Intelligent Systems' Holistic Evolving Analysis of Real-Life Universal Speaker Characteristics

Björn Schuller[1,2], Yue Zhang[2], Florian Eyben[2], Felix Weninger[2]

[1]Department of Computing, Imperial College London, London, U. K.
[2]Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Munich, Germany
{schuller,y.zhang,eyben,weninger}@tum.de

Abstract

In this position paper we present the FP7 ERC starting grant project iHEARu (Intelligent systems' Holistic Evolving Analysis of Real-life Universal speaker characteristics). This project addresses several fundamental shortcomings in state of the art methods for computational paralinguistics, by introducing holistic analysis, evolving learning of features and models, and collection of real-life, large-scale data annotated in multiple dimensions ('universally'). We discuss the first aspect of the project, holistic analysis, in more detail, and give benchmark results using state of the art multi-target learning methods on the INTERSPEECH 2012 Speaker Trait Challenge dataset (Likability Sub-Challenge). The results clearly indicate the need for improved machine learning methods and data collection to learn holistic speaker classification.

Keywords: Computational Paralinguistics, Holistic Analysis, Multi-target Learning

1. Introduction

With recent technology advances, automatic speech recognition and synthesis have matured to the degree that they are used on a daily basis by millions of people, e.g., on their smart phones or in call services. During the next years, it is expected that speech processing technology will move to a new level of social awareness to make interaction more intuitive, speech retrieval more efficient, and lend additional competence to computer-mediated communication and speech analysis services in the commerce, health, security, and further sectors. To reach this goal, rich speaker traits and states such as age, height, personality and physical and mental states as carried by the tone of the voice and the spoken words must be reliably identified by machines.

The **iHEARu** project aims to push the limits of **i**ntelligent systems for computational paralinguistics by considering **H**olistic analysis of multiple speaker attributes at once, **E**volving and self-learning, deeper **A**nalysis of acoustic parameters - all on **R**ealistic data on a large scale, ultimately progressing from individual analysis tasks towards **u**niversal speaker characteristics analysis, which can be easily learnt about and can be adapted to new, previously unexplored characteristics.

In this paper, the state of the art in the field is described in Section 2. Next, we will introduce our long-term goals and describe the methodologies of the iHEARu project in Section 3. An in-depth discussion of holistic analysis of multiple speaker attributes is given in Section 4. Further, a first attempt on multi-target classification to improve on three paralinguistics tasks by jointly learning age, gender, and subjective likability of the voice, is presented and evaluated in Sections 5 and 6. We conclude with a summary and outlook on future research topics in Section 7.

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu) and No. 289021 (STREP ASC-Inclusion).

2. State of the Art

Analysing 'the voice behind the words' has been an active topic in many fields of research for more than two decades now (Wu and Childers, 1991; Cowie et al., 2001; Schuller and Batliner, 2014). Early studies have emerged from research in phonetics and automatic speech recognition (ASR), and have focussed on simple characteristics such as gender (Wu and Childers, 1991). Research on recognizing human emotion from speech started at the beginning of this century (Cowie et al., 2001). As a matter of fact, the related paradigm of 'affective computing', that focusses on emotional aspects of natural human-machine interaction, has driven speech technology research throughout the last decade. In the recent years, a new major field of speech recognition research investigating the speaker characteristics beyond affective states is evolving: 'computational paralinguistics' (Schuller and Batliner, 2014). Research in this field has delivered highly promising results and tools for the community including the first widely used open-source affect analysis toolkit openEAR (Eyben et al., 2009) and its large-scale acoustic feature extractor openSMILE (Eyben et al., 2013) which both have become standard tools and references in the field. Furthermore, researchers from all over the world have reviewed their speech analysis systems in the light of the INTERSPEECH Challenges that have targeted a multitude of tasks such as emotion (2009), interest, age and gender (2010), sleepiness and alcohol intoxication (2011), as well as the OCEAN five personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism), voice pathology and likability (2012), emotion, autism, and social signals (2013), and cognitive and physical load (2014). An overview of the evaluation campaigns up to 2012 is given in (Schuller, 2012).

From a methodological point of view, today's speaker characteristics recognition mostly relies on standard machine learning techniques that have been proven successful for various audio recognition tasks including speech and

speaker recognition. Most established techniques are static modelling with Support Vector Machines (SVMs) and dynamic modelling with Hidden Markov Models (HMMs). Generally, one starts with standard low-level descriptors (LLDs) such as (Mel-frequency) spectrum, Cepstrum, pitch, or voicing probability, extracted from short overlapping frames of fixed length. Static modelling is then performed by computing statistics of the LLD contours. Combining static modelling of utterances with context knowledge, Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) have successfully been introduced for affect recognition (Eyben et al., 2010). As a more recent approach to machine learning from unsupervisedly generated features, Deep Belief Networks (DBNs) have been applied to affect and likability recognition (Stuhlsatz et al., 2011; Brueckner and Schuller, 2012; Le and Mower, 2013). Despite the manifold work done for a plethora of speaker characteristics, the methodology has converged to a degree of standardisation, and major breakthroughs have been lacking in the past years. For many studies, it remains largely unclear to what extent their findings can be transferred to actual systems 'in the wild', for reasons outlined below.

Most importantly, today's studies consider speaker characteristics in isolation, i. e., single or only few speaker characteristics are considered at once (cf. Figure 1). There is very little exploitation of the interplay and synergies between different characteristics, yet in reality, strong interdependencies between bits of paralinguistic information exist. For example, it is intuitively clear that acoustic models for gender classification (male vs. female) should be different by age, since arguably the most important feature, pitch, is also influenced by age. Still, before interdependencies can be exploited in a more generic fashion, i.e., be learnt from data, richly annotated data sets will have to be created: at present, databases provide labels for only one or a few speaker characteristics at the same time. Another significant limitation of today's systems can be seen in their usage of acoustic features. These are mostly chosen ad-hoc because 'they seem to work well', and are often simply borrowed from neighbouring disciplines in audio processing such as ASR, instead of being tailored to the modelling of speaker characteristics. Apart from features, the limited transferability of most of today's studies to real-life applications is a more generic issue. First of all, this is because they are mostly carried out on hand-segmented, often manually transcribed utterances recorded from noise-free channels or in the presence of artificial noise and reverberation, and often prompted speech. To cope with real-life conditions in retrieval applications, however, robust single-channel automatic speech detection, segmentation and enhancement of spontaneous utterances in real acoustic environments, transmitted over arbitrary channels, must be addressed. Furthermore, all but a very few studies overlook the issue of potential malicious system use, such as faking of age, alcohol intoxication, or affective states; in fact, this phenomenon has only lately received some attention in speaker verification (Alegre et al., 2013). Finally, meaningful confidence measures (i. e., beyond simple posterior probabilities or distances in the feature space) have only been attempted recently (Deng and Schuller, 2012) despite them being crucial for real-life applications such as retrieval, dialogue systems and computer-mediated human-to-human conversation.

All these shortcomings are the starting point for the research envisioned in the iHEARu project.

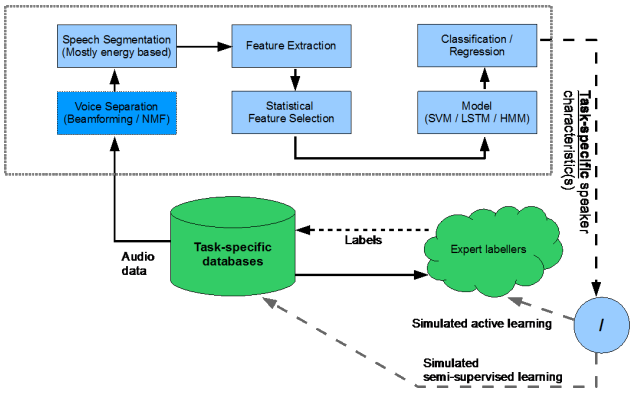

Figure 1: State of the art method for recognition of individual speaker characteristics. A standard machine learning pipeline is applied, consisting of pre-processing (voice separation and segmentation), feature extraction and selection, and classification/regression. Labels for (rather) small task specific databases are supplied by expert labellers. Simulated active and simulated semi-supervised learning are only considered by omitting labels from those expert labelled databases.

3. Methodology

To realise its ambitious goals, the iHEARu project aims to leverage novel techniques for multi-target (multi-task), semi-supervised and unsupervised learning. It is envisioned to overcome today's sparseness of annotated realistic speech data by large-scale speech and meta-data mining from public sources such as social media, crowdsourcing for labelling and quality control, and shared semi-automatic annotation. Furthermore, by utilising feedback, deep, and evolutionary learning methods, all stages from pre-processing and feature extraction to the statistical modelling can be subject to 'life-long learning' according to new data. Finally, human-in-the-loop system validation and novel perception studies are expected to help understanding both of system behaviour and human interpretation in a large variety of speaker classification tasks.

3.1. Holistic Processing

The iHEARu project intends to advance the state of the art by investigating novel methodology for holistic analysis of established speaker attributes, such as age and gender, in conjunction with currently under-researched characteristics, such as speech in different physiological and mental states. Large-scale speech and meta data mining from public sources (e.g., social media), combined with semi-automatic annotation methods (e.g., active learning) will be an essential means for building large, realistic, richly annotated and transcribed data sets.

3.2. Evolving "Life-Long" Learning for Self-Improvement

Self-learning and self-improvement in the iHEARu project will not be limited to iterative data collection. Rather, iHEARu will consider self-optimising feature extraction and self-organising classifiers: The whole process of speaker characteristics learning and analysis shall be self-optimising, as depicted in the flow chart above. For realising these ambitious goals, deep learning (Hinton et al., 2012) combined with neuroevolutionary methods and non-parametric Bayesian learning will play an essential role. This provides promising means for creating self-optimising statistical models and hierarchical input representations with very little amount of supervision.

3.3. Analysis with Deeper Understanding and Context-Dependent Speech Features

The iHEARu project approaches the acoustic feature generation and selection issue by trying to understand human reasoning in challenging conditions, from very low SNR, application of voice conversion algorithms, and speech compression, all the way to deliberate faking of voice or speaker states by the subjects. As a consequence, the iHEARu project will not only address environmental (technical) robustness, but more importantly also robustness against fraud.

3.4. Real-Life

To automatically obtain robust speech detection and segmentation into meaningful units, the iHEARu project aims to improve all of the pre-processing algorithms including speech separation, noise reduction, voice activity detection, and segmentation in a loop with the subsequent analysis algorithms and the confidence scores given by these (cf. Fig. 2). Further, dealing with real-life data also means coping with various transmission channels.

3.5. Universal Analysis

The iHEARu project addresses the automatic recognition of speaker attributes and speaking styles that can be clearly identified by humans. However, the iHEARu approach to universal analysis is not to simply define more and more new recognition tasks that are chosen 'ad hoc'; conversely, it is aimed at developing data-driven methods for a framework which is able to automatically identify characteristics of interest by looking at crowd-sourced resources, such as tag collections, opinions in textual comments, or explicitly collected annotations from paid click-workers.

4. Holistic Speaker Analysis with Multi-Task Learning

Integrating the concept of holistic analysis into automatic systems demands enhanced machine learning methods for context-aware learning. The first step toward a holistic analysis of speaker attributes is to consider multiple speaker attributes simultaneously and jointly in existing learning methods. One encounters many terms and buzz-words in this respect in the literature, which all refer to different concepts: multi-class, multi-label, multi-target, multi-task, multi-instance, and others. Therefore, it is important to

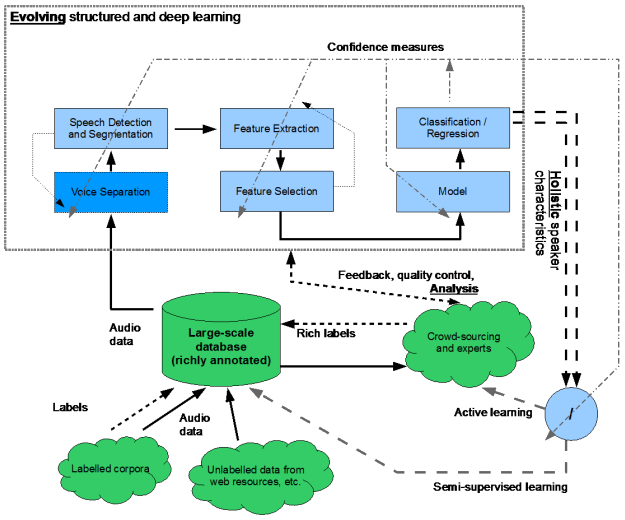

Figure 2: Flowchart of the proposed concept for holistic evolving analysis of realistic universal speaker characteristics. A large-scale collection of richly annotated data is created and extended by semi-supervised and active learning. Confidence measures of system components as well as humans in the loop are used to give feedback to components in the processing chain in order to implement evolving holistic learning.

first clarify the definitions of these terms at this point. Traditional *single-label* or *single-target* learning is concerned with learning from examples, where each example is associated with a single label l from a set of disjoint labels $L, |L| > 1$. For $|L| = 2$, the learning problem is called a binary classification problem or filtering in the case of textual and web data, while for $|L| > 2$, it is referred to as a multi-class problem (Tsoumakas and Katakis, 2007; Madjarov et al., 2012). In contrast, *multi-label learning* is concerned with learning from examples, where each training example is associated with zero, one, or more labels taken form a finite set of labels $Y \subseteq L$ (Zhang and Zhou, 2013).

During the past decade, the multi-label problem has received significant attention due to its wide variety of applications including text categorization, automatic annotation for multimedia contents (e.g., images, music, video), bioinformatics, web and rule mining, information retrieval, tag recommendation, etc. (Zhang and Zhou, 2013). Tsoumakas and Katakis (Tsoumakas and Katakis, 2007) were the first to group the multi-label learning approaches into two main categories: a) problem transformation methods, and b) algorithm adaption methods. The problem transformation methods refer to methods which transform the multi-label classification problem into either one or more single-label classification problems, for which there exits a plethora of machine learning algorithms. The algorithm adaption methods refer to multi-label methods where an existing machine learning algorithm is adapted, extended and customised in order to handle multi-label data directly. Furthermore, besides these two categories of methods for multi-label learning, Madjarov et al. (Madjarov et al., 2012) have introduced a third category: en-

semble methods. The most well known problem transformation ensemble methods are the RAKEL system by Tsoumakas et al. (Tsoumakas and Vlahavas, 2007), ensembles of pruned sets (EPS) (Read et al., 2008) and ensembles of classifier chains (Read et al., 2011) (ECC). The ECC method iteratively trains a multi-target classifier (or regressor) $(y_1, \ldots, y_{|L|}) = h(\mathbf{x})$, where $\mathbf{x}$ is a feature vector. For $l = 1, \ldots, |L|$, a single-target base classifier $y_l = h_l(\mathbf{x}, y_1, \ldots, y_{l-1})$ is trained, i.e., the estimates of the other targets are included as features. Since the order of labels clearly affects the results, bagging is performed to create an ensemble of classifiers using different label orders (and instance weights). An advantage of ECC over multi-task methods based on regularization (Evgeniou and Pontil, 2004), which presumes task similarity, is that not only correlations among labels but also correlations of labels with label-feature combinations can be effectively exploited, and that the method does not saturate with large amounts of training data (Read, 2010).

In a broad sense, multi-label learning can be regarded as a special case of *multi-target* learning, i. e., multi-dimensional learning. In multi-target learning, an example (a data instance) is associated with more than one target variable (as opposed to single-target learning, where only one target value is associated). Each target variable can take multiple numeric (regression) or nominal values (discrete classes). The multi-label case can now be seen as a special case of multi-target learning, where all target variables are binary and each target variable corresponds to a label being present or not.

Multi-target learning is often also referred to as multi-task learning. Besides learning multiple tasks/targets in parallel, information of related tasks is used as an inductive bias to improve the generalization performance of other tasks (Caruana, 1997).

Going back to multi-label learning, the differences between multi-label and multi-task learning are not conceptually based, but given by the different nature of the problems and use-cases addressed. Thus, in multi-label learning often a large space of labels is handled while in standard multi-task or multi-target learning a small set of labels is handled. For the holistic analysis in the iHEARu project both methods will be considered and investigated. Given the fact that they are closely related might result in novel, beneficial combinations of algorithms from both areas (Mencía, 2010).

Completely different from the problems of multi-label and multi-task learning, is *multi-instance learning*, where label sparseness is the core issue: for a bag of multiple instances, only one label exists for the whole bag and information on labels for the individual instances is lacking (Maron and Lozano-Pérez, 1998). In the most primitive case the label is only a binary label (positive and negative instances) and positively labelled bags have to contain at least one instance with a positive label, and negatively labelled bags contain only instances with negative labels (Maron and Lozano-Pérez, 1998). In the context of computational paralinguistics, potential applications of multi-instance learning can be found, e.g., in emotion detection: For example, if a speaker displays negative emotion, this usually affects a few short-time observations, while the remaining observations are similar to a 'neutral' state; in turn, manual annotation of each short-time observation is too cumbersome to perform on a large scale, in contrast to labelling whole utterances.

Table 1: *Partitioning of Speaker Likability Database (L: likable / NL: non-likable); Age (Y: young / A: adult / O: old); Gender (M: male / F: female)*

Task	SLD #	Train	Devel	Test	Σ
Likability	L	189	94	117	400
	NL	205	84	111	400
Age	Y	116	47	70	233
	A	131	58	76	265
	O	147	73	82	302
Gender	M	195	89	113	397
	F	199	89	115	403

5. Experimental Setup

5.1. Selected Database

This section introduces multi-task learning experiments for the joint classification of speaker age, gender, and the average subjective likability of the speaker's voice by others. For that purpose, we use the database of the *Likability Sub-Challenge* of the INTERSPEECH 2012 Speaker Trait Challenge and perform multi-task learning with the MEKA toolkit, which is an extension to the WEKA machine learning framework by adding support for multi-label and multi-target classification (Hall et al., 2009).

In the *Likability Sub-Challenge*, the "Speaker Likability Database" (SLD) was used (Burkhardt et al., 2011). The SLD is a subset of the German Agender database (Burkhardt et al., 2010), which was originally recorded to study automatic age and gender recognition from telephone speech. The speech is recorded over fixed and mobile telephone lines at a sample rate of 8 kHz. The database contains 18 utterance types taken from a set listed in detail in (Burkhardt et al., 2010). An age and gender balanced set of 800 speakers is selected. While the annotation provides likability in multiple levels, the classification task is binarised into 'likable' (L) and 'non-likable' (NL). The data are partitioned into a training, development, and test exactly as in the INTERSPEECH 2012 Speaker Trait Challenge (cf. Table 1).

5.2. Feature Extraction

The acoustic feature set used in this experiment corresponds to the baseline feature set of the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012). The open-source openSMILE feature extractor is used (Eyben et al., 2013) to 'brute-force' a high-dimensional feature set by applying statistical functionals to frame-wise LLDs, which comprise energy, spectral and voicing related low-level descriptors (LLDs). The chosen set of LLDs is shown in Table 2. Regarding functionals, we aim at a compromise between a broad variety of functionals, and careful selection so as not to include meaningless features, such as the arithmetic mean of delta coefficients, which is expected to be zero. The set of applied functionals is given in detail

Table 2: *64 provided low-level descriptors (LLD).*

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
54 spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0–8 kHz)
MFCC 1–14
Spectral energy 250–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity
6 voicing related LLD
F0 by SHS + Viterbi smoothing, Probability of voicing
logarithmic HNR, Jitter (local, delta), Shimmer (local)

Table 3: *Applied functionals.* [1]*: arithmetic mean of LLD / positive* Δ *LLD.* [2]*: only applied to voice related LLD.* [3]*: not applied to voice related LLD except F0.* [4]*: only applied to F0.*

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
1 % percentile ($\approx$ min), 99 % percentile ($\approx$ max)
position of min / max
percentile range 1 %–99 %
arithmetic mean[1], root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above / below 25 / 50 / 75 / 90% range
rel. duration LLD is rising / falling
rel. duration LLD has positive / negative curvature[2]
gain of linear prediction (LP), LP Coefficients 1–5
mean, max, min, std. dev. of segment length[3]
Functionals applied to LLD only
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks – arithmetic mean
mean / std.dev. of rising / falling slopes
mean / std.dev. of inter maxima distances
amplitude mean of maxima / minima
amplitude range of maxima
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames[4]

in Table 3. Altogether, the 2012 Speaker Trait Challenge feature set contains 6 125 features, which is roughly a 40 % increase over previous year's feature set.

5.3. Single- and Multi-Target Learning

To assess the potential of multi-target learning, we compare the following learning schemes, all of which can be found in MEKA.

- Single-target learning (ST), i.e., independent training of single-target classifiers – linear support vector machines (SVMs) trained by sequential minimal optimization (SMO) are chosen;
- Multi-target learning by the ECC method, using SMO-trained SVMs as the base classifier;
- Multi-target learning by the Ensembles of Class Relevance (ECR) method, using SMO-trained SVMs as the base classifier – this corresponds to bagging of single-target SVM classifiers;
- 'Oracle' multi-target learning with SMO-trained SVMs (OMT), where each single-task classifier uses the correct labels for the other tasks as features.

Table 4: Classification results for likability, age, and gender targets for single target classification (ST), multi-target classification by Ensembles of Classifier Chains (ECC) or Class Relevance (ECR), and "oracle" single target classification with the other two labels included as features (OMT). SVM with SMO training, complexity C optimised on the development set between 0.0001 and 1.0.

UAR [%]	**ST**	**ECC**	**ECR**	**OMT**
Development set				
Likability	58.9	55.4	54.9	**60.0**
Age	49.7	**51.9**	**51.9**	50.2
Gender	94.4	94.9	94.9	**95.5**
Test set				
Likability	**58.1**	52.8	57.5	57.3
Age	**46.9**	46.0	45.3	**46.9**
Gender	**96.9**	**96.9**	96.0	96.9

In contrast to ECC, ECR breaks down the multi-target learning problem by considering each l independently, i.e., $y_l = h_l(\mathbf{x})$. However, in contrast to ST, an ensemble of classifiers is trained with different instance weights (bagging). Finally, the OMT method can be written as $y_l = h_l(\mathbf{x}, \hat{y}_1, \ldots, \hat{y}_{l-1}, \hat{y}_{l+1}, \ldots, \hat{y}_{|L|})$, where $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{|L|})$ is a vector of ground truth labels.

For the parameter instantiation, we choose the complexity parameter $C \in \{10^{-4}, 10^{-3}, \ldots, 1\}$ for the SMO algorithm that achieves best UA recall on the development set, while the rest of the parameters are set as default values recommended by MEKA.

As evaluation measure, we use unweighted average (UA) recall (UAR) as used in the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012).

6. Results and Discussion

Table 4 shows the results obtained for single and multi-task classification, as well as for the oracle single-task experiment where the ground truths of the other labels are included as features in the training and development/test sets. Let us first look at the results of the oracle experiment, which hint at the performance attainable by the ECC approach, which is based on iterative classification using *estimated* class labels for the other tasks. It can be seen that only a few slight (statistically insignificant, according to a

z-test) performance improvements on the development set are obtained when including the ground truth labels for the other two tasks (OMT). Unsurprisingly, this greatly limits the performance of the ECC multi-task learning approach. Comparing with the ECR results, the slight performance improvement observed in age classification by the ECC approach might as well be attributed to bagging, not multi-target learning as such. On the test set, none of the multi-target methods can improve over the single-target baseline (ST).

Overall, but particularly for the likability task, we found that performance heavily depended on the complexity parameter, and parameter selection on the development set did not generalise to the test set. As the complexity parameter controls the feature weights in the SVM, this indicates that the features deemed most important on the development set do not model well the test set. For instance, if we tuned the complexity for the likability task on the test set, we could attain 61.4 % UAR with ECC and 61.0 % with ECR, instead of 52.8 / 57.5 %.

7. Conclusions

In this paper, we introduced the iHEARu project, which addresses some of the shortcomings of current research in computational paralinguistics, one of them being looking at speaker attributes in isolation. A few initial experiments with state of the art multi-target learning methods could not demonstrate improvements over conventional methods. As there are clear signs of overfitting, poor performance can also be attributed to very limited amounts of training data, and failure to extract features that generalise across different speakers. Furthermore, since even the inclusion of ground truth labels from other tasks could not improve performance, it is obvious that there is still large room for improvement in existing machine learning methods for multi-target learning, as foreseen in the iHEARu project. For example, the combination of large-scale, continuous valued feature sets with small-scale, discrete valued label sets in a linear or kernel feature space is arguably sub-optimal; a more suited alternative could lie in novel architectures of Bayesian networks or decision forests. Besides, it seems that multi-target learning can only be successful if considerable progress is also made in the other research challenges addressed by the iHEARu project: large-scale data collection with truly multi-dimensional ('universal') labels, but also unsupervised and semi-supervised feature learning, as well as features inspired by human perception, which are expected to lead to better generalisation. For example, to address the scarcity of multi-target databases (where all instances are labelled in multiple dimensions), and alleviate overfitting, we can investigate large-scale unsupervised feature learning followed by discriminative fine-tuning, using semi-supervised learning to determine missing labels.

8. References

F. Alegre, A. Amehraye, and N. Evans. 2013. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *Proc. of ICASSP*, pages 3068–3072, Vancouver, Canada. IEEE.

R. Brueckner and B. Schuller. 2012. Likability Classification-A not so Deep Neural Network Approach. In *Proc. of INTERSPEECH*, Portland, OR, USA.

F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann. 2010. A database of age and gender annotated telephone speech. In *LREC*.

F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger. 2011. 'Would You Buy A Car From Me?'—On the Likability of Telephone Voice. In *Proc. Interspeech*.

R. Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S.K.W. Fellenz, and J. Taylor. 2001. Emotion Recognition in Human-computer Interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80.

J. Deng and B. Schuller. 2012. Confidence Measures in Speech Emotion Recognition Based on Semi-supervised Learning. In *Proc. INTERSPEECH*.

T. Evgeniou and M. Pontil. 2004. Regularized multi-task learning. In *Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117.

F. Eyben, M. Wöllmer, and B. Schuller. 2009. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. ACII*. HUMAINE Association, IEEE, September.

F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie. 2010. On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues. *Journal on Multimodal User Interfaces, Special Issue on Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots*, 3(1–2):7–12, March.

F. Eyben, F. Weninger, F. Groß, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia*, Barcelona, Spain. ACM.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. 2009. The WEKA Data Mining Software: an Update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

D. Le and E. Mower. 2013. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic.

G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski. 2012. An Extensive Experimental Comparison of Meth-

ods for Multi-label Learning. *Pattern Recognition*, 45(9):3084–3104.

O. Maron and T. Lozano-Pérez. 1998. A Framework for Multiple-instance Learning. *Advances in neural information processing systems*, pages 570–576.

E.L. Mencía. 2010. Multilabel Classification in Parallel Tasks. *Working Notes*, page 29.

J. Read, B. Pfahringer, and G. Holmes. 2008. Multi-label Classification Using Ensembles of Pruned Sets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 995–1000. IEEE.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. 2011. Classifier Chains for Multi-label Classification. *Machine learning*, 85(3):333–359.

J. Read. 2010. *Scalable Multi-label Classification*. Ph.D. thesis, The University of Waikato, New Zealand.

B. Schuller and A. Batliner. 2014. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, November. to appear.

B. Schuller, S. Steidl, A. Batliner, E. Nth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. 2012. The INTERSPEECH 2012 speaker trait challenge. In *Proc. of INTERSPEECH*, Portland, OR, USA. ISCA.

B. Schuller. 2012. The Computational Paralinguistics Challenge. *IEEE Signal Processing Magazine*, 29(4):97–101, July.

A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller. 2011. Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. In *Proc. ICASSP*. IEEE.

G. Tsoumakas and I. Katakis. 2007. Multi-label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

G. Tsoumakas and I. Vlahavas. 2007. Random k-labelsets: An Ensemble Method for Multilabel Classification. In *Machine Learning: ECML 2007*, pages 406–417. Springer.

K. Wu and D. Childers. 1991. Gender Recognition from Speech. Part I: Coarse Analysis. *The Journal of the Acoustical society of America*.

M.L. Zhang and Z.H. Zhou. 2013. A Review On Multi-Label Learning Algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, Preprints(99):1–43.