COMPARISON OF FOUR APPROACHES TO AGE AND GENDER RECOGNITION FOR TELEPHONE APPLICATIONS

Florian Metze, Jitendra Ajmera, Roman Englert, Udo Bub ♣*;*
Felix Burkhardt, Joachim Stegmann ❖*;*
Christian Müller ♦*; Richard Huber* ♠*;*
Bernt Andrassy, Josef G. Bauer, Bernhard Littel ♥

♣ Deutsche Telekom Laboratories; Berlin; Germany
❖ T-Systems Enterprise Services GmbH, SSC ENPS; Berlin/ Darmstadt; Germany
♦ Deutsches Forschungszentrum für Künstliche Intelligenz; Saarbrücken; Germany
♠ Sympalog Voice Solutions GmbH; Erlangen; Germany
♥ Siemens AG, CT IC 5; München; Germany

{florian.metze|jitendra.ajmera|roman.englert|udo.bub}@telekom.de;
{felix.burkhardt|joachim.stegmann}@t-systems.com;
christian.mueller@dfki.de; huber@sympalog.de;
{bernt.andrassy|josef.bauer|bernhard.littel}@siemens.com

ABSTRACT

This paper presents a comparative study of four different approaches to automatic age and gender classification using seven classes on a telephony speech task and also compares the results with Human performance on the same data. The automatic approaches compared are based on *(1)* a parallel phone recognizer, derived from an automatic language identification system; *(2)* a system using dynamic Bayesian networks to combine several prosodic features; *(3)* a system based solely on linear prediction analysis; and *(4)* Gaussian mixture models based on MFCCs for separate recognition of age and gender. On average, the parallel phone recognizer performs as well as Human listeners do, while loosing performance on short utterances. The system based on prosodic features however shows very little dependence on the length of the utterance.

***Index Terms*—** speech processing, acoustic signal analysis, speaker classification, age, gender

1. INTRODUCTION

Interactive voice response (IVR) systems are one of the most mature applications of automatic speech recognition (ASR) today and are widely deployed for customer care and service applications. ASR research is currently moving from mere "speech-to-text" (STT) systems towards "rich transcription" (RT) systems, which annotate recognized text with non-verbal information such as speaker identity, emotional state. In IVR systems, this approach is already being used to identify dialogs involving angry customers, which can then be analyzed with the goal of automatically identifying problematic dialogs, transferring unsatisfied customers to an agent, and other purposes.

Also, the first adaptive dialogs are now appearing, particularly in systems exposed to inhomogeneous user groups. These can adapt degree of automation, order of presentation, waiting queue music, or other properties to properties of the caller such as age or gender. As an example, it would be possible to offer different advertisements to children and adults in the waiting queue.

In non-personalized services, speaker classification will be based on the caller's speech data. While classifier performance is only one factor influencing the utility of the above approach in an IVR system, it is certainly a major factor.

We therefore initiated an evaluation of different approaches on age and gender recognition on data which resembles "real-world" telephony channel data, in order to measure current classification performance and also incite development through collaboration and friendly competition. This effort will possibly result in a series of evaluations being organized. We also compared our results to a Human baseline experiment conducted on the same data.

1.1. Related Work

While the general influence of speaker age on voice characteristics is being studied since the late 1950s [1] and sustained continuous attention since then (see e.g. [2]), the first actual systems estimating the age and the gender of the speaker were developed only recently [3, 4, 5, 6]. However, the quality of these systems is difficult to compare, as they vary considerably regarding the number and distribution of speaker age as well as the types of speech material.

The variability of IVR system use patterns across age and gender is investigated in [7], indicating that dialog strategies tailored to specific age and gender groups can be very useful in improving overall service quality.

1.2. Paper Organization

The paper is laid out as follows: Section 2 describes the data used in this experiment and the conditions of the evaluation. Section 3 describes the individual systems submitted. Section 4 presents the results of the system evaluation, while Section 5 describes a Human

baseline experiment for comparison. An interpretation of the results is offered in Section 6.

2. DATABASE AND EVALUATION CONDITIONS

In order to evaluate different approaches under controlled conditions, we conducted a benchmark using the following procedure: one of the sites, which did not participate in the evaluation itself, sent the participants training and development test data, which was selected according to previously agreed upon conditions. During one month, participants then developed a system optimized on this set using only the data provided. At the end of this period participants were sent the test set and returned a list of age and gender labels for the evaluation data one week later, which were then scored by the organizing site.

For this evaluation, we used the following 7 groups and labels:

- Children: $\leq$ 13 years (C)
- Young people: 14-19 years, male (YM) and female (YF)
- Adults: 20-64 years, male (AM) and female (AF)
- Seniors: $\geq$ 65 years, male (SM) and female (SF)

While somewhat arbitrary, these classes stem from an IVR application we are currently developing. Evaluation data was taken from the German SpeechDat II corpus [8], 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. Except for children and seniors, for whom these numbers were not available, we selected 80 speakers of each age and gender group for training and 20 for testing, thereby gaining a weighted age and gender structure. Training data consisted of the whole utterance set of each person, up to 44 utterances.

For further analysis, we created a sub-set of short utterances "SpeechDat_short" (SpeechDat II corpus identifiers "a" and "o") and another set of longer sentences "SpeechDat_long" (identifier "s").

In order to evaluate the performance on data that originates from a different domain, we also tested the systems on "VoiceClass" data. This data was collected in-house and consists of 660 native speakers of German, which called a voice recorder and freely talked for about 5 to 30 seconds on the topic of their favorite dish. The age structure is not controlled, the data consists of many children and youth but almost no seniors.

3. SYSTEM DESCRIPTIONS

3.1. System A

The underlying system was originally developed for ASR and automatic acoustic Language Identification (LID). It is based on Parallel Phoneme Recognizers (PPR) using Continuous Densities Hidden Markov Models (CDHMMs) and phoneme bi-grams. Feature extraction consists of computation of Mel Frequency Cepstral Coefficients (MFCCs) and a linear transformation based on Linear Discriminant Analysis (LDA), retaining 24 components for the final feature vectors. Figure 1 explains the system architecture.

For each of the 7 age/ gender categories a specific phoneme recognizer with category specific HMM and phoneme bi-gram is used. Neg-log scores for each respective category are computed using a Viterbi decoder. In a final step the classified category is determined by a minimum decision with regard to the 7 neg-log category scores.

To build the PPR system, we first created category specific mono-phone HMMs using maximum likelihood estimation as used for standard ASR system generation. For the following LDA the category specific mono-phone HMM states served as the LDA classes. Based on the retained LDA matrix optimized for age/ gender classification the final category specific mono-phone HMMs were build. In a final step the phoneme recognizers were applied to the training material to estimate category specific phoneme bi-grams also based on the maximum likelihood criterion.

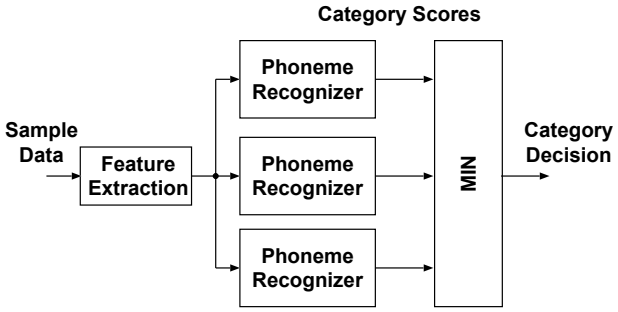

Fig. 1. Example age/ gender System based on Parallel Phoneme Recognizers (PPR) for 3 categories. The "MIN" decision selects the category pertaining to the phoneme recognizer yielding best (lowest) score.

3.2. System B

This system computes the following prosodic features: *(1)* **jitter** (micro-variations of the fundamental frequency F_0); *(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 (APQ) for shimmer; *(3)* the mean and the stddev of the **harmonics-to-noise-ratio**, *(4)* some statistical derivatives of the **fundamental frequency** F_0 including mean, stddev and mean average slope (MAS). All together, for each training utterance a 17 dimensional feature vector was calculated.

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 (MLP) C_1, C_2, C_3 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, forming the *first layer*.

The *second layer* performs post processing using Dynamic Bayesian Networks (DBNs). The DBN is used to model the classification-inherent uncertainty by introducing *(1)* three observable nodes $O_{i=1,\ldots,3}$ each representing the result of one classifier C_i with states corresponding to the classes (e.g. MALE and FEMALE); *(2)* the nodes AGE and GENDER representing the actual speaker class; *(3)* links between $O_{1,\ldots,3}$ and AGE/ GENDER representing a causal relationship. In this vein, the uncertainty of that relationship can be expressed in terms of the conditional probability table (CPT) attached to O_i. The CPT-values were optimized on the cross-validation set of the respective classifier C_i. The DBN also fuses the results of the classifiers by letting the nodes AGE and GENDER both be parents of each O_i. Appropriate CPTs then provide the precedence of one age classifier over the other depending on the result of the gender classifier.

3.3. System C

This system exploits the dependency of age and gender on the linear prediction (LP) envelope of a windowed speech signal using the following distance measure between signal spectrum and LP smoothed spectrum at signal harmonics (formants) [9].

$$d = \frac{1}{2 * W * N_f} \sum_{m=1}^{N_f} \sum_{w=-W}^{W} |\log \frac{P(\omega_m + w)}{\hat{P}(\omega_m + w)}| \qquad (1)$$

where N_f is the number of formants estimated by LP analysis, ω_m is the position of the m^{th} formant, $P(\omega)$ is the original power spectrum and $\hat{P}(\omega)$ is the estimated spectral envelope. The distance is computed over a small window of size $2 * W$, around the formant positions to avoid localization errors.

This choice of distance measure for this purpose was motivated by two characteristic properties of LP analysis: *(1)* error cancellation property which makes it select an envelope other than the only one which passes through all spectral points, and *(2)* poles estimated by LP analysis generally move in the direction of pitch harmonics. Thus, we can expect that voices with higher pitch frequencies (e.g. females and children), should exhibit higher values of d when compared to lower pitch frequencies.

Gaussian distributions of this distance were estimated using training data. A test sample was classified as the one corresponding to the distribution with maximum likelihood, the whole utterance was assigned to the most frequent class.

At the time of training, the distribution of distance d for *young* speaker class (both male and female) was found to have significant overlap with *adult* speaker class, not only making it difficult to discriminate from other classes but also increasing confusion between *children* and *adult* speaker classes. Therefore, this class was merged with *adult* speaker class for this evaluation and cannot be identified using this technique.

3.4. System D

This approach uses two independent, but similar, frame-wise classifiers for age and gender classification, whose decisions are combined at the utterance level.

Four age classes were recognized using a Gaussian mixture model classifier with 256 independent Gaussian densities per class, where every age class was divided in a male (M) and a female (F) sub-class, which were trained using 128 Gaussians each on the respectively labeled portions of the training data.

Gaussians were trained on 12 MFCC features and their first and second order derivatives per frame. To avoid using feature vectors belonging to pauses or other non-verbal parts like breathing which hold no or only little information about the age (or the gender) of the speaker, only feature vectors belonging to voiced frames were used in training and testing. These were determined using a power-based criterion.

For the classification of whole utterances, we used a two step approach. In a first step all (voiced) feature vectors are classified with the Gaussian mixture classifier, selecting the class with the best score. The result of the first step is a sequence of (internal) class labels for an utterance, at which point we ignore the gender distinction and assign the utterance to the age class with the highest count.

To improve discrimination, gender was determined separately using a dedicated Gaussian mixture classifier modeling every class (F and M) with 128 independent Gaussian mixture models using the same basic setup as above, also using pitch as an additional feature, resulting in a 37-dimensional feature space.

System	SpeechDat II	SD_short	SD_long	VoiceClass
System A	54% 55%	45% 46%	61% 61%	60% 58%
System B	40% 52%	38% 51%	42% 62%	52% 50%
System C	27% 50%	23% 44%	31% 56%	53% 69%
System D	42% 46%	38% 40%	45% 52%	64% 65%

Table 1. Precision (left) and recall (right) on the different data sets for the individual systems.

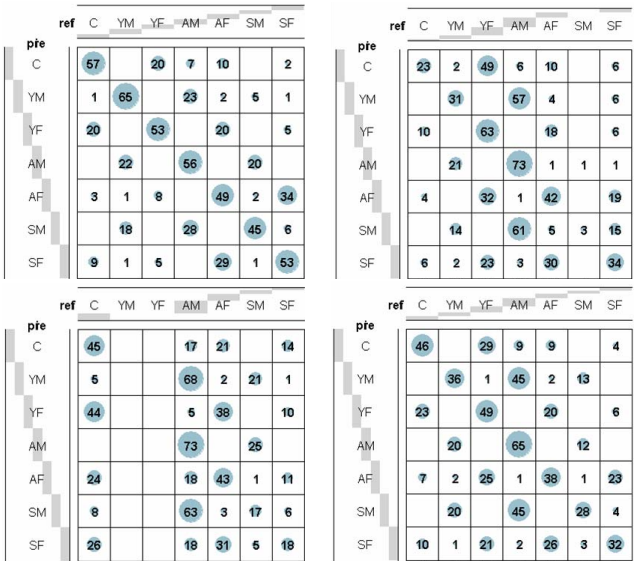

Fig. 2. Relative confusion matrices (rows sum to 100%) on SpeechDat II data: System A (top left, 54% accuracy overall), System B (top right, 40%), System C (bottom left, 27%), System D (bottom right, 42%). Class symbols are defined in Section 2; columns contain hypothesized classes (output labels), rows contain reference classes.

4. EVALUATION RESULTS

Evaluation results are tabulated in Table 1. Accuracy on SpeechDat II ranges between 27% and 54% for the approaches which distinguished all 7 classes, while recall is between 46% and 55%.

Overall normalized confusion matrices are shown in Figure 2. Of all approaches, System A (based on class-specific phone recognizers) reaches the best performance and also shows the most balanced confusion matrix. Performance however drops for the "short" utterances, presumably due to the temporal structure realized in the phone bi-grams. System B on the other hand is based on multiple prosodic features computed on the entire signal and its accuracy shows very little dependence on the length of the utterance.

Results on the VoiceClass task are similar to, or even better then the SpeechDat_long results. The robustness of the approaches against data from different domains and channels seems good.

We also tried combining the three systems separating seven classes into one system by majority voting on the SpeechDat II task, however the combined system has an accuracy lower than the best individual system. We suspect that errors of System B and System D are highly correlated, as their confusion matrices are similar (see Figure 2).

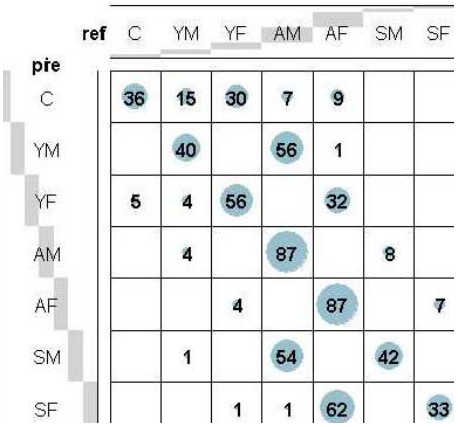

Fig. 3. Confusion matrix of Human comparison experiment on SpeechDat II. Class symbols are defined in Section 2; columns contain hypothesized classes, rows contain reference classes.

	SpeechDat II	SpeechDat_long	SpeechDat_short
Correct	2609	595	441
Total	4772	991	867
Precision	54.7%	60.0%	50.9%
Recall	69.3%	72.5%	67.2%

Table 2. Human labeling experiment on SpeechDat II data.

5. COMPARISON WITH HUMAN PERFORMANCE

The results of a baseline experiment involving Human listeners labeling the data using the same classes are shown in Figure 3. For this experiment, 30 members of our respective working groups listened to 100 randomly chosen audio files each over headphones and annotated them, covering about half of our evaluation corpus. Subjects tagged utterances which contained semantic context information (e.g. an age) which had helped them to determine age or gender of the caller. We found that only 1% of SpeechDat II utterances contain this information, while 96% of VoiceClass utterances contain context information. Discarding these utterances, the Human baseline experiment can be compared to the automatic classifier evaluation on the SpeechDat II corpus.

The overall classification accuracy on the (near complete) SpeechDat II eval set is 55%, with a precision of 69% (see Table 2). Comparing automatic and Human results, the overall of the best automatic system is comparable to Human performance, while the recall is significantly lower. The difference between long and short sentences also exists for Human labelers, although Human labelers do not perform that much worse on short sentences.

Comparing these results with other results on telephony speech [10], we find the same "centralization" trend for the perceived age and a similar performance of our Human labelers on longer utterances, even though the average sentence length of SpeechDat_long utterances is below the 40s measured in [10].

6. SUMMARY AND CONCLUSION

This paper presented a comparison of different approaches to age and gender classification on telephone speech. We find that the best automatic system performs on average comparably to Human listeners, although the performance of our classifiers is worse on short utterances. A simple "majority voting" combination study did not improve classification accuracy, presumably due to the systematic nature of confusions, which we hope to overcome in further experiments by combining System B and System C at the feature stage.

The results of a user study [7] shows that use patterns of IVR systems for senior citizens differ significantly from those of adults or young callers. Given that most confusions appear between neighboring classes, we believe the overall acceptance of IVR systems can be increased significantly by providing tailored versions of such systems, which adapt characteristics such as the degree of automation in a caller pre-selection scenario, order of presentation of options, or waiting music and advertisement to the caller.

We are therefore currently working on design guidelines and a development framework in order to easily derive user group specific versions of a baseline IVR system. In these scenarios, age and gender classification is not used to limit access (e.g. as in protection of minors), but to increase user satisfaction by providing individualized services even in the absence of knowledge about the caller's identity.

7. ACKNOWLEDGMENTS

The evaluation described in this paper was funded by Deutsche Telekom AG and Siemens AG through the "Speech Based Classification" and "Adaptive Speech Dialogs" research projects.

8. REFERENCES

[1] Edward D. Mysak, "Pitch duration characteristics of older males," *Journal of Speech and Hearing Research*, vol. 2, pp. 46–54, 1959.

[2] Sue E. Linville, *Vocal Aging*, Singular Publishing Group, San Diego, CA; USA, 2001.

[3] Christian Müller, Frank Wittig, and Jörg Baus, "Exploiting speech for recognizing elderly users to respond to their special needs," in *Proc. Eurospeech 2003*, Geneva; Switzerland, Sept. 2003, ISCA.

[4] Nobuaki Minematsu, Mariko Sekiguchi, and Keikichi Hirose, "Automatic estimation of one's age with his/ her speech based upon acoustic modeling techniques of speakers," in *Proc. ICASSP 2002*, Orlando, FL; USA, May 2002, IEEE.

[5] Izhak Shafran, Michael Riley, and Mehryar Mohri, "Voice signatures," in *Proc. ASRU 2003*, U.S. Virgin Islands, Dec. 2003, IEEE.

[6] Susanne Schötz, "Automatic prediction of speaker age using CART," Term paper for course in Forensic Phonetics, Göteborg University.

[7] Thomas Hempel, "Usability of a telephone-based speech dialogue system as experienced by user groups of different age and background," in *Proc. 2nd ISCA/DEGA Tutorial & Research Workshop on Perceptual Quality of Systems*, Berlin; Germany, Sept. 2006, ISCA.

[8] European Language Resources Association (ELRA), "http://www.speechdat.org/," http://www.elra.info/.

[9] Jitendra Ajmera, "Effect of age and gender on LP smoothed spectral envelope," in *Proc. Speaker Odyssey*. 2006, IEEE.

[10] Loredana Cerrato, Mauro Falcone, and Andrea Paoloni, "Subjective age estimation of telephonic voices," *Speech Communication*, vol. 31, no. 2–3, pp. 107–102, 2000.