The INTERSPEECH 2010 Paralinguistic Challenge*

Björn Schuller[1], *Stefan Steidl*[2], *Anton Batliner*[2], *Felix Burkhardt*[3], *Laurence Devillers*[1], *Christian Müller*[4], *Shrikanth Narayanan*[5]

[1]CNRS-LIMSI, Spoken Language Processing Group, Orsay, France
[2]FAU Erlangen-Nuremberg, Pattern Recognition Lab, Germany
[3]Deutsche Telekom AG Laboratories, Berlin, Germany
[4]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[5]University of Southern California, SAIL, Los Angeles, CA, USA

{schuller|devil}@limsi.fr, {batliner|steidl}@informatik.uni-erlangen.de,
Felix.Burkhardt@telekom.de, cmueller@dfki.de, shri@sipi.usc.edu

Abstract

Most paralinguistic analysis tasks are lacking agreed-upon evaluation procedures and comparability, in contrast to more 'traditional' disciplines in speech analysis. The INTERSPEECH 2010 Paralinguistic Challenge shall help overcome the usually low compatibility of results, by addressing three selected sub-challenges. In the Age Sub-Challenge, the age of speakers has to be determined in four groups. In the Gender Sub-Challenge, a three-class classification task has to be solved and finally, the Affect Sub-Challenge asks for speakers' interest in ordinal representation. This paper introduces the conditions, the Challenge corpora "aGender" and "TUM AVIC" and standard feature sets that may be used. Further, baseline results are given.

Index Terms: Paralinguistic Challenge, Age, Gender, Affect

1. Introduction

Most paralinguistic analysis tasks resemble each other not only by means of processing and ever-present data sparseness, but by lacking agreed-upon evaluation procedures and comparability, in contrast to more traditional disciplines in speech analysis. At the same time, this is a rapidly emerging field of research, due to the constantly growing interest on applications in the fields of Human-Machine Communication, Human-Robot Communication, and Multimedia Retrieval. In these respects, the INTERSPEECH 2010 Paralinguistic Challenge shall help bridging the gap between excellent research on paralinguistic information in spoken language and low compatibility of results, by addressing three selected tasks. The "aGender" and the "TUM AVIC" corpora are provided by the organizers. The first consists of 46 hours of telephone speech, stemming from 954 speakers, and serves to evaluate features and algorithms for the detection of speaker age and gender. The second features 2 hours of human conversational speech recording (21 subjects), annotated in 5 different levels of interest. The corpus further features a uniquely detailed transcription of spoken content with word boundaries by forced alignment, non-linguistic vocalizations, single annotator tracks, and the sequence of (sub-)speaker-turns. Both are given with distinct definition of train, develop, and test partitions, incorporating speaker independence, as needed in most real-life settings. Benchmark results of the most popular approaches are provided. Three sub-challenges are addressed:

In the *Age Sub-Challenge*, the four 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 female, male, and children.

Finally, the *Affect Sub-Challenge* features the related state of interest in ordinal representation. Thus, regression is used for this task. Here, participants may include linguistic features by incorporating automatic speech recognition. To this end, transcription of the train and development partitions, including non-linguistic vocalizations, are known. Contextual knowledge may be used, as the sequence of 'sub-speaker-turns' is known.

All Sub-Challenges allow contributors to find their own features with their own classification algorithm. However, a standard feature set is given for each corpus that may be used. Participants have to stick to the definition of train, develop, and test partitions. They may report on results obtained on the develop partition, but have only two trials to upload their results on the test partitions, whose labels are unknown to them. The use of well-known and obtainable further language resources, e. g. for speech recognition, is permitted.

In the following we introduce the Challenge corpora (Section 2), features (Section 3), and baselines (Section 4) before concluding (Section 5).

2. Challenge Corpora

2.1. aGender

In the *Age* and *Gender Sub-Challenge* the "aGender" corpus serves for analyses and comparison [1].

An external company was employed to identify possible speakers of the targeted age and gender groups. The subjects received written instructions on the procedure and a financial reward. They were asked to ring up the recording system six times. Each time they were prompted by an automated Interactive Voice Response system to repeat given utterances or produce free content. The speakers obtained individual prompt sheets containing the utterances and additional instructions. Between each session a break of one day was scheduled to ensure more variations of the voices. Each subject's six calls had to be done with a mobile phone alternating indoor and outdoor to obtain

* The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE). The authors would further like to thank the sponsors of the challenge, the HUMAINE Association and Deutsche Telekom Laboratories. The responsibility lies with the authors.

different recording environments. The caller is connected by mobile network or ISDN and PBX to the recording system, which consists of an application server hosting the recording application and a VoiceXML telephony server (Genesys Voice Platform). The utterances were stored on the application server as 8 bit, 8 kHz, A-law. To validate the data, the associated age cluster was compared with a manual transcription of the self stated date of birth.

Four age groups – CHILD, YOUTH, ADULT, and SENIOR – were defined. The choice was not motivated by any physiological aspects that might arise from the development of the human voice with increasing age, but solely on market aspects stemming from the application as dialog control in call centers. Since children are not subdivided into female and male, this results in seven classes as shown in Table 1.

Table 1: *Age and gender classes of the aGender corpus, where f and m abbreviate female and male, and x represents children w/o gender discrimination. The last two columns represent the number of speakers/instances per partition (Train and Develop).*

class	group	age	gender	# Train	# Develop
1	CHILD	7-14	x	68 / 4 406	38 / 2 396
2	YOUTH	15–24	f	63 / 4 638	36 / 2 722
3	YOUTH	15–24	m	55 / 4 019	33 / 2 170
4	ADULT	25–54	f	69 / 4 573	44 / 3 361
5	ADULT	25–54	m	66 / 4 417	41 / 2 512
6	SENIOR	55–80	f	72 / 4 924	51 / 3 561
7	SENIOR	55–80	m	78 / 5 549	56 / 3 826

Note that the given age in years might differ one year due to birthdays near the date of speech collection. Also there are six cases where youth stated an incorrect age. Nonetheless, the (external) speaker recruiter assured that these n speakers indeed are youth.

The following requirements were communicated to the company assigned with the speaker recruitment: at least 100 German speakers for each class acquired from all German Federal States without perfect balance of German dialects. Multiple speakers from one household were allowed. The ability to read the given phrases was a precondition for the participation of children. As further minimum requirement we defined age sub-clusters of equal size: to account for the different age intervals of the groups, CHILDren and YOUTH should be uniformly distributed within 2 year clusters and ADULTs and SENIORs in 5 year clusters. This means, for example, that 25 children from seven to eight years and 20 young-aged females between 17 to 18 years should participate. All age groups, including the CHILDren, should have equal gender distribution.

The content of the database was designed in the style of the Speech Dat corpora. Each of the six recording sessions contained 18 utterances taken from a set of utterances listed in detail in [1]. The topics of these were *command words*, *embedded commands*, *month*, *week day*, *relative time description*, *public holiday*, *birth date*, *time*, *date*, *telephone number*, *postal code*, *first name*, *last name*, *yes/no* with according free or preset inventory and according 'eliciting' questions as *"Please tell us any date, for example the birthday of a family member."*.

On an accompanied instruction sheet all items, relevant for the specific recording session, were listed. Within the set of the preset words it was taken care of that the content for each speaker did not recur.

In total, 47 hours of speech in 65 364 single utterances of 954 speakers were collected, Note that not all volunteers completed all six calls, and there were cases where some called more often than six times, resulting in different numbers of utterances per speaker. The mean utterance length was 2.58 sec.

We selected randomly for each of the seven classes 25 speakers as a fixed Test partition (17 332 utterances, 12.45 hours) and the other 770 speakers as a Training partition (53 076 utterances, 38.16 hours), which was further subdivided into Train (32 527 utterances in 23.43 hours of speech of 471 speakers) and Develop (20 549 utterances in 14.73 hours of speech of 299 speakers) partitions. Overall, this random speaker-based partitioning results in roughly 40 %/30 %/30 % Train/Develop/Test distribution.

Table 1 lists the number of speakers and the number of utterances per class in the Train and Develop partitions, Figure 1 depicts the number of speakers as a histogram over their age.

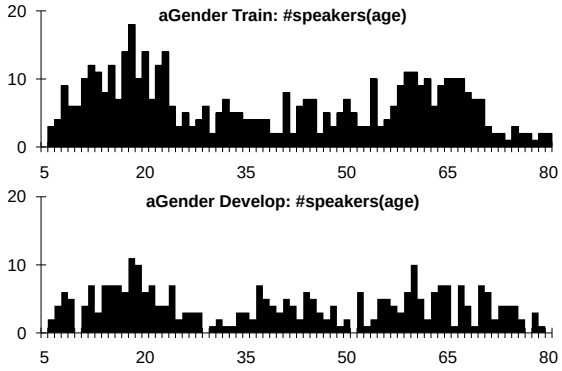

Figure 1: Age (in years) histograms for the Train and Develop partitions of aGender.

Decisive for the *Age Sub-Challenge* is the age group $\{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$ by classification as indicated in Table 1, not the age in years. The age group can be handled either as combined age/gender task by classes $\{1, \ldots, 7\}$ or as age group task independent of gender by classes $\{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$. For the official comparison of results though, only the age group information is used for the competition in the *Age Sub-Challenge* by mapping $\{1, \ldots, 7\} \rightarrow \{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$ as denoted. For the *Gender Sub-Challenge* the classes $\{f, m, x\}$ have to be classified, as gender discrimination of children is considerably difficult, yet we decided to keep all instances for both tasks.

2.2. TUM AVIC

For the *Affect Sub-Challenge* we selected the Audiovisual Interest Corpus recorded at the Technische Universität München ("TUM AVIC") as described in detail in [2]. In the scenario setup, an experimenter and a subject are sitting on opposite sides of a desk. The experimenter plays the role of a product presenter and leads the subject through a commercial presentation. The subject's role is to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and actively interact with the experimenter considering his/her interest in the addressed topics. The subject was explicitly asked not to worry about being polite to the experimenter, e. g. by always showing a certain level of 'polite' attention. Visual and voice data was recorded by a camera and two microphones, one headset and one far-field microphone, in this situation. In the Challenge the lapel microphone recordings at 44.1 kHz, 16 bit are used. 21 subjects took part in the recordings, three of them Asian, the remaining European. The language throughout ex-

periments is English, and all subjects are non-native, yet very experienced English speakers. More details on the subjects are summarized in Table 2.

Table 2: *Details on subjects contained in the TUM AVIC database. Further details in the text.*

Group	# subjects	mean age	rec. time [h]
All	21	29.9	10:22:30
Male	11	29.7	5:14:30
Female	10	30.1	5:08:00
Age <30	11	23.4	5:13:10
Age 30–40	7	32.0	3:37:50
Age >40	3	47.7	1:31:30

To acquire reliable labels of a subject's 'Level of Interest' (LOI), the entire video material was segmented into speaker- and sub-speaker-turns and subsequently labeled by four male annotators, independently from each other. The annotators were undergraduate students of psychology. The intention was to annotate observed interest in the common sense. A speaker-turn is defined as continuous speech segment produced solely by one speaker – back channel interjections ("*mhm*", etc.) are ignored, i. e. every time there is a speaker change, a new speaker turn begins. This is in accordance with the common understanding of the term 'turn-taking'. Speaker-turns thus can contain multiple and especially long sentences. In order to provide Level of Interest analysis on a finer time scale, the speaker turns were further segmented at grammatical phrase boundaries: a turn lasting longer than 2 seconds is split by punctuation and syntactical and grammatical rules, until each segment is shorter than 2 seconds. These resulting segments are referred to as sub-speaker-turns.

The LOI is annotated for every such sub-speaker turn. In order to get an impression of a subject's character and behavior prior to the actual annotation, the annotators had to watch approximately 5 minutes of a subject's video. As the focus of interest based annotation lies on the sub-speaker turn, each of those had to be viewed at least once to find out the LOI displayed by the subject. Five Levels of Interest were distinguished:

LOI-2 – *Disinterest* (subject is tired of listening and talking about the topic, is totally passive, and does not follow)

LOI-1 – *Indifference* (subject is passive, does not give much feedback to the experimenter's explanations, and asks unmotivated questions, if any)

LOI0 – *Neutrality* (subject follows and participates in the discourse; it cannot be recognized, if she/he is interested or indifferent in the topic)

LOI+1 – *Interest* (subject wants to discuss the topic, closely follows the explanations, and asks questions)

LOI+2 – *Curiosity* (strong wish of the subject to talk and learn more about the topic).

Additionally, the spoken content has been transcribed, and *long pause*, *short pause*, and non-linguistic vocalizations have been labeled. These vocalizations are *breathing* (452), *consent* (325), *hesitation* (1 147), *laughter* (261), and *coughing, other human noise* (716). There is a total of 18 581 spoken words, and 23 084 word-like units including 2 901 non-linguistic vocalizations (19.5%). In summary, the overall annotation contains per sub-speaker-turn segment spoken content, non-linguistic vocalizations, individual annotator tracks, and mean LOI.

For the Challenge, ground truth is established by shifting to a continuous scale obtained by averaging the single annotator LOI. The histogram for this mean LOI is depicted in figure 2. As can be seen, the subjects still seemed to have been somewhat polite as almost no negative average LOI is found. Note that here the original LOI scale reaching from LOI-2 to LOI+2 is mapped to $[-1, 1]$ by division by 2 in accordance with the scaling adopted in other corpora, e. g. [3]. Apart from higher precision, this representation form allows for subtraction of a subject's long-term interest profile. Note that the Level of Interest introduced herein is highly correlated to arousal. However, at the same time there is an obvious strong correlation to valence, as e.g. boredom has a negative valence, while strong interest is characterized by positive valence. The annotators however labeled interest in the common sense, thus comprising both aspects.

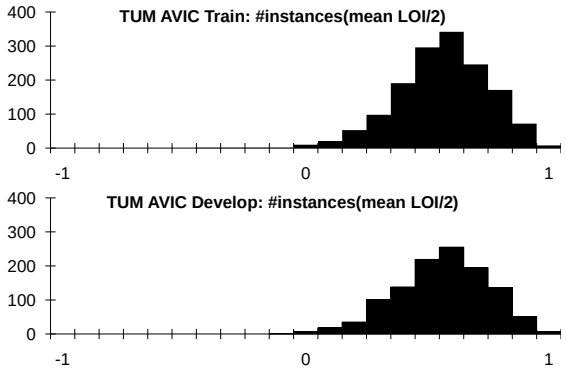

Figure 2: Mean Level of Interest (LOI, divided by 2) histograms for the Train and Develop partitions of TUM AVIC.

As before, we partitioned the 21 speakers (and 3 880 sub-speaker-turns) speaker-independently in the best achievable balance with priority on gender, next age, and then ethnicity into three partitions for Train (1 512 sub-speaker-turns in 51:44 minutes of speech of 4 female, 4 male speakers), Develop (1 161 sub-speaker-turns in 43:07 minutes of speech of 3 female, 3 male speakers), and Test (1 207 sub-speaker-turns in 42:44 minutes of speech of 3 female, 4 male speakers).

3. Challenge Features

In this Challenge an extended set of features compared to the INTERSPEECH 2009 Emotion Challenge [4] is given to the participants, again choosing the open-source Emotion and Affect Recognition toolkit's feature extracting backend openSMILE [5]. This extension intends to better reflect a broader coverage of paralinguistic information assessment [6, 7].

1 582 acoustic features are obtained in total by systematic 'brute-force' feature (over)generation in three steps: first, the 38 low-level descriptors shown in Table 3 are extracted at 100 frames per second with varying window type and size (Hamming, 25 ms for all but pitch with Gaussian, 60 ms) and smoothed by simple moving average low-pass filtering with a window length of 3 frames. Next, their first order regression coefficients are added in full HTK compliance. Then, 21 functionals are applied (cf. Table 3) per instance in the databases. However, 16 zero-information features (e. g. minimum F0, which is always zero) are discarded. Finally, the 2 single features F0 number of onsets and turn duration are added.

Due to the size of the aGender corpus, a limited set is provided consisting of 450 features (missing descriptors and functionals are marked by $^-$ in Table 3). This is reached by reducing the number of descriptors from 38 to 29 and that of functionals

from 21 to 8. However, the configuration file to extract the same features as for TUM AVIC with openSMILE is provided.

Table 3: *Provided feature sets: 38 low-level descriptors with regression coefficients, 21 functionals. Details in the text. A $^-$ indicates those only used for the TUM AVIC baseline. Abbreviations: DDP: difference of difference of periods, LSP: line spectral pairs, Q/A: quadratic, absolute.*

Descriptors	**Functionals**
PCM loudness$^-$	position$^-$ max./min.
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0–7]$^-$	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff.$^-$ 1/2
F0 by Sub-Harmonic Sum.	lin. regression error Q/A$^-$
F0 Envelope	quartile$^-$ 1/2/3
Voicing Probability	quartile range$^-$ 2–1/3–2/3–1
Jitter local	percentile 1/99
Jitter DDP	percentile range 99–1
Shimmer local	up-level time$^-$ 75/90

4. Challenge Baselines

For the baselines we exclusively exploit acoustic feature information. Linguistic information can be used for the *Affect Sub-Challenge*, yet, the word level transcription is given exclusively for the Train and Develop partitions of TUM AVIC, as the Challenge aims at 'real-life' conditions as if for a running system 'in the wild' [8]. Spoken content of the Test partition thus needs to be recognized by automatic speech recognition rather than using perfect transcription – in the end even recognition of affective speech may be a challenge . However, for evaluation of best suited textual analysis methods for interest determination the Develop partition providing perfect transcription may be used. For transparency and easy reproducibility reasons we use the WEKA data mining toolkit for classification and regression [9].

Table 4: *Age and Gender Sub-Challenge baseline results by Sequential Minimum Optimization learned pairwise Support Vector Machines with linear Kernel.*

Sub-Ch.	**Task**	**% UA**	% WA
Train vs. Develop			
–	$\{1,\ldots,7\}$	44.24	44.40
Age	$\{1,\ldots,7\} \rightarrow \{\mathrm{C,Y,A,S}\}$	47.11	46.17
	$\{\mathrm{C,Y,A,S}\}$	46.22	45.85
Gender	$\{1,\ldots,7\} \rightarrow \{x,f,m\}$	77.28	84.60
	$\{x,f,m\}$	76.99	86.76
Train + Develop vs. Test			
–	$\{1,\ldots,7\}$	44.94	45.60
Age	$\{1,\ldots,7\} \rightarrow \{\mathrm{C,Y,A,S}\}$	48.83	46.71
	$\{\mathrm{C,Y,A,S}\}$	**48.91**	46.24
Gender	$\{1,\ldots,7\} \rightarrow \{x,f,m\}$	**81.21**	84.81
	$\{x,f,m\}$	80.42	86.26

Table 4 shows results for the *Age* and *Gender Sub-Challenge* by unweighted and weighted accuracy on average per class (UA/WA, weighting with respect to number of instances per class). As the distribution among classes is not balanced, the competition measure is UA. Visibly, the 'blind' Test partition shows better results, likely due to the now larger training set.

Table 5: *Affect Sub-Challenge baseline results. Unpruned REPTrees (25 cycles) in Random-Sub-Space meta-learning (500 Iterations, sub-space size 5 %).*

Sub-Challenge	**CC**	MLE
Train vs. Develop		
Affect	0.604	0.118
Train + Develop vs. Test		
Affect	**0.421**	0.146

Table 5 depicts the results for the *Affect Sub-Challenge* baseline. The measures for this task are cross correlation (CC) and mean linear error (MLE) as found in other studies (e. g. [3]), whereby CC is the primary measure. Here, a clear downgrade is observed for the apparently more 'challenging' Test condition.

5. Conclusions

This INTERSPEECH 2010 Paralinguistic Challenge represents an extension of last year's Challenge [4] by broadening the scope, including several different tasks such as age, gender, and level of interest. Again, we attach importance to both realism of the tasks and strict comparability of results. Hopefully, this is a next step towards defining and using standards within the field of paralinguistics, and a first step towards integrating different aspects of paralinguistics within the same or similar frames of modeling – a longer-range objective being a combined modeling of, e. g., static speaker characteristics and dynamic speaker states.

6. References

[1] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A Database of Age and Gender Annotated Telephone Speech," in *Proc. 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010.

[2] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, pp. 1760–1774, 2009.

[3] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, Hannover, Germany, 2008, pp. 865–868.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 312–315.

[5] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. ACII*, Amsterdam, Netherlands, 2009, pp. 576–581.

[6] C. Ishi, H. Ishiguro, and N. Hagita, "Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction," in *Proc. of Speech Prosody 2006*, Dresden, 2006, pp. 883–886.

[7] C. Müller, "Classifying speakers according to age and gender," in *Speaker Classification II*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Heidelberg - New York - Berlin: Springer, 2007, vol. 4343.

[8] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP)*, Pittsburgh, PA, 2006, pp. 801–804.

[9] I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques, 2nd Edition*. San Francisco: Morgan Kaufmann, 2005.