Combining Five Acoustic Level Modeling Methods for Automatic Speaker Age and Gender Recognition

Ming Li[1], *Chi-Sang Jung*[2], *Kyu J. Han*[1]

[1]Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,
University of Southern California, Los Angeles, USA
[2]School of Electrical and Electronic Engineering,
Yonsei University, Korea

mingli@usc.edu, jtoctos@dsp.yonsei.ac.kr, kyuhan@usc.edu

Abstract

This paper presents a novel automatic speaker age and gender identification approach which combines five different methods at the acoustic level to improve the baseline performance. The five subsystems are (1) Gaussian mixture model (GMM) system based on mel-frequency cepstral coefficient (MFCC) features, (2) Support vector machine (SVM) based on GMM mean supervectors, (3) SVM based on GMM maximum likelihood linear regression (MLLR) matrix supervectors, (4) SVM based on GMM 'Tandem' supervectors, and (5) SVM baseline system based on the 450-dimensional feature vectors including prosodic features at the utterance level provided by the challenge organizing committee. To improve the overall classification performance, fusion of these five subsystems at the score level is performed. The proposed fusion system achieves 52.7% unweighted accuracy for the joint age-gender classification task and outperforms the GMM-MFCC system and SVM baseline, respectively, by 9.6% and 8.2% absolute improvement on the 2010 Interspeech Paralinguistic Challenge aGender database.

Index Terms: Age, Gender, Gaussian mixture models, Support vector machine, Maximum likelihood linear regression, Tandem, Supervector, Score level fusion

1. Introduction

Automatic recognition of paralinguistic information, such as speaker identity, gender, age range, emotional state, etc., can guide human computer interaction systems to automatically adapt to different user needs. Identifying the age and gender information of a speaker given a short speech utterance is a challenging task and has gained significant attention recently.

Four approaches for age and gender recognition from telephone speech have been compared [1]; namely, a parallel phoneme recognizer system to compare the Viterbi decoding scores for each category specific phoneme recognizer, a system using dynamic Bayesian networks to combine several prosodic features, a system based solely on linear prediction analysis, and a GMM system based on MFCCs. It was reported in [1] that the parallel phone recognizer system performs as well as human listeners but loses performance on short utterances while the system based on prosodic features, such as F0, jitter, shimmer and harmonics-to-noise-ratio, has shown relative robustness to the variation of the utterance duration. More recently, novel acoustic level features [2, 3] and lexical level features [4] have been proposed to improve the recognition performance. Furthermore, techniques from speaker verification applications, such as GMM-SVM mean supervector systems [5], nuisance attribute projection (NAP), and anchor models [6], have been successfully applied to speaker age and gender identification tasks to enhance the performance of acoustic level modeling. Due to the different aspects of modeling, combining different classification methods together can significantly improve the overall performance [7, 8].

In this paper, we focus on acoustic-level approaches for speaker age and gender identification. The GMM-SVM mean supervector method is extended by two kinds of supervectors, namely maximum likelihood linear regression (MLLR) matrix supervector [9] and Tandem posterior probability (TPP) supervector [10, 11]. Generally, in the GMM-SVM mean supervector method, maximum a posteriori (MAP) adaptation is used to adapt the means of a GMM Universal Background Model (UBM), and the corresponding feature vectors are the Gaussian supervectors (GSVs) which consist of the stacked adapted means. The idea of MLLR is to estimate an affine transformation to adapt the means of a speaker independent large vocabulary speech recognition (LVCSR) system to a given speaker. Thus the MLLR matrix itself contains speaker specific characteristics and the entries of this affine transformation matrix can be used as feature supervectors for speaker modeling [9] and age-gender recognition. MLLR can also be used to adapt the means of a GMM UBM to generate feature supervectors which is more efficient [12]. For the TPP method, a UBM is trained as an age and gender independent model; thus each component of the UBM can be considered to be modeling some underlying phonetic sounds [11]. It is also shown in [11] that the utterances from different speakers should get different average Tandem posterior probability (TPP) on the same gaussian component. This inspired us to explore the potential to consider the TPP supervector as a histogram describing the characteristics of different age and gender groups. Thus, we concatenate these average Tandem posterior probabilities from all the components of the UBM into a TPP supervector for SVM modeling.

In this work, we use MAP adaptation, MLLR adaptation, and TPP feature extraction to map each age and gender specific input utterance into three different supervectors for SVM classification. All three methods share the same framework of using a GMM UBM as a front end and thus combining these approaches is efficient in terms of computational cost. As shown in Figure 1, score level fusion of the three methods together with the GMM baseline method is performed to utilize the com-

This work was supported by National Science Foundation and United States Army.

plementary information of each method. Moreover, combining these four MFCC based methods with the SVM baseline system using prosodic features can further enhance the performance.

The remainder of the paper is organized as follows. Description of the corpus and classification task is provided in Section 2. In Section 3, each subsystem as well as the score level fusion method is explained. Section 4 presents experimental results and Section 5 provides conclusions.

2. Corpus and Classification Task

The database used to evaluate the proposed approach is aGender database [13]. The task is to classify a speaker's age and gender class which is defined as follows: children $<$ 13 years (C), young people $14-19$ years (YF/YM), adults $20-54$ years (AF/AM), and seniors $>$ 55 years (SF/SM). The mean and standard deviation of speech duration after voice activity detection (VAD) in the training and development data sets of the aGender database are 1.13 ± 0.86 seconds and 1.14 ± 0.87 seconds, respectively. Thus it is indeed a short length speech database. The training data set of the aGender database (471 speakers) was divided into 2 parts: data from the last 20 speakers in the alphabet order of each age and gender class was used for MAP and MLLR adaptation (140 speakers) and the rest of the data (331 speakers) was used for UBM training. In this paper, these 2 partitions are denoted as the training set and the UBM set, respectively. The development data set from the aGender database (20548 utterances) is adopted as the evaluation set in this paper. Finally, the testing data set from the aGender database (17332 utterances) is evaluated and the results are reported in the end of Section 4. The details about the aGender database and the evaluation methods are provided in [13].

3. Methods

The overview of the proposed approach is demonstrated in Figure 1. In this section, we present the details of each subsystem as well as the score level fusion method.

3.1. GMM baseline system

The features in our system are 13 dimensional MFCCs (including C0) and their first and second order derivatives, which are in total 39 dimensional coefficients per frame. After voice activity detection, non-speech frames were eliminated and the 39 dimensional MFCC features were extracted. Cepstral mean subtraction and variance normalization were performed to normalize the MFCC features to zero mean and unit variance on a per-utterance basis. In the proposed work, since the training data for each age and gender class is too limited to train a good GMM, a UBM in conjunction with a MAP model adaptation approach [14] was used to model different age and gender classes in a supervised manner. All the data in the UBM set was adopted to train a 256-component UBM, and MAP adaptation was performed using the training set data for each age and gender class. A relevance factor of 12 was used for the MAP adaptation.

3.2. GMM-SVM mean supervector system

The feature extraction and UBM training were done in the GMM baseline system. Means of Gaussian components were adapted by MAP adaptation for each UBM set, training set, and evaluation set utterance. Then the corresponding GMM supervectors, created by concatenating the mean vectors of all the Gaussian components, were modeled by SVM. The supervector was normalized by the corresponding standard deviation and weight to fit the supervector kernel [15]. We arbitrarily added one dummy dimension with value 1 at the head of each mean supervector so that all the support vectors can be collapsed down into a single model vector and each target score can be calculated by a simple inner product which makes this framework computationally efficient [15]. The dimension of the GMM mean supervector is $1 + (39 \times 256) = 9985$. In addition, there are more than 30000 utterances in the training and UBM sets which makes the SVM training data set too large to be handled efficiently. Instead of directly training a multi-class SVM classifier using all the high dimensional supervectors, we adopted a two stage framework [16] which can solve the practical limitation of computer memory requested by large database training. First, based on the supervector samples from the UBM set, multiple binary age and gender group based discriminative classifiers (in our case, 21 1vs1 classifiers plus 7 1vsRest classifiers) were trained by our modified version of SVMTorch [17] and employed to map the supervectors into discriminative aGender characterization score vectors (DACSV) [16]. Since the scoring function for each binary model is just an inner product, the mapping from supervectors to DACSV vectors is computationally efficient. Furthermore, a back end SVM classifier was trained using LIBSVM [18] to model the probability distribution of each target age and gender class in the DACSV space using training set supervector samples.

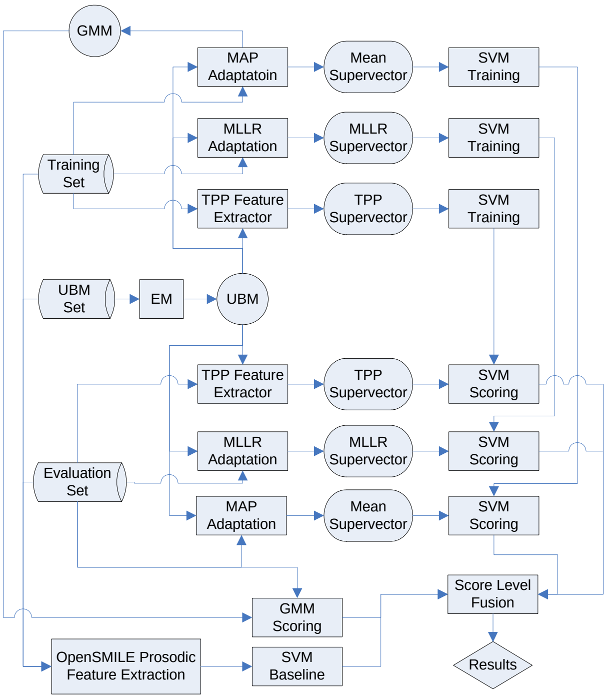

Figure 1: *System overview.*

3.3. GMM-SVM MLLR supervector system

For each utterance in the training set and the evaluation set, MLLR adaptation on the UBM was performed [9, 12]. The corresponding MLLR matrix supervector was used for SVM modeling. Since the dimension of the MLLR matrix supervector is $39 \times 40 = 1560$ which is considerably smaller than the dimension of GMM mean supervector, we used all the supervectors from the training set to train a multi-class SVM classifier and performed scoring on the evaluation set. Linear Discriminant Analysis (LDA) was employed to perform dimension reduction on the MLLR supervector space, and the linear kernel multi-class SVM classifier was trained by LIBSVM [18].

3.4. GMM-SVM TPP supervector system

For each utterance in the training and evaluation sets, TPP feature extraction is performed on the UBM. Given a frame-based MFCC feature x_t and the GMM-UBM λ with M Gaussian components (each component is defined as λ_i),

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, \cdots, M, \tag{1}$$

the posterior probability is calculated as follows:

$$P(\lambda_i|x_t) = \frac{w_i p_i(x_t|\mu_i, \Sigma_i)}{\Sigma_{j=1}^{M} w_j p_j(x_t|\mu_j, \Sigma_j)}. \tag{2}$$

This posterior probability can also be considered as the normalized likelihood ratio. The larger the posterior probability, the better this Gaussian component can be used to represent this feature vector. Thus the TPP supervector is defined as follows:

$$TPP_{supervector} = [b_1, b_2, \cdots, b_M] \tag{3}$$

$$b_i = \frac{1}{T}\Sigma_{t=1}^{T} P(\lambda_i|x_t). \tag{4}$$

Since the TPP supervector is a probability distribution over all the Gaussian components, it is appropriate to use KL-divergence when measuring the similarity between vectors. Figure 2 shows that the symmetric KL Divergence between TPP templates and supervectors which are from the same class is statistically lower than the one between mismatched supervectors. Thus, based on the discriminative information from KL divergence, TPP supervectors do contain age and gender specific information. However, because a matrix of kernel distances directly based on symmetric KL divergence does not satisfy the Mercer conditions, a linear kernel multi-class SVM classifier was trained by LIBSVM [18] using TPP supervector samples from the training set.

3.5. SVM baseline system

The SVM baseline system [13] is provided by the 2010 Paralinguistic Challenge, which is based on 450 dimensional acoustic features per utterance. The details of feature extraction and SVM modeling are presented in [13]. Since different kinds of prosodic features, such as F0, F0 envelop, jitter, and shimmer, are also included, this system can capture age and gender information from a prosodic level. Combining this system with our MFCC feature based systems can further improve the performance.

3.6. Score level fusion

Let there be $\mathbb{K}$ input subsystems (as shown in Figure 1, $\mathbb{K} = 5$ in this work) where the k^{th} subsystem outputs its own posterior probability vector $l_k(\boldsymbol{x}_t)$ for every trial (log-likelihood ratio (LLR) is used for the GMM baseline subsystem). Then the fused score vector $\acute{l}(\boldsymbol{x}_t)$ is given by:

$$\acute{l}(\boldsymbol{x}_t) = \sum_{k=1}^{\mathbb{K}} \beta_k l_k(\boldsymbol{x}_t) \tag{5}$$

The weight, β_k, can be determined either manually or automatically by using the inverse entropy as the weight [19].

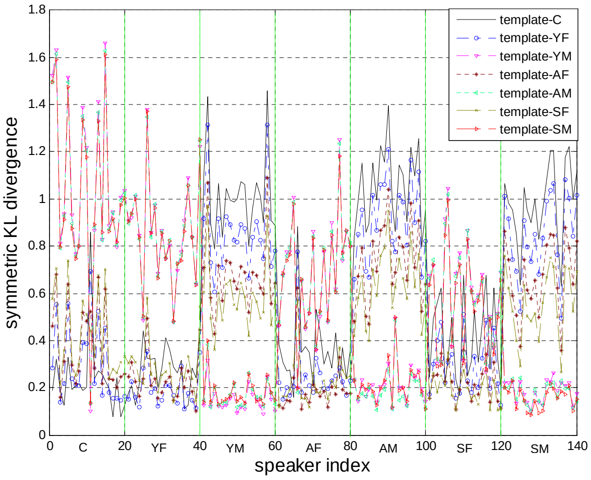

Figure 2: *Symmetric KL divergence between TPP templates and TPP supervectors in the training set.*

4. Experimental Results

The evaluation set is used to evaluate the performance of each subsystem as well as the fusion approach. In Table 1, both unweighted and weighted accuracy on average per class (UA/WA, weighting with respect to number of instances per class) for each of the 3 different classification tasks (7 class age & gender {C,YF,YM,AF,AM,SF,SM}, 4 class age {C,Y,A,S} and 3 class gender {C,F,M}) are presented. The details of these 3 tasks as well as the evaluation method are provided in [13].

Table 1: Performance of each method and the fusion approach (MFuse and AFuse denote manually and automatically tuned weight in score level fusion, respectively.)

	age & gender		age		gender	
ID & System	UA	WA	UA	WA	UA	WA
1.GMM base	43.1	42.4	47.0	45.0	76.0	81.7
2.mean supervector	42.6	43.1	46.0	45.5	75.6	82.8
3.MLLR supervector	36.2	36.1	40.5	41.1	68.4	75.8
4.TPP supervector	37.8	38.0	41.6	41.5	71.2	79.6
5.SVM base	44.6	45.0	47.4	46.7	77.6	85.2
MFuse 1+2	45.2	45.3	47.4	48.7	77.6	83.7
MFuse 3+4	40.3	40.3	44.1	43.5	73.4	81.0
MFuse 1+2+3+4	50.4	50.5	53.7	51.8	83.7	89.0
MFuse 1+2+3+4+5	**52.7**	**53.0**	**55.5**	**54.0**	**84.7**	**90.3**
AFuse 1+2+3+4+5	**51.2**	**51.4**	**54.6**	**52.5**	**84.7**	**89.7**

It is shown in Table 1 and Figure 3 that GMM baseline, SVM baseline, and GMM-SVM mean supervector systems outperform the other 2 systems (GMM-SVM MLLR supervector system and GMM-SVM TPP supervector system) in all 3 tasks. However, by combining different methods together, significant improvements in both UA and WA are achieved for all the 3 classification tasks. The automatically tuned weight based score level fusion (AFuse) approach also increased the classification accuracy dramatically compared to each individual subsystem. Thus combining these 5 methods together for speaker age and gender classification is useful.

In Figure 3, we can see that the improvement is bigger for short utterances which might be due to the dominant role of

Table 2: Confusion matrix for 7 class age and gender task

	C	YF	YM	AF	AM	SF	SM
C	**67.0**	13.9	5.8	5.2	1.1	6.4	0.7
YF	16.4	**59.0**	0.5	16.0	0.1	7.7	0.2
YM	0.2	0.4	**57.1**	1.8	17.6	3.3	19.7
AF	4.0	25.0	0.9	**37.9**	0.2	31.8	0.1
AM	0.2	0.0	34.5	1.5	**25.4**	1.3	37.1
SF	4.7	8.9	0.9	27.9	0.3	**56.4**	0.9
SM	0.1	0.0	15.3	0.5	15.6	2.6	**65.9**

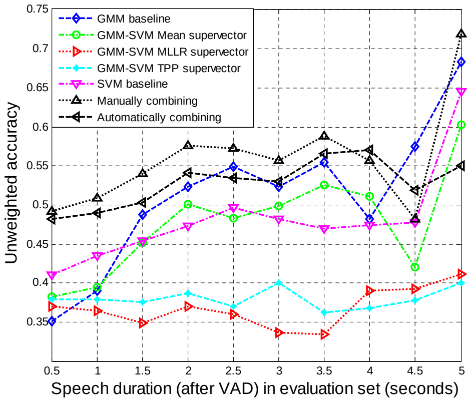

Figure 3: *Accuracy for different valid speech durations in 7 class age and gender task. (The mean and standard deviation of valid speech duration in evaluation set are* 1.14 *and* 0.87 *seconds.)*

short duration speech in this data set. Moreover, according to the increase of utterance duration, the performance of MLLR and TPP supervector based systems did not improve dramatically. It might be because the training utterances (1.13 seconds average) were too short to map a good quality supervector. Validating these 2 methods on a longer duration database is important for further work. The small decrease at 4.5 seconds duration (Figure 3) is consistent with the results in [8] which might be due to the sparse data in that duration interval.

The confusion matrix for the 7 class age and gender task is shown in Table 2. We can see that children, youth and senior groups perform better than adult group and the main confusion comes from speakers with the same gender of other age groups. This result is consistent with the big gap between age classification accuracy and gender classification accuracy in Table 1.

By using the official testing data set from the aGender database as the evaluation data, our AFuse system achieved 50.12% UA (47.52% WA) and 82.38% UA (86.27% WA) on the age recognition and gender recognition tasks, respectively.

5. Conclusions

In this work, we proposed a novel automatic speaker age and gender identification approach which combines five different acoustic level modeling methods. GMM baseline, GMM-SVM mean supervector, GMM-SVM MLLR supervector, GMM-SVM TPP supervector, and SVM baseline subsystems are complementary with each other and fusing these five methods together on the score level improves the classification accuracy significantly. Future work includes investigating the GMM-SVM Constrained MLLR supervector method, combining other prosodic or phonetic level methods, and validating the results with a relatively larger and longer duration database.

6. References

[1] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. Bauer, and B. Littel, “Comparison of four approaches to age and gender recognition for telephone applications,” in *Proc. ICASSP*, vol. 4, 2007, pp. 1089–1092.

[2] J. Ajmera and F. Burkhardt, “age and gender classification using modulation cepstrum,” in *Proc. Speaker Odyssey*, 2008.

[3] W. Spiegl, G. Stemmer, E. Lasarcyk, V. Kolhatkar, A. Cassidy, B. Potard, S. Shutn, Y. Song, P. Xu, P. Beyerlein, J. Harnsberger, and E. Nöth, “Analyzing features for automatic age estimation on cross-sectional data,” in *Proc. INTERSPEECH*, 2009.

[4] M. Wolters, R. Vipperla, and S. Renals, “Age recognition for spoken dialogue systems: Do we need it?” in *Proc. INTERSPEECH*, 2009, pp. 1435–1438.

[5] 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*, vol. 1, 2008, pp. 1605–1608.

[6] G. Dobry, R. Hecht, M. Avigal, and Y. Zigel, “Dimension reduction approaches for SVM based speaker age estimation,” in *Proc. INTERSPEECH*, 2009, pp. 2031–2034.

[7] C. Müller and F. Burkhardt, “Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age,” in *Proc. INTERSPEECH*, 2007, pp. 2277–2280.

[8] C. van Heerden, E. Barnard, M. Davel, C. van der Walt, E. van Dyk, M. Feld, and C. Müller, “Combining regression and classification methods for improving automatic speaker age recognition,” in *Proc. ICASSP*, 2010, pp. 5174–5177.

[9] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, “MLLR transforms as features in speaker recognition,” in *Proc. INTERSPEECH*, 2005, pp. 2425–2428.

[10] H. Hermansky, D. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMMsystems,” in *Proc. ICASSP*, vol. 3, 2000, pp. 1635–1638.

[11] X. Zhang, H. Suo, Q. Zhao, and Y. Yan, “Using a kind of novel phonotactic information for SVM based speaker recognition,” *IEICE TRANSACTIONS on Information and Systems*, vol. 92, no. 4, pp. 746–749, 2009.

[12] A. Stolcke, S. Kajarekar, L. Ferrer, E. Shrinberg, S. Int, and M. Park, “Speaker recognition with session variability normalization based on MLLR adaptation transforms,” *IEEE Trans Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1987–1998, 2007.

[13] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller, and S. Narayanan, “The INTERSPEECH 2010 paralinguistic challenge,” in *Proc. INTERSPEECH*, 2010.

[14] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[15] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *Proc. ICASSP*, vol. 1, 2006.

[16] M. Li, H. Suo, X. Wu, P. Lu, and Y. Yan, “Spoken language identification using score vector modeling and support vector machine,” in *Proc. INTERSPEECH*, 2007, pp. 350–353.

[17] R. Collobert and S. Bengio, “SVMTorch: Support vector machines for large-scale regression problems,” *The Journal of Machine Learning Research*, vol. 1, p. 160, 2001.

[18] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[19] H. Misra, H. Bourlard, and V. Tyagi, “New entropy based combination rules in HMM/ANN multi-stream ASR,” in *Proc. ICASSP*, vol. 3, 2003, pp. 1–5.