Noname manuscript No.
(will be inserted by the editor)

Personalized Models for Speech Detection from Body Movements Using Transductive Parameter Transfer

Ekin Gedik · **Hayley Hung**

Received: date / Accepted: date

Abstract We investigate the task of detecting speakers in crowded environments using a single body worn triaxial accelerometer. Detection of such behaviour is very challenging to model as people's body movements during speech vary greatly. Similar to previous studies, by assuming that body movements are indicative of speech, we show experimentally, on a real world dataset of 3 hours including 18 people, that transductive parameter transfer learning [35] can better model individual differences in speaking behaviour, significantly improving on the state-of-the-art performance. We also discuss the challenges introduced by the in the wild nature of our dataset and experimentally show how they affect detection performance. We strengthen the need for an adaptive approach by comparing the speech detection problem to a more traditional activity (i.e. walking). We provide an analysis of the transfer by considering different source sets which provides a deeper investigation of the nature of both speech and body movements, in the context of transfer learning.

Keywords Social signal processing · Wearable sensors · Social actions · Transfer learning · Human behaviour

An extended abstract version of this paper is published in UBICOMP 2016 with the title of "Speaking Status Detection from Body Movements Using Transductive Parameter Transfer" [9]. In addition to preliminary results presented in the UBICOMP paper, current paper presents an analysis of connection between speech and body movements, provides comparisons with the state-of-the-art methods and different implementations of TPT, analyses source quality in transfer learning and presents an analysis of effects of gender in transfer.

Ekin Gedik
TU Delft, Mekelweg 4, Delft, The Netherlands
Tel.: +31-15-2786224
Fax: +31-15-2786224
E-mail: e.gedik@tudelft.nl

Hayley Hung
TU Delft, Mekelweg 4, Delft, The Netherlands
E-mail: h.hung@tudelft.nl

1 Introduction

This research addresses the analysis of social behaviour in crowded mingling events. Such events contain a large number of people interacting with each other closely. These scenarios are interesting since they are concentrated moments for people to interact, make new contacts, renew existing ones, or even influence each other.

In this paper, we focus on the detailed analysis of how to automatically detect whether someone is speaking in these dense crowded scenarios using just a single wearable triaxial accelerometer hung around the neck. Different challenges are introduced with the dense nature of such events, like the high non-stationary background noise from the audio and the heavy occlusion of people in the video. On the other hand, wearable sensors such as accelerometers are less affected by these challenges and their easy scalability makes their use appealing for such scenarios. Moreover, perceptions of privacy are often more sensitive to the recording of audio during conversations, even if the signal is immediately converted into privacy-sensitive features. In this paper, we focus on the use of accelerometers that could be embedded in a smart badge such as a conference badge and hung around the neck.

The use of accelerometers to detect speaking status is generally under-explored in the literature. However, limited amount of studies have shown that it is possible to detect whether someone is speaking based on just a single worn accelerometer [12, 13] by exploiting findings in behavioural psychology that speakers move (e.g. gesture) during speech [23]. One of the biggest challenges, which has not been addressed in the literature before, is accounting for the huge variation in ways in which people move while speaking. This person specific connection between movement and speech requires special approaches for detection, since relying on a single unified model to predict the speaking behaviour of ev-

Fig. 1 A snapshot from the event

eryone leads to large estimation errors as the size of the test population increases. We have chosen speech as the focus as our study since it is a vital unit of behaviour to analyse social behaviour between people at the conversation level [32]. Some examples of further, higher level understanding that may follow from speech detection are the evaluation of an individual's social activeness, detection of conversing groups [12], dominance and group hierarchy [15,21] and, cohesion [14]. In this paper, we propose to use transfer learning to enable the adaptation of a learnt ensemble model of speaking behaviour to a new unseen subject, based only on unlabelled data. The proposed method, Transductive Parameter Transfer [35], has never been used for this problem. With this method, we provide a solution that can generalise over large populations without requiring personal labelled data.

The key contributions of our work are: (i) we provide a study of speech detection through accelerometers, in a real world event (a snapshot is shown in Figure 1), with 18 participants (three hours worth of data); to our knowledge, no similar study at such scale exists. (ii) we delve deep into the connection between body movements and speech, showing how this problem differs from the traditional action recognition (e.g. walking) by providing results that compares the person dependent and independent models. (iii) we propose a transfer learning approach, which can generalise over large populations without requiring personal labelled data, overcoming the restrictions introduced by the person specific nature of speech. (iv) we present a detailed analysis of the parameter transfer that connects detection performance to personality which provides insight into the nature of both speech and transfer learning in this context.

2 Related Work

2.1 Action recognition with accelerometers

Most research that has involved the detection of behaviour from worn accelerometers have tended to focus on the detection of daily activities. In 2004, Bao and Intille used five accelerometers worn on different body locations to detect 20 different actions which include activities like walking, sitting, running and vacuuming [2]. The data for the experiment was collected in a lab environment for 20 different participants. Statistical and spectral features extracted from acceleration data were used and different classifiers were compared for performance. Their results shown that, even without using person specific data, high recognition performance was possible for such actions.

The following year, Ravi et. al. presented their work that aims to detect similar eight daily activities with single worn accelerometer only [27]. The data collection was semi-controlled where the ordering of the activities was random. Their study showed that one accelerometer worn around the thigh area was sufficient for detecting many actions. With the rapid development of this domain, many different feature extraction techniques and classifiers are considered and compared with each other, providing a solid knowledge base for the detection of such activities [26].

Another research area that benefits from the utilisation of wearable sensors is health care, where people presented their work on automatic fall detection [36,8]. As expected, also in these experiments, the data collection was carried out in a controlled environment where participants imitate falling. Both studies reported nearly perfect recognition scores. We show later in "Comparing Controlled & In-The-Wild Settings" section that there are significant differences in the nature of the data collected in controlled and acted settings compared to less controlled ecologically valid ones. Moreover, since such high accuracy was already obtained across a number of different participants, we can conclude that the nature of these tasks is much less sensitive to person-specific variations.

Unfortunately, none of these studies focuses on addressing the challenges of real life crowded environments or a social action like speech.

2.2 Transfer learning for behaviour recognition

Transfer learning is also used in some studies that focus on activity recognition for better performance but generally the setup of the transfer differs from our work. In their survey, Cook et. al. [6] grouped existing transfer learning studies with respect to the modalities used: video sequences [34], wearable [11,37] and ambient sensors [16].

Some of these studies aim to transfer knowledge between different data acquisition setups, like van Kasteren et.al. [16]. This study is somewhat close ours, since they used transfer learning to exploit existing labelled data sets to learn the parameters of a model applied in a new home. This was done to eliminate sensor placement and individual behaviour differences in each house. However, the sensors utilised (ambient sensors such as pressure mats, mercury contacts and passive infrared) and the detected actions

(daily activities such as going to bed, brushing teeth, etc.) were entirely different than ours.

Another concept studied before is the transfer between actions. For example, Hu et. al. proposed a method, which focused on cross domain activity recognition [11]. They transferred the information from an available labelled data of a set of existing activities to a different yet still related set of activities. This was done by learning a similarity function between activities using Web search where web pages related to these activities are extracted and further processed to obtain a similarity measure (Maximum Mean Discrepancy). Similar to the former study, this study also presented its results on daily activities and used multimodal data streams as input.

Perhaps the closest study to ours was published by Zhao et. al. [37]. In this study, the authors presented a transfer learning based personalized activity recognition method. They used accelerometers embedded in mobile phones to gather data from different people while performing daily activities such as standing, walking, running and going upstairs or downstairs. In their method, they integrated decision trees (DT) and k-means clustering where decision trees were used to learn optimal parameters for labelled source data. Then, the DT model was transferred to a new user by classification and the initial parameters for k-means were set with respect to it. Finally, non terminal nodes of the DT was adapted to the new user, resulting in a personalized model. We discuss and experimentally show in our paper that the mentioned activities are less affected from interpersonal differences when compared to speech. Also, this method could only utilise a single source set for transfer while our approach can exploit multiple sources simultaneously. However, this study shows that transfer learning could be a good candidate for eliminating interpersonal differences.

2.3 Social computing with wearables

There are some studies in the literature that focus on analysing social phenomena using wearable sensors but most of them differ from ours in some aspects like the different modalities used as input, analysis of less crowded scenarios and lack of focus on fine time scale detection of social actions such as speech.

2.3.1 Large scale long-term studies

One of the first studies that utilises a wearable sensor for analysis of social phenomena was presented by Choudhury et. al. in 2003 [5]. Authors presented an automated method of analysing social network structures with the so-called sociometer, a wearable multimodal sensor that has a microphone, IR transceiver and two accelerometers. The data collection was done in two stages. In the first stage, 8 subjects from the same research group wore the sociometer during working hours for 10 days. The second stage included 23 participants from four different study groups wearing the badge for 11 days. In the study, audio data is used to detect speaking status, IR transceiver data was utilised for detecting interactions but acceleration information was not used. Using the frequency and duration of interactions detected, a social network of participants are formed. It is shown that by analysing these network, higher level information about the group structures, such as centrality of a participant, can be obtained.

Olguin et. al. obtained high level descriptions of human behaviour like physical and speech activity, face-to-face interaction, proximity and social network attributes using the sociometric badge mentioned earlier [25]. With this high level information, the authors classified the personality traits of participants, with respect to the "Big Five" model. The dataset included 67 participants and was collected for 27 days. Microphones and accelerometers were used to measure speech and physical activity, respectively. Although the study presented an excellent analysis of social phenomena throughout time, it did not focus on fine time grained detection of any action and aims to provide a higher level overview of social phenomena.

In a similar study conducted by Wyatt [33], social ties and collective behaviour of groups were investigated using a multimodal sensing device with 8 different modalities. Conversational characteristics of 24 people were analysed over 6 months. Similar to the former study, speech detection was applied to microphone data. Since social phenomena in a longer period of time is analysed in these studies, we expect the speech detection results to be quite rough. We believe participants current environment will greatly affect the actual detection performance. Such results are satisfactory for obtaining general statistics throughout time but if a fine grained analysis of speech and interaction is required, an approach that can provide more robust detection results of a fine scale is needed (e.g. over just a few seconds).

Apart from specialised sensor devices, some studies use mobile phones as social sensors like Madan et.al [19]. They used proximity, call data records and cellular-tower identifiers to investigate activities and interactions of individuals aiming to detect social behaviour changes with respect to illness. With the development of smart phones, this may eliminate the need for special sensing devices and makes scaling to bigger populations much easier.

2.3.2 Studies of short-term dense crowded social events

There are also studies that aim to analyse social behaviour in crowded mingling settings at a short-term level (i.e. minutes or hours rather than weeks or days). A recent study from Alameda-Pineda et.al. [1] showed that by combining sensor

data from distributed cameras and wearable sensors, it was possible to obtain head and body pose estimation of people in a real life crowded event, with a fine time scale. The proposed method combined visual input from four cameras with noisy estimates of binary speaking status and proximity input obtained from wearable sensors and estimated the behaviour by learning from noisy incomplete observations using a matrix completion method. They went on to show that their automatically extracted head and body poses could be used to infer high level information such as detecting conversing groups or social attention attractors.

Cattuto et. al. [3] used conference badges equipped with RFID to analyse face-to-face interactions in crowded social gatherings. The exchange of radio packets between these badges were used to measure proximity and ultimately detect face-to-face interactions. The mentioned method was highly scalable and tested in three different events that include 25 to 575 people. Their analysis of the dynamics of interaction networks in these events showed a super-linear behaviour between the number of connections and their durations which can be used to define super connectors. However, this study automatically labelled interactions when two people came in close proximity but the accuracy of this was never evaluated. While such methods tend to have a very high precision, the recall is often poor, particularly when the density of the crowd is high.

Martella et. al. used accelerometers to predict implicit responses of an audience to a real life dance performance [20]. 32 spectators of the event were fitted with accelerometers hung around the neck. Aside from analysing their direct responses to the performance they also analysed the effects of the dance performance on the mingling behaviour of participants before and after the event using proximity sensing. Although the sensor pack was fitted with an accelerometer, no speech detection was carried out.

2.4 Speech detection with accelerometers

Although it is hard to find studies where wearable sensors were used for detecting speech and/or other social actions, there exists a few. Matic et. al. [22] used accelerometers for speech detection where accelerometers were tightly attached to the chest of participants in order to detect acoustic phenomena from speech. This methodology requires accelerometer to have a sample rate high enough to detect acoustic speech-based utterances and demands strict placement of the sensor which is impractical for many real life scenarios.

More similar to our work, Hung et. al.[13] presented their method for predicting social actions such as speaking, drinking, gesturing and laughter in a crowded environment with a single accelerometer hung around the neck. Spectral features were used to model these actions and HMMs were used for classification in a non adaptive learning approach. In a follow up to this study in 2014 [12], random forests were considered for classification and proved to perform better. In both studies, no detailed analysis to show variations of performance with respect to interpersonal differences were presented.

3 The Nature of Speech and Body Movements

In this section, we show how the person specific connection between speech and body movements shows itself in accelerometer readings by providing simple statistics computed from accelerometer readings of speech and non-speech intervals. These statistics, by proving the existence and personal nature of this connection, acts as a basis for our choice of an adaptive method that can eliminate interpersonal differences.

Similar to [12,13], we aim to use movement information, obtained from accelerometers hung around the neck, as the proxy for speech. Fortunately, this assumption is partially backed by existing studies. Prior work has shown that it is possible to automatically classify conversing participants with an acceptable performance using acceleration information only [12,13]. The connection between body movements and social behaviour is also extensively studied in social psychology [17,4,23]. For example, McNeill discussed that speakers tend to move noticeably more when compared to listeners [23]. It was discussed that gestures and speech are integrated parts of communication where gestures are used to complement the content of speech by providing visual stimuli acting as "symbols". Multiple studies also showed that there is a strong correlation and synchrony between speech and body movements in conversing groups [17,4].

However, the connection between speech and body, specifically torso, movements is not theoretically well defined. Previous studies pointed to the existence of this connection but none made a precise description of the torso movement that can be exploited for automated detection that can generalise over large populations. We believe that this connection is highly personal and should be detectable from accelerometer readings. To test this assumption, we calculated the variation of accelerometer magnitudes over a sliding window (3 seconds with 1 second shift) of speech and non-speech intervals for 18 different people wearing accelerometers in a real life, crowded mingling event (see Section 5 for details).

Figure 2 shows the median values of the variation in accelerometer magnitudes for speech and non-speech intervals. Each axis of raw acceleration is normalised using z-score standardization before computing the magnitude and extracting the variance values with sliding windows of the same length and shift size. We see huge differences between participants. One can easily see that one participants median variation of accelerometer magnitude for speech inter-

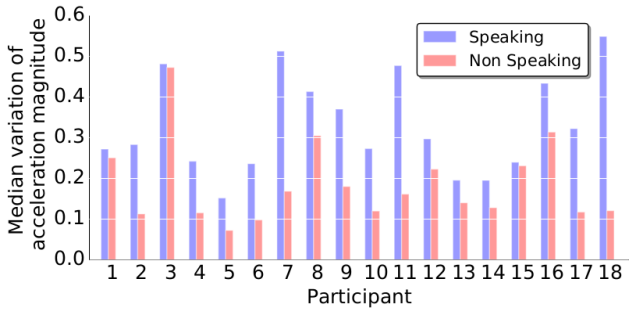

Fig. 2 Median variance of acceleration magnitudes for speech & non-speech intervals for 18 people.

vals can be closer to another participants non-speech feature. One-tailed t-tests applied to this feature during speech intervals for all pairwise combinations of participants showed that nearly 50% of these couples have significantly different distributions.

We also see that, for nearly all participants, the median of the variance in acceleration magnitude tends to significantly differ for speaking and non-speaking intervals. However, it can be also seen that the amount of this difference varies greatly per person. These two observations show that there is definitely a connection between speech and body movements but the nature of this connection is quite person specific.

This personal connection between speech and torso movement makes the problem entirely different and more challenging than traditional approaches to speech detection using audio. The connection between speech and audio is physically well defined via articulation of the vocal folds leading directly to resonances in the vocal tract. Of course, different speakers will have different spectral characteristics depending on their physiology[28] but satisfying speech detection results are already possible with person independent models [7].

With these findings, a traditional learning approach where the data of different subjects is amalgamated into a single training set will perform poorly since the decision surface obtained in this way will not be optimal. In our study, we propose to use Transductive Parameter Transfer [35,30], an adaptive approach which uses transfer learning to overcome this issue by computing a personalised decision surface for each subject based on the similarity of a test subject's data distribution with those of multiple individuals in a training set.

4 The Transductive Parameter Transfer Method

With the findings of the last section, we propose to use an adaptive transfer learning approach, Transductive Parameter Transfer, presented in [30,35]. The authors of [30,35] used their method to compute personalized models for facial expression analysis from video input. To our knowledge, we present the first example of application of this method to action recognition and more specifically, speech detection from wearable sensors task. Although the main theory of the method stays the same, we have some different implementation choices than [30,35] which we elaborate on below.

In this approach, with feature space X and label space Y, N source datasets with label information and the unlabelled target dataset are defined as $D_1^s,...,D_N^s$, $D_i^s = \left\{x_j^s, y_j^s\right\}_{j=1}^{n_i^s}$ and $X^t = \{x_j^t\}_{j=1}^{n_t}$, respectively. It is assumed that samples X_i^s and X^t are generated by marginal distributions P_i^s and P^t, where $P^t \neq P_i^s$ and $P_i^s \neq P_j^s$. P_t and P_i^s are presumed to be drawn from ρ, the space of all possible distributions over X, with respect to meta distribution Π.

This approach aims to find the parameters of the classifier for the target dataset X^t, without using any label information of X^t, by learning a mapping between the marginal distributions of the source datasets and the parameter vectors of their classifiers. Main steps of the Transductive Parameter Transfer approach are shown in Algorithm 1 and each step is explained in detail below.

Algorithm 1 Transductive Parameter Transfer approach [30]

Input: Source sets $D_1^s,...,D_N^s$ with labels and the target set X^t
1: Compute $\{\theta_i = (w_i, c_i)\}_{i=1}^N$ using (1).
2: Create training set $\tau = \{X_i^s, \theta_i\}_{i=1}^N$.
3: Compute the kernel matrix K where $K_{ij} = \kappa(X_i^s, X_j^s)$ using (8).
4: Given K and τ, compute $\hat{f}(.)$ solving (6).
5: Compute $(w_t, c_t) = \hat{f}(X^t)$ with (7).
Output: $\mathbf{w_t}, c_t$

4.1 Obtaining personalized hyperplane parameters

First, person specific classifiers are trained on each source dataset individually to obtain the best performing parameter set θ. Instead of a Linear SVM used in [35,30], we have selected the well known binary class L2 penalized logistic regression classifier which minimizes Equation (1). Since both are linear classifiers and the format of the resulting parameters is similar, this selection does not require any extra steps.

$$\min_{(w,c)} \frac{1}{2} w^T w + C \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1) \quad (1)$$

We have used Stochastic Average Gradient descent [31] to solve this optimization problem, obtaining the optimal parameter sets $\{\theta_i = (w_i, c_i)\}_{i=1}^N$ for each subject. The optimal

regularization parameter C is found through k-fold cross validation and the model is trained on the complete dataset of the participant with this C value.

4.2 Mapping from distributions to hyperplane parameters

The second step aims to learn the relation between the marginal distributions P_i^s and the parameter vectors θ_i. The assumption here is that for each participant, the hyperplane whose parameters are defined by θ_i are dependent on the underlying distribution P_i. By learning this relation, the optimal hyperplane parameters for the target dataset can be computed without any label information. The actual underlying distributions are not known, neither for the source datasets $\mathbf{P_i^s}$ nor the target P^t, however they can be approximated using the samples X_i^s and X^t. Thus, the method aims to learn a mapping from samples to the parameters, $\hat{f} : 2^x \rightarrow \theta$, using the training set $\tau = \{X_i^s, \theta_i\}_{i=1}^N$, formed after the first step of the algorithm.

Since we assume that elements in θ are correlated, we employ Kernel Ridge Regression(KRR), instead of the multiple, independent regressors proposed in [35]. The primal problem for ridge regression is defined as follows [24]:

$$min((y - Xw)^T(y - Xw) + \|w\|^2) \tag{2}$$

where the optimal solution is given as:

$$w = (X^T X + \lambda I_D)^{-1} + X^T y = (\sum_i x_i x_i^T + \lambda I_D)^{-1} X^T y \tag{3}$$

The formulation for ridge regression can be kernelized with the following steps. First, Equation (3) is rewritten as

$$w = X^T (XX^T + \lambda I_N)^{-1} y \tag{4}$$

Term XX^T in Equation (4) can be directly replaced with the Gram Matrix K, partially kernelizing the equation. In order to eliminate term X^T and completely kernelize the formulation of ridge regression, following dual variables are introduced:

$$\alpha \equiv (K + \lambda I_N)^{-1} y \tag{5}$$

With the introduction of dual variables, Equation (4) becomes

$$w = X^T \alpha = \sum_i^N \alpha_i x_i \tag{6}$$

After solving for w, the solution for any variable x can be found as:

$$\hat{f}(x) = w^T x = \sum_i^N \alpha_i x_i^T x = \sum_i^N \alpha_i \kappa(x, x_i) \tag{7}$$

It can be seen from Equations (5) and (7), a kernel κ that can define the similarities between two distributions is needed. Instead of the density estimate kernel defined in [35], we have selected an Earth Mover's Distance[29] based kernel which is discussed in [30]. In our implementation, each sample is treated to be a signature where all samples have uniform weights. The EMD kernel is defined as

$$\kappa_{EMD} = e^{-\gamma EMD(X_i, X_j)} \tag{8}$$

where $EMD(X_i, X_j)$ corresponds to the EMD between two datasets X_i and X_j, the minimum cost needed to transform one into another. γ, a user defined parameter, is set to be the average distance between all possible pairs of datasets and experimentally shown to perform well.

4.3 Classification

By solving (6) for the source datasets, we learn the mapping $\hat{f} : 2^x \rightarrow \theta$. For any new target dataset, we can compute the parameter vector θ_t by plugging X^t into the mapping function $\hat{f}$. Classification of the samples in the target dataset is then obtained by $y = sign(w_t x + c_t)$.

5 Dataset & Feature Extraction

5.1 Dataset

We recorded data in a real pub with 16 male and 16 female volunteers during a speed dating social event. The first phase involved having three minute dates with each member of the opposite sex. After this, participants could get to know each other better in a mingling session. This phase has the characteristics of a crowded mingling scenario which we needed for our experiments. All throughout the event, participants wore a specialised sensor pack around their necks which collects acceleration and proximity information. The accelerometer in the sensor pack provides 20 samples per second. In our experiments, we only used accelerometer data. The area was fitted with multiple video cameras facing down on the scene, covering all the area participants were present. The video footage was used for labelling the ground truth.

5.2 Annotations & Features

5.2.1 Annotation procedure

In this study, we will be focusing on the mingling phase. The mingling session lasted for approximately an hour. Due to hardware malfunctions, only 28 of the sensor packs recorded data in this session. Although we would have preferred to use all the data we have for the classification experiments,

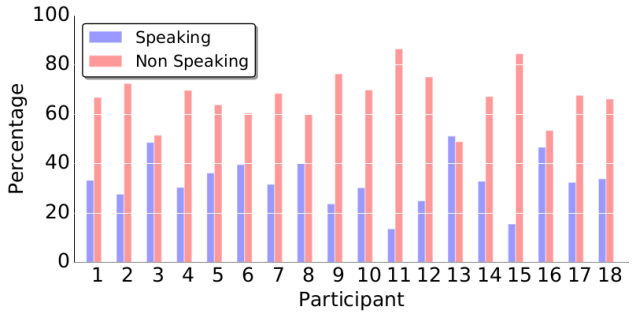

Fig. 3 Percentages of speaking-non speaking samples

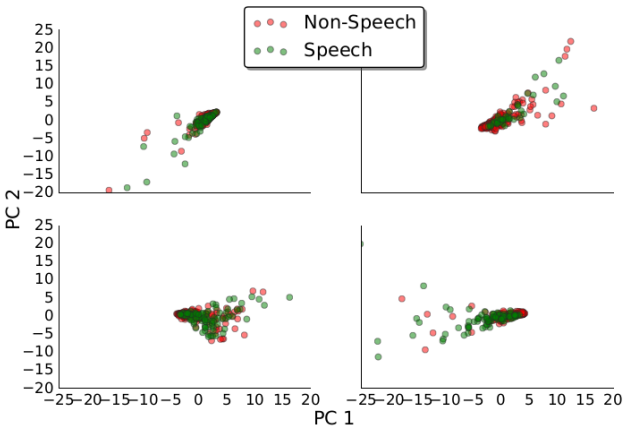

Fig. 4 First two principal components of four participants (18,17,10,11)

the annotation of social actions (in our case, speech) is extremely time consuming and costly. Also, some of the participants were at the blind spots of our cameras for the majority of the event, making robust annotation of their data extremely challenging. These factors forced us to use a subset of 18 participants for our experiments. This is in keeping with the numbers of test subjects typically used for studies in activity recognition, where datasets of varying sizes from 1 to 24 participants are reported [2, 13].

Thus, speaking status for these 18 participants were carefully labelled using the video for 10 minutes of the mingling phase with a time resolution of one twentieth of a second. A qualitative inspection revealed a rich dataset including participants with differing levels of expressiveness, interacting in dyads, larger groups or hardly interacting with someone at all, covering different types of personal characteristics and interactions possible in such an event. Detailed inspection of the annotations also showed that the speaking turn lengths per person vary greatly, from few seconds to more than half a minute, further showing the variety captured in the dataset.

5.2.2 Feature extraction

Before feature extraction, each axis of the acceleration input is standardised to have zero mean and unit variance. We selected our features from the literature and ensuring that were as simple as possible so as to avoid overfitting the data of the participants. The selected features can be grouped into two categories; statistical and spectral. As our statistical features, we calculated mean and variance values. As the spectral features, the power spectral density (PSD) was computed in the same way as [13], using 8 bins with logarithmic spacing from 0-8 Hz. These were extracted from 3s windows with on third overlap for each axis of the raw acceleration, absolute value of the acceleration, and magnitude of the acceleration. The length of the window was selected to be big enough to capture the speaking action while preserving a fine temporal resolution. All features were concatenated to obtain a 70-dimensional feature vector per window.

5.2.3 Dataset analysis

Using the annotations and acceleration from this 10 minute interval, we have extracted features for each participant. This resulted in the total of 18 feature vectors, each having 299 samples with 70 dimensions, with varying class distributions. The class distributions for each participant is shown in Figure 3. The mean percentage of the positive samples (speech) across all participants are found to be 33, with a standard deviation of 10%. Participant 11 had the least number of positive samples (14%) whereas, person 13 had the highest percentage (51%). This imbalance in class distribution, which is also person specific, introduces a new difficulty for robust detection of speech.

In order to see how person specific nature of speech affects the distribution of samples in feature space, we have applied dimensionality reduction to samples of four participants and plotted them for the first two principal components. To standardize the plots, samples from the four participants were collectively normalized with z-score standardization. We can see from Figure 4 even after preprocessing, distributions are close to each other in the feature space, while the distribution of samples and the characteristics of the data still vary greatly between different participants.

6 Experimental Results

In this section, we will discuss and compare the performance obtained with different classification setups and approaches. When presenting classification performance, we have specifically selected Area Under Curve (AUC) since it provides a more valid performance estimate in the presence of our imbalanced binary classification problem. Also, while training any classifier, class weights are set to be inversely proportional to the number of samples in the class so as to remove any bias caused by imbalanced class sizes. Of the all setups

Fig. 5 Performance in terms of AUC for speech detection. Person dependent setup uses data from the same participant for training and testing and expected to act as an upper bound for the performance. Person independent and TPT setups use data from other participants in a leave-one-subject-out manner.

discussed in this section, only person dependent one uses the data from a single participant, for training and testing, in a leave-one-sample out manner. Other setups, person independent and TPT, use data from other participants. Thus, person dependent setup is expected to act as an upper bound on the performance since it is a personalised setting by nature.

6.1 Person dependent performance

In the person dependent setup, each participant is trained and tested on their own data. Since we don't have enough data to come up with distinct training and test sets, we applied Leave-One-Sample-Out cross validation scheme for performance evaluation. Based on the findings reported in [10], we made sure that training set is not contaminated. This means for each fold, any adjacent samples to the test sample are eliminated from the training set. With this elimination, we aim to provide an unbiased performance estimate. We have used a logistic regressor as classifier where the optimal regularization parameter C in Equation (1) is found by nested k-fold cross validation.

The procedure is applied to each participants' data separately, obtaining performance evaluations for each. This resulted in varying performance scores, ranging from an AUC score of 55% to 79%. The mean performance across all participants is $68\% \pm 6$. Individual scores for each participant are shown in Figure 5.

The variation in performance scores can be linked to two different factors we have already discussed. The first is the personal connection between speech and body movements read through the accelerometer. As expected, the problem becomes harder for people with more subtle movements, resulting in lower performance. Still, each participants' performance score is higher than random (50% AUC), proving that our features are still discriminative.

Second factor is related to the class distributions. As shown in Figure 5, some participants' class distributions are highly skewed towards the negative class. We can not say that such imbalance always guarantees low performance, since it may still be possible to train robust models from small numbers of highly informative samples. However, we already see negative effects of this imbalance in our results. The participants with the lowest performance scores have small number of positive samples. There are only two participants with AUC scores lower than 60% (P12: 55% and P15: 57%) and they have the second and third lowest percentages of positive samples (25% and 16%, respectively) in the whole dataset. So, for these two participants, we can not be sure if the low performance is caused by subtle movement while speaking or the small number of positive samples.

We expect these results to act as an upper-unbiased limit for speech detection performance.

6.2 Person independent performance

In the person independent setup, we have used Leave-One-Subject-Out cross validation for performance evaluation, where each participants' samples are classified with the model obtained from other participants' data. So, the training set is formed by concatenating and standardising all other participants data. Similar to the person dependent setup, logistic regressor is used as the classifier and optimal regularization is then found on the training set with cross validation.

With this setup, we obtained an average AUC score of 58%, with a standard deviation of 7%. The individual scores for participants varied from 45 to 60%. The individual scores obtained with the person independent setup are also shown in Figure 5, together with the results of other setups. Apart from two participants (7 and 8), where the person independent setup yielded slightly better AUC scores than the dependent one, the person dependent setup always outperforms the independent setup. We compared the performances of person dependent and independent setups per person using a paired one-tailed t-test. As expected, the result of the t-test showed that the person dependent setup yields significantly better results than the independent one ($p < 0.01$).

In the ideal learning paradigm, training with more samples should yield a better, more robust model, contradicting what we see. However, it is also assumed that the samples in the dataset are coming from the same independent and identically distributed (i.i.d.) probability distribution. From what we see from Figures 3 and 4, it is more likely that every participant has their own probability distribution that their samples are drawn from. Thus, concatenating the data of all participants and training a model on this dataset results in an unreasonable and impractical decision boundary. These person independent results strengthen our claim of the personal nature of connection between speech and body movements and motivate the requirement of an adaptive model.

6.3 Transductive parameter transfer performance

Our TPT experiments also employed a Leave-One-Subject-Out setup, where each participant is treated to be the target dataset while all other participants acted as source sets. This setup is similar to the person independent one, since the labels of only other participants are used for classification. With TPT, an average AUC of $65\% \pm 6$ is obtained. Individual performance values are included in the Figure 5, in addition to those of the person dependent and independent setups.

It is clearly seen that TPT outperformed the person independent setup for majority of the participants (16 out of 18), providing an AUC score close to the person dependent setup. One-tailed t-test between the TPT and the person independent scores showed that TPT is significantly better than the other ($p < 0.01$). For few cases, TPT even outperforms the person dependent setup (participants 2, 7, 8, 11), however, the person dependent results are still significantly better than TPT ($p < 0.02$). This result is quite interesting and might be caused by different factors. When the performance for participants 7 and 8 are inspected, it can be seen that even the person independent setup outperforms that of the person dependent one. This suggests that for these participants, using more data (even belonging to other participants) provides a better estimation of the decision boundary. In such a case, we may expect TPT to outperform all other setups. Although the same pattern is not present for participants 2 and 11, we might still argue that these participants have benefited from the use of the data of other participants, most probably the ones having a similar distribution.

These results prove that it is still possible to generalise over unseen data, with an acceptable performance, if an adaptive method like TPT is employed. In 10 minutes one might argue that there is relatively little variation in an individuals' behaviour. However, assuming that between-person variation remains fairly high over this interval, as it can be seen from Figures 2 and 3, it is particularly interesting that we get good results, showing the robust generalisation ability of our method even with a limited amount of data. With the proposed transfer learning approach, performance results that are always better than the random baseline are obtained and statistical significance tests showed that our proposed method guarantees to perform better than traditional non-adaptive person independent learning.

6.4 Comparison with the state-of-the-art

This section compares the performance of our Transductive Parameter implementation with the state-of-the-art approaches. Firstly, we present the person independent results obtained with Random Forests (RF) and Hidden Markov Model (HMM) based approaches proposed in [12]. Secondly, we present the results obtained with the TPT implementation given in [35] and discuss in detail how our different choices affected the final performance. Individual performance scores obtained with all four methods, including ours, can be seen in Figure 6.

6.4.1 Non-adaptive person independent methods

We have implemented the methods presented in [12]. We have used the exact same setup they defined which includes the features they used (PSD 0-8Hz), window sizes for feature extraction (5s for RF, 3.5s for HMM), number of trees in Random Forest classifier (500) and number of states in HMM (2). We compare with the Leave-One-Subject-Out cross validation setup reported in [12].

With the RF, we obtained an average AUC score of $55\% \pm 6$. The HMM performed slightly better, providing an average AUC of $59\% \pm 6$. When compared to our person independent results obtained with logistic regression, neither RF nor HMM provided a significantly better result. This is an interesting finding since it shows that a linear model is as powerful as a nonlinear model for the speech detection problem, in a Leave-One-Subject-Out setup. Our proposed TPT method, on the other hand, significantly outperforms both of these methods. There are only 3 participants that have better performance scores than our proposed implementation of TPT; participants 1 and 3 for RF and participants 1 and 11 for HMM. One tailed t-tests between our TPT results and both RF and HMM showed TPT performs significantly better ($p < 0.01$ for both RF and HMM). The authors of [12] applied their non-adaptive method on a limited dataset that includes only 9 people. We believe, with the increasing number of participants, the person specific nature of speech is magnified and the requirement for adaptive methods increases.

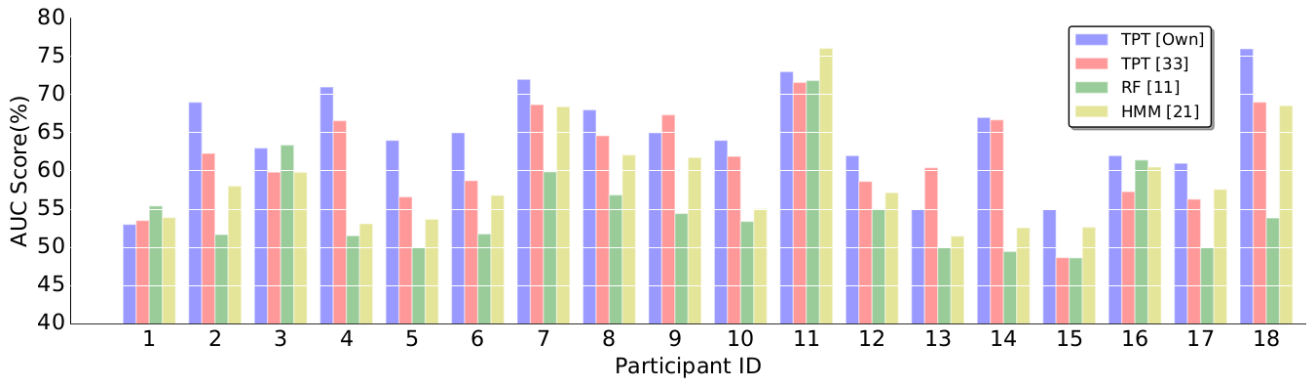

Fig. 6 Comparison with the state-of-the-art as presented in [11] (RF and HMM) & [33] (TPT)

Table 1 Performance and significance of the four modified TPT implementations compared to ours, which had an average AUC score of $65\% \pm 6$ (**(p<0.01),*(p<0.05)).

	Modification (Our implementation)			
AUC	SVM (LR)	SVRs (KRR)	DK (EMD)	SV (WD)
$\pm$ Std	60 ± 4 **	63 ± 7 *	65 ± 7	61 ± 5**

6.4.2 Detailed comparison with state-of-the-art TPT implementation

Our proposed TPT implementation improves upon that presented in [35]. Although the basic framework of the method remains, our implementation choices made the method more suitable to the nature of our problem, as demonstrated by the performance results. We have used the implementation provided by [35] and obtained performance results with that setup, resulting in an AUC of $62\% \pm 6$. Our implementation outperforms it for 15 out of 18 participants.The paired one-tailed t-test between performance scores shows that our implementation is significantly better than [35] ($p < 0.01$).

There are four main differences between our implementation and the one in [35]. TPT implementation in [35] uses: (i) a SVM instead of logistic regression (LR), (ii) independent Support Vector Regressors (SVRs) instead of KRR, (iii) a density kernel (DK) instead of EMD kernel, (iv) support vectors (SV) instead of the whole data (WD) to estimate distributions of source sets. To investigate which modification affected the performance most, we carried out four follow-up experiments. In these experiments, we replaced one of our choices with the original one in [35]. Table 1 shows the average AUC and standard deviation over all participants obtained with each of these modifications. One tailed t-tests were used to quantify differences between our full implementation and one of the modified approaches.

Table 1 shows that the most effective change uses a logistic regressor instead of a Linear SVM. The two setups where our logistic regressor is replaced by a SVM (SVM and SV in Table 1) have the lowest performances. It is an unexpected result since the two classifiers are quite similar. However, the logistic regressor was more successful than the Linear SVM when person specific classifiers were being trained which we believe resulted in this performance difference.

Since our features are often correlated with each other, we preferred to use a KRR instead of the SVRs which is also supported by [30]. The performances shown in Table 1 backs our decision since our method with KRR performed significantly better than the SVRs method. The average performance difference between two methods could be low but our method provides significantly better results.

Finally, we can see that replacing EMD with a density kernel (DK) does not affect the performance at all. For our data, a density kernel was as successful as the EMD kernel in estimating the similarities of distributions. This is quite different than the findings in [30] but we believe it is related to the distribution characteristics of our data.

7 Comparing Speech Detection with Walking

We investigated how different the nature of the speech detection problem was compared to other more traditional actions in the action detection literature to see if speech detection from body motion really requires a different approach. To address this question, we have conducted a follow-up experiment where we compared the speech detection results to an action which is widely studied in the action detection literature, walking.

Here, we used the same setup from our speech detection experiments. Similar to the former section, we obtain two performance scores for each participant; one for each of the person dependent and independent setups. We used a subset of the participants from 9 people who had enough walking samples. In order to obtain an acceptable number of samples, we only included participants that continuously walked more than 3 seconds with at least 15 seconds total walking time. To make the problem similar to our speech detection experiments, we added a random number of non-walking samples to each participant, creating possibly imbalanced

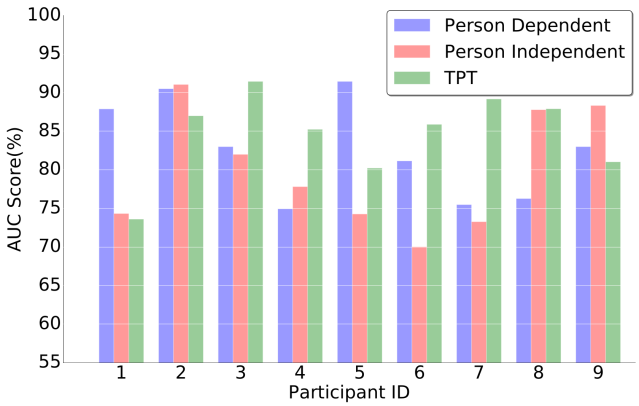

Fig. 7 Performance in terms of AUC for walking

distributions. The performances for this experiment are shown in Figure 7.

The person dependent setup yielded an average AUC of $83\% \pm 6$. With the person independent setup we have obtained an average AUC of $80\% \pm 7$. We have also applied TPT to the walking data with the same leave-one-participant-out setup of the former experiments where data from other participants acted as sources for the transfer. TPT obtained an average AUC of $84\% \pm 7$. The pairwise t-tests between setups showed that no single setup is significantly better than the others and all might provide better performance for an unseen participant. From Figure 7, we can see that the pattern here is entirely different than the speech detection one. First, both person independent and person dependent setups yielded relatively high performances, when compared to performances of the speech detection experiments reported in Section 6 (average AUC of $68\% \pm 6$ versus $83\% \pm 6$ for the person dependent setup and $58\% \pm 7$ versus $80\% \pm 7$ for the person independent one). This is an expected result, since the connection between speech and body movements are not as universally characterizable as the connection between walking and body movements. Secondly, in many cases, better performances than the person dependent setup are actually obtained with the independent one.

Interestingly, the best overall performance score is obtained with the TPT, resulting in an average score slightly higher than the person dependent one. This is definitely different than the speech detection problem where the person dependent setup and TPT performed significantly better than the person independent one. We can still argue that the relatively smaller sample sizes compared to the speech detection experiments might have caused the person dependent setup to perform sub-optimally, explaining the cases where person independent and TPT setups outperformed the dependent one. Yet, these experimental results show that the detection of walking is less challenging, is not influenced by personal differences as much as speech-related body movements and it is still possible to achieve high performance with a non-adaptive model, unlike our speech detection task. In addition, high performances obtained with the TPT, even for a problem that seemed to be less person specific, show that the proposed method is quite robust and still preferable to the traditional person-independent setup in such cases.

8 Comparing Controlled & In-the-Wild Settings

To experimentally demonstrate the restrictions introduced by a real event, we organised a small controlled experiment where one participant imitated speaking, walking and standing in a structured way while wearing an accelerometer. The participant alternated between actions where each action is performed for at least 15 seconds, resulting in a dataset that has 125, 139 and 110 seconds of standing, speaking and walking, respectively. The participant did not exaggerated any action to make them distinguishable from others. It should be noted that, the standing parts also include the imitation of listening, where head-hand gestures and body shifts natural to listening were randomly acted by the participant.

We have used the same experiment setup of the person dependent experiments discussed in former section. Thus, the logistic regressor is used as the classifier, the same set of features and Leave-One-Sample-Out evaluation scheme is utilised. Even though we had three different classes, we treated the problem as a binary classification task, where the samples corresponding to walking and standing formed the negative class. This results in roughly one third of the samples being positive.

Using this controlled data, we achieved an AUC score of 84%. More detailed analysis shows that 4% of walking samples and one third of standing samples are misclassified as speech. This is consistent with former experiments, showing that distinguishing between speech and walking is relatively easy. On the other hand, listening-standing is often confused, probably because similar gestures occur in both. Still, the majority of standing samples are classified correctly. Also, the trained model is quite robust in detecting speech, only misclassifying 8% of speech samples as non-speech.

The performance score obtained in a controlled environment outperforms all our previous experiments with real in-the-wild data. We believe this is related to the two main differences between the setups. First, in the controlled environment we have precise annotations for each action. The noise introduced in the annotation procedure tends to affect the learning procedure. Even it is not guaranteed, since we don't have a robust way of measuring the quality of the annotations we have for the real life event; better annotations may increase performance. However, the annotation quality may also not be related to the essence of the difference between the real life and controlled events.

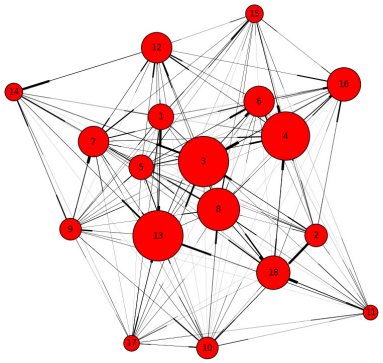

Fig. 8 Visualisation of optimal source sets for each person.

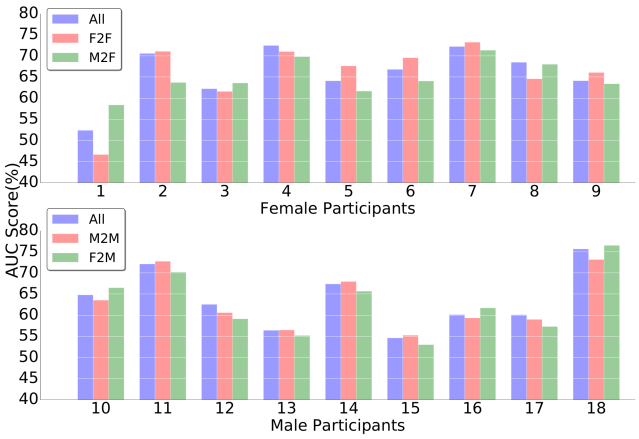

Fig. 9 Performance scores of TPT for gender based transfer (Participant IDs are same with the ones shown in previous figures)

Secondly, the actions performed by the participant in the controlled experiment is highly structured and limited. However, the actions of participants in a real life event is completely unstructured. There is no limit to the type of actions they may perform and transition between. Participants can even perform multiple actions at the same time. It is nearly impossible to cover all the possibilities that may happen in a real life event in a controlled environment. So, we believe that the results obtained from controlled experiments will be always positively biased and would not reflect the true phenomena as it occurs in the wild.

9 Analysis of Transfer Source Quality

While using TPT, we employed a Leave-One-Subject-Out learning scheme where data of all other participants acted as sources. Some source sets might be more informative than others. Conversely, some source sets may negatively affect the mapping function learned, dropping the final performance. Thus, we hypothesised that there might be optimal source subsets for each participant. To check this hypothesis, we classified each participant with every possible triads of source sets. Then, we selected the top 10 best performing triads for each participant. We should note that, all of these setups were somewhat optimal, performing better than the setup where all sources were used.

Figure 8 visualises links between the best performing subset where the size of each node indicates the number of times it was in one of the best performing source sets. A directed edge from node A to B (where end of the edge is slightly wider) means that participant B was at least in one of participant A's best performing source sets. The width of the edges are proportional to the number of times B was in A's source sets.

From the Figure 8, we can see that participants 3, 4, 8 and 13 are the optimal sources for the majority of others. Still, the directed edges show that there is no single perfect source for everyone, meaning multiple sources are needed to cover a larger population. When we inspected the person dependent performances and class distributions for these participants, we did not see any distinguishing features to indicate their quality as sources. Closer inspection of the video of the event confirmed no spatial connection or presence of interaction was necessary for one participant to act as a good source for another. We believe these findings show that the success of these participants as sources comes from something more inherent, most probably related to connection between speech and torso movements.

We analysed whether being a good source might be related to personality. Each participant filled in the HEXACO personality inventory [18] before the event. The HEXACO scale measures personality in 6 dimensions and can broadly be considered similar to the more well-known Big Five personality traits except with an additional sixth dimension measuring humility or honesty. The dimensions are mapped onto a 5-point likert scale. We observed that all these four participants have relatively high extroversion (3.8, 3.6, 3.9, 3.6) and openness (4.1, 3.6, 4.2, 3.4) scores which may contribute to them being good sources. Further analysis of the connection between personality and transfer is left for future work.

10 Analysis of Gender Differences in Transfer

One interesting aspect we haven't investigated in the former sections is how gender specific attributes affect the proposed method. In all of the former TPT experiments, we either used all remaining participants as sources or fetched all possible triads without considering the gender of the participants. In a traditional speech detection setup, where audio recordings are used as input, gender is expected to be a distinctive feature because of the frequency differences in male and female voices. In this section, we present a detailed analysis to see if such a difference exist for our method which relies on accelerometer readings instead of sound. Luckily,

we have a balanced dataset in terms of gender, 9 females (Participant IDs 1-9 in Figure 5 and 6) and 9 males (Participant IDs 10-18 in Figure 5 and 6).

In order to check if there are any gender specific characteristics affecting our method, we devised three different experiment setups. The first setup is entirely same as the one presented in Section 6.3, where we use all other participants as sources. For the second setup, we only use participants as sources who are the same sex with the target participant. The last setup is the reverse of the second one where all sources are the opposite sex of the target participant. Figure 9 shows the performances for all these three setups applied on male and female participants. The items in the legend correspond to all three setups where All corresponds to setup 1, F2F (female sources, female targets) and M2M (male sources, male targets) are setup 2 (same gender transfer) and M2F (male sources, female targets) and F2M (female sources, male targets) are setup 3 (transfer from the opposite gender).

As it can be seen from the Figure 9, there seems to be no significant difference between any of these setups. When we use the all participants as sources, the average AUC scores for female and males are $66\% \pm 6$ and $64\% \pm 6$, respectively. For the same gender transfer, F2F and M2M setups, the average AUC scores are $66\% \pm 8$ and $63\% \pm 6$. Finally, for the opposite sex transfer, M2F and F2M, we have obtained average AUC scores of $65\% \pm 4$ and $63\% \pm 7$, respectively. The individual performances of participants seem to be slightly changing with respect to the setups, however, there is no apparent pattern suggesting a convincing effect of gender on the transfer quality. This is further proved by the t-tests between all different pairs of setups that showed no significant difference.

These results are somehow expected since we are not trying to infer speech from vibrations in the chest which might be strongly affected by the frequency differences of sound between genders. Our method is based on the connection between body (mostly torso) movements and speech which is expected to be affected less by any gender specific differences. These results are also on par with the analysis of the last section where we identified optimal sources for transfer. Three out of four optimal sources were found to be females in this analysis, however, they were good sources for participants from all genders. Thus, we can conclude that even though there might be some gender specific gesturing, we haven't seen any strong effects of it on the success of transfer in our data.

11 Conclusion and Future Work

In this study, we presented a transfer learning approach for detecting speech in real world crowded environments, using accelerometers. By comparing speech detection task to a traditional action recognition problem (e.g walking), we have shown the requirement for a specialised approach that can address the person specific nature of the speech and body movements. As a novel contribution, for the first time, Transductive Parameter Transfer [35] was used to address the person specific patterns of estimating speech from body acceleration. We also analysed the parameter transfer in detail by considering different source sets, providing insights into the nature of transfer and the task of speech detection.

Results obtained with the proposed method outperformed the state-of-the-art, providing performance scores close to person dependent setups. We discussed the challenges that are introduced by a more ecologically valid setting when compared to controlled experiments and experimentally showed how they affected the detection performance. Analysis of transfer quality demonstrated that an optimal subset of sources could be identified for each target set. Moreover, we found that some participants generally acted as good sources for subsets of the population in our data. We observed that this connection was not related to the spatial distance or to their corresponding interaction partners but something more inherent in the individuals.

As future work, we plan to explore automated methods of selecting source sets for each target. Another direction we would like to pursue is testing our method in a different environment, for example, a seated scenario where different variety of actions can be examined.

Acknowledgements This publication was supported by the Dutch national program COMMIT. We would also like to thank Andrew Demetriou (VU Amsterdam), Dr. Leander van der Meij (VU Amsterdam) and Laura Cabrera Quiros (TU Delft) for their support in the data collection process.

References

1. Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., Sebe, N.: Analyzing free-standing conversational groups: a multimodal approach. In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, pp. 5–14. ACM (2015)
2. Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. Pervasive Computing pp. 1–17 (2004). URL http://link.springer.com/chapter/10.1007/978-3-540-24646-6_1
3. Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.F., Vespignani, A.: Dynamics of person-to-person interactions from distributed rfid sensor networks. PloS one **5**(7), e11,596 (2010)
4. Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception–behavior link and social interaction. Journal of personality and social psychology **76**(6), 893 (1999)
5. Choudhury, T., Pentland, A.: Sensing and modeling human networks using the sociometer. In: Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWC03), vol. 1530, pp. 17–00 (2003)
6. Cook, D., Feuz, K.D., Krishnan, N.C.: Transfer learning for activity recognition: A survey. Knowledge and information systems **36**(3), 537–556 (2013)
7. Dines, J., Vepa, J., Hain, T.: The segmentation of multi-channel meeting recordings for automatic speech recognition. In: Int.

Conf. on Spoken Language Processing (Interspeech ICSLP), LIDIAP-CONF-2006-007 (2006)
8. Doukas, C., Maglogiannis, I., Tragas, P., Liapis, D., Yovanof, G.: Patient fall detection using support vector machines. In: Artificial Intelligence and Innovations 2007: from Theory to Applications, pp. 147–156. Springer (2007)
9. Gedik, E., Hung, H.: Speaking status detection from body movements using transductive parameter transfer. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 69–72. ACM (2016)
10. Hammerla, N.Y., Plötz, T.: Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1041–1051. ACM (2015)
11. Hu, D.H., Zheng, V.W., Yang, Q.: Cross-domain activity recognition via transfer learning. Pervasive and Mobile Computing **7**(3), 344–358 (2011)
12. Hung, H., Englebienne, G., Cabrera Quiros, L.: Detecting conversing groups with a single worn accelerometer. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 84–91. ACM (2014)
13. Hung, H., Englebienne, G., Kools, J.: Classifying social actions with a single accelerometer. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pp. 207–210. ACM (2013)
14. Hung, H., Gatica-Perez, D.: Estimating cohesion in small groups using audio-visual nonverbal behavior. Multimedia, IEEE Transactions on **12**(6), 563–575 (2010). DOI 10.1109/TMM.2010.2055233
15. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations using nonverbal activity cues. Audio, Speech, and Language Processing, IEEE Transactions on **17**(3), 501–513 (2009). DOI 10.1109/TASL.2008.2008238
16. van Kasteren, T., Englebienne, G., Kröse, B.J.: Transferring knowledge of activity recognition across sensor networks. In: Pervasive computing, pp. 283–300. Springer (2010)
17. Kendon, A.: Conducting interaction: Patterns of behavior in focused encounters, vol. 7. CUP Archive (1990)
18. Lee, K., Ashton, M.C.: Psychometric properties of the hexaco personality inventory. Multivariate Behavioral Research **39**(2), 329–358 (2004)
19. Madan, A., Cebrian, M., Lazer, D., Pentland, A.: Social sensing for epidemiological behavior change. Proceedings of the 12th ... (2010). URL http://dl.acm.org/citation.cfm?id=1864394
20. Martella, C., Gedik, E., Cabrera-Quiros, L., Englebienne, G., Hung, H.: How was it?: Exploiting smartphone sensing to measure implicit audience responses to live performances. In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, pp. 201–210. ACM (2015)
21. Mast, M.S.: Dominance as expressed and inferred through speaking time. Human Communication Research **28**(3), 420–450 (2002). DOI 10.1111/j.1468-2958.2002.tb00814.x. URL http://dx.doi.org/10.1111/j.1468-2958.2002.tb00814.x
22. Matic, A., Osmani, V., Mayora, O.: Speech activity detection using accelerometer. In: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, pp. 2112–2115. IEEE (2012)
23. McNeill, D.: Language and gesture, vol. 2. Cambridge University Press (2000)
24. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
25. Olguín, D.O., Waber, B.N., Kim, T., Mohan, A., Ara, K., Pentland, A.: Sensible organizations: Technology and methodology for automatically measuring organizational behavior. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **39**(1), 43–55 (2009)
26. Preece, S.J., Goulermas, J.Y., Kenney, L.P., Howard, D.: A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE Transactions on Biomedical Engineering **56**(3), 871–879 (2009)
27. Ravi, N., Dandekar, N., Mysore, P., Littman, M.: Activity recognition from accelerometer data. AAAI pp. 1541–1546 (2005). URL http://www.aaai.org/Papers/IAAI/2005/IAAI05-013
28. Reynolds, D.: An overview of automatic speaker recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), S 4072, p. 4075 (2002)
29. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)
30. Sangineto, E., Zen, G., Ricci, E., Sebe, N.: We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In: Proceedings of the ACM international conference on multimedia, pp. 357–366. ACM (2014)
31. Schmidt, M., Roux, N.L., Bach, F.: Minimizing finite sums with the stochastic average gradient. arXiv preprint arXiv:1309.2388 (2013)
32. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schr¿der, M.: Bridgind the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. IEEE Transactions on Affective Computing (2012)
33. Wyatt, D.: Collective Modeling of Human Social Behavior. AAAI Spring Symposium: Human Behavior Modeling (2009). URL http://www.aaai.org/Papers/Symposia/Spring/2009/SS-09-04/SS09-04-015.pdf
34. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: Proceedings of the 15th international conference on Multimedia, pp. 188–197. ACM (2007)
35. Zen, G., Sangineto, E., Ricci, E., Sebe, N.: Unsupervised domain adaptation for personalized facial emotion recognition. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 128–135. ACM (2014)
36. Zhang, T., Wang, J., Liu, P., Hou, J.: Fall detection by embedding an accelerometer in cellphone and using kfd algorithm. International Journal of Computer Science and Network Security **6**(10), 277–284 (2006)
37. Zhao, Z., Chen, Y., Liu, J., Shen, Z., Liu, M.: Cross-people mobile-phone based activity recognition. In: IJCAI, vol. 11, pp. 2545–250. Citeseer (2011)