INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speaker age classification and regression using i-vectors

Joanna Grzybowska, Stanisław Kacprzak

AGH University of Science and Technology, Poland
{gjoanna, skacprza}@agh.edu.pl

Abstract

In this paper, we examine the use of i-vectors both for age regression as well as for age classification. Although i-vectors have been previously used for age regression task, we extend this approach by applying fusion of i-vectors and acoustic features regression to estimate the speaker age. By our fusion we obtain a relative improvement of 12.6% comparing to solely i-vector system.

We also use i-vectors for age classification, which to our knowledge is the first attempt to do so. Our best results reach unweighted accuracy 62.9%, which is a relative improvement of 16.7% comparing to the best results obtained in age classification task at *Age Sub-Challenge* at Interspeech 2010.

Index Terms: speaker age recognition, regression, classification, computational paralinguistics

1. Introduction

The cues about the age of a person can be observed in the voice due to numerous anatomical and physiological aspects that change during our lifetime [1]. Learning about the age of the speaker fulfills the profile of a speaker [2–5]. Automatic speaker age recognition can have forensic and commercial applications, e. g. narrowing down suspects or adjusting the performance of an Interactive Voice Response system to a specific age group.

In recent years many methods have been applied to recognize the age of the speaker. We can distinguish two approaches within age recognition: age estimation and age classification.

For age estimation different types of regression methods have been used. The most recent are based on the i-vector extraction [6, 7]. The i-vectors were introduced in [8] and were initially used for speaker recognition. In [6] the authors use Within-Class Covariance Normalization (WCCN) for session variability compensation. They also apply the Least Squares Support Vector Regression (LSSVR) to estimate the age of the speaker. In [7] the authors examine the Artificial Neural Network (ANN) back-end for an i-vector based age estimation system. They report that the change of back-end for age estimation did not affect the accuracy significantly and focusing more on front-end processing is therefore advised.

The age classification problem was previously examined in [9–12]. In the *Age Sub-Challenge* at Interspeech 2010 the task was to assign a correct age class to an unseen utterance. The Support Vector Machine (SVM) baseline for age and gender classification is provided in [9]. In [11] the authors present an automatic speaker age and gender identification approach which combines seven different methods at both acoustic and prosodic levels. In [10] the authors propose a fuzzy SVM for age and gender classification. In [12] the authors use GMMs with SVM followed by the linear Gaussian back-ends and the logistic regression-based fusion. The best results are obtained with the fusion of several sub-systems.

In our paper we examine both approaches to age recognition, namely age regression and age classification, and we report our results for the same database. To our knowledge, the i-vectors have not been previously used for the task of age classification. For the age regression task, we focus on front-end processing by applying two approaches, the i-vectors and acoustic features, and their fusion. We also combine age regression with age classification by mapping the results from the regression system to age classes and we compare this approach with age classification based on cosine distance scoring. We also examine the effect of WCCN on our data.

2. Speaker age regression

Speaker age regression is the task of predicting from an unseen utterance X_{test} the age of the speaker y_{test} as closely matched to the speaker's true age as possible. The prediction system is trained on a set of N training utterances X_n with their corresponding age labels y_n.

To provide the reference level of accuracy (prior) for our database we use the average age of the training data as an estimation function, $g(X_{\text{test}}) = \frac{1}{\text{N}} \sum_n y_n$, where N is the number of training segments and y_n is the actual age of nth training speaker.

2.1. Front-end

In our first subsystem, we use Mel Frequency Cepstral Coefficients (MFCC) as input features for the i-vector extraction. We calculate 19 MFCCs with their energy, deltas and double deltas. 60-dimensional feature vectors are normalized using Cepstral Mean and Variance Normalization (CMVN). We use those features both in regression and classification tasks. In our second subsystem we use a set of 450 acoustic features briefly described in section 2.3.

2.2. I-vectors for age estimation

In an i-vector extraction, a low dimensional total variability space T accounts for both channel and speaker variabilities. I-vectors are widely used for different tasks in the speech processing domain. The details concerning the i-vector extraction can be found in [8]. In practice, we use the Matlab MSR Identity Toolbox [13] for the i-vector extraction. In our work, the i-vector space dimensionality is 400, like in [6] and [7].

2.3. Acoustic features for age estimation

Feature sets were extracted with openSMILE [14] and consist of 450 features for each utterance in aGender corpus [15]. Those acoustic, prosodic and voice quality features are low-level descriptors (e. g. MFCCs, LSP Frequency, F0, voicing probability,

http://dx.doi.org/10.21437/Interspeech.2016-1118

Jitter, Shimmer) and their statistics (e. g. mean, standard deviation, skewness, kurtosis, percentile 1/99) [9]. Further in our paper we refer to those features as *acoustic features*.

We standardize our training and testing data by subtracting mean and dividing by standard deviation of the aGender testing set (17 332 utterances).

2.4. Back-end

To compensate for session and channel variability we use WCCN for the speakers as classes. It was previously shown in [6] that this approach improves the age regression accuracy. The details concerning WCCN can be found in [16].

As shown in [6], the LSSVR estimates speakers' age more accurately than the SVR. It is also faster in model training and faster to tune because of fewer hyper-parameters. Thus, in our work we use the LSSVR. In practice, we use the LS-SVMlab toolbox [17] with the radial basis function (RBF) kernel. It was shown in [18] that it gives better age predictions than the linear kernel. We use 10 fold cross-validation on a subset of training set to tune the hyperparameters: the regularization parameter γ and the Gaussian width σ^2. In our experiments, the training data comprises of females, males and children.

2.5. I-vectors and acoustic features fusion

In this experiment we combine the results from the i-vector and the acoustic subsystems. In this new approach, the predicted age from both subsystems is used as an input to higher level regression model. For this purpose we use the Generalized Regression Neural Network (GRNN) [19] designed using Matlab *newgrnn* function.

3. Speaker age classification

Speaker age classification is the task of determining from an unseen utterance X_{test} the age class of the speaker y_{test} corresponding to the speaker's true age. The classification system is trained on a set of N training utterances X_n with their corresponding age class labels y_n. The reference level of accuracy (prior) for a classification system with N possible age classes is $g(X_{\text{test}}) = \frac{1}{\text{N}}$.

3.1. Back-end

Our speaker age classification system is based on the i-vectors. We extract one i-vector for each speaker in the aGender database. Then, we use two different approaches for reporting our results:

1. Classification based on the cosine distance scoring (CDS classification).
2. Classification by mapping the results from the regression system (mapped classification).

In the first approach, we use the cosine distance scoring

$$\text{CDS} = \frac{w_{\text{test}}^{\text{T}} w_{\text{tar},a}}{\|w_{\text{test}}\| \|w_{\text{tar},a}\|} \tag{1}$$

to measure the similarity between two i-vectors – the test i-vector (w_{test}) and the i-vectors of the target age class a ($w_{\text{tar},a}$). This score is used to make the final classification decision. The i-vector of the target age class is the average of the i-vectors for all training speakers in the age class a, i. e.

$$w_{\text{tar},a} = \frac{1}{\text{N}_a} \sum_{i=1}^{N_a} w_{i,a}, \tag{2}$$

where N_a is the number of the training speakers in the age class a and $w_{i,a}$ is the i-vector of the training speaker i from the age class a.

In the second approach we use the predicted age labels from the regression system y_{test} and the gender labels to map the speakers to seven age classes described in section 4.1.

4. Experimental setup

To train the Universal Background Model (UBM) and T-matrix we use about 90 hours of speech from YouTube, mostly in English. We extend this database with other databases in German [15, 20–24] – they contain approximately 19 hours of speech. An extension of only English speech was performed to reduce the language mismatch between training and testing databases. Such a mismatch significantly decreases the performance of the i-vector based age estimation system [6]. Furthermore, the authors in [25] report that for an accent recognition system, hyperparameters (UBM and T-matrix) should be trained on as closely matched data as possible. The UBM consists of 1024 mixture components.

4.1. Database description and modifications

We used the aGender database to evaluate the performance of our approaches. This database was used in the *Age Sub-Challenge* at Interspeech 2010 where the task was to classify an unseen utterance to one of seven age and gender classes:

ch children (age 7-14)
yf young females (age 15-24)
ym young males (age 15-24)
af adult females (age 25-54)
am adult males (age 25-54)
sf senior females (age 55-80)
sm senior males (age 55-80)

The aGender database is comprised of about 47 hours of speech (free wording and fixed phrases) recorded in up to 6 sessions for each speaker. Each speaker produced on average 68 utterances. The average duration of an utterance is 2.58 seconds. The aGender database is divided into training (471 speakers), development (299 speakers) and testing set (184 speakers). The age range for speakers in aGender database is 7-80.

Whereas acoustic features described in section 2.3 are calculated for each utterance for each speaker, the i-vectors were extracted from concatenated utterances per speaker. This concatenation was motivated by the fact that the i-vector extraction should be performed on rather longer utterances. Mean utterance length after concatenation is 88 seconds.

In our work, we also examined the influence of WCCN on the recognition accuracy. The authors in [12] did not observe any significant gain due to the use of the the channel compensation technique for utterances in the aGender database. They assume it is caused by the short duration of testing utterances. For this reason, we partition the previously concatenated utterances to obtain longer utterances and then perform WCCN on the extracted i-vectors. We chose to partition utterances of at least 30 seconds. For such an utterance we create N new segments such as the new segment comprises of random 70% of baseline concatenated utterance. As a result, we obtain N+1 utterances per speaker, which we call segments in this paper.

4.2. Performance metric

For the age regression evaluation we use Mean Absolute Error (MAE) and Pearson's correlation coefficient. MAE is defined as

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{n=1}^{N} |\hat{y}_n - y_n| \tag{3}$$

where N is the number of testing segments, $\hat{y}_n$ is the predicted age and y_n is the actual age of nth speaker. The Pearson's correlation coefficient between true age y_n and the age predicted by a regression model $\hat{y}_n$ is our second performance metric:

$$\rho = \frac{1}{\mathrm{N}-1} \sum_{n=1}^{N} \left(\frac{\hat{y}_n - \mu_{\hat{y}}}{\sigma_{\hat{y}}} \right) \left(\frac{y_n - \mu_y}{\sigma_y} \right) \tag{4}$$

where $\mu_{\hat{y}}$ and μ_y are mean values of estimated and true ages, respectively; $\sigma_{\hat{y}}$ and σ_y are standard deviation values of estimated and true ages, respectively.

We also calculate the relative improvement of MAE to the prior system in order to provide more objective measure of comparison to the baseline system in [6]. Direct MAE comparison would not be reliable because the age range used in our work was wider than in [6]. We define the relative improvement $i_{\mathrm{MAE\%}}$ as follows:

$$i_{\mathrm{MAE\%}} = \frac{\mathrm{MAE}_{\mathrm{prior}} - \mathrm{MAE}}{\mathrm{MAE}_{\mathrm{prior}}} \cdot 100\% \tag{5}$$

For the age classification evaluation we use unweighted accuracy (%UA), which is not weighted with respect to the number of instances per class. This measure was also used to evaluate the *Age Sub-Challenge* results at Interspeech 2010, because the distribution of speakers among classes in the used aGender database is not balanced.

5. Results and discussion

All our results are reported jointly for all the speakers: children, females and males.

5.1. Baseline regression results

The authors in [6] and [7] report their results separately for males and females. The MAE in [6] is 6.53 for males and 5.78 for females, ρ is 0.73 for males and 0.81 for females. The MAE in [7] is 6.35 for males and 5.49 for females, ρ is 0.73 for males and 0.81 for females. The authors in [6] provide MAE values for prior estimator for their training data. Based on those values, we calculated the corresponding relative improvement $i_{\mathrm{MAE\%}}$, which is 35% for males and 45% for females. Jointly for all speakers, MAE is improved by 41% relative to the prior for their database.

5.2. Our regression results

5.2.1. I-vector based subsystem

We perform 15-fold cross-validation on a database comprising of the aGender training and development set. In the i-vector subsystem this database consists of 770 utterances concatenated per speakers. Table 1 presents MAE, ρ and $i_{\mathrm{MAE\%}}$ for the i-vector based subsystem with and without WCCN jointly for males, females and children. With the WCCN transformation we were able to decrease the MAE by 0.28, thus we use this configuration in the fusion of the i-vector and the acoustic features based subsystems. In table 2 we also show results with the gender division in the testing phase. Comparing our results to the results in [6] and [7], MAEs are bigger in our case, which is caused by wider age range and thus greater prior MAE for our database. In terms of $i_{\mathrm{MAE\%}}$ for all speakers our system reaches 50%, while the one in [6] reaches 41%.

Table 1: *Regression scores for (1) reference estimation function, (2) the i-vector based subsystem with and without WCCN transformation, (2) acoustic features based subsystem with and without feature standardization, (3) and GRNN fusion of the i-vector subsystem with WCCN and the acoustic features subsystem with standardization.*

Configuration	MAE	ρ	$i_{\mathrm{MAE\%}}$
prior	19.33	0.00	0%
i-vectors	9.96	0.81	48%
i-vectors (WCCN)	**9.68**	**0.83**	**50%**
Acoustic features (no standardization)	14.46	0.56	25%
Acoustic features (standardization)	**12.96**	**0.75**	**33%**
GRNN fusion	**8.46**	**0.86**	**56%**

Table 2: *Regression scores for the i-vector based subsystem with WCCN transformation for males, females and children.*

speakers	MAE	ρ	$i_{\mathrm{MAE\%}}$
males	10.63	0.76	39%
females	9.77	0.80	45%
children	6.47	0.18	77%

5.2.2. Acoustic features based subsystem

In the acoustic features subsystem we use 53 074 utterances for 770 speakers. We predict the age label $\hat{y}_n$ for each utterance and the final age of each speaker is the average age of predicted age labels for the corresponding utterances. In this subsystem we also perform 15-fold cross-validation. The results for all speakers are presented in table 1. With features standardization we were able to decrease the MAE by 1.5, thus we use this configuration further in the fusion of the i-vector and the acoustic features based subsystems. For standardized acoustic features based subsystem, MAE for males is 13.51 and $i_{\mathrm{MAE\%}}$ is 22%. For females, MAE is 12.44 and $i_{\mathrm{MAE\%}}$ is 29%. For children, MAE is 12.91 and $i_{\mathrm{MAE\%}}$ is 55%. Pearson's correlation coefficient ρ for males, females and children is 0.65, 0.73 and 0.46 respectively.

5.2.3. Fusion of i-vector and acoustic features based subsystems

We combined the results obtained from the i-vector and the acoustic subsystems by using predicted age as input to another regression model. Table 1 presents MAE, ρ and $i_{\mathrm{MAE\%}}$ obtained by GRNN from 15-fold cross-validation. Fusion of both subsystems (the i-vectors with WCCN and the acoustic features with standardization) allows us to achieve better results. MAE of 8.46 was achieved with GRNN model. MAE for males is 9.19, for females it is 8.90 and for children it achieves 4.55. With this fusion we obtain the improvement of MAE of 1.22 comparing to our subsystem based solely on the i-vectors (which is a relative improvement of 12.6%) and 4.5 compared to our subsystem based solely on the acoustic features. Jointly for all speakers, MAE is improved by 10.87 comparing to the prior.

5.3. Baseline classification results

In [26] the authors compare their age and gender classification results with human performance on the same database. The overall classification accuracy for human listeners was 55%. Human perception of age is influenced by numerous phonetic and non-phonetic factors: speaker, listener, speech-sample and task-related [1]. The results for baseline systems are reported for the aGender development set. The %UA result for SVM baseline system for the age and gender classification task (classification to one of seven classes) was 44.2% [9]. In [11] the %UA result for the age and gender classification reached 50.3% for the system consisting of the fusion of seven proposed methods. In [10] the obtained %UA for age and gender classification task is 45.2% using fuzzy SVM method. The best results in *Age Sub-Challenge* at Interspeech 2010 were obtained in [12], the %UA is 53.9%, where the authors use fusions of several subsystems.

5.4. Our classification results

For the task of classification, we use the aGender training set as our training data and the aGender development set as our testing data, thus we can directly compare our results to other results reported in the literature. Table 3 presents the results for our two approaches to classification.

Table 3: *Classification results based on (1) cosine distance scoring between test i-vector and i-vectors of target age class and (2) based on mapping the results from i-vector based regression system to age classes.*

Classification approach	%UA
CDS classification	**62.9%**
mapped classification	54.9%

With the CDS classification, we obtain %UA = 62.9% while the best result obtained in the age classification task at the *Age Sub-Challenge* at Interspeech 2010 on the aGender database was 53.9%. This is a relative improvement of 16.7%. Also, our score outperforms the classification accuracy for human listeners (55%) reported in [26].

In our second approach we examined the influence on accuracy of mapping the results from the regression system to the age classes. We obtained %UA = 54.9%, thus the %UA decreased by relative 12.7%. This result indicate that the age classification approach based on comparing testing utterances to previously trained age class models is more accurate in determining the age class of the speaker than mapping the regression results to the age classes.

Table 4 shows types of confusions made by our CDS age classification system for the age classes described in section 4.1. We can see that the classifier detects gender with high accuracy. For females the accuracy is 96.1%. The remaining 3.9% of females were always recognized as children, none of the females were recognized as males. Moreover, most of the females missclassified as children were young (70%). The rest (30%) were adult, none of the missclassified women were senior. These results can be explained by the similarities in frequencies between children and young female voices (high fundamental frequency for children and young females).

For males the accuracy is 96.0%. The remaining 4% of males were always recognized as senior females. The majority of mistakes (78%) were made for young males, and the rest (22%) for senior males. These results can also be explained by the similarities between frequencies for male and senior female voices, for whom the fundamental frequency decreases with age.

Children were classified with 57.9% accuracy. The most miss-classifications (21.1%) were made with young females.

Table 4: *Confusion matrix for age classification system based on cosine distance scoring*

Class. acc.(%)		Predicted class ch	yf	ym	af	am	sf	sm
	ch	**57.9**	21.1	5.3	7.9	0	7.9	0
	yf	5.4	**70.3**	0	21.6	0	2.7	0
True	ym	0	0	**81.3**	0	12.5	6.3	0
class	af	2.3	13.6	0	**34.1**	0	50.0	0
	am	0	0	33.3	0	**31.0**	0	35.7
	sf	0	2.0	0	16.0	0	**82.0**	0
	sm	0	0	1.8	0	16.1	1.8	**80.4**

6. Conclusions

In this paper we use the i-vectors modeling for age estimation as well as for age classification. We report our results on the aGender database, for which the age range of the speakers is wider and thus the prior Mean Absolute Error (MAE) is higher than for the databases used previously in the literature for the age regression task. To compare our results with the baseline results we use the relative improvement to the prior value.

In age regression, by performing WCCN on partitioned utterances in a way described in the paper, we decrease MAE by 0.28. By the GRNN fusion of the i-vectors and the acoustic features we obtain the best relative improvement 56%, while for our subsystem based solely on i-vectors the relative improvement is 50%, and our baseline relative improvement in [6] is 41%. In our regression approach, higher recognition rate was obtained for females than for males. To our knowledge, our work is the first attempt to estimate the age of children using i-vectors.

In age classification using i-vectors the accuracy reached 62.9%, which is the relative improvement 16.7% over the best result at *Age Sub-Challenge* at Interspeech 2010. In our classification approach, we obtained higher accuracy for males than for females, differently than in age regression approach, which suggest that those two approaches are complementary. Age classification based on the CDS scoring gave better results than mapping the results from the regression system. But age classification has also drawbacks over the age regression approach, i. e. two speakers with small age difference can be placed in different age groups. To overcome this problem we plan to investigate the age classification approach with narrower and overlapping age intervals.

We also plan to: (1) fuse the i-vectors and the acoustic features for the age classification task, (2) examine the influence of WCCN on the acoustic features, (3) use the a priori knowledge about the age distribution in the society to model the age classes.

7. Acknowledgements

The project was supported by the AGH University of Science and Technology granted by decision 15.11.120.885.

8. References

[1] S. Schötz, "Perception, analysis and synthesis of speaker age," Ph.D. dissertation, Lund University, 2006.

[2] M. Witkowski, M. Igras, J. Grzybowska, P. Jaciòw, J. Gałka, and M. Ziòłko, "Caller identification by voice," in *Pacific Voice Conference (PVC), 2014 XXII Annual*, April 2014, pp. 1–7.

[3] J. Gałka, J. Grzybowska, M. Igras, P. Jaciòw, K. Wajda, M. Witkowski, and M. Ziòłko, "System supporting speaker identification in emergency call center," in *Proceedings of Interspeech 2015*. International Speech Communication Association, September 2015, pp. 724–725.

[4] J. Grzybowska and M. Ziòłko, "I-vectors in gender recognition from telephone speech," in *Proceedings of the Twenty-First National Conference on Applications of Mathematics in Biology and Medicine 2015*. Institute of Applied Mathematics and Mechanics, University of Warsaw, September 2015, pp. 57–62.

[5] M. Witkowski, J. Gałka, J. Grzybowska, M. Igras, P. Jaciòw, and M. Ziòłko, "On-line caller profiling solution for call centre," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, to appear.

[6] M. H. Bahari, M. McLaren, H. Van hamme, and D. A. van Leeuwen, "Speaker age estimation using i-vectors," *Eng. Appl. Artif. Intell.*, vol. 34, no. C, pp. 99–108, Sep. 2014.

[7] A. Silnova, O. Glembek, T. Kinnunen, and P. Matějka, "Exploring ann back-ends for i-vector based speaker age estimation," in *Proceedings of Interspeech 2015*, vol. 2015, no. 09. International Speech Communication Association, 2015, pp. 3036–3040.

[8] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[9] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *In Proc. Interspeech*, 2010.

[10] P. Nguyen, T. Le, D. Tran, X. Huang, and D. Sharma, "Fuzzy support vector machines for age and gender classification," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 2806–2809.

[11] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 151–167, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2012.01.008

[12] M. Kockmann, L. Burget, and J. Cernocky, "Brno university of technology system for interspeech 2010 paralinguistic challenge," in *Proceedings of INTERSPEECH 2010*, 2010, pp. 2822–2825.

[13] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1.0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, November 2013. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=205119

[14] F. Eyben, M. Wllmer, and B. Schuller, "Openear 2014: Introducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, Sept 2009, pp. 1–6.

[15] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proceedings of the Language and Resources Conference (LREC)*, 2010.

[16] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. of ICSLP*, 2006, pp. 1471–1474.

[17] "LS-SVMlab toolbox," http://www.esat.kuleuven.be/sista/lssvmlab/, [Online; accessed March-2016].

[18] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1975–1985, Sept 2011.

[19] D. F. Specht, "A general regression neural network," *Neural Networks, IEEE Transactions on*, vol. 2, no. 6, pp. 568–576, 1991.

[20] "Corpus ZIPTEL," http://www.bas.uni-muenchen.de/forschung/Bas/BasZIPTELeng.html, [Online; accessed March-2016].

[21] "PhonDat 1," http://www.bas.uni-muenchen.de/forschung/Bas/BasPD1eng.html, [Online; accessed March-2016].

[22] R. Cole and Y. Muthusamy, "Ogi multilanguage corpus," *Linguistic Data Consortium, Philadelphia, USA*, 1994.

[23] "www.voxforge.org," http://www.voxforge.org/, [Online; accessed March-2016].

[24] "Open acoustic models and speech data for German speech recognition," https://www.nist.gov/itl/iad/mig/ivec.cfm, [Online; accessed March-2016].

[25] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C. Lee, "i-vector modeling of speech attributes for automatic foreign accent recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 1, pp. 29–41, 2016. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2015.2489558

[26] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–1089–IV–1092.