Age and gender classification from speech using decision level fusion and ensemble based techniques

Florian Lingenfelser, Johannes Wagner, Thurid Vogt, Jonghwa Kim, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Lingenfelser, Florian, Johannes Wagner, Thurid Vogt, Jonghwa Kim, and Elisabeth André. 2010. "Age and gender classification from speech using decision level fusion and ensemble based techniques." In *INTERSPEECH 2010: 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, edited by Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, 2798–2801. ISCA. https://www.isca-archive.org/interspeech_2010/lingenfelser10_interspeech.html.

Age and Gender Classification from Speech using Decision Level Fusion and Ensemble Based Techniques

Florian Lingenfelser, Johannes Wagner, Thurid Vogt, Jonghwa Kim, Elisabeth André

Multimedia Concepts and their Applications, University of Augsburg, Germany
<name>@informatik.uni-augsburg.de

Abstract

In this contribution to INTERSPEECH 2010 Paralinguistic Challenge we explore the capabilities of decision level fusion and ensemble based techniques for classification tasks on the provided AGENDER corpus. Ensemble members are generated by providing multiple feature sets generated by feature selection, and novel fusion methods (developed in order to give special support to under-represented classes) are applied for decision making. Results are compared to standard classification approaches and possible benefits are discussed.

Index Terms: age and gender recognition from speech, decision fusion, ensemble classification

1. Introduction

Humans are used to adapt their behaviour in dependence of age and gender of their communication partner(s). Likewise, a system could involve this contextual information about its users in the decision making process to respond in a more adequate way. Using paralinguistic information conveyed in speech offers a possibility to automatically detect the speakers' age and gender, and has evolved to an own sub-discipline in the field of speech analysis contributing to speaker classification [1]. Yet, there is no standard regarding the evaluation procedures and comparability. In response to this deficit, the INTERSPEECH 2010 Paralinguistic Challenge addresses age and gender classification in two out of three sub-challenges [2]. In this paper we contribute to both and explore the capabilities of decision level fusion and ensemble based techniques for classification. All experiments are carried out on the standard feature set provided by the organizers.

Age and gender recognition from speech has often been treated as a combined problem. For example, Müller [1] uses a variety of paralinguistic features including pitch, voice quality, speech rate and pauses to recognise 8 classes (4 age classes separated by gender). Bocklet et al. [3] compare Gaussian Mixture Models (GMMs) and Support Vector Machines (SVM) on cepstral features only to recognise 7 gender/age classes. Both approaches can exceed chance level many times over.

In terms of extracted features and classification methods, affect recognition from paralinguistic information is closely related to the task of gender and age recognition. Therefore, we apply here an ensemble approach called Cascading Specialists and variations to the classification of age and/or gender that we have previously used for affect recognition [4] and where the final decision on the classification is obtained from fusing results of sub-tasks.

There are also other approaches to emotion recognition that have explored decision level fusion as strategy to improve recognition performance. Schuller and colleagues [5] apply Bagging, Boosting and Stacking, three popular methods of ensemble classification, to avoid data overfitting in situations with a relative small number of training samples compared to a high number of features (also known as curse of dimensionality). Using the latter method the authors could achieve a slight improvement from 70.27% to 71.62% on an 8-class problem compared to the best single classification achieved with Support Vector Machines (SVM).

The approach most closely related to the one presented in this paper is reported by Lee and colleagues [6]. With their proposed framework the authors were able to improve the unweighted recall in a 5-class problem by 3.37% compared to a single SVM classifier. The authors propose a hierarchical tree of binary classifiers, select the individual feature sets for each tree node from the same global feature set and derive the tree structure from the difficulty of the classification subtasks. However, the order is chosen in a way that the easier recognition tasks are found in the top levels of the tree. This differs from our strategy, which starts with the most difficult problem.

2. Decision Level Fusion

While common classification approaches focus on the establishing of a single and potent classifier that deals with a high dimensional feature vector, the term decision level fusion sums up a variety of methods that rely on the usage of some small classifiers and their combination. Instead of using all calculated features as training set for a sole classifier, the available feature set is somehow grouped and these partitions are used to form several small classifiers. The outcomes of these slim classifier models are taken into account for certain decision making processes, that lead to a final classification result.

2.1. Building the Ensemble

As the term decision level fusion describes the establishment and evaluation of an ensemble based system, decisions are made by combination of a certain amount of classifiers, i.e. classifier ensemble.

The classifiers used in creating the ensemble for all discussed decision making algorithms stem from different feature sets generated by modified feature selection processes. More precisely, for n given classes of a classification problem $n+1$ ensemble members are generated. Generally said, the first n feature sets are chosen with preferential treatment of the respective n classes. In addition one subset is meant for equally distributed classification accuracies among all classes. The reason for this partitioning of available features will be further explained in the following sections.

2.2. Feature Selection

In order to discard irrelevant features, reduce the search space and to provide different feature sets (resulting in diverse ensemble members), we carry out feature selections. The n sets meant to support each of the n classes are generated by respectively labelling the supported class against the other categories in the data set. Resulting feature sets should be able to recognize associated classes with preference. The feature set aimed at balanced classification results undergoes the standard feature selection process without manipulation of labels.

2.3. Diversity

As mentioned above an ensemble based system consists of a set of different classification models. Neither must these classifiers provide perfect performance on some given problem, nor do their outputs need to resemble each other. It is preferable that the chosen classifiers make mistakes, at best on different instances. A base idea of decision level fusion is to reduce the total error rate of classification by strategically combining the members of the ensemble and their errors. Therefore the single classifiers need to be diverse from one another. In the given case this crucial requirement is achieved by the described variation of available features through modified feature selection appliance.

3. Applied Fusion Methods

3.1. Mean Rule (MEAN)

The Mean Rule is a standard decision level fusion method, meant to combine the continuous outputs of all available ensemble members. By averaging the support given to each class ω_n the total support μ_n for class n throughout the whole ensemble can be calculated as:

$$\mu_n(x) = \frac{1}{T}\sum_{t=1}^{T} s_{t,n}(x)$$

T denotes the total amount of classifiers (therefore $\frac{1}{T}$ serves as normalization factor), $s_{t,n}$ describes the support given to class n by the t_{th} classifier in the ensemble. Finally the ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.

The appliance of this method is meant as a first appraisal of fusion potential and delivers some clues about compositions of the datasets, sample distributions and resulting characteristics in classification accuracies for single classes.

3.2. Cascading Specialists (CS)

In many cases, accurate investigation of confusion matrices for ensemble classifiers shows two phenomena concerning performance on single classes: A good true positive classification rate - which commonly serves as measure for single class performance - goes hand in hand with a high false positive classification rate, because many samples are simply put into the inspected dominant class. In return classes with mediocre performance mostly also show low false positive rates, as a small amount of samples ever gets put into that category. The cascading specialists method bases on choosing experts for single classes and brings them in a special sequence in order to soften the mentioned phenomena. In a preparation step experts for each of the n classes are selected by finding the classifier with best true positive rating for every class of the classification problem. These choices are based on evaluation of the training phase. Then the classes are rank ordered, beginning with the worst classified class across all classifiers and ending with the best one.

Given the preparation step and a test sample, the algorithm works as follows: The first class in the sequence is chosen and the corresponding expert is asked to classify the sample. If the output matches the currently observed class, this classification is chosen as ensemble decision. If not, the sample is passed on to the next weaker class and corresponding expert whilst repeating the strategy. Sometimes the case occurs that none of the experts classifies its connected class and the sample remains unclassified at the end of the sequence. Then the classifier with the best overall performance on the training data is selected as final instance and is asked to label the sample as ensemble decision.

This strategy aims at a more uniformly distributed accuracy among classes. Weakly recognized classes are treated with priority and the belonging samples are more unlikely to end up falsely classified as a more dominating class later on. This results in a flattening effect that will at best improve overall classification performance.

The build-up of this ensemble fusion method explains the choice of sub-sets generated by feature selection described in Section 2.2. Feature sets supporting single classes are expected to serve as base for respective specialists while the single set covering all classes is meant to result in a classifier to be used for labelling samples which passed the process unclassified.

3.3. Cascading Specialists - One versus Rest (CS-OvR)

In the simple Cascading Specialists approach, specialists are automatically chosen among a set of classifiers trained to recognize all observed classes. These experts are rated based on their performance on the single classes. A more adapted approach can be developed by training classifiers specialised in separating their connected classes from $n-1$ remaining classes. They are no longer chosen based on evaluation of training performance but determined right from the start. This results in n classifiers dealing with two-class classification problems. The established concepts of bringing the specialists in ascending order and intercepting unclassified samples with a well-balanced classification model are maintained.

The theoretical advantage of this strategy lies in the generation of best possible specialists for recognizing the associated classes. While this specialisation implies the potential of enhancing classification accuracies for single classes, there are some major drawbacks involved. The strict association of specialists with their belonging classes can lead to problems whenever training and test samples vary greatly, because there is no flexibility in specialist selection. This disadvantage also applies to the simple Cascading Specialists approach to a lesser extent, as its classifiers feature a broader classification potential than the one-versus-rest classification models.

3.4. Cascading Specialists - Multiple Specialists (CS-MS)

The last presented variant of the Cascading Specialists approach deals with the question of how to become less dependent on the evaluation of training data. A possible answer is to choose more than one specialist for every class and to execute internal fusion steps. This way the specialist assortment becomes more flexible and less conditioned by unreliable training performance, as the risk of choosing one wrong expert becomes less important.

In detail $\lceil \frac{n}{2} \rceil$ classifiers among the ensemble are chosen to become specialists for each of the n classes. Whenever the sequence demands a new decision concerning class-belonging the corresponding specialists generate a decision via the Weighted Average Rule. Again, the concepts of ascending class order and an intercepting classification model stay untouched.

The Weighted Average Rule is an extension of the Mean Rule by adding weights to the classifiers and their continuous outputs.

$$\mu_n(x) = \frac{1}{T}\sum_{t=1}^{T} w_t s_{t,n}(x)$$

The formula is simply adjusted by the weight (deduced from overall performance on training data) w_t of the t_{th} classifier.

4. Experimental Setup

The experimental setup is realised with Smart Sensor Integration (SSI), a framework for multi-modal signal processing in real-time published under public domain [7]. SSI supports the building of online recognition systems by offering all necessary tools to assemble a machine learning pipeline and apply it to a certain recognition problem. In particular, this also involves the possibility to train and evaluate a classifier (or an ensemble of classifiers) on a set of training samples. Once a classifier has been trained it can be integrated into a pipeline for online classification. Usually, such a pipeline contains a sequence of filter components, which pre-process the raw input signal in a suitable way, and a set of feature blocks, which extract the final feature vector. However, for the accomplishment of the following experiment the pipeline is shrunken to a single component, which simply reads in the pre-calculated features from a file.

As already mentioned, the aim of this paper is to investigate the potential of decision fusion on the Age and Gender Sub-Challenge. The corpus provided by the organizer is divided into a training and development set containing 32 527 samples and 20 549 samples, respectively. For each sample the age in years and the gender (m=male, f=female, x=child) is given. Based on the age, samples are divided in 7 different age groups (1-7), which are further aggregated into four general groups (C=Child, Y=Youth, A=Adult, S=Senior). To maintain comparability with other submissions following a similar approach, we decided to use the set of standard features provided by the organizer composed of 450 prosodic features [2].

The recognition component in SSI supports a hierarchical architecture. It is composed of a fusion strategy, which sits on top of the classifier ensemble. For this experiment we implemented four different fusion strategies as discussed in Section 3. The classifiers in the ensemble are not restricted to a certain classification method, but can be chosen independently from each other. This allows us to test the same fusion strategy with different types of classifiers. Due to the massive number of training samples we decided to start with the Naive Bayes classifier, which is a simple classification scheme and extremely fast in training and test, even for high-dimensional feature vectors and large training databases. It is therefore especially suited for real-time processing and commonly used in our real-time emotion recognition system EmoVoice [8], which is part of the SSI system. As a second classifier scheme we used the SVM classifier provided by LibSVM [1], which is also available from the SSI framework.

As feature selection strategy we use correlation-based feature subset selection (CFS) from Weka[2], a selection technique that searches for subsets with features that are highly correlated with the class but not with each other.

5. Results

Table 1 shows the results obtained for experiments carried out with the Naive Bayes classifier. The three blocks describe different label-assignments, i. e. GENDER classifies the given samples in three classes describing probands as children, male or female adults. AGE4 groups them into four subsequent age-groups and AGE7 further sub-divides these partitions. Respective feature fusion results are presented without (NO) and with (SEL) suitable feature selection in the first two rows and can be interpreted as Bayes-related baseline results. The following lines sum up the applied fusion experiments and show possible gains in classification accuracy compared to feature fusion. Classification performance is given class-wise (CW), as well as the unweighted average recall (UA). Note that decision level fusion methods (especially cascading specialists with multiple specialists - CS-MS) tend to outperform feature level fusion by about 5% without feature selection and around 1% with feature selection, whilst establishing a more balanced accuracy distribution among the single classes.

Table 1: *Classification results using **Naive Bayes** classifier.*

method	% CW	% UA
GENDER $\{m, f, x\}$		
NO	90.2 61.0 47.3	66.15
SEL	89.5 70.5 51.9	70.62
MEAN	89.2 75.3 49.5	71.34
CS	86.5 79.0 49.8	**71.75**
CS-OvR	87.8 62.1 64.3	71.40
CS-MS	88.0 74.0 52.1	71.31
AGE4 $\{C, Y, A, S\}$		
NO	58.6 13.0 67.8 08.2	36.88
SEL	63.4 20.6 44.9 36.2	41.28
MEAN	62.1 26.6 43.0 37.2	42.21
CS	50.1 48.0 25.2 40.8	41.04
CS-OvR	47.7 57.7 10.4 45.4	40.29
CS-MS	62.3 23.9 40.2 43.5	**42.47**
AGE7 $\{1, 2, 3, 4, 5, 6, 7\}$		
NO	32.6 63.0 27.8 20.1 58.0 05.8 18.7	32.28
SEL	41.9 63.2 45.4 26.4 32.2 19.1 29.3	36.79
MEAN	41.8 62.4 42.8 35.8 36.0 11.8 27.9	36.93
CS	46.5 56.3 50.6 30.6 17.5 23.2 33.7	36.89
CS-OvR	31.1 01.3 02.2 30.4 06.1 57.4 80.8	29.91
CS-MS	44.2 56.4 45.9 34.8 30.4 19.0 29.6	**37.18**

6. Discussion

Because of diverse fusion results with the SVM classifier the following discussion at first bases solely on observations made whilst experimenting with Bayes classifiers. Afterwards we present a possible explanation regarding the performance drop observed with SVM.

The simple cascading specialists approach performs stable on all tested datasets, as the aim of supporting under-represented or weakly recognized classes is nearly always achieved. A gain in classification accuracy for these classes of course is attained at the expense of stronger classes. Unfortunately the One versus Rest variant pushes this behaviour beyond justifiable limits, so that other classes accuracies get lowered in a disproportionate way. While simple cascading specialists mostly lead to enhanced overall performance - because of the appreciated flattening effect - the One versus Rest ap-

[1] http://www.csie.ntu.edu.tw/

[2] http://www.cs.waikato.ac.nz/ml/weka/

proach tends to lower the decision level fusion result in comparison to feature fusion. Both methods suffer from disparities between training data and actual test data as only one specialist is chosen for respective classes and the selection of classification models as well as their structure heavily depends on evaluation of training. For example the similarities between child and female classes in the GENDER dataset seem to provoke the selection of wrong specialists. Furthermore the child class is heavily under-represented in the dataset. This fact together with implicit inhomogeneous characteristics of children's voices (e.g. as the class sums up individuals before and after puberty vocal change) bears special risks for wrong specialist-selections between female subjects and children.

The strategy of choosing multiple specialists for each class lowers among others these mentioned risks. It results in the most beneficial flattening effect of the three presented variants and therefore leads to best overall performance in most cases. Being less reliant on evaluation of training data further affirms the impression of multiple specialists being the best choice across given datasets.

So far, we used the Naive Bayes classification model for made experiments. It consumes dramatically less computation time (and therefore is well suited for on-line classification – our main field of interest) compared to the SVM classification models used in generating baseline results for the Paralinguistic Challenge. Due to it's simplicity, the bayesian model generally tends to result in worse classification accuracies. Results discussed in this work evaluate *Train vs. Develop* datasets and are compared to final competition results on respective *Train vs. Test* sets as well as the more time consuming baseline results in Table 2.

Table 2: *Train vs. Develop and Train vs. Test classification results with Naive Bayes compared to SVM baseline.*

Train vs. Develop	Train vs. Test	Baseline
GENDER $\{m, f, x\}$		
71.75	73.9	76.99
AGE4 $\{C, Y, A, S\}$		
42.47	41.82	46.22
AGE7 $\{1, 2, 3, 4, 5, 6, 7\}$		
37.18	-	44.24

But why no competitive results with SVM classification? Table 3 shows first results for GENDER classification with SVM on *Train vs. Develop* datasets. Classification without feature selection does already equal the respective challenge baseline but in contrast to the Naive Bayes classifier, feature selection in combination with the SVM classifier turns out to have a negative effect on the classification performance. However, this consequently must lead to a negative impact on our proposed fusion scheme, as the design of the cascading specialist is based on the idea to find for each class a subset of features that yields a significant better recognition result for the particular class compared to the whole feature set. To verify this assumption Table 4 lists the evaluation results for each of the subsets, which are used by the cascading specialist to determine the specialist for each class. In fact, the improvements in terms of the associated classes are marginal or even worse (like in the case of child). A possible solution for problems with the SVM classification model and the chosen feature selection could be the appliance of a different selection algorithm. Strategies like sequential backward search (SBS) or sequential forward search (SFS) inherently use the SVM classifier and therefore should be capable of producing more fitting feature sets.

Table 3: *Classfication results using **SVM** classifier.*

method	% CW	% UA
GENDER $\{m, f, x\}$		
NO	93.5 90.1 48.1	**77.24**
SEL	91.9 89.2 38.6	73.21
CS	91.9 89.8 38.6	73.43
CS-MS	91.2 90.6 37.6	73.15

Table 4: *Classwise evaluation results in % for the GENDER $\{m, f, x\}$ subsets used by the cascading specialist to determine class specialists.*

subset	Naive Bayes	SVM
SEL-M	**88.9** 69.4 52.0	**92.4** 88.1 38.9
SEL-F	84.7 **81.1** 45.1	91.1 **88.1** 31.2
SEL-X	87.9 71.1 **55.7**	91.1 86.7 **40.1**
SEL-ALL	**88.9 69.0 54.9**	**92.4 87.6 42.4**

7. Conclusions

Present evaluation of a decision level fusion method with cascading specialists (that was originally developed for multi-sensor data fusion in [4]) and its further variations shows encouraging results on the INTERSPEECH 2010 Paralinguistic Challenge AGENDER dataset. The intended uprating of under-represented classes can be achieved and the resulting flattening effect increases overall classification performance by up to 5% compared to feature level fusion results obtained by the same classification model. All in all these fusion methods yield reliable classification rates with the huge benefit of a very balanced accuracy on all observed classes. They are capable of very precise detection for weakly recognised emotion-classes and therefore they can always be considered as possible fusion method for practical applications.

8. Acknowledgements

This work was partially financed by the EU in the CALLAS Integrated Project (IST-34800) and the IRIS Network of Excellence (Reference: 231824).

9. References

[1] C. Müller: "Automatic recognition of speakers' age and gender on the basis of empirical studies", INTERSPEECH (2006).

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller and S. Narayanan: "The Interspeech 2010 Paralinguistic Challenge", INTERSPEECH (2010).

[3] T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, E. Nöth: "Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines", ICASSP (2008).

[4] J. Kim, F. Lingenfelser "Ensemble Approaches To Parametric Decision Fusion For Bimodal Emotion Recognition", Biosignals (2010).

[5] B. Schuller, R. Müller, M. Lang, and G. Rigoll: "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles", INTERSPEECH (2005).

[6] C.-C. Lee, E. Mower, C. Busso, S. Lee and S. Narayanan: "Emotion Recognition Using a Hierarchical Binary Decision Tree Approach", INTERSPEECH (2009).

[7] J. Wagner, E. André and F. Jung, "Smart sensor integration: A framework for multimodal emotion recognition in real-time", ACII (2009).

[8] T. Vogt, E. André, N. Bee, "EmoVoice - A framework for online recognition of emotions from voice", PIT (2008)