# Investigations of features and estimators for speech-based age estimation

Toshiya Wada* , Takahiro Shinozaki† , Sadaoki Furui‡
* Tokyo Institute of Technology, Japan. wada@ks.cs.titech.ac.jp
† Tokyo Institute of Technology, Japan. shinot@furui.cs.titech.ac.jp
‡ Tokyo Institute of Technology, Japan. furui@cs.titech.ac.jp

*Abstract*—**Age estimation approaches using a discrete support vector machine (SVM) and continuous support vector regression (SVR) are systematically compared. Along with the two types of estimators, several speech-based features including a maximum a posteriori (MAP) adapted Gaussian mixture model (GMM) supervector and our proposed maximum likelihood linear regression (MLLR) transform vector are investigated. Experiments are performed using speech data from the Corpus of Spontaneous Japanese (CSJ). Experimental results show that the SVR-based estimator using MAP adapted GMM supervector features give the highest estimation performance.**

## I. Introduction

Speech-based age estimation is useful for various applications. For example, it could be used for determining the target of an advertisement for marketing purposes. For an automatic dialogue system, it could be used to improve dialogue strategy based on an estimated age [1]. Several approaches have been proposed for speech-based age estimation in terms of features and estimators. These include combinations of mel-frequency cepstral coefficient (MFCC) features and Gaussian mixture model (GMM)-based classifiers [2], [3], MFCC features and hidden Markov model (HMM)-based classifiers [4], and fixed length features consisting of the median of a MFCC sequence and support vector machine (SVM)-based classifiers [2]. These age estimation frameworks are similar to those used in speaker recognition.

While these GMM and SVM-based systems predict a discrete age class, continuous regression-based systems are more popular in the image processing area since they can naturally model the continuous nature of age. Recently, support vector regression (SVR) [5]-based age estimation systems using speech features have been proposed [6]. However, there has been no systematic comparison on these continuous and discrete modeling approaches using speech features.

In terms of features, maximum likelihood linear regression (MLLR) transform [7]-based features have been proposed in the speaker recognition area [8]. Since an MLLR transform is estimated so as to adapt a speaker-independent model to a speaker-dependent model, the transform can be regarded as a compact representation of speaker characteristics. When it is used as features for speaker recognition, it has the advantage that it can extract speaker characteristics more directly than short-term features such as the MFCC. While it is expected that the MLLR transform-based features can also be used for age estimation, there has been no such study.

In this study, we applied the MLLR transform-based features for age estimation and compared them with maximum a posteriori (MAP) [9]-based GMM supervector features [10] that were also based on speaker adaptation. As the age estimator, both discrete SVM and continuous SVR-based systems were systematically evaluated using these features with the same evaluation criteria.

This paper is organized as follows. In Section II, features for age estimators used in this study are explained. Age estimators using discrete and continuous estimators are explained in Section III. Evaluation criteria are explained in Section IV, and experimental setups are described in Section V. Experimental results are shown in Section VI, and a summary and future work are given in Section VII.

## II. Features for age estimation

Three types of features are investigated for SVM and SVR-based age estimation. The first type of feature is the MAP-GMM supervector that has been used in [10]. The second and third types of features are our proposed MLLR transform and MLLR-GMM supervector features, respectively. These MAP and MLLR adaptation-based features are described below. In this study, as in studies [6], [10], [11], age is estimated using multiple utterances.

### *A. MAP-GMM supervector features*

MAP-GMM supervector features are based on MAP adaptation of a GMM. Given utterances from a speaker, a speaker-independent initial GMM is first adapted by MAP adaptation to that speaker. Then, a supervector is formed by concatenating mean vectors of Gaussian components of the adapted GMM, and the supervector components are used as fixed-size features for age estimation.

### *B. MLLR transform features*

MLLR transform features are based on MLLR speaker adaptation using a GMM or HMM initial model. Equation 1 shows a model-space MLLR transform for a mean vector $\mu$ of a Gaussian component, which is an affine transform specified by a matrix $A$ and a vector $b$.

$$\hat{\mu} = A\mu + b \tag{1}$$

The transform is estimated so as to maximize the likelihood of the transformed model for a set of adaptation utterances.

*Proceedings of the Second APSIPA Annual Summit and Conference*, pages 470–473,
Biopolis, Singapore, 14-17 December 2010.

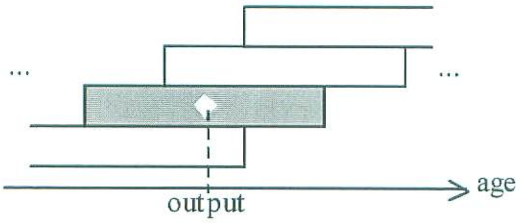

Fig. 1. Age estimation based on overlapping age window

To estimate an age using a small number of utterances in this study, a single global transform is used to convert all Gaussian components. Moreover, to further reduce the number of free parameters in the transform estimation for robust parameter estimation, $A$ is restricted to be a block-diagonal matrix.

When phone HMM is used as an initial model, phone labels are required to estimate an MLLR transform. In this study, a manual transcript is used for this purpose. No transcript is required when a GMM initial model is used. In the following of this paper, MLLR transform features based on the GMM initial model are referred to as MLLR-GMM transform, and the HMM initial model-based transform is referred to as MLLR-HMM transform.

### *C. MLLR-GMM supervector features*

Similar to the MAP-GMM supervector features, MLLR-GMM supervector features are formed by re-arranging the mean vectors of an MLLR adapted GMM. Differences from the MLLR-GMM transform features are that transformed mean vectors are used rather than the MLLR transform itself. The dimension of the MLLR-GMM supervector is the same as the MAP-GMM supervector when the same initial GMM is used.

## III. AGE ESTIMATORS

The SVM-based discrete age estimator and the SVR-based continuous estimator are investigated. For comparison purposes, the MFCC-based GMM classifier is also evaluated.

### *A. SVM-based age estimator*

Age estimation using SVM is based on predicting an age class. For precise age estimation, the number of age classes needs to be large. On the other hand, enough training samples need to be assigned for each age class for better SVM estimation. To accommodate these requirements, an overlapping age class approach is used as shown in Figure 1. To use the same evaluation criteria as continuous age estimators, the center value of the age for each class is output rather than the range of the class. To handle more than two age classes, multi-class SVM [12] is used.

### *B. SVR-based age estimator*

SVR is an extension of SVM for regression problems [5]. The cost function for training SVR gives zero error if the absolute difference between the prediction and the target is less than $\epsilon$ where $\epsilon > 0$, which makes SVR depend only on a subset of the training data. For age estimation, SVR is estimated so as to directly predict the speaker's age rather than a discrete age class.

### *C. GMM-based age estimator*

The GMM-based age estimation system uses the same overlapping age classes as the SVM-based system. The training sample is a sequence of MFCC vectors of multiple utterances. The GMM is trained for each age class. For age estimation, likelihood values of all the age-class GMMs are calculated for a test sample consisting of multiple utterances, and the age class that gives the highest likelihood is identified. As for the SVM-based system, a center value of the age for each class is output as a result.

## IV. EVALUATION CRITERIA

Two evaluation criteria are used for all the types of age estimators. One is the mean absolute error (MAE), and the other is the cumulative match score (CS). The MAE indicates a mean estimation error across all speakers and is defined as Equation (2):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |a_i - \hat{a}_i| , \qquad (2)$$

where $N$ is the size of the dataset, $a_i$ is the true age of the $i$-th test speaker, and $\hat{a}_i$ is the estimated age.

The CS plots the relationship between an acceptable level of estimation error and estimation accuracy, and is defined as Equation (3):

$$CS(j) = \frac{N_{e\leq j}}{N} \times 100, \qquad (3)$$

where $N_{e\leq j}$ is the number of test speakers whose absolute estimation error $e$ is within $j$ years.

## V. EXPERIMENTAL SETUPS

Training and evaluation of the age estimation systems were performed by 5-fold cross-validation using 300 speakers (197 males, 103 females) from the Corpus of Spontaneous Japanese [13]. Ten utterances were aggregated from each speaker to form a single sample for feature extraction. Each utterance was about 3 seconds long. Cross-validation was performed so that the evaluation was done for many test speakers avoiding overlap between training and test speakers. Each cross-validation model was trained using 2400 samples from 240 speakers and it was tested for held-out data having 600 samples from 60 speakers. The evaluation was performed for five cross-validation folds and the final result was obtained by testing a total of 300 ($= 60 \times 5$) speakers.

Speech waveform was digitized with 16-kHz sampling and 16-bit quantization. All GMM and HMM were based on MFCC features with 39 elements comprising of 12 mel-frequency cepstral coefficients, log energy, their deltas, and their delta deltas. GMM and HMM used as initial models for feature extraction were trained by speaker adaptive training based on constrained MLLR [14]. The HMM had 3000 states and 128 Gaussian components per state. All GMM including the one used as age estimators had 128 Gaussian components. For MLLR-based features, a block-diagonal transform matrix was used, and a block size of 13 was chosen based on preliminary experiments. For SVM and GMM-based age estimators, age classes were equally arranged between 15 to 74 years old. The window size of each age class was 15 years, and the shift was 5 years. Therefore, there were a total of 10 age classes. Based on a previous observation that gender-independent age estimation give close performance as gender-dependent estimation [11], all the estimators were gender independent. Both SVM and SVR used a linear kernel.

## VI. Results

Figure 2 shows age estimation performance measured by the MAE using various combinations of the estimators and the features. When the GMM classifier with MFCC features was compared to the SVM-based systems with various features, the SVM-based systems gave better performance. When an SVM-based system and an SVR-based system using the same features were compared, the SVR-based system gave better performance than the SVM-based system irrespective of the features. This was probably because SVR-based systems could model ages continuously. Even though overlapping age class strategy was used, it is unavoidable for SVM-based systems to have some errors in the discretely estimated results [1].

MLLR-GMM and MLLR-HMM transform features gave similar results when the SVM-based classifier was used. When SVR was used, MLLR-HMM transform features gave slightly better performance than MLLR-GMM transform features. However, considering that the MLLR-HMM transform was estimated using a phone transcript, this improvement was minor. MLLR-GMM supervector features gave slightly better performance than MLLR-GMM transform features for SVR. The lowest MAE of 7.3 years was obtained by the combination of the MAP-GMM supervector features and the SVR estimator[2]. It was even better than that obtained by the MLLR-HMM transform features estimated using a phone transcript. The differences of the MAE values with this system and the other SVR-based systems were statistically significant at 1% significance level by the t-test.

Figures 3 and 4 show the cumulative match scores for SVM and SVR-based systems, respectively. When SVM was used, it can be seen that the MLLR-GMM transform features generally gave the best performance and about 80% of speakers' ages were correctly estimated within an error margin of 15 years. When SVR was used, MAP-GMM supervector features gave the best results and about 90% of the speakers' ages were correctly estimated within an error margin of 15 years.

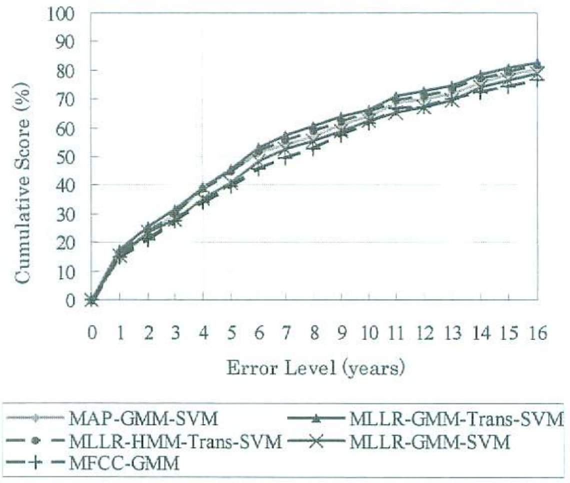

Fig. 3. Cumulative scores of SVM and GMM-based estimators using different features.

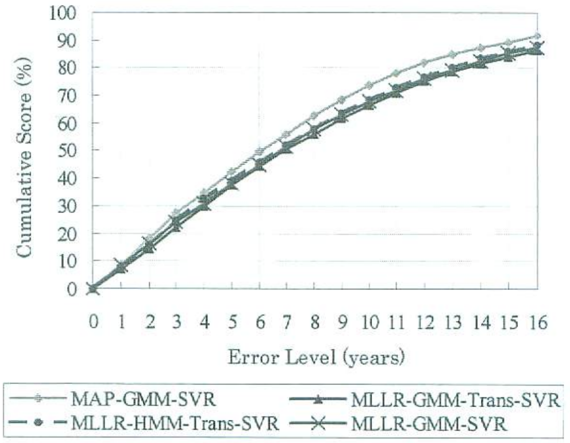

Fig. 4. Cumulative scores of SVR-based estimators using different features.

In speech recognition, it is said that MLLR adaptation is more advantageous than MAP adaptation when a smaller amount of adaptation data is available. To see whether this applies to age estimation, supplemental age estimation experiments were performed using a smaller number of test utterances per sample. Figure 5 shows the MAEs for SVR-based systems using 10, 8, 6, 4 and 2 utterances. It was found that MAP-GMM supervector features consistently gave better performance than both MLLR-GMM supervector and MLLR-GMM transform features for SVR-based age estimation using smaller utterances.

[1] We have run an additional experiment and have confirmed that the overlapping age range strategy gave small improvement over a non-overlapping strategy.

[2] Not in the Figure but when two MLLR transforms were used with a regression tree to extract MLLR-GMM supervector features as an supplemental experiment, the MAE was 7.6 which was lower than when a single transform was used. Although, it was still higher than the MAP-GMM supervector features.

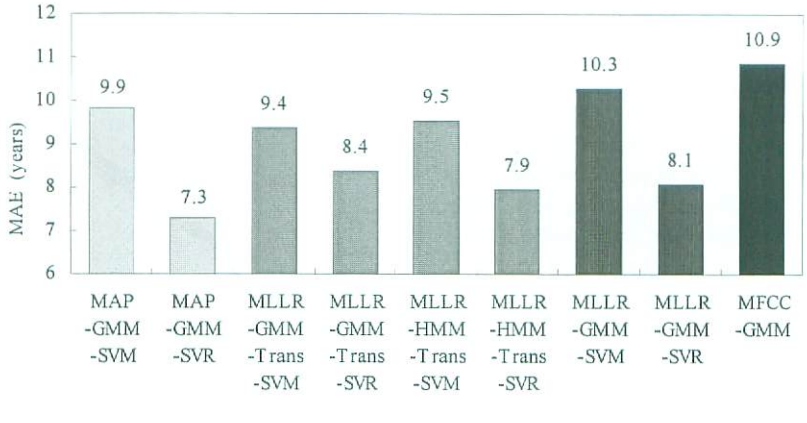

Fig. 2. Comparison of MAE using various combinations of features and estimators. MAP-GMM: MAP adaptation-based GMM supervector features, MLLR-GMM: MLLR adaptation-based GMM supervector features, MLLR-GMM-Trans: MLLR transform-based features using GMM initial model, and MLLR-HMM-Trans: MLLR transform-based features using HMM initial model. SVM: SVM-based estimator, and SVR: SVR-based estimator. MFCC-GMM: GMM-based estimator with MFCC features.

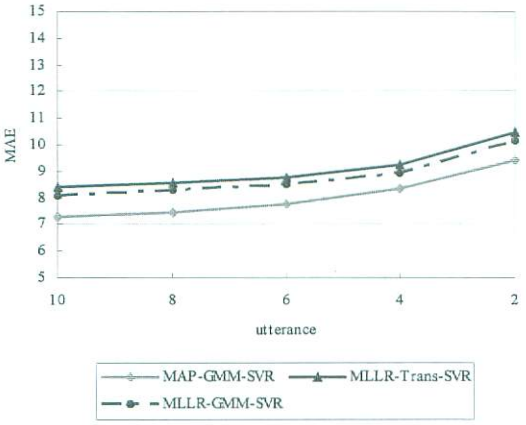

Fig. 5. Number of utterances and MAE using SVR based estimators with different features.

## VII. Conclusion

Age estimation approaches using a discrete support vector machine (SVM) and continuous support vector regression (SVR) were systematically compared using several features that were based on MAP and MLLR adaptations. It has been shown that the combination of MAP-GMM supervector features and SVR estimator gave the best performance and the MAE was 7.3 years. Future work includes investigating kernels that utilize the position of elements in a matrix for MLLR transform-based features to improve SVM and SVR-based age estimators.

## References

[1] T. Hempel, "Usability of telephone-based speech dialog systems as experienced by user groups of different age and background," in *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, 2006, pp. 76–78.

[2] C. Müller and F. Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," in *Proc. ICSLP*, Antwerp, 2007, pp. 2277–2280.

[3] R. Nishimura, S. Miyamori, K. Suzuya, H. Kawahara, and T. Irino, "Web-based adult and child voice collection to develop a voice-oriented web filtering service," Tech. Rep. 77-19, IPSJ SIG Technical Report, 2009.

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

[5] H. Drucker, C. J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *NIPS*, 1996, pp. 155–161.

[6] C. Heerden, E. Barnard, M. Davel, C. Walt, E. Dyk, M. Feld, and C. Muller, "Combining regression and classification methods for improving automatic speaker age recognition," in *Proc. ICASSP*, 2010, pp. 5174–5177.

[7] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[8] 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.

[9] Jean luc Gauvain and Chin hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[10] T. Bocklet, A. Maier, J. G. 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, pp. 1605–1608.

[11] W. Spiegl, G. Stemmer, E. Lasarcyk, V. Kholhatkar, A. Cassidy, B. Potard, S. Shum, Y. C. Song, P. Xu, P. Beyerlein, J. Harnsberger, and Elmar Nöth, "Analyzing features for automatic age estimation on cross-sectional data," in *Proc. INTERSPEECH*, 2009, pp. 2923–2926.

[12] Koby Crammer and Yoram Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 2002.

[13] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.

[14] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, vol. 2, pp. 1137–1140.