Age and Gender Classification using Fusion of Acoustic and Prosodic Features

Hugo Meinedo[1], *Isabel Trancoso*[1,2]

[1]L2F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
[2]Instituto Superior Técnico, Lisboa, Portugal
{hugo.meinedo, isabel.trancoso}@l2f.inesc-id.pt

Abstract

This paper presents a description of the INESC-ID Spoken Language Systems Laboratory (L2F) Age and Gender classification system submitted to the INTERSPEECH 2010 Paralinguistic Challenge. The L2F Age classification system and the Gender classification system are composed respectively by the fusion of four and six individual sub-systems trained with short and long term acoustic and prosodic features, different classification strategies (GMM-UBM, MLP and SVM) and using four different speech corpora. The best results obtained by the calibration and linear logistic regression fusion back-end show an absolute improvement of 4.1% on the unweighted accuracy value for the Age and 5.8% for the Gender when compared to the competition baseline systems in the development set.

Index Terms: Paralinguistic Challenge, Age, Gender, Fusion of Acoustic and Prosodic Features

1. Introduction

Paralinguistic analysis, where age and gender detection are two of its tasks, is a rapidly emerging field of research due to the constantly growing interest in applications in the fields of Multimedia Retrieval and Human-Machine Interaction. Gender detection is a very useful task for a wide range of applications. In the Spoken Language Systems lab of INESC-ID, the Gender Detection module is one of the basic components of our audio segmentation system [6], where it is used prior to speaker clustering, in order to avoid mixing speakers from different genders in the same cluster. Gender information is also used for building gender-dependent acoustic modules for speech recognition. In our fully automatic Broadcast News subtitling system, deployed at the national TV channel since March 2008 [6], gender information is also used to change the color of the subtitles, thus helping people with hearing difficulties to detect which speaker the subtitle refers to, a useful hint that partially compensates the small latency of the subtitling system. Gender Detection is also a prominent part of our participation in the VIDIVIDEO European project, aiming at the semantic search of audio-visual documents [11]. In this application, the audio concept "male-voice" may be much easier to detect than the corresponding video-concept "male-speaker".

Most gender classification systems are trained for distinguishing between male and female adult voices alone. In fact, in some applications like Broadcast News (BN) transcription, children's voices are relatively rare, hence justifying their non-inclusion. The difficulties in collecting large corpora of children's voices may also be one of the reasons why most detectors do not attempt a three class distinction. In some applications such as the automatic detection of child abuse videos on the web, however, the detection of children's voices is specially important. This is the target of our participation in the European I-DASH project.

The goal of the INTERSPEECH 2010 Paralinguistic Challenge is to help bridging the gap between the research on paralinguistic information in spoken language and low compatibility of results, because of lacking of agreed-upon evaluation procedures and comparability, in contrast to more traditional disciplines in speech analysis. The Paralinguistic Challenge addresses these issues by implementing an open competition with three selected tasks (Age, Gender and Affect) and also by supplying appropriate train and test resources. Our laboratory participated in the Age and in the Gender sub-challenges. In the first of these sub-challenges, four age groups children, youth, adults and seniors have to be discriminated. In the gender sub-challenge a three-class classification task has to be solved, separating children, female and male.

In the following sections we introduce the corpora (Section 2), the system developed for the Age sub-challenge (Section 3), the system developed for the Gender sub-challenge (Section 4) before concluding (Section 5).

2. Corpora

Four different corpora were used to train and to evaluate the performance of the developed age and gender detection systems. The aGender corpus [5], the CMU Kids corpus [3], the PF STAR children corpus [4] and the BN ALERT corpus [6]. All corpora were pre-processed in order to boost the energy levels and remove unwanted silence. All aditional corpora were down-sampled from 16kHz to 8kHz to match the sampling frequency of the aGender corpus audio files.

2.1. aGender

The aGender corpus [5] was supplied by the InterSpeech 2010 Paralinguistic Challenge organization to assist in the development of speaker age and gender detection systems. It consists of 49 hours of telephone speech, stemming from 795 speakers, which are divided into train (23h, 471 speakers), development (14h, 299 speakers) and test sets (12h, 175 speakers). In our work, this partitioning was respected with the train set being used for training of age and gender systems and the development set being used for the calibration, fusion and evaluation. The classification results obtained in the test set were sent to the organizers for the competition evaluation.

2.2. CMU Kids

The CMU Kids corpus [3] is comprised of sentences read aloud by children from 6 to 11 years old. It was recorded in a controlled environment. It consists of 24 male and 52 female speakers totaling approximately 9 hours. All the available speech data

was used as training material both for age and gender systems.

2.3. PF STAR Children

The PF STAR Children corpus [4] was provided by the KTH Research group. Similar to CMU-Kids, this corpus was also recorded in a controlled environment, but includes more diversity of speakers (108 male and 91 female children). This corpus has 2 types of recordings, each with approximately 9 hours of speech, one recorded with headset and the other with a desktop microphone. As expected, the energy level of the second type of recordings is much lower and some reverberation effects can be perceived. Both types of recordings were used for training age and gender systems.

2.4. BN ALERT

The BN ALERT corpus [6] was the first European Portuguese Broadcast News corpus. It is composed of recordings from the RTP public TV station. This corpus was used for training gender detection systems since it is labelled according to the gender of the speakers but no age information is available. We used as training data three different sets (train, pilot and devel) consisting of 57 hours with 1182 male and 508 female speakers.

3. Age Sub-Challenge

In this sub-challenge, it was requested to detect the age of the speakers in four separate classes Child, Young, Adult, Senior $\{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$. The training and development data from the aGender corpus is labelled according to these age groups, but additionally distinguishes each of the Young, Adult and Senior classes between gender. This gives a total of seven combined age-gender classes $\{1, \ldots, 7\}$. The developed age detection systems output the results in one of these seven classes which are then combined to produce the required four age classes. This is achieved by adding the output probability scores of female and male for each of the age classes, Young, Adult and Senior.

$$\begin{aligned} p(C) &= p(1) \\ p(Y) &= p(2) + p(3) \\ p(A) &= p(4) + p(5) \\ p(S) &= p(6) + p(7) \end{aligned} \tag{1}$$

Our approach for the age detection system (Figure 1) uses several separate age detection front-end systems, that take advantage of different features, classification paradigms, and different training datasets. The output scores of each of these front-end systems is then calibrated and combined to produce the final system output. The motivation for having several front-ends with different properties is that the diversity will improve the combination and ultimately will lead to a more robust age detection system.

3.1. Front-Ends

In this section we describe the four developed front-end age detection systems. The first two systems used as training data the provided challenge features [5], here denoted as "arff450", obtained in the aGender training set. Two different classification paradigms were used, that take as input these "arff450" features. Support Vector Machines (SVM) for which we used the toolkit LibSVM [1], and Multi-Layer Perceptrons (MLP) for which we used our own implementation [6]. The SVM front-end uses a linear kernel, and the MLP front-end uses a fully

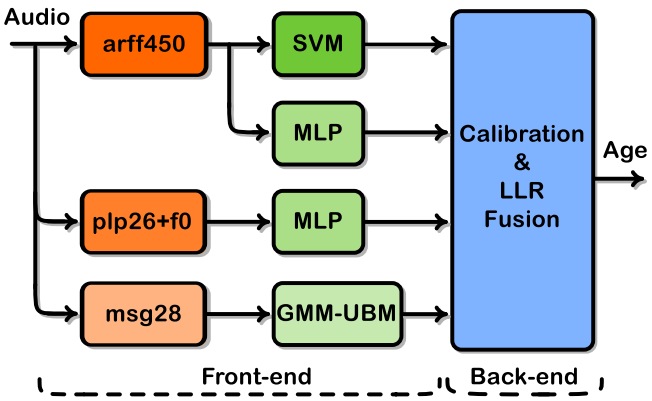

Figure 1: *Age detection system.*

connected feed-forward architecture with two hidden layers of 50 sigmoidal units and softmax outputs.

The other two front-ends developed for detecting age used all available training data that as age labels, that is, the aGender training set plus the CMU Kids, PF STAR children head-mount and desktop sets. These last three sets are herein simply referred to as "child" data. By using the additional child data, we are promoting diverseness and ultimately better combination scores. The third front-end extracts from the audio every 10ms a frame with 12th order Perceptual Linear Prediction (plp) [7] coefficients plus energy plus deltas plus pitch (f0). The slower average speaking rate of children and senior relative to adults is the motivation for including delta, plp and other temporal modeling coefficients in the feature set. Experiments with higher order plp did not lead to improved results possibly because of the small quantities of training data when compared with usual speaker recognition evaluation campaigns which typically use thousands of hours of speech material. The same applies to the use of double-deltas in our feature set. This front-end takes advantage not only of acoustic but also of the pitch prosodic feature both at frame level (short term features) and at utterance level (long term functional arff features). For the third front-end, the classification paradigm used was the MLP which takes as input context seven contiguous frames of features and has two hidden layers of 100 and 50 units. This configuration of hidden units was the one that achieved the better classification scores.

The fourth front-end extracts from the audio a frame of 28 static modulation spectrogram [9] features. The classification paradigm used was also different from the others, Gaussian Mixture Models - Universal Background Model (GMM-UBM) with 1024 mixtures. After training the UBM, each of the age class GMMs was created by performing five iterations of Maximum a Posteriori (MAP) adaptation.

Other pairings of features and machine learning methods were tested but the reported configurations were the ones which lead to better combination results in the aGender development set. The motivation for having front-ends with different properties is that the diversity improves the combination and ultimately will lead to a better final system.

3.2. Calibration and Fusion Back-End

Linear logistic regression fusion and calibration of the four front-end systems has been done with the FoCal Multiclass Toolkit [2]. The output log-likelihood ratio (llr) scores from

this fusion back-end were later converted into probabilities, as requested by the competition results format. This was achieved by scaling the scores to produce confidence values with the expression (2).

$$p(score(t)) = \frac{e^{score(t)}}{\sum_k e^{score(k)}} \quad (2)$$

3.3. Results

Table 1 summarizes the results obtained in the aGender development set by the four different front-ends individually and by the combination of them using the calibration and llr fusion back-end. Results are expressed in terms of Unweighted and Weighted Accuracy on average per class (% UA and %WA). The former (%UA) is the challenge measure since the distribution among different classes is not well balanced [5].

Table 1: *Age Results obtained in the aGender Development set.*

Systems $\{1,\ldots,7\} \rightarrow \{C, Y, A, S\}$	% UA	%WA
SVM - arff450 - aGender	47.4	47.0
MLP - arff450 - aGender	46.7	48.1
MLP - plp26+f0 - aGender+child	49.2	47.5
GMM-UBM - msg28 - aGender+child	39.7	44.5
Fusion	51.2	50.6

The best individual front-end in terms of % UA combines short term acoustic and prosodic features (plp26+f0) with an MLP classifier and uses an expanded training set with additional child data. Fusion of these four systems represents an improvement of 2% absolute over the best front-end and 4.1% over the reported challenge baseline [5]. Our final competition result in the test set is 48.7% UA.

Table 2: *Age confusion matrix results for the final fusion system obtained in the aGender Development set.*

	C	Y	A	S
C	**57.5**	24.0	11.1	7.4
Y	8.5	**50.7**	24.4	16.4
A	2.6	24.2	**39.9**	33.3
S	2.5	12.6	28.1	**56.7**

An analysis of Table 2 shows that there are some confusions between neighboring classes especially in Adult / Young and Adult / Senior. This mix is somewhat expected, since in these age groups it is more difficult to establish a clear border and as such will ultimately lead to overlaps.

4. Gender Sub-Challenge

In the Gender sub-challenge, the requirement was to detect the gender of the speakers in three separate classes, child, female and male $\{x, f, m\}$. Since this is a much more well behaved task with a smaller number of classes and clearer borders between them one should expect to obtain higher accuracies. Another factor that might lead to better performance of our gender detection system is that we have available a larger training set with gender labels. For this task we used in some of the front-ends the BN ALERT corpus [6], which not only has more speaker variability but also has more diverse audio background conditions. This increased variability may lead to a more robust gender detection system.

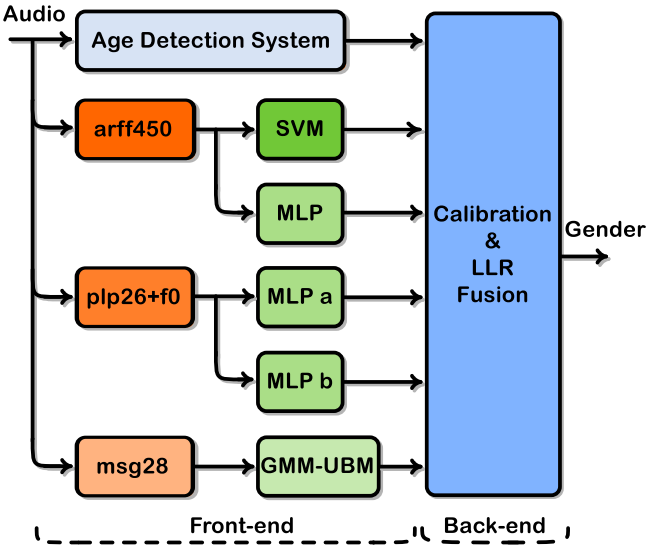

Figure 2: *Gender detection system.*

Our developed approach (Figure 2) for the gender detection system uses several separate front-ends which again take advantage of different features, classification paradigms and different training datasets. The output scores of each of these front-end systems is then calibrated and combined to produce the final system output. The following section describes in detail each of the developed front-end systems.

4.1. Front-Ends

For the gender detection system, six independent front-ends were developed. The first one is the complete age detection system where its seven class output probability scores were converted into the three class gender scores $\{1,\ldots,7\} \rightarrow \{x, f, m\}$ by summing the female age probability scores together and the male age probability scores together. Again, the child class score is directly the score from age class 1.

$$\begin{aligned} p(x) &= p(1) \\ p(f) &= p(2) + p(4) + p(6) \\ p(m) &= p(3) + p(5) + p(7) \end{aligned} \quad (3)$$

The other five developed front-end systems output directly the scores in the required three classes. The second and the third systems used as training data the provided challenge features [5], denoted "arff450", which were obtained in the aGender training set. Two different classification paradigms were used: SVM (linear kernel), and MLP. The later was used without input context since a single frame of features represents the whole audio file [5]. Our experiments used a fully connected feed-forward architecture with two hidden layers of 100 and 50 sigmoidal units, and softmax outputs.

The other three front-ends developed for detecting speakers gender used additional training data besides the aGender corpus. The fourth front-end also used the other "child" corpora (CMU Kids and PF STAR children head-mount and desktop sets) (denoted MLP a in Table 3). The fifth (denoted MLP b in Table 3) and the sixth front-ends used all available training

data that has gender labels, that is, all of the above plus the BN ALERT training, pilot and development sets.

In terms of feature extraction and classification methods, the third and the fourth front-ends extract from the audio every 10ms a frame with 12th order plp [7] coefficients plus energy plus deltas plus pitch (f0). Both use MLP classifiers with seven input context frames and two hidden layers of 350 units.

The sixth front-end extracts from the audio a frame of 28 static msg [9] features. A GMM-UBM with 1024 mixtures is employed. After training the UBM, each of the gender class GMMs was created by performing five iterations of Maximum a Posteriori (MAP) adaptation.

4.2. Calibration and Fusion Back-End

Similar to the age detection system, linear logistic regression fusion and calibration of the six independent front-end systems have been made with the FoCal Multiclass Toolkit [2]. The output scores from this fusion back-end were converted to probability confidence values in order to fulfill the competition results format using the equation 2.

4.3. Results

Table 3 summarizes the results obtained in the aGender development set by the six independent gender front-ends and by their combination using the calibration and llr fusion back-end. Results are expressed in terms of accuracy (% UA which is the challenge measure and % WA [5]).

Table 3: *Gender Results obtained in the aGender Development set.*

Systems $\{x, f, m\}$	% UA	%WA
Age Detection $\{1, \ldots, 7\} \rightarrow \{x, f, m\}$	80.6	89.3
SVM - arff450 - aGender	77.0	86.4
MLP - arff450 - aGender	76.5	86.5
MLP a - plp26+f0 - aGender+child	78.9	89.5
MLP b - plp26+f0 - aGender+child+BN	82.2	88.2
GMM-UBM - msg28 - aGender+child+BN	75.9	84.1
Fusion	83.1	86.9

The best individual front-end combines acoustic with prosodic features (plp26+f0) and uses an expanded training set with additional child and Broadcast News male and female speech data. Fusion of all six systems represent an improvement of 1% absolute over this best single front-end and represents a significant 5.8% over the reported challenge baseline [5]. Our final competition result in the test set is 84.3% UA.

Table 4: *Gender confusion matrix results for the final fusion system obtained in the aGender Development set.*

	x	f	m
x	**70.5**	20.4	9.1
f	14.9	**83.8**	1.3
m	1.7	3.3	**95.0**

The inspection of Table 4 reveals that the biggest misclassifications come from the child (x) and female (f) classes. This is somewhat expected, since the two gender types are more similar than the male gender. In fact, the male class (m) detection has excellent results. We suspect that this is also because the training sets are not balanced and there is more male training data.

5. Conclusions

In this paper we presented the INESC-ID Spoken Language Systems Laboratory (L2F) Age and Gender classification systems that were submitted to the INTERSPEECH 2010 Paralinguistic Challenge. These Age and Gender classification systems are composed respectively by the fusion of four and six individual sub-systems trained with short and long term acoustic and prosodic features, different classification paradigms (GMM-UBM, MLP and SVM) and different speech corpora. The complementary nature of these different approaches boosted their combined performance. The best results obtained by the calibration and linear logistic regression fusion back-end show an absolute improvement of 4.1% on the unweighted accuracy value for the Age task and 5.8% for the Gender task when compared to the competition baseline systems [5].

More important than the results obtained, our participation in the INTERSPEECH 2010 Paralinguistic Challenge resulted in the development of a new age detection system, an area where we had no prior experience, and resulted in the development of a much improved gender detection system. This system together with the experience gained will be relevant to our participation in the European I-DASH project.

6. Acknowledgements

The authors would like to thank Mats Blomberg and Daniel Elenius for letting us use the KTH PF STAR children corpus. This work was partly funded by the European project I-DASH.

7. References

[1] C.-C. Chang, and C-J Lin, "LibSVM - A Library for Support Vector Machines," URL: http://www.csie.ntu.edu.tw/cjlin/libsvm/index.html.

[2] N. Brummer, "FoCal Multiclass Toolkit," URL: http://niko.brummer.googlepages.com/focalmulticlass.

[3] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus," in *Linguistic Data Consortium*, Philadelphia, USA, 1997.

[4] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF STAR Childrens Speech Corpus," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005.

[5] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mueller and S. Narayanan, "The Interspeech 2010 Paralinguistic Challenge," in *Proc. Interspeech 2010*, Makuhari, Japan, 2010.

[6] H. Meinedo, "Audio Pre-Processing and Speech Recognition for Broadcast News," Ph.D. dissertation, IST, Lisbon, Portugal, 2008.

[7] H. Hermansky, and N. Morgan, "RASTA Processing of Speech,", in *IEEE Transactions on Speech and Audio Processing*, Vol. 2(4), pp 578-589, Oct 1994.

[8] R. Martins, I. Trancoso, A. Abad, and H. Meinedo, "Detection of Children's Voices," in *Proceedings of the I Iberian SLTech 2009*, Lisbon, Portugal, 2009.

[9] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust Speech Recognition using the Modulation Spectrogram," in *Speech Comunication*, vol. 25, pp. 117-132, 1998.

[10] C. Mueller, "Automatic Recognition of Speakers' Age and Gender on the basis of Empirical Studies," in *Proceedings Interspeech 2006*, Pittsburgh, USA, 2006.

[11] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting Audio Events for Semantic Video Search," in *Proceedings Interspeech 2009*, Brighton, UK, 2009.