Age and Gender Recognition Based on Multiple Systems - Early vs. Late Fusion

Tobias Bocklet[1], Georg Stemmer[2], Viktor Zeissler[3], Elmar Nöth[1]

[1] Chair of Pattern Recognition, University Erlangen-Nuremberg, Germany
[2] SVOX, Munich, Germany
[3] Elektrobit Germany, Erlangen, Germany

tobias.bocklet@informatik.uni-erlangen.de

Abstract

This paper focuses on the automatic recognition of a person's age and gender based only on his or her voice. Up to five different systems are compared and combined in different configurations: three systems model the speaker's characteristics in different feature spaces, i.e., MFCC, PLP, TRAPS, by Gaussian mixture models. The features of these systems are the concatenated mean vectors. System number 4 uses a physical two-mass vocal model and estimates in a data-driven optimization procedure 9 glottal features from voiced speech sections. For each utterance the minimum, maximum and mean vectors form a 27-dimensional feature vector. The last system calculates a 219-dimensional prosodic feature set for each utterance based on voice and unvoiced speech segments. We compare two different ways to fuse the different systems: First, we concatenate the system on feature level. The second way of combination is performed on score level by multi-class logistic regression. Despite there are just minor differences between the two approaches, late fusion is slightly superior. On the development set of the Interspeech Agender challenge we achieved an unweighted recall of 46.1 % with early fusion and 47.8 % with late fusion.

Index Terms: acoustic analysis, classification, Gaussian mixture models

1. Introduction

Speech segments uttered by humans do not only contain the (speaker-independent) semantics of the spoken text but also non-verbal (speaker-dependent) characteristics.
This paper addresses the automatic recognition of non-verbal characteristics, i.e., a person's gender and age, by means of automatic speech processing. The problem of age- and gender recognition is treated as combined classification problem of seven classes. These classes and the datasets used for training and evaluation of our system(s) are described in [1]. The task of age and gender classification based on a person's voice became quite popular within the last few years. The most successful systems so far model each speaker by Gaussian Mixture Models of short-time spectral features and use either Gaussian classifiers or SVMs for classification [2, 3].
In this paper we compare five different systems for age and gender recognition. Three systems are based on spectral features with short- and longer temporal context. These systems model each utterance with *Gaussian Mixture Models* (GMMs). System number four models each utterance by a prosodic feature vector. The last system estimates 9 glottal features estimated by a physical two-mass vocal model. These systems can be used standalone or can combined by early or late fusion.
The outline of this paper is as follows: First we describe our five different systems and the early and late fusion in Sec. 2. In Sec. 3 we show the results for the five individual systems and the results achieved by combination. We finish with a conclusion in Sec. 4.

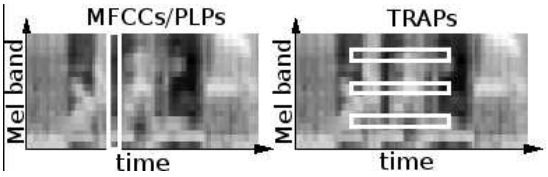

Figure 1: Difference between MFCC/PLP and TRAPS feature calculation

2. System Description

In this section our different subsystems are described. Our five systems can be divided into three different groups: GMM-based, glottal and prosodic. The first group of systems uses supervectors, that contain concatenated mean vectors of Gaussian Mixture Models (GMMs). Three different systems belong to this group: GMM-MFCC, GMM-PLP, and GMM-TRAPS. They are described in Sec. 2.1. The prosodic system is described in Sec. 2.2 and the glottal excitation system is described in Sec. 2.3. Since the glottal

2.1. GMM-based Systems

Three of our subsystems model specific features by *Gaussian Mixture Models* (GMMs). Two systems use short-time spectral features, namely Mel Frequency Cepstrum Coefficients (MFCCs) and Perceptual Linear Prediction (PLPs). The third system is based on Temporal Patterns (TRAPS).

2.1.1. Mel Frequency Cepstrum Coefficents

A Hamming window with a size of 16 ms and a time shift of 10 ms is applied to the speech signal. Afterwards the Mel-spectrum with 25 triangle filters is calculated. We take the first 12 Mel-frequency cepstral coefficients, substitute the first coefficient by the log energy and calculate the first-order derivatives of these features. The derivative covers a context of five frames, two to the left of the current frame and tow to the right of this frame. In the end a 24-dimensional feature vector is created.

2.1.2. Perceptual Linear Prediction

The second set of features is revised PLP (RPLP) [4], a simplified and improved variant of PLP [5] employing the Mel filter-bank instead of the Bark filter-bank. We took the first 13 cepstral coefficients of the PLP model spectrum and their first-order derivatives which results in a feature dimension of $D = 26$. Since we use the same filter-bank for MFCCs and RPLP, the sole difference between these two feature types is that RPLP performs an additional spectral smoothing step by applying linear prediction (LP) on the Mel filter-bank spectrum and obtaining the cepstral coefficients from the resulting model spectrum [4].

2.1.3. Temporal Patterns

Our *TempoRAl PatternS* (TRAPS) in this work are quite similar to the original approach of Hermansky [6]. The main difference of our approach is the different processing within the time trajectories. The time trajectories consider a long temporal context (310 ms) in 18 bands. These bands are generated by a Mel-filter-bank. Each trajectory is smoothed by a Hamming window and transformed into frequency domain. A detailed explanation can be found in [7].
Figure 1 illustrates the difference between the short-time processing of MFCCs or RPLP and the long temporal context used for TRAPS. In a fusion step we concatenate the 31 coefficients of each band together to a high-dimensional feature vector ($D = 558$). This vector is then transformed by a *Linear Discriminant Analysis* (LDA) to a 24-dimensional vector. The LDA is trained on a 578-speaker subset of the German Verbmobil database [8], which was down-sampled to 8 kHz. 46 German phonetic classes serve as labels for this transformation.

2.1.4. GMM Modeling

After extraction of the long- or short-term spectral features an *Universal Background Model* (UBM), i.e., a speaker-independent GMM, is trained on the whole training set. This UBM is adapted by *Maximum A Posteriori* (MAP) adaptation to the recordings of train and development set. MAP adaptation adjusts the UBM to the speaker dependent training data in a single iteration step and combines these new densities with the UBM parameters. Finally, for each recording a single GMM is created. The mean vectors of each Gaussian are extracted and concatenated to a big vector with a dimension of $D * 64$, where D is the dimension of the acoustic features, and 64 is the number of Gaussian densities. These so-called GMM-supervectors are then used for classification with *Support Vector Machines* (SVM). This approach is command in the field of speaker identification and has been applied to age recognition in [2].

2.2. Prosodic System

The prosodic system is not based on any speech recognition output or forced time alignments. Thus, the prosodic features are calculated whenever a voiced speech segment is found. The voiced-unvoiced (VUV) decision is based on the zero crossing rate, the normalized energy of the signal and the maximum energy.
Prosodic base features are calculated on the whole utterance. These are, fundamental frequency (F_0), energy, VUV segments and pitch periods. The structured prosodic features are calculated on the voice segments. Adjacent segments are merged, when they are separated less than 50 ms; the corresponding F_0 contour is interpolated to make the segmentation more robust. Context segments, that merge two adjacent segments together, are used additionally. All in all 73 features are calculated for each segment. They model F_0, energy, duration, pauses, jitter and shimmer. A detailed description of the whole feature set is given in [9]. Finally, we compute mean, minimum and maximum of these 73 segments features. This forms our 219-dimensional prosodic feature vector.

2.3. Glottal Excitation System

A number of voice features that listeners often consider to be characteristic for aged speakers, like increased harshness or hoarseness, increased strain, higher incidence of voice breaks, vocal tremor and increased breathiness[10], may be related to physiological changes of the larynx. This motivates an approach that applies a physical model of the glottis to represent age-related changes of the voice quality. The parameters of the physical model are adapted to each speaker and contain information about the speaker's age. Here we decided to use these parameters directly for age classification.
The glottis model applied to represent the characteristics of the speaker's voice is a physical mass-spring vocal fold model introduced by Stevens in [11]. A detailed description of this model and the algorithm that has been proposed for parameter estimation can be found in [12]. Here we shortly summarize the iterative glottal inversion procedure.
For each voiced 25 ms speech frame an optimization loop is run starting from an initial set of nine glottis model parameters. These parameters determine the physical properties of the model, including the masses, the compliances of the springs, etc. Given the parameter values the glottis model generates an excitation signal. The similarity of the generated excitation signal is compared to the LPC residue of the original speech signal. As a distance measure we use the weighted sum of the Euclidean distance between the log spectra of the two signals and the difference between the generated and the original pitch for the frame. The distance measure is handed over to the simplex parameter optimization

System	UR	WR	C	YF	YM	AF	AM	SF	SM
MFCC-GMM	**42.4**	**42.4**	**56.3**	39.1	**40.4**	**37.9**	33.2	**39.9**	50.0
PLP-GMM	41.2	41.2	53.7	41.1	39.7	35.8	31.5	38.0	48.7
TRAPS-GMM	39.3	39.4	50.2	37.5	39.1	33.7	30.3	37.6	46.9
Prosodic	39.9	40.6	44.2	36.0	32.7	36.9	**37.4**	37.8	**54.2**
Glottis	36.3	37.3	39.1	**43.6**	26.9	29.3	26.1	35.2	53.9

Table 1: Results of the single systems on the development set (in %)

algorithm [13, 14], generating a new set of parameters. Thus, an optimization loop is formed that passes the new parameters again to the excitation model until the distance between synthesized and original speech has been minimized.

2.4. System Fusion

We intentionally use quite low dimension GMM system in order to compare different ways of combination, i.e., early and late fusion. An early fusion would not be possible in an appropriate time with GMM dimensions of up to 2048 like it is common in speaker id. The goal of this paper is to compare early and late fusion of different systems with the different results of the stand-alone systems for age and gender recognition.

2.4.1. Early Fusion

The early fusion is performed on meta-feature level. Therefore the three supervectors of the three GMM systems, the 219-dimensional prosodic vector and the 27-dimensional glottis system are concatenated to a high-dimensional vector. The feature dimension with an early combination of all of our five systems is 3878. This vector is then used for classification with SVMs.

2.4.2. Late Fusion

Late calibration and fusion is based on multi-class logistic regression as it is implemented in the *FoCal* toolkit [15]. In a training step weights are assigned to each of our five systems. The actual combination is a weighted sum of the different system scores. Note that the late fusion was trained on the development set. That means the calculated weights lead to the optimal combination on the development set.

3. Experiments and Results

The data used for system training and evaluating are sufficiently described in [1]. We first show the results of our stand-alone systems in Section 3.1. After that we summarize our results achieved with early and late system fusion in Section 3.2.

3.1. Results of single systems

For these experiments we created five different system, while each is using a *Support Vector Machine* for classification. The results on the five stand-alone systems are summarized in Table 1. The column UR denotes the unweighted recall and WR denotes the weighted recall in %. We also present the recall for each of the seven classes ('C', 'YF', 'YM', 'AF', 'AM', 'SF', 'SM'). The best stand-alone system, w.r.t. UR and WR, is the MFCC-GMM system with 42.4 % in both cases. This system achieves also the best result for the classes 'C', 'YM', 'AF' and 'AM'. 56.3 % of all children, 40.4 % of young men, 37.9 % of all adult females and 39.9 % of senior females are classified correctly. Within the cepstral based systems the MFCC system is the best stand alone system regarding each of the seven classes. The prosodic system has significantly better results regarding the classes AM and SM with rates of 37.4 % and 54.2 % respectively. The glottal excitation system achieves the best results for the class 'YM'.
Focusing on the UR and WR results of the different systems one can see, that the UR and WR results of the GMM-based system are perfectly matched. UR and WR of the prosodic system differs by an absolute value of 0.7 %. The difference within the glottal system is even higher; 1 % absolute. We expect this difference to be based on the different types of features, i.e., GMM-based v.s. non-normalized min-max-mean features. We don't have a rational explanation for this fact yet. This needs a deeper investigation.

3.2. Results of combined systems

In this section we describe the results achieved with the two different ways of system combination, i.e., early and late fusion. The results are summarized in Table 2. The first part of the table contains the results achieved with early fusion, the second part contains the results with late fusion. In both cases three different combination are used: the MFCC-GMM system in combination with the prosodic system (denoted by MFCC+PROS), the combination of MFCC-GMM, PLP-GMM and prosodic system (denoted by MFCC+PLP+PROS) and a combination of all five systems denoted by ALL.

3.2.1. Early Fusion

When comparing the results of the stand-alone MFCC-GMM and the stand-alone prosodic system vs. the early combination of both of them, we achieve a significant improvement of 6 %.
The best UR result of 46.1 % with early fusion is achieved by the MFCC+PLP+PROS system, a combination of the MFCC-GMM, PLP-GMM and prosodic system. This system is also the best one in terms of WR. Note that these results are not significantly different to the results of the early combination of all five systems.

System	UR	WR	C	YF	YM	AF	AM	SF	SM
Early Fusion									
MFCC+PROS	45.0	45.0	57.2	44.5	42.7	40.2	36.6	42.8	51.1
MFCC+PLP+PROS	46.1	46.0	**59.2**	45.6	46.5	39.0	37.2	42.6	52.7
ALL	45.9	45.6	57.4	46.1	**47.7**	40.1	36.7	42.5	50.6
Late Fusion									
MFCC+PROS	45.0	45.6	50.7	42.8	38.8	44.3	38.9	46.5	53.4
MFCC+PLP+PROS	47.0	47.7	53.4	45.3	40.4	**45.2**	39.1	47.4	58.2
ALL	**47.8**	**48.9**	51.6	**47.9**	38.2	44.4	**39.4**	**51.6**	**61.6**

Table 2: Results of different combined systems on the development set (in %)

3.2.2. Late Fusion

The late fusion is optimized on the development set, meaning that the values shown in the table are optimal for the development set. We achieve a maximum UR of 47.8 % with the ALL system combination. This result is not significant compared to our second best late fusion system, the combination of MFCC-GMM, PLP-GMM and prosodic system. WR is also highest for the ALL combination system: 48.9 %.

3.2.3. Comparison of Early and Late Fusion results

When focusing on the class-wise results it can be seen, that the early combined systems outperform the systems combined at score level for the class 'C'. The late fusion systems on the other hand have a clear advantage on elderly speakers (SF and SM). Focusing on the UR and WR results of the different ways of comparison shows, that the early combination is more balance, i.e., there is almost no difference between UR and WR. Late fusion takes the a posterior probability of the development set into account, which is not the case for the early combination.

4. Conclusions

In this paper we compared five system and their combination for the task of age and gender recognition. Three systems modeled spectral features of different temporal context, one system used prosodic features, and one used glottal features estimated by a two-mass model. The best stand-alone system was the GMM-UBM with an unweighted recognition rate of 42.4 %. In case of early system combination we achieved 46.1 %. Combination on score level achieved an additional improvement of 4 %. The best results was a UR of 47.8 %, achieved by score level combination of all five systems.

5. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller, and C. Narayanan, "The Interspeech 2010 Paralinguistic Challenge," in *Proc. Interspeech (2010)*, 2010, p. no pagination.

[2] T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, and E. Nöth, "Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines," in *Proc. ICASSP 2008*, vol. 1, 2008, pp. 1605–1608.

[3] G. Dobry, R. Hecht, M. Avigal, and Y. Zigel, "Dimension Reduction Approaches for SVM-Based Speaker Age Estimation," in *Proc. Interspeech 2009*, 2009, pp. 2031–2034.

[4] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising Perceptual Linear Prediction (PLP)," in *Procof the 9th European Conference on Speech Communication and Technology*, ISCA, Ed., Bonn, 2005, pp. 2997–3000.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[6] H. Hermansky and S. Sharma, "TRAPS – classifiers of temporal patterns," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998.

[7] C. Hacker, *Automatic Assessment of Children Speech to Support Language Learning*. Berlin: Logos Verlag, 2005.

[8] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*. New York, Berlin: Springer, 2000.

[9] A. Maier, F. Hönig, V. Zeissler, A. Batliner, E. Körner, N. Yamanaka, P. D. Ackermann, and E. Nöth, "A language-independent feature set for the automatic evaluation of prosody," in *Proc. Interspeech 2009*, Brighton, England, 2009, pp. 600–603.

[10] S. E. Linville, "The Sound of Senescence," *The Journal of Voice*, vol. 10, no. 2, pp. 190–200, 1996.

[11] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA 02141: The MIT Press, 1998.

[12] P. Beyerlein, A. Cassidy, V. Kholhatkar, E. Lasarcyk, E. Nöth, B. Potard, S. Shum, Y. C. Song, W. Spiegl, G. Stemmer, and P. Xu, "Vocal aging explained by vocal tract modelling: 2008 JHU summer workshop final report," Tech. Rep., 2008.

[13] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[14] D. Olsson and L. Nelson, "The Nelder-Mead simplex procedure for function minimization," *Technometrics*, vol. 17, no. 1, pp. 45–51, 1975.

[15] N. Brümmer, *FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores*. available online: http://sites.google.com/site/nikobrummer/focalmulticlass, 2007.