# Maximum Likelihood and Maximum Mutual Information Training in Gender and Age Recognition System*

Valiantsina Hubeika, Igor Szöke, Lukáš Burget, Jan Černocký

Speech@FIT, Brno University of Technology, Czech Republic,
xhubei00@stud.fit.vutbr.cz, {szoke, burget, cernocky}@fit.vutbr.cz

**Abstract.** Gender and age estimation based on Gaussian Mixture Models (GMM) is introduced. Telephone recordings from the Czech SpeechDat-East database are used as training and test data set. Mel-Frequency Cepstral Coefficients (MFCC) are extracted from the speech recordings. To estimate the GMMs' parameters Maximum Likelihood (ML) training is applied. Consequently these estimations are used as the baseline for Maximum Mutual Information (MMI) training. Results achieved when employing both ML and MMI training are presented and discussed.

## 1 Introduction

Estimation of gender and age is an open topic in the speech processing field. When gender estimation is a simple task with two classes, age estimation is a lot more complicated due to non-linearity in changing of voice during aging. It is difficult to define precisely a border between two adjoined age groups. In this work, age groups were created experimentally according to limited amount of the available data. Nevertheless, the achieved results are optimistic.

Previously carried out studies [1], [3], [5] proved that it is possible to estimate gender and age of an unknown speaker only by listening to a low quality recording of his/her voice, such as from an analogue telephone line. Subjective gender estimation by human listeners shows very high performance. The accuracy of estimation is almost 100 %. However, estimation is not as accurate in case of children and elderly people.

During aging, changes of voice are continuous, therefore precise age estimation is unfeasible even by human listeners. Studies [1], [5] show that the accuracy in case of subjective age estimation by human listeners depends on several factors. Estimation is more precise using long sentences instead of isolated words.

* This work was partly supported by European projects AMIDA (IST-033812) and Caretaker (FP6-027231), by Grant Agency of Czech Republic under project No. 102/05/0278 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under projects No. 119/2004, No. 162/2005 and No. 201/2006. Lukáš Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383.

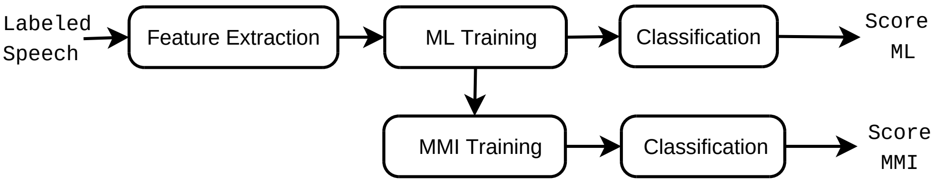

**Fig. 1.** Structure of the recognizer

An important fact is that voice of an atypical speaker can seem to be far younger or far older than he or she actually is. When using whole sentences in case of typical speakers, the error is mostly not greater than 10 years. In case of short isolated words from atypical speakers, the error can rise up to 40 years [5].

This work presents automatic gender and age estimation from telephone speech recordings based on Gaussian Mixture Models (GMM) which are proven to be a powerful tool often employed in text-independent classification tasks. The GMM parameters are estimated using ML training [6] and following MMI training [4]. The paper is organized as follows: Section 2 introduces the approach. Experiments are described in sections 3 and 4. Finally, the results are summed up in section 5.

## 2 Architecture of the Recognizer

The basic structure of the recognizer is shown in figure 1. The HTK toolkit [6] and STK toolkit from Speech@FIT [1] for HMM-based speech processing are used. Perl, C Shell and Awk scripts are used to process the data and evaluate the results.

The Czech SpeechDat-East database [2], used in the experiments, contains telephone speech recordings (8 kHz / 8 bit) from 1052 Czech speakers. 12 phonetically rich phrases from each speaker are used. The data are divided into training and test sets, that are mutually disjoint. The training set amounts to 81% of all data and consists of recordings from speakers aged 9 to 79 years. The remaining 19 % is the test set which consists of recordings from speakers aged 12 to 75 years. Distribution of single ages in the database is presented in figure 2. In both, training and test set, gender is covered equally. Altogether, 10207 recordings are used as training set and 2397 as test set. According to the transcription files, a lot of data contain speaker, background and channel noises. Only about 12 % of all the available data are considered as clear.

[1] http://www.fit.vutbr.cz/research/groups/speech/stk.html
[2] http://www.fee.vutbr.cz/SPEECHDAT-E

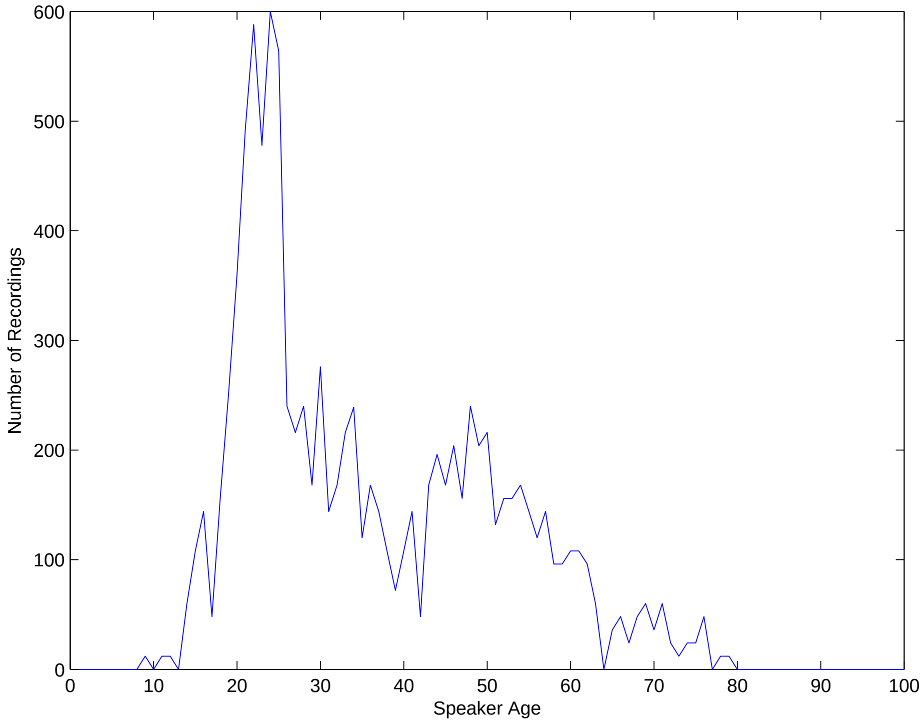

**Fig. 2.** Distribution of Single Ages in the Czech SpeechDat-East Database

### 2.1 Feature Extraction – Mel-Frequency Cepstral Coefficients

Speech is divided to frames with a sampling window of 25 ms with a shift of 10 ms. From every frame, 12 MFCC and either the energy (age estimation) or the log-energy (gender estimation) are extracted [6]. The first order and the second order time derivatives are concatenated with the base static coefficients. The final feature vector has 39 coefficients.

### 2.2 Models' Training

Gaussian mixture models are used to represent distributions of cepstral features of gender and age classes. When using GMM, the recognition process is divided to two subproblems: estimation of the parameters of GMM using a set of training samples and following classification using trained models [6]. First, models' parameters (means and covariance matrices) are estimated using Maximum Likelihood (ML) training technique [6]. ML training determines GMM's parameters that maximize the likelihood of the given data samples by estimating means and covariance matrices from all the data for given class. When models are ML

**Table 1.** Age groups with spans of 25 years and the amount of used training and test recordings.

| | Young | Middle Aged | Elderly |
|---|---|---|---|
| Range | 9..30 | 31..55 | 56..79 |
| Training Set | 4259 | 3333 | 969 |
| Test Set | 1125 | 984 | 276 |

trained, they are used as the starting point for discriminative training [4]. Discriminative training is an approach used to maximize the probability of correct decision and to minimize the classification error. MMI objective function is:

$$F_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(O_r|s_r)^{K_r} P(s_r)}{\sum_{\forall s} p_\lambda(O_r|s)^{K_r} P(s)} \tag{1}$$

where $p_\lambda(O_r|s_r)$ is likelihood of $r$-th training segment, $O_r$, given the correct transcription (gender or age) of the segment, $s_r$, and model parameters, $\lambda$. $R$ is the number of training segments and the denominator is the overall probability density, $p_\lambda(O_r)$. The prior probabilities, $P(s_r)$ and $P(s)$, are considered to be equal for all classes and are dropped. Usually, segment likelihood, $p_\lambda(O_r|s)$, is computed as multiplication of frame likelihoods incorrectly assuming statistical independence of feature vectors. The factor $0 < K_r < 1$ can be considered as a compensation for underestimating segment likelihoods caused by this assumption. This compensation factor is experimentally set to 0.01. MMI objective function (1) can be increased by re-estimating model parameters using extended Baum-Welch algorithm [2].

## 3 Gender Estimation

This work shows that automatic estimation proves to be almost as accurate as in case of subjective estimation by human listeners (see the introduction). When using all the available data and 30 Gaussian components (further adding of Gaussians shows no improvement of the result) in each gender GMM trained by ML, the accuracy is 94.64 %. With MMI re-estimation of the models' parameters, the accuracy went up to 97.41 %. Further improvement was achieved by filtering the training data. When utterances containing noise, mis-pronunciations or other defects (according to the transcription files) are discarded, the accuracy increased up to 98.25 %.

## 4 Age Estimation

Age groups are formed according to the limited size of the database. The first experiment is performed to estimate which age category the speaker belongs to.

**Table 2.** Age groups with spans of 5 years and the amount of used training and test recordings.

| | Group | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| Age from | 9 | 16 | 21 | 26 | 31 | 36 | 41 | 46 | 51 | 56 | 61 | 66 | 71 |
| Age to | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 79 |
| Training Set | 84 | 999 | 2507 | 1113 | 838 | 599 | 720 | 1020 | 755 | 514 | 287 | 202 | 192 |
| Test Set | 48 | 237 | 624 | 240 | 252 | 96 | 336 | 204 | 72 | 144 | 84 | 24 | 24 |

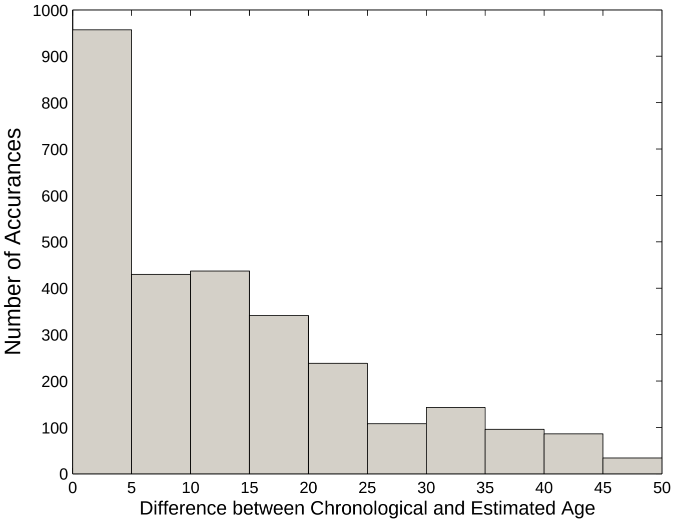

**Fig. 3.** Estimation Error Rate when Using Groups with Spans of 5 Years

Three groups are defined with spans of 25 years (table 1). The aim of the second experiment is to estimate the age more precisely. Classification was done using 13 age groups with 5 year spans (table 2).

When models are ML trained, the accuracy using groups of 25 years is 49.60 %. The most accurate estimation is obtained for speakers belonging to the group of young people (56.62 %). The age of elderly people is estimated with the greatest error (only 28.26 % accuracy). The estimation accuracy in case of the middle aged speakers is 47.56 %. This difference in accuracy is caused by non-uniform distribution of single ages in the database, where 50 % of all recordings belong to young people, 39 % belongs to middle aged people and only 11 % belongs to elderly people.

When the models' parameters are MMI re-estimated, the accuracy of classification is 60.13 %. Correct classification in case of young speakers is done for

78.49 % of all utterances. In case of old people, the accuracy decreased to 17.39 % (due to low amount of training data). For middle aged people, the estimation was correct in 52.13 % of all cases.

After, the data were divided to groups with the ranges of 5 years (Tab. [3]) and 13 GMMs were ML trained. The average difference between chronological age and estimated age is 13.71 years. After MMI training, this difference went down to 11.38 years. Maximum difference between chronological and estimated age is 50 years (1 % of all cases). For 48 % of all cases, the estimation error is not greater then 10 years. A histogram of the estimation error is presented in figure 3.

## 5 Conclusion

An acoustic recognition system for gender and age estimation was presented. For gender estimation, the accuracy is high and satisfies the expectations.

The age is estimated with errors comparable to subjective human age estimation (errors of 10 years is commonly supposed as standard) although the models of some groups are trained on relatively small amount of data. The training is negatively influenced by large amount of disturbed data contained in the training set. Also, data from atypical speakers affect correct parameter estimation of the models which impairs correct estimation of models' paramteres. A possible solution would be an iterative reduction of outliers in the training data, we are however limited by the its relatively small size.

We have shown that the MMI training increased the accuracy. While the ML training tends to cover the whole regions uniformly by Gaussians, MMI probably concentrates less on the border regions (for example 10 and 11 years) which can not be reliably distinguished anyway, and models better the central parts of age groups.

## References

1. L. Cerrato, M. Falcone, and A. Paoloni. Subjective age estimation of telephonic voices. *"Speech Communication"*, 31(2):107–112, June 2000.
2. P. Matejka, L. Burget, P. Schwarz, and J. Černocký. Brno University of Technology System for NIST 2005 Language Recognition Evaluation. In *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop*, pages 57–64, 2006.
3. K. Sikeguchi N. Minematsu and K. Hirose. Performance improvement in estimating subjective ageness with prosodic features. *"Speech Prosody"*, April 2002.
4. D. Povey. *Discriminative Training for Large Vocabulary Speech Recognition.* PhD thesis, Cambridge University, July 2004.
5. S. Schotz. A perceptual study of speaker age. Technical report, Lund University, 2001.
6. S. Young, G. Evermann, M. Gales, T. Hain, and D. Kershaw. The HTK book, 2005.