Estimation of Gender and Age from Recorded Speech

Bc. Valiantsina Hubeika [*]
xhubei00@stud.fit.vutbr.cz

Abstract: This document deals with gender and age estimation from recorded speech based on Gaussian Mixture Models (GMMs). Estimation process is composed of training of models representing appropriate speaker groups and following speaker classification using these trained models. Records from Czech SpeechDat(E) database are used as training and test data set. In order to reduce data size, Mel-Frequency Cepstral Coefficients (MFCC) are extracted speech. Additionally, discriminative training is applied to the trained models to provide further improvement of the results. Finally, obtained results are discussed and explained. Mistakes accompanying the estimation procedure are compared with the mistakes of subjective gender and age estimation by human listeners.

Keywords: gender estimation, age estimation, speech recognition, Gaussian Mixture Model (GMM), Czech SpeechDat(E), Maximum Liklihood (ML) training, discriminative training

1 Introduction

Human beings use speech as communication medium. The information speech carries is not only what the speaker wants to express but also hidden information which is present in the speaker's voice. This information is related to the speaker's personal characteristics. Some of these characteristics refer to speaker's cultural, social and geographic background, such as spoken language, level of education, accent and emotional and physical state. There are also characteristics which are speaker dependent, such as gender, age, weight and height. Previously carried out studies [5] proved that it is possible to estimate certain characteristics of an unknown speaker only by listening to a low quality recording of his/her voice, such as from an analogue telephone line. This paper deals with automatic gender and age estimation from speech records using a recogniser based on Gaussian Mixture Model (GMM).

The second section describes the extraction of speech features. The third section gives an introduction to speech classification. The experimental setup is introduced in the section 4. The fifth section describes the architecture of the recogniser. The sixth section deals with gender estimation. Age estimation is described in the seventh section. The eighth section presents the final result obtained by applying the technique called discriminative training and section 9 concludes the paper.

2 Feature Extraction

Speech carries not only information which is relevant for speech recognition. It is convenient to extract only speaker independent information from the speech. Speech sig-

[*] Faculty of Infromation Technology, Brno University of Technology, Božetěchova 2, 612 00 Brno

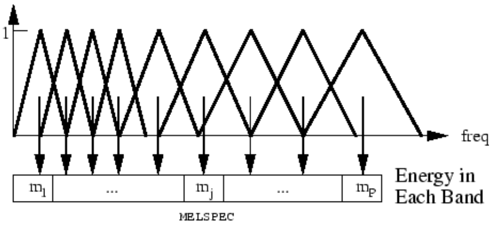

Fig. 1: Mel-Scale Filter Bank

nal parametrisation (also called feature extraction) [6] is based on assumption of quasi-stationary of speech signal within a short time segment. The whole segment can be represented by a parameter which is either scalar or vector. Methods used to parameterize an acoustic signal are called *short-time analysis methods.* In this paper, Mel-Frequency Cepstral Coefficients are extracted. Before parameterization, digital preprocessing [6] (pre-emphasis, segmentation and windowing) is done.

2.1 Mel-Frequency Cepstral Coefficients (MFCC)

The human ear resolves frequencies non-linearly across the audio spectrum. It is less sensitive to higher than to lower frequencies. It is convenient to adapt the signal to the frequency range of the human ear. This is done by applying the filterbank. Its structure is shown in the figure 1. The window of the speech data is transformed using a Fourier transformation in which the magnitude is taken. The magnitude coefficients are then calculated by correlating them with the corresponding triangular filter and the results are accumulated. Cepstral features are calculated from the log filter bank amplitudes m_j using the discrete cosine transformation:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi j}{N}(j - 0.5)\right) \tag{1}$$

where N is the number of filter bank channels which is set to 24. The lowest order cepstrum c_0 is short-term energy (the speech power). The final MFCC feature vector is composed of 39 parameters and has the following format:

c_1	...	c_{12}	E	Δc_1	...	Δc_{12}	ΔE	$\Delta^2 c_1$	...	$\Delta^2 c_{12}$	$\Delta^2 E$

where E is the energy, either c_0 or the log energy:

$$E = \log \sum_{n=1}^{N} s_n^2 \tag{2}$$

where s_n is the speech segment in time n. The log energy is normalized to the range $-E_{min}..1.0$. The first order regression coefficients (Δ) and second order regression coefficients (Δ^2) are time derivatives of the basic static parameters (c_i and the energy) and are added to model inter-frame dependencies in speech.

3 Speech Classification using Gaussian Mixture Models (GMM)

The aim of speech recognition is to arrange a mapping between the sequences of speech vectors and the appropriate transcription. In case of gender and age estimation the transcription describes the whole speech utterance (voice cues do not change within the

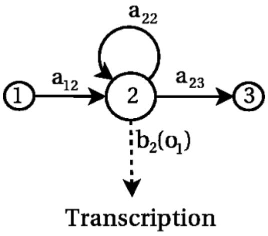

Fig. 2: Gaussian Mixture Model

record). In order to perform the mapping, for text-independent speaker recognition GMM are used (Fig. 2). A GMM is a density model which comprises a number of Gaussian mixture components (GMC). Here, GMM is three-state (where only the second state is emitting. The entrance and the exit state is added to facilitate the principle of the model.) left-right (index of the next entered state either increases or persists unchanged) finite state machine which changes its state once every time unit and every time t the second state is entered, a speech vector $\mathbf{o_t}$ is generated from the probability density $b(\mathbf{o_t})$ represented by the Gaussian Mixture Density:

$$b(\mathbf{o_t}) = \sum_{m=1}^{M} c_m N(\mathbf{o_t}; \mu_m, \Sigma_m) \tag{3}$$

where M is a number of mixture components, c_m is the weight of the m'th component and $N(\cdot; \mu, \Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ (here, the covariance matrix is diagonal which is allowed by global decorrelation of cepstral coefficients extracted from the speech):

$$N(\mathbf{o}; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \left|\Sigma\right|}} e^{-\frac{1}{2}(\mathbf{o}-\mu)} \tag{4}$$

where n is the dimensionality of $\mathbf{o}$ (length of the parameter vector, here 39).

When using GMMs, the recognition is divided to two subproblems: estimation of the parameters of GMMs using a set of training examples and when maximum likelihood is calculated, the recognition can be performed (the likelihood of each model representing certain group is calculated and the most likely model identifies the group). To determine the parameters of GMMs iterative *Baum-Welch re-estimation algorithm* [6] is applied. True model evaluation is performed using *Veterbi probability* [6].

3.1 Experimental Setup

Records from Czech SpeechDat(E) database [1] are used to train and test GMMs. Czech SpeechDat(E) is a Czech telephone speech database (sampled at 8 kHz) which contains records from 1052 callers. 12 phonetically rich phrases from each speaker are used. The total coverage is 50% for male speakers and 50% for female speakers. The whole data are divided into training and test sets. None of them contains records of speakers from the other group. The training set amounts to 81% of all data and consists of records from speakers aged 9 to 79 years. The remaining 19% are the test set which consists of records from speakers aged 12 to 75 years (the gender in both groups is covered (nearly) at the same rate). Altogether 10207 utterances are used as training set and 2397 as test set.

[1] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/czech.html

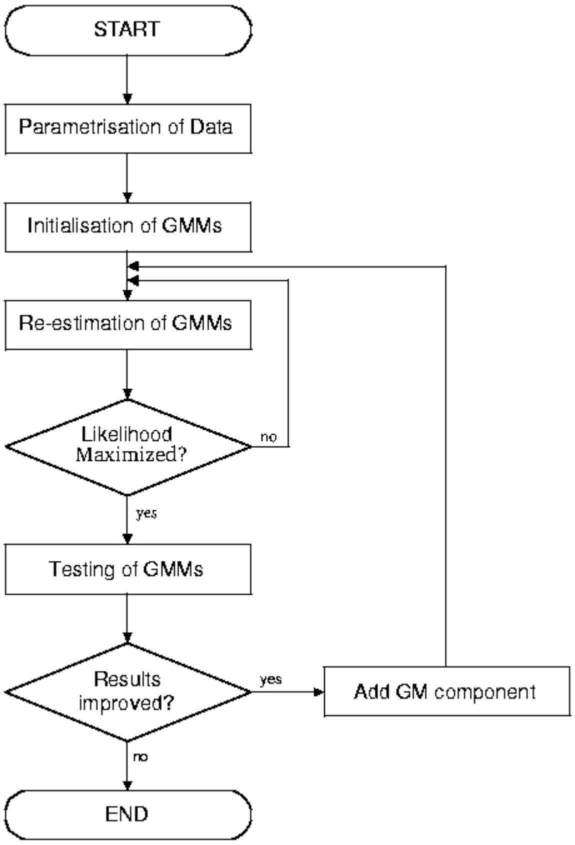

Fig. 3: Sequence of Recogniser States

4 Architecture of the Recogniser

The HTK toolkit [7] and STK toolkit from Speech@FIT [2] are used for HMM-based speech processing. Furthermore, STK provides Maximum Mutual Information (MMI) training [3] of models. In addition Perl, C Shell and Awk scripts are used to process the data and evaluate the results.

The basic structure of the recogniser is shown in figure [3]. First, MFCC feature extraction is applied to all data. The second step is the estimation of the GMM parameters. Estimation is performed in two steps: At the beginning, the parameters of the models are initialized to be identical and have state means and variances equal to the global speech mean and variance (within a certain group). After initialisation the Maximum Likelihood (ML) re-estimation [4] of the GMM means, variances and weights follows in several iterations. All parameters of GMMs are re-estimated incrementally. The progression starts with one single GMC and follows by splitting of GMC. After splitting the GMC, re-estimation is done until the likelihood is fully maximized. In case of gender estimation, models comprise 20 GMC. For age estimation, the number of GMC is increased to 30. When further splitting is does not provide any additional improvement, the MMI re-estimation of the parameters is performed using the STK toolkit.

[2] http://www.fit.vutbr.cz/research/groups/speech/stk.html

Feature Vector Label	Feature Vector	Accuracy (in %)
MFCC_E_D_A	$(12 \text{ MFCC} + E) + \Delta + \Delta^2$	95.99
MFCC_0_D_A	$(12 \text{ MFCC} + c_0) + \Delta + \Delta^2$	96.16
USER	$12 \text{ MFCC} + (\Delta + \Delta E) + (\Delta^2 + \Delta^2 E)$	95.70

Tab. 1: The Accuracy of Gender Estimation obtained using Different Coefficients.

5 Gender Estimation

The accuracy of the gender estimation was examined in many previous studies [1]. The results show that it is possible to estimate the gender of a speaker by only listening to the voice with an accuracy of almost 100%. The following experiments prove that it is possible to estimate the gender automatically with results close to subjective gender estimation by human listeners.

5.1 Experimental Results

Experiments using different types of feature vectors are performed. Different sets of records for training and testing of GMMs are used. As result, the percentage of correct estimation per utterance is evaluated. All training data are divided to 2 groups: the group of female speakers F (425) and the group of male speakers M (426). In order to find out which coefficients show the best result, 3 experiments are performed using different feature vectors. The results are presented in the table [1]. The feature vector including c_0 brings the best result and is used in all following experiments. In order to estimate how noise influences the estimation, all disturbed utterances (according to the transcription file) are omitted. The training data contains 1249 utterances (559 utterances from 154 female speakers, 690 utterances from 189 male speakers) and the testing data contains 307 utterances (129 utterances from 40 female speakers, 178 utterances from 44 male speakers). Hence the filtered training set contains 12.25 % of all available training data and the filtered test set 12.81 % of all available test data.

When the models are trained and tested on the filtered data the result is 100%. When the models are trained on the unfiltered data and tested on the filtered data the result is 99.67%. When the models are trained on the filtered data and tested on the unfiltered data the result is 96.08%. This experiment brought the lowest result. It was supposed that the reason is the small amount of used data. In order to prove this, the same amount of the unfiltered data (1249 randomly selected utterances composed of 624 utterances from 53 female speakers and 625 utterances from 53 male speakers) were used to train new models. The obtained result was lower than the result obtained from the models trained on filtered data. The accuracy is only 95.05 %. This proves that the models have to be trained on huge amount of noiseless data.

6 Age Estimation

Age estimation is more complicated than gender estimation. Precise age estimation is unfeasible even by human listeners. The estimation is always disturbed by certain deviation between the chronological age and the estimated age. Previous studies [1] and [5] show that the accuracy in case of subjective age estimation by human listeners depends on

	Young	Middle Aged	Old
Range	9..30	31..55	56..79
Train.	4259	3333	969
Test	1125	984	276

Tab. 2: Age Groups with Ranges of 25 Years and the Amount of used training and test Records.

several factors. The estimation is more precise using long sentences instead of only single words. An important fact is that voice of an atypical speaker seems to be far younger or older than it actually is. When using whole sentences in case of typical speakers, the mistake is mostly not greater than 10 years. When using short words in case of atypical speakers, the mistake can be up to 50 years [5]. The following voice cues are the most relevant in age estimation. The most dominant is the fundamental frequency (pitch) but the vocal intensity (loudness), jitter and shimmer (roughness), the formant frequencies and the spectral scope (voice quality), the duration and pausation (speech rhythm and timing) are also important. In this paper, only spectral envelope, represented by MFCC coefficients is used.

6.1 Age Groups

Groups of ages are formed because the database used in this work does not contain high enough amount of data for every individual age. This is done for all data (training data and testing data). The first experiment is performed to estimate which age group the speaker belongs to. Three groups are created with ranges of 25 years to estimate whether the speaker is young, middle aged or old (Tab. [2]). To estimate the age more accurately, groups with 5 year ranges are defined. The border age groups span more than 5 years. Altogether, 13 groups are created (Tab. [3]).

6.2 Scoring

An efficient way to evaluate the result is to calculate the average deviations between the chronological age and the estimated age. Using groups of 25 years, a one level deviation (young speakers are estimated as middle aged and vice-verse, middle aged speakers are estimated as old and vice-verse) and a two level deviation (young speakers are estimated as old and vice-verse) are calculated. In the detailed experiments, the accuracy is calculated separately for the elements from the middle and from the borders of each group.

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

Tab. 3: Age groups with Ranges of 5 Years and the Amount of used training and test Records.

6.3 Experimental Results

The accuracy using groups of 25 years is 49.60 %. The most accurate result is obtained for the speakers belonging to the group of young people (56.62 %). The age of old people is estimated with the greatest error (only 28.26 % success). The correct estimation of the middle aged is 47.56 %.

Furthermore, two groups of all elements of each testing group are created. Five border elements of each group are defined as unstable group of elements. The rest (middle elements) is considered as stable group. Utterances belonging to the stable group are estimated with a higher accuracy than the utterances from the unstable group. The result for testing the models (trained on data excluding the border elements) on the stable data is 52.92 % and for testing on the unstable data it is 40.56 %. After the data are divided to appropriate groups with the ranges of 5 years, 13 GMMs are trained. The accuracy is 22.43 %. The average deviation between the estimated age and chronological age of a speaker is 13.71 years.

7 Discriminative Training

Discriminative training (DT) is based on Maximum Mutual Information (MMI) estimation of the model parameters. The aim of MMI estimation is to make the correct hypothesis more likely and at the same time make incorrect hypothesis less likely. Information about discriminative training can be found in [3].

In this paper, discriminative training is applied to trained models which provide the best results for both, gender and age estimation. In both cases used data are unfiltered. All models are re-estimated in 20 iterations of Maximum Mutual Information (MMI) training. Further re-training does not improve these values.

7.1 Improvement of Results

The accuracy of gender estimation increases to 97.41% (from 96.16 % using the original ML-trained models). In age estimation, using groups with a range of 25 years, the accuracy improves to 60.13% (from 49.60%). A one level deviation decreases to 35.18% (from 40.84%) and a two level deviation decreases to 4.70% (from 9.56%). Using groups with a range of 5 years, the result improves to 23.85% (from 22.43%). The mean deviation decreases to 11.38 years (from 13.72 years).

8 Conclusion

This paper described the problem of gender and age estimation. A recognition system was designed and built to estimate speaker's gender and age. For gender estimation, the accuracy of the obtained results is high and satisfies the expectations. The gender is estimated with a success of 97.41 % from records of relatively low quality.

The age is estimated with errors comparable to subjective human age estimation. Errors of 10 years is commonly supposed as standard for subjective human age estimation. In this work, the average deviation between chronological and estimated age is 11.38 years. The age is estimated correctly for 23.85% of all utterances using groups of a 5 years range and for 60.13% using groups of a 25 years range. A disadvantage is the lack of suitable

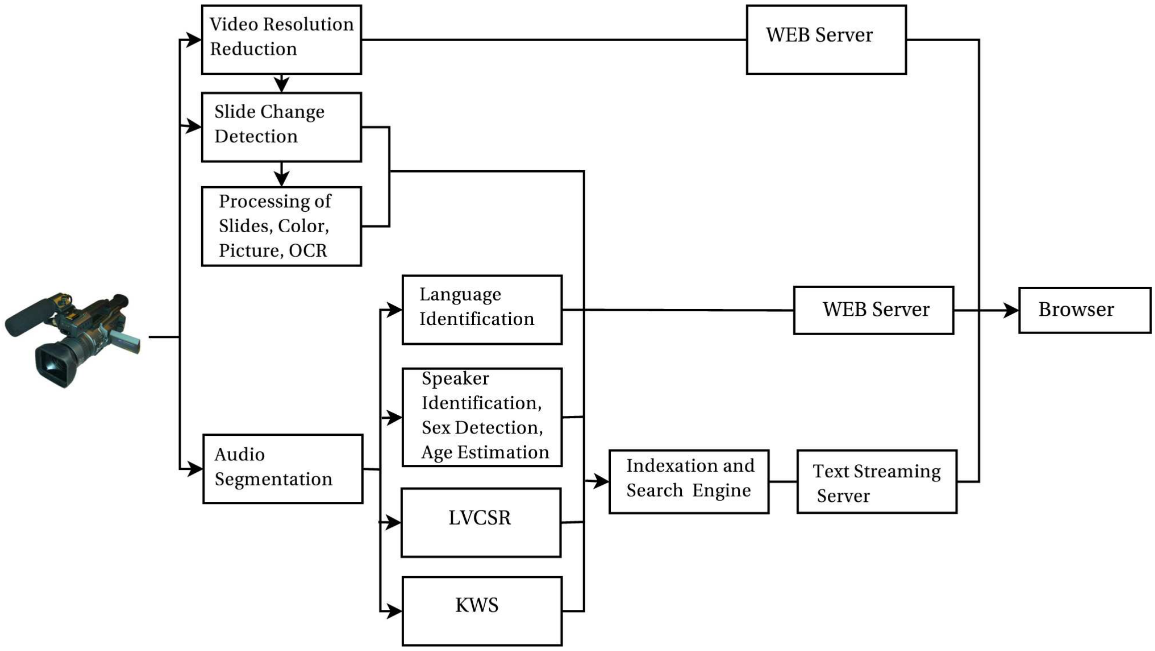
Video Resolution Reduction
Slide Change Detection
Processing of Slides, Color, Picture, OCR
WEB Server
Language Identification
WEB Server
Browser
Speaker Identification, Sex Detection, Age Estimation
Audio Segmentation
Indexation and Search Engine
Text Streaming Server
LVCSR
KWS

Fig. 4: Multi-Modal Recognition, Indexing and Search System

data. Some groups' models are trained on relatively small amounts of data. The training set contains a large amount of disturbed data. Records from atypical speakers affect the training of the models which makes correct evaluation less likely.

Both estimations will be used in a multi-modal recognition, indexing and search system developed at the Faculty of Information Technology of BUT (Fig. 4).

9 Acknowledgement

I would like to thank my supervisor Igor Szöke for his great help, Lukáš Burget for providing me valuable information and Jan Černocky for proof-reading the paper.

Bibliography

1. CERRATO L., FALCONE M., PAOLONI A. *Subjective Age Estimation of Telephonic Voices* [online]. Speech Communication 31, p. 107 – 112. Available: `<http://www.speech.kth.se/loce/papers/specom\_29-3.doc>`
2. MINEMATSU N., SEKIGUCHI K., HIROSE K. *Performance Improvement in Estimating Subjective Ageness with Prosodic Features* [online]. Speech Prosody 2002, Apr. 2002. Available: `<http://www.lpl.univ-aix.fr/sp2002/pdf/minematsu-etal.pdf>`
3. POVEY D. *Discriminative Training for Large Vocabulary Speech Recognition.* PhD. Thesis, Cambridge University, July, 2004.
4. RABINER L.R. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition* [online]. Available: `<http://www.cs.ubc.ca/murphyk/Bayes/rabiner.pdf>`
5. SCHÖTZ S. *A perceptual Study of Speaker Age* [online]. 2001. Available: `<http://www.ling.lu.se/disseminations/pdf/49/bidrag35.pdf>`
6. YOUNG S., EVERMANN G., KERSHAW D., MOORE D., ODELL G. J., OLLASON D., VALTCHEV V., WOODLAND D. *The HTK Book*[online]. Cambridge University Engineering Department, 2002. Available: `<http://www.ee.columbia.edu/labrosa/doc/HTKBook>`