaker age recognition are proposed.

The remainder of this article is organized as follows. Section 2 summarizes the training and evaluation criteria for the approaches proposed here. Particularly, the underlying database, the age class definitions as well as the reference system are described. Section 3 comprises the description of the systems as well as the comparison with each other and the reference system. Section 4 presents the conclusions discusses directions for future work.

2. Evaluation Criteria

The Evaluation data was taken from the German SpeechDat II corpus, which is annotated with age and gender labels as given by callers at the time of recording. This database consists of 4000 native German speakers, who called a recording system over the telephone and read a set of numbers, words and sentences. Eighty speakers of each age and gender group were selected for training and twenty speakers for testing. The training data consisted of the whole utterance set of each speaker, which was up to 44 utterances. The age classes were defined as follows: **C** children $\leq$ 13 years, **YF** young people (female) 14 – 19 years, **YM** young people (male), **AF** adults (female) 20 – 64 years, **AM** adults (male), **SF** seniors (female), $\geq$ 65 years, **SM**, seniors (male).

The reference system uses the following features: (1) jitter (micro-variations of the fundamental frequency F0); (2) shimmer (micro-variations of the amplitude), for each of which multiple algorithms were used including the Relative Average Perturbation (RAP) and the Period Perturbation Quotient (PPQ) for jitter as well as the Three-, Five- and Eleven-Point Amplitude Perturbation Quotient (AQP) for shimmer [5]; (3) the mean and the stddev of the harmonics-to-noise-ratio (quantifying the amount of additive noise in the voice signal [6]); (4) some statistical derivatives of the fundamental frequency F0 (pitch) including mean, stddev and mean average slope (MAS). The features were all calculated on the basis of an entire utterance (below referred to as utterance-based features). All together, a 17 dimensional feature vector was calculated for each training sample. The individual results were analyzed manually to rate the discriminative power of each feature on the basis of the class-specific Gaussian probability density. On the basis of this analysis, three multi-layer perceptron networks (MLPs) with one hidden layer each and sigmoid activation functions were trained on three different sets of features to determine (1) the gender, (2) the age class for female speakers and (3) the age class for male speakers. The performance of the system in the evalu-

	C	YF	YM	AF	AM	SF	SM
C	23	2	49	6	10	0	6
YF	0	31	0	57	4	0	6
YM	10	0	63	0	18	0	6
AF	0	21	0	73	1	1	1
AM	4	0	31	1	42	0	19
SF	0	14	0	61	5	3	15
SM	6	2	23	3	30	0	34

Table 2: Performance of the reference system. The average recall is 39 %.

ation is summarized in table 2 (only the rounded values have been provided by the organizers of the benchmark workshop). This approach is described in detail in [4]. The performances of the other systems in the test in terms of the average recall were 54 %, 26 % and 42 %[1]. To be able to compare the performances of the systems with the accuracy by which human listeners are able to judge speaker age, [3] also performed a listening experiment on the same test data. They report an average recall of 51 %.

3. Novel Approaches Combining Cepstral and Long-Term Features

3.1. Support-Vector-Machines Using Phone-Conditioned MFCCs plus Utterance-based Pitch Combined on the Feature-Level.

System A uses an open loop phone recognizer as a segmentation front-end. For each sample, a set of segment-based features is calculated and combined with utterance-based features. The resulting feature vector is used to train seven binary support-vector-machines (SVMs), one for each class. The final category decision in achieved by choosing the model with the highest likelihood. For each segment (phoneme), the medians of the mel-frequency-cepstrum-coefficients (MFCCs) one to twelve were calculated (the 0th coefficient was left out as it is related to the amplitude). The median was used instead of the mean because it is less sensitive to outliers, which are likely to occur in the segment because (1) the border area of a phoneme is influenced by the transitions from adjacent phonemes and (2) the alignment is likely to be imprecise. It should be noted though, that the transitions between the phonemes should not be omitted in any case since they are likely to contain speaker-specific information. However, this has to be examined in a separate study. The mean could be used instead of the median if it is calculated on the basis of a portion taken from the center of the segment. A respective test yielded slightly worse results compared to the median.

MFCCs are features commonly used for example in speaker recognition, because they accurately characterize the vocal tract configuration of the speaker. The source of the speech signal containing speaker-dependent characteristics such as pitch can complement the vocal tract information. Therefore, the MFCC features were combined with utterance-based pitch features calculated in the same way as in the reference system. To avoid effects caused by the differences in magnitude between the pitch values and MFCC values, a mean/variance normalization was performed[2].

With the phone recognizer's underlying set of 50 phonemes, the combination of twelve segment-based MFCC features with seven utterance-based statistical derivates of pitch (min, max, mean, quantile, stddev, mean average slope, slop without octave jumps) yields a 607-dimensional feature vector (some segments like pauses and undefined speech were discarded). However, the vector contains a varying portion of missing values depending on the number of phonemes actually occurring in the given utterance. With the training set at hand, 78 % of the segment-based features was missing on average. To obtain a better instantiated feature vector, attempts have been made to use broad phonetic classes and other phoneme-groupings as segments instead of the actual phonemes. In another experiment, the missing values have been replaced by the class-specific mean values. However, the results could not be improved with any of the these variants.

To cope with this problem, support-vector-machines (SVMs) were used for classification. Especially the SVM-LITE system [7] which is commonly used for speech related classification problems, was appropriate for this task, because it is able to handle a very large feature vector and provides a mechanism to deal with missing values. Since SVMs are binary classifiers, a separate model for each of the seven speaker classes was trained using samples from the respective class as positive training cases and samples from all other classes as negative ones. The bias which results from a larger number of negative cases was compensated by choosing an appropriate "j-factor" – a cost-factor by which training errors on positive examples outweigh errors on negative examples. Various methods of balancing the training data beforehand were tested as well. However, adapting the cost-factor yielded the best results.

The final category decision was obtained by choosing the model with the highest likelihood. The results of system A (with normalized features) are presented in table 4. Although the average recall of 39.9 % is not significantly higher than with the reference system, the matrix is more balanced in the sense that the difference between the lowest and the highest recall is smaller. Note, that if the application requires a higher recall in a certain cell, say adults, this can be done explicitly by adapting the decision thresholds of the binary classifiers accordingly. This feature was not supported by the reference system. The average recall of system A with non-normalized data was 39.1 %.

	C	YF	YM	AF	AM	SF	SM
C	33.99	38.69	3.64	11.84	2.88	8.5	0.46
YF	15.05	51.18	0.54	21.08	0	11.72	0.43
YM	1.18	1.5	60.64	2.9	23.82	0.43	9.23
AF	12.68	19.58	3.68	37.32	0.74	23.07	2.94
AM	0.3	0	41.6	0.4	36.84	0.3	20.55
SF	10.01	18.25	7.16	26.5	1.47	31.7	4.91
SM	0.53	1.31	38.9	2.37	24.97	4.34	27.6

Table 4: Performance of system A. The average recall is 39.9 %.

[1] A non-disclosure-agreement between the participants prohibits a detailed itemization of the results here.

[2] In a separate test, the normalized version performed in fact slightly better than the one using the original values.

	C	YF	YM	AF	AM	SF	SM
C	42.94	20.03	18.21	7.28	6.07	2.58	2.88
YF	19.89	46.45	2.26	23.66	0.11	6.24	1.4
YM	0.43	0	28.11	3	59.66	1.29	7.51
AF	8.73	16.73	2.48	51.19	0.18	19.49	1.19
AM	0.1	0	1.72	2.83	90.18	1.11	4.05
SF	11.19	16.39	2.94	40.43	0.2	27.67	1.18
SM	0.13	0	11.56	2.37	63.34	3.02	19.58

Table 6: Performance of system B.The average recall is 43.7 %

3.2. Support-Vector-Machines Using Phone-Conditioned MFCCs Plus Utterance-based Pitch Combined on the Score-level.

System B is a variant of system A where separate models are built for each segment and later combined on the score-level. By training a model for each phoneme, the problem of missing values could be obviated since the training data only contained those samples in which the respective phoneme actually occurred. At test time, again only those models corresponding to the phonemes occurring in the test sentence are applied. In system B, they are combined with a single utterance-based pitch model. The cost-factors balancing the uneven distribution of positive and negative examples were calculated for every model separately. This was necessary because of the different amount of training data.

An obvious difference compared to system A is the considerably larger number of models. While system A only needs seven models corresponding to the seven speaker classes, system B comes to a total of 7 classes * 50 segments = 350 models. However, some of them could be discarded because of their poor performance in the individual evaluations (see below). The use of multiple models for each class required the introduction of additional parameters to control the fusion of the results. Since full search in an almost fifty-dimensional space was too expansive and experiments with a hill-climbing method yielded no satisfactory results, the models were combined using a weighted sum in which the weights are derivated from the individual performances of each model. Each of the intervals was assigned to a weight. The models with a performance of lower than 50 % were virtually discarded. The authors are aware of the fact that the described weighting method is suboptimal. If system B is further developed in the upcoming project phase, it should be replaced by a more sophisticated one.

Figure 6, which summarizes the performance of system B, shows that the average recall is higher compared to system A as well as to the reference system. However, the improvement has been achieved at the cost of the balance. Although the potential of this approach was most likely not fully exploited due to the manual configured weighting, the complexity of the system makes it unlikely to be the best choice for an actual application. Aside from the hard-to-control number of parameters, the evaluation of a large number of models for a given test utterance leads to longer response times and makes the system is harder to maintain because the consistency of the set of models has to be ensured.

	C	YF	YM	AF	AM	SF	SM
C	41.73	42.64	0.3	7.89	6.37	0.76	0.3
YF	13.12	62.26	0.32	16.56	0	7.74	0
YM	3.65	1.39	54.94	3	28.76	1.07	7.19
AF	6.16	21.72	3.03	43.29	0.92	23.99	1.19
AM	0.61	0.1	18.12	0.3	70.75	0.1	9.82
SF	4.32	21.2	4.32	35.43	0.49	31.99	2.26
SM	4.6	1.97	20.24	1.84	27.6	5.12	38.63

Table 8: Performance of system C. The average recall is 49.11 %

3.3. Gaussian Mixture Models Using Frame-based MFCCs plus Support-Vector-Machines Using Utterance-based Pitch Combined on the Score-level.

System C represents a streamlined approach getting by with no more than two models per class. One is the utterance-based pitch features SVM from system A, the other model is still based upon MFCCs but uses the actual frame-by-frame values instead of the segment means (or medians). Hence, rather than SVMs, Gaussian-mixture-models (GMMs) have been applied, a method which is commonly used for frame-wise classification in speaker recognition and related fields (see for example [8], [9], [10]). With the frame-by-frame processing of the features, system C requires no segmentation front-end which is beneficial in terms of the system's complexity and classification speed.

When designing a GMM for a particular task, a major choice is the appropriate number of Gaussians. More Gaussians usually model the training data better but bear the risk of overfitting i.e. not finding a suitable generalization to classify unseen material properly. While in speaker recognition systems often 1024 and more Gaussians are used, the classification problem at hand requires a higher generalization and therefore a much smaller number of Gaussians was expected to be optimal. Experiments were conducted using 16, 32, 64, 96, 128 and 256 mixtures. The best results were obtained with 96.

The frame-based GMM and the utterance-based SVM were combined on the score level. The weights were trained using a class-specific (i.e. seven-dimensional) full search within the interval 0 – 3 and an increment of 0.1. The optimal weights were (GMM + 0.3 * SVM) for all classes. A hill-climbing method optimizing the weights for each class separately was implemented as well. However, it yielded a weight constellation which performed slightly worse.

As depicted in figure 8, the average recall of system C is of 49.11 % is considerably higher compared with any of the other systems. At the same time, the individual recall values are fairy balanced. The average recall of the GMM alone constitutes 43.66 %; the unweighted combination of the GMM and the SVM yielded an average recall of 45.99 %.

4. Conclusions and Outlook

The lessons learned from analyzing the results of the study presented here is, that to some extent information about the content of the utterance ("what was said") is necessary to achieve information about the speaker ("who said it"). All three proposed approaches take into account the nature of the phonemes uttered, either explicitly as with the segment-based approaches or implicitly as with the frame-based ones. The most success-

ful approach (system C) represents a combination of frame-based cepstral features (MFCC) and utterance-based pitch features. We believe that this is due to the fact that MFCCs are powerful features to characterize the vocal tract configuration of the speaker whereas pitch features contain complement speaker-dependent characteristics representing the source of the speech signal. This corresponds to findings in the related field of speaker recognition where 1. the combination of short-term cepstral features with long-term prosodic features was found to improve the results in general and [11] and 2. pitch was identified as the single best long-term feature [12].

The most incising limitations of the work presented here can be seen in the data that were used for training. Not only the total amount of data was to small to build accurate models but the age structure was not optimally balanced. An important aspect of future work will to be compare the various approaches to the recognition of speaker age on the basis of large databases. Particularly, the corpora Switchboard II, Fisher, and Mixer could be used for this task. Switchboard [13] is a corpus of spontaneous telephone conversations collected at Texas Instruments with funding by DARPA. It comprises 240 hours of recorded speech spoken by over 500 speakers of both sexes. The Fisher corpus [14] represents a collection of conversational telephone speech that was created at the Linguistic Data Consortium (LDC) during 2003. It contains 11700 audio files, each one containing a full conversation of up to 10 minutes adding up to almost 2000 hours of speech. Mixer [15] is a multi-language conversational telephone speech corpus collected by LDC to support the 2004 and subsequent NIST speech recognition evaluations. The database comprises 600 speakers who completed twenty conversations and 1150 speakers who completed ten. With an average conversation being six minutes long, the corpus contains nearly 3000 hours of speech.

The advantage of using these databases (and especially the large Fisher and Mixer) is twofold. First of all, using more data is always beneficial for a machine learning system. Secondly, and even more important, results achieved on an internationally approved database that is available for the community of researchers in the field will have a larger impact. At the same time, this constitutes an important prerequisite of another application of age recognition: investigating the question whether or not the information about the age of a speaker can be used as an additional source of information to recognize the speaker's identity. This aspect has been investigated by a only a few studies before, all leaving the question unanswered. In a study with human raters, [16] conclude for example that "whether the perceived age can [...] lead to a correct speaker identification remains an open question".

5. References

[1] N. Minematsu, K. Yamauchi, and K. Hirose, "Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques," in *Proc. Eurospeech '03)*, Geneva, Switzerland, 2003, pp. 3005 – 3008.

[2] C. Müller, F. Wittig, and J. Baus, "Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs," in *Proc. Eurospeech '03)*, Geneva, Switzerland, 2003, pp. 1305 – 1308.

[3] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications," in *Proc. ICASSP '07)*, Honolulu, Hawaii, 2007, to appear.

[4] C. 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. dissertation, Computer Science Institute, University Saarland, Germany, 2005.

[5] R. Baken and R. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. San Diego, Ca, USA: Singular Publishing Group, 2000.

[6] C. T. Ferrand, "Harmonics-to-Noise Ratio: An Index of Vocal Aging," *Journal of Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[7] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT-Press, 1999.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] D. Sturim, D. Reynolds, R. Dunn, and T. Quatieri, "Speaker Verification using Text-Constrained Gaussian Mixture Models," in *Proc. ICASSP '02)*. Orlando, FL: IEEE, 2002, pp. I: 677–680.

[10] C. Miyajima, Y. Hattori, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker identification using Gaussian mixture models based onmulti-space probability distribution," in *Proc. ICASSP '01*. Salt Lake City, Utah: IEEE, 2001.

[11] E. Shriberg, "Higher-Level Features in Speaker Recognition," in *Speaker Classification I*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Heidelberg / Berlin / New York: Springer, 2007, vol. 4343, to appear.

[12] E. Shriberg, S. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication*, vol. 46, pp. 455–472, 2005.

[13] D. Graff and S. Bird, "Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies," in *Proc. LREC '00)*, Athens, Greece, 2000.

[14] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proc. LREC '04)*, Lisbon, Portugal, 2004.

[15] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybocki, and K. Walker, "The mixer and transcript reading corpora: Resources for multilingual, cross-channel speaker recognition research," in *Proc. LREC '06)*, Genoa, Italy, 2006.

[16] E. Eriksson, J. Green, M. Sjöstrom, K. Sullivan, and E. Zetterholm, "Perceived age: a distracter for voise disguise and speaker identification ?" in *Proc. of Fonetik 2004*, Stockholm, Sweden, 2004.