ESTIMATION OF GENDER AND AGE

Valiantsina Hubeika
Bachelor Degree Programme (1), FIT BUT
E-mail: xhubei00@stud.fit.vutbr.cz

Supervised by: Igor Szöke
E-mail: szoke@fit.vutbr.cz

ABSTRACT

G ender and age estimation based on Gaussian Mixture Models (GMM) is introduced. Records from the Czech SpeechDat(E) database are used as training and test data set. In order to reduce the data size, Mel-Frequency Cepstral Coefficients (MFCC) are extracted from the speech recordings. Maximum Likelihood (ML) training is applied to estimate the models' parameters and additionaly discriminative training (DT) is applied to the trained models to provide further improvement of the results.

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. Previously carried out studies [1], [2], [4] 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. In this approach automatic gender and age estimation from speech recordings is performed based on Gaussian Mixture Models (GMM) which is proven to be a powerful tool often employed in text-independent classification tasks. The GMM parameters are estimated using ML training [5] and DT training [3]. The paper is oranized as follows: Section 2 introduces the approach. Experiments are described in sections 3 and 4. Finally, the results are sumed up in section 5.

2 ARCHITECTURE OF THE RECOGNISER

The basic structure of the recognizer is shown in figure 2. The HTK toolkit [5] and STK toolkit from Speech@FIT (http://www.fit.vutbr.cz/research/groups/speech/stk.html) are used for HMM-based speech processing. In addition Perl, C Shell and Awk scripts are used to process the data and evaluate the results.
The Czech SpeechDat(E) database (http://www.fee.vutbr.cz/SPEECHDAT-E/sample/czech.html), used in experiments, contains telephone speech recordings (8 kHz) from 1052 Czech callers. 12 phonetically rich phrases from each speaker are used. The data are divided into training and test mutually disjoint sets. The training set amounts to 81% of all data and consists of recordings from speakers aged 9 to 79 years. The remaining 19 % are the test set which consists of recordings from speakers aged 12 to 75 years.

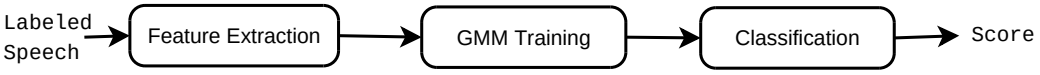

Figure 1: Structure of the Recognizer

2.1 FEATURE EXTRACTION – MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Feature extraction [5] is based on assumption that speech is quasy-stationary within a short time segment. Speech is divided to frames with a sampling window of 25 ms weighted by a Hammimg window and consequentely transformed to MFCC [5] (a standard choise for most of recognition systems).

2.2 GAUSSIAN MIXTURE MODELS (GMM)

The aim of speech recognition is to arrange mapping between a sequence of speech vectors and the appropriate transcription. GMM are used as a classification tool for the reason that distribution of the data belonging to a certain group is identical to the Gaussian distribution. For every speech vector $\mathbf{o}_t$, transcription is estimated from the probability density $b(\mathbf{o}_t)$ represented by the Gaussian Mixture Density:

$$b(\mathbf{o_t}) = \sum_{m=1}^{M} c_m \frac{1}{\sqrt{(2\pi)^n |\Sigma_m|}} e^{-\frac{1}{2}(\mathbf{o}_t - \mu_m)} \tag{1}$$

where M is a number of mixture components, c_m is the weight, μ_m is the mean and Σ_m is the covariance matrix of the m'th component, n is the dimensionality of $\mathbf{o}$.
When using GMMs, the recognition is divided to two subproblems: estimation of the parameters and of GMMs using a set of training examples and recognition itself [5]. Maximum likelihood estimation determines GMM's parameters that maximize the probability of the given data samples [5]. Discriminative training is an approach used to maximize the right estimation and to minimize the classification error [3].

3 GENDER ESTIMATION

Previous studies [1] show that it is possible to estimate gender of a speaker by only listening to the voice with an accuracy of almost 100 %. This work shows that automatic estimation proves to be almost as accurate. When using all the available data and 30 Gaussian components in each GMM, the accuracy is 97.41 %. In case of utterances containing noise, mis-pronunciations or other deffects are discarded the accuracy improved upto 99.4 %.

4 AGE ESTIMATION

Age estimation is a more complex process. 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], [4] show that the accuracy in case of subjective age estimation by human listeners depends on 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 far older than it actually is. When using whole sentences in

	Young	Middle Aged	Old
Range	9..30	31..55	56..79

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

	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

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

the 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 40 years [4].
Age groups are formed due to limited size of the database. The first experiment is performed to estimate which age group the speaker belongs to. Three groups are defined with spans of 25 years (Tab. [2]). The accuracy of classification is 60.13 %. The most accurate estimation is achieved for the speakers belonging to the group of young people (78.49 %) due to the highest amount of training data. To estimate age more accurately, 13 groups with 5 year spans are defined (Tab. [3]). An efficient way to evaluate the result is to calculate the average deviations between the chronological age and the estimated age which is 11.38 years.

5 CONCLUSION

The problem of gender and age estimation was discussed. For gender estimation, the accuracy of the achieved results 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 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 techniques can be advantageously used in speech search system currently developed in Speech@FIT group.

REFERENCES

[1] L. Cerrato, M. Falcone, and A. Paoloni. Subjective age estimation of telephonic voices. *"Speech Communication"*, 31(2):107–112, jun 2000.

[2] K. Sikeguchi N. Minematsu and K. Hirose. Performance improvement in estimating subjective ageness with prosodic features. *"Speech Prosody"*, apr 2002.

[3] Dan Povey. *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, jul 2004.

[4] Susanne Schotz. A perceptual study of speaker age. Technical report, Lund University, 2001.

[5] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, and Dan Kershaw. The htk book, 2005.