COMBINING REGRESSION AND CLASSIFICATION METHODS FOR IMPROVING AUTOMATIC SPEAKER AGE RECOGNITION

Charl van Heerden, Etienne Barnard, Marelie Davel, Christiaan van der Walt Ewald van Dyk

Human Language Technologies
Meraka Institute, CSIR, South Africa
cvheerden@csir.co.za

Michael Feld, Christian Müller

Intelligent User Interfaces Dept.
German Research Center for AI
Germany
{michael.feld, christian.mueller}@dfki.de

ABSTRACT

We present a novel approach to automatic speaker age classification, which combines regression and classification to achieve competitive classification accuracy on telephone speech. Support vector machine regression is used to generate finer age estimates, which are combined with the posterior probabilities of well-trained discriminative gender classifiers to predict both the age and gender of a speaker. We show that this combination performs better than direct 7-class classifiers. The regressors and classifiers are trained using long-term features such as pitch and formants, as well as short-term (frame-based) features derived from MAP adaptation of GMMs that were trained on MFCCs.

Index Terms— Age classification, gender classification, support vector machines

1. INTRODUCTION

Speech is the most natural means of communication among humans. We not only share semantic concepts through our speech but also convey so-called paralinguistic information, such as the identity of the speaker, the gender, the approximate age, emotional state etc. This information guides the way in which we communicate with each other and is also considered to be important for human-computer communications, such as in interactive voice response (IVR) systems. Angry customers can, for example, be detected and approached differently by connecting them to a sympathetic human operator. In much the same way, dialogues, as well as background music or advertisements played while waiting for an operator, can be adapted based on other paralinguistic cues such as the speaker's age or gender [1].

Age and gender classification from speech has been a topic of interest from as early as the 1950's [2]. More recently, workshops have been organized to compare existing approaches to age and gender classification on a common database (German SpeechDat II corpus) [1] and the age classification task formalized as the classification of a speaker according to 7 age/gender groups (described in more detail in section 2). Approaches that have been employed successfully include classification based on phone recognition and direct age classification. For the latter case, two main classes of features have been most popular: long-term (mostly prosodic) features and short-term features based on Mel frequency cepstral coefficients (MFCCs). Extensive work has been done on refining and measuring the significance of the long-term features [3], as well as on ways to optimally combine the two feature classes [4].

Regression to estimate speaker ages has recently been suggested as an alternative to age-category classification [5]. Since the focus in [5] was to compare different feature types, the relative performance of regression-based and classification-based approaches was not investigated. We perform such a comparison (using support vector machine regression), and also show how regression can be combined with gender classification to perform the standardized 7-class task mentioned above. In fact, this regression-based approach is somewhat more accurate than 7-class classifiers trained on either of the above-mentioned feature classes.

In section 2, we describe the speech corpus used and provide background on the 7-class age classification task. We provide an overview of the features used in section 3 and describe our regression-based approach in section 4. In section 5, the regression / gender classification combination is compared to state-of-the-art 7-class classifiers, and a straightforward combination proposed. Our interpretation of the results, as well as possible future work is discussed in section 6.

2. CORPUS AND CLASSIFICATION TASK

The corpus used for this study contains the voices of approximately 700 German speakers, each taking part in 18 turns of up to 6 sessions. The audio was recorded at 8000 Hz (telephone quality). The selection of speakers is approximately evenly distributed over the 7 target classes (see Table 1), with

class 1 also being balanced for gender. The majority of utterances are between 1 and 6 seconds in duration (the distribution of durations is shown in Figure 2). In total, the corpus consists of 47 hours of speech.

Of all the speakers in the corpus, 90% were labeled and used for the experiment. Three working sets were created: a training set (40%), a development test set (30%) and an evaluation set (30%), each with non-overlapping speakers. The test and evaluation sets were balanced to a degree, based on the amount of available material.

Class	C	YF	YM	AF	AM	SF	SM
Gender	M+F	F	M	F	M	F	M
Ages task 1	< 13	$13-19$		$20-64$		> 64	
Ages task 2	< 15	$15-24$		$25-54$		> 54	

Table 1. *Age division by class for task 1 as determined in [1] and for task 2 as used in this experiment. The age boundaries for task 1 are motivated by earlier literature, while those for task 2 were selected based on the design criteria of the main application scenario of the corpus used. The abbreviations "C", "Y", "A" and "S" refer to "children", "youths", "adults" and "seniors", respectively.*

The age classification task closely follows a setup which was formalized at the first workshop of this nature, organized by Deutsche Telekom [1]. The age classes that are prevalent in the corpus on which our study is conducted are listed in Table 1. The original age-class scheme (*task 1*) was introduced by [6] and later applied in other studies on speaker age recognition. Due to the (acoustic) gender-indiscriminability of children before puberty, a 7-class task is obtained. Following the requirements of the IVR application for which the classifiers were trained, these boundaries were adjusted to match the *task 2* setup. The change of both class definitions and the underlying corpora between the two tasks needs to be taken into account when comparing system performance involving different setups. In general, task 2 can be considered more difficult due to shorter speech samples.

3. FEATURE EXTRACTION

Two classes of features were used to perform the age classification task: long-term (mostly prosodic) features (LTF) and short-term (frame-based), supervector features (SPV) derived from MAP adaptation of Gaussian mixture models (GMMs).

Long-term features were taken from a set of 22 common features covering the feature families: pitch, jitter, shimmer, and intensity. These features are known from the literature to carry various types of cues related to speaker characteristics, in particular age. For example, a high average micro-variation in voice frequency (jitter) may be due to an age-related deterioration of the glottis. The actual features include values aggregated over an utterance such as mean, minimum, maximum, standard deviation, and deltas. They were extracted using Praat[7] employing a cross-correlation method for pitch period analysis with a step width of $10ms$. A detailed description of all features is provided in [6].

In addition to these features, formants were also extracted from the voiced sections of the spoken audio. Praat was used to extract the first four formants using Burg's algorithm. A sliding window with a length of $25ms$ and a stepsize of $20ms$ was used. The maximum formant frequency was specified to be $5500Hz$, a common choice for adult females. One would typically choose a lower value for adult males and a much higher value for young children if the gender was known beforehand. The mean and standard deviation was then calculated for each formant, as well as its first derivative. The pitch corresponding to the period over which the formants were extracted was also added to create a 20-dimensional vector.

As **short-term features**, MFCCs were extracted from all utterances using the HTK toolkit with a stepsize of $5ms$, a window length of $30ms$, and a dimension of 12. A 128-mixture GMM was then trained to model the coefficients, with MAP adaptation applied to update the means and weights for all mixtures, given a new utterance. The resulting adapted means were then concatenated to form a $1,536$ dimensional supervector (12 coefficients $*$ 128 mixtures).

4. CLASSIFICATION DESIGN

In order to perform regression, the utterance vectors of both feature classes were annotated with the true ages of the speakers, as provided by the speakers during data collection. A support vector machine was then trained, with the objective of finding a function f that predicts the target ages with at most ϵ years deviation, while being as flat as possible [8]. Given these models, regression was then performed by mapping the test vectors into a high dimensional feature space, computing dot products with the transformed training vectors and adding the results using precomputed weights to obtain the final age estimate. Regressors were trained for both feature classes with LIBSVM, using the radial basis kernel function (RBF) [9]. It was found that the performance of these regressors, as well as of the classifiers described below, depends strongly on the parameters employed during training. For the regressors, these parameters are ϵ (the regression error that is allowed before a particular sample is penalized), C (which controls the trade-off between margin width and training-set error) and γ (the width of the RBF kernel), whereas classification involves C and γ only. Both regressors and classifiers were optimized in terms of ϵ, C and γ using 10-fold cross validation and grid searches on the training set. All folds contained data from distinct speakers and were balanced based on the number of speakers per fold.

Since the task of age classification in the commonly-used format requires a distinction between for example young males and young females, and since the regression estimate is insensitive to the gender of a speaker, it was necessary to train classifiers for distinguishing children, males and females

from each other. Two gender classifiers were thus trained to estimate the posterior probabilities of an utterance originating from children, males or females, using the LTF and SPV feature classes respectively.

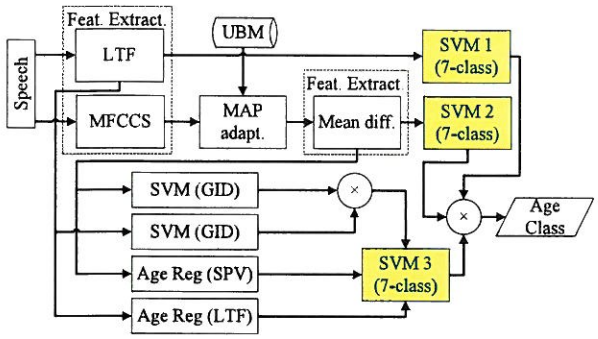

Fig. 1. *Flowchart illustrating the complete classification process. The 3 7-class classifiers are highlighted in yellow.*

A second level of classification was necessary to combine the outputs from the gender classification and the regressors. The posteriors from the two gender classifiers were multiplied and together with the two regression outputs, a 5-dimensional vector was created. These vectors were then used to train a 7-class classifier (SVM 3 in Figure 1).

Since the combination described above entails using the output of classification and regression results, we had to "generate" training data. This was accomplished by dividing the training set into 10 folds (having distinct speakers) and then following a round-robin approach to train gender classifiers and regressors on 90% and repeatedly classifying the remaining 10%. All classifiers used in this round-robin approach used the same parameters for a particular feature set. Another grid search was then used to optimize the 7-class classifier using cross validation.

In order to benchmark our regression-based classifier against existing techniques, we trained 7-class classifiers on both feature classes. The SVMs were trained using an RBF kernel, and grid searches combined with 10-fold cross validation were again employed to search for the optimal values of C and γ for each of the classifiers.

5. RESULTS

The performance of our *regressors* is measured in two ways: root mean squared error (RMSE), which reflects the average squared difference between the predicted and target ages, and correlation coefficients. By both measures, SPV-based regression was somewhat more accurate than LTF-based regression, the RMSE / correlation values being 18.4 and 0.50 (SPV) compared to 19.1 and 0.45 (LTF) on the independent test set.

For the three-class gender ID problem, the relative performances of the two feature types were inverted: here, the test set classification accuracies were 83.9% (SPV features) and 86.4% (LTF features), respectively. Combining these two classifiers, by multiplying the posterior probabilities of the respective classes, further improved the accuracy to 88.5%. The confusion matrix obtained with this classifier is shown in Table 2. This matrix contains recall values – that is, the number of true-class / estimated-class conjunctions normalized by the number of true-class occurrences. As expected, the children and females are most confusable in this set, and the relative preponderance of adult data causes a low level of recall for the utterances produced by children.

	C	F	M
C	**0.487**	0.413	0.100
F	0.032	**0.922**	0.045
M	0.002	0.025	**0.973**

Table 2. *Confusion matrix for the combined LTF and SPV feature classes when used for gender ID. Rows indicate the actual classes of the utterances, whereas columns correspond to the outputs produced by the classifier.*

Using the method described above, the regressors and gender ID system were combined to perform the standard 7-class classification problem (see SVM 3 in Figure 1) . This resulted in a 7-class accuracy of 48.4% on the test set. By way of comparison, the two direct 7-class classifiers (SVMs 1 and 2 in Figure 1) achieved accuracies of 45.7% (LTF features) and 45.2% (SPV features). As with gender identification, multiplying the class posterior probabilities produces a combined classifier that outperforms each of the individual classifiers, and an overall accuracy of 50.7% is measured with this combined classifier. Combining only the direct classifiers yielded a classification accuracy of 46.9%.

On comparable but different data sets, direct classification accuracies of 40% and 42% were reported for LTF and SPV features, respectively [1]. We therefore believe that our results are amongst the best that have been achieved with such transcription-free approaches, although they are somewhat worse than the 54% that was reported with a phone-based system (which, of course, required the availability of transcriptions during training) [1].

It is interesting to see how classification accuracy depends on utterance duration for the various approaches. As can be seen in Figure 2, all the approaches benefit from longer utterances, with the accuracy of the combined classifier increasing from less than 50% for utterances around 1 second in length, to over 60% for utterances longer than 6 seconds. The SPV features appear to be somewhat more sensitive to utterance length than the LTF features, but the differences are quite small.

The confusion matrix of the combined system is shown in Table 3. We see that the majority of confusions are either within the same gender, or between children and young females. Of the age groups, the adults are the hardest to clas-

sify (for both males and females), since they are confusable with both youths and seniors of the same gender.

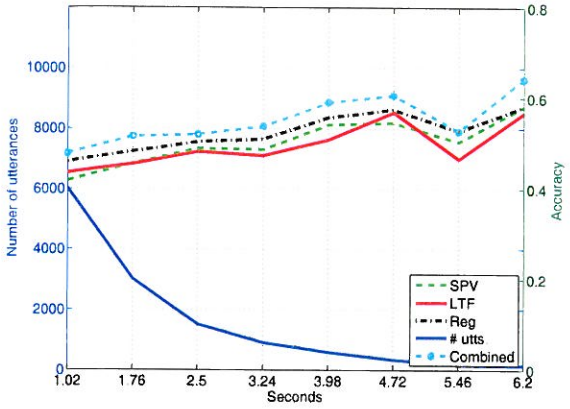

Fig. 2. *Accuracy of the different classifiers vs utterance duration, along with the distribution of utterance durations in the corpus employed.*

	C	YF	YM	AF	AM	SF	SM
C	**0.593**	0.156	0.059	0.077	0.011	0.080	0.025
YF	0.130	**0.559**	0.001	0.198	0.000	0.111	0.003
YM	0.003	0.000	**0.467**	0.003	0.205	0.035	0.287
AF	0.051	0.184	0.040	**0.382**	0.026	0.299	0.018
AM	0.006	0.000	0.263	0.014	**0.308**	0.026	0.384
SF	0.055	0.126	0.006	0.242	0.002	**0.560**	0.009
SM	0.003	0.000	0.089	0.001	0.156	0.026	**0.726**

Table 3. *Confusion matrix for the combination classifier, with an overall accuracy of 50.7%. Conventions as in Table 2*

6. CONCLUSION

This paper presents a novel approach to the age/gender classification task. We used support vector machine regression to estimate speaker ages directly and then combined these age estimates with the posterior probabilities of well-trained gender classifiers in order to present results comparable to the standard 7-class classification task. The regressors performed reasonably well, with mean prediction errors of approximately 18 and 19 years for the two feature classes, respectively. This is is a useful level of accuracy, when calibrated by the fact that classes 4 & 5 cover a range of 44 years.

It was also shown that the combination of a regressor and a gender classifier performs consistently better than either of the two direct 7-class classifiers (Figure 2). It was interesting to note a consistent and approximately equal relative improvement of all classifiers as the length of the test utterances increases. This is in contrast to what was found in [1]. The fact that the 7-class LTF classifier significantly outperforms the SPV classifier on very short utterances is also notable. This may be due to the fact that the LTF features were extracted for segments containing voiced speech only, whereas the SPV features were extracted for the whole utterance.

Future work should include the optimization and further investigation of formants as a feature for age classification; of all the features described here, these have received the least attention. Another promising area of investigation is feature selection, especially for the high-dimensional supervector features. It will also be interesting to see whether significantly improved results can be obtained by combining the strategies that we have used with transcription-based approaches where such transcriptions are available (or by bootstrapping such transcriptions, where they are not).

7. REFERENCES

[1] 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 *ICASSP*, Honolulu, Hawaii, April 2007, pp. 1089–1092.

[2] Edward D. Mysak, "Pitch duration characteristics of older males," *Journal of Speech and Hearing Research*, vol. 2, pp. 46–54, 1959.

[3] C. Müller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," in *Interspeech*, Pittsburgh, Pennsylvania, September 2006.

[4] C. Müller and F. Burkhardt, "Combining short-term cepstral and long-term prosodic features for automatic recognition of speaker age," in *Interspeech*, Antwerp, Belgium, August 2007, pp. 2277 – 2280.

[5] W. Spiegl, G. Stemmer, E. Lasarcyk, V. Kolhatkar, A. Cassidy, B. Potard, S. Shum, Y. Chol Song, P. Xu, P. Beyerlein, J. Harnsberger, and E. Nth, "Analyzing features for automatic age estimation on cross-sectional data," in *Interspeech*, Brighton, England, Sept 2009, pp. 2923 – 2926.

[6] Christian Müller, *Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht*, Ph.D. thesis, Computer Science Institute, University of the Saarland, Germany, 2005.

[7] P. Boersma, *Praat, a system for doing phonetics by computer*, Amsterdam: Glott International, 2001.

[8] Alex J. Smola and Bernhard Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 1573–1375, August 2004.

[9] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.