Likability Estimation of Call-center Agents by Suppressing Annotator Variability

Hosana Kamiyama, Atsushi Ando, Ryo Masumura, Satoshi Kobashikawa and Yushi Aono
NTT Corporation, NTT Media Intelligence Laboratories, Japan
E-mail: hosana.kamiyama.tf@hco.ntt.co.jp Tel: +81-46-859-4624

***Abstract*—This paper proposes a effective likability estimation technique for call-center agents. Most likability estimation models need numerous annotated speech samples to obtain high-quality training labels since the likability annotations often vary due to annotator disagreement. The performance of conventional likability estimation models is often poor since they do not adequately account for annotator variability which is the difference between each annotator's assessment reliability. Our approach suppresses the effect of annotator variability by taking into account the individual annotator's reliability, which is the probability of correctly assessing likability. To estimate target annotator-independent likability, we introduce a graphical model with annotator reliability and optimize the model by using the EM-algorithm. We also propose a new neural network architecture to improve the model's performance. The architecture has a layer that takes as input the probability of target annotator-independent likability and the probability of annotator reliability. To propagate the loss of likability estimation independent from annotator reliability, our proposal processes annotations via the proposed layer. Given just two annotations per call, our proposal yields better accuracy than either the baseline or conventional methods.**

I. Introduction

Evaluation of likability of agents in telephone conversations is important for improving call-center performance. Existing automatic techniques of evaluating call-center agents fail to sufficiently assess likability, and instead usually assess just the linguistic information (ex. words) of the conversations such as opening script, name confirmation, etc. [1], [2]. Therefore, the aim of this paper is to estimate agent likability, a subjective measure of the quality of agent's attitudes, from calls.

Estimation of subjective characteristics such as likability requires the determination of reliable ground truths since subjective values often differ with the annotator. Most studies settle the ground truth as unanimous assessments from many annotations. Interspeech 2012 Speaker Trait Challenge [3], [4] also used unanimous likability labels; they are included in the "Speaker Likability Database (SLD)" [5]. In SLD, 32 people annotated the likability label of each speech utterance (each a few seconds long). Most likability estimation studies based on machine learning use the unanimous labels of SLD as ground truth, which are binary labels such as "*likable*" or, "*non-likable*" [6], [7], [8], [9]. Soft-label was proposed to handle the case of unanimous annotations not being available, [10], [11]. Assuming that even annotations that are not unanimous also have information important for estimating class, soft-label uses the rate of annotated classes as an objective function available for machine learning. Since soft-label can use all annotations, it is more efficient than using only unanimous labels and is more reliable when collecting annotations for machine learning.

Though the conventional method improves the performance by using the rate of annotated classes, the problem of annotator variability, which is the difference between each annotator reliability. Each annotator has different experience in evaluating agent likability, so some assessments given by annotators who have limited experience are likely to include errors. To suppress annotator variability when using the conventional method, we need numerous annotators to label each call. Soft-labels of numerous annotations reduce the impact of annotation errors. However, this approach is not feasible due to annotation costs, i.e. likability evaluation of calls takes more time than that of short utterances and annotators must be experts (supervisor) in call-center operation. In evaluating agent likability according to telephone responses gathered in actual call-centers, the likability labels assigned by annotators will have the following characteristics:

- Only a few likability labels per call are available (one or two labels per call).
- Annotators are randomly assigned to calls.

Although approaches that identify the reliable annotators by annotation testing have been proposed [12], [13], determining ground truths of ambiguous calls by collecting many annotations is also not feasible. This resulting evaluation variability significantly degrades the accuracy of machine learning.

This paper proposes two training methods for likability estimation by using annotations that are assumed to contain annotator variability. Both proposals suppress annotator variability which is the difference in each annotator's reliability based on the probability of each annotator correctly assessing likability. The first proposed method estimates target annotator-independent likability and each annotator's reliability by utilizing all available annotations. To estimate target annotator-independent likability, we introduce a graphical model that assumes that annotated labels are generated from the latent variable of annotator-independent likability and annotator reliability. The expectation of annotator-independent likability given by the EM-algorithm based on [14], [15], [16] is used for likability estimation model training. The second proposed method structures a neural network so as to suppress each annotator's variability by marginalization. The network has a

layer that takes as inputs the probability of target annotator-independent likability and the probability of annotator reliability, and outputs the probability of label annotation as marginal probability. To propagate losses of the likability estimation model independent of annotator reliability, annotated labels are processed via the proposed layer. We conduct experiments to rank our proposals against previous techniques.

II. Likability Estimation

This section describes the task description, the baseline and the conventional methods of likability estimation.

A. Task description

The task of this paper is call-level likability estimation of call-center agents. The classes are *likable* and *non-likable*. *likable* means a call with better impression than usual telephone responses. *non-likable* means a call with usual or worse impression. Estimation of likability uses features of conversations between agents and customers as follows:

$$\hat{y} = \arg\max_{y \in C} p(y|\mathbf{X}, \mathbf{\Theta}) \tag{1}$$

where $\hat{y}$ is estimated likability, C is the set of likability classes, and we set $C = \{0, 1\}$ (Class 1 means *likable* and Class 0 means *non-likable*), $\mathbf{X}$ represents the features of the call (for example, turn-wise acoustical features such as mean of fundamental frequency F_0 of agents utterances, dialogue features such as the frequency of backchannel of agents while customers are uttering and language features such as count of thank word) , and $p(y|\mathbf{X}, \mathbf{\Theta})$ is a likability estimation model that outputs the posterior probability of likability, y, given $\mathbf{X}$ and $\mathbf{\Theta}$. In this paper, the training methods of model $p(y|\mathbf{X}, \mathbf{\Theta})$ are different between the baseline, conventional and our two proposed methods. The models are based on neural networks.

Our task is to find those calls that are either *likable* or *non-likable* for everyone in order to improve call center performance. Therefore, we define ground truths of likability estimation as unanimous labels by all annotators when many labels (for example, larger than or equal to four labels) per call are available. However, many calls have only few labels (for example, one or two labels), and the labels may differ depending on the annotators in actual call centers. The baseline, the conventional and the proposed methods differ in how to use such existing labels in actual call centers.

B. Baseline method

The baseline method trains labels that match from labels per call, similar to the determination of ground truths. To train the likability estimation model, the parameter set of model $\mathbf{\Theta}$ is calculated so as to minimize the following formula based on cross-entropy loss:

$$\mathcal{L} = -\sum_{c \in C} t_{i,c} \log p(y_i = c|\mathbf{X}_i, \mathbf{\Theta}) \tag{2}$$

where i is sample number, $\mathcal{L}$ is a loss function, c is the class of likability, and $t_{i,c}$ is the target value of model $p(y_i = c|\mathbf{X}_i, \mathbf{\Theta}$, and the c-th element of one-hot vector of objective function y_i (for example, when $y_i = 1$, $t_{i,0} = 0$ and $t_{i,1} = 1$).

To obtain high-quality labels, previous studies accepted only labels that received unanimous decisions [5], i.e., when all annotators said one call was *likable*, the label of the call was *likable* and when all annotators said *non-likable*, the label was *non-likable*. When the annotators set different labels, the call was excluded from the training data. Label determination can be formulated as follows:

$$t_{i,c} = \begin{cases} 1 & (\text{if } \forall j \in J_i, y_i^{(j)} = c) \\ 0 & (\text{otherwise}) \end{cases} \tag{3}$$

where j is annotator number, J_i is the set of annotators who annotated the i-th sample and $y_i^{(j)}$ is the label decided by the j-th annotator, for the i-th sample. When all annotators j unanimously assessed the i-th sample to have likability c, the target $(t_{i,0}, t_{i,1})$ is a one hot vector based on class c. When the annotators assess different likability labels of the i-th sample, target $t_{i,c}$ is equal to 0 regardless of class c and the i-th sample are not used for training due to loss $\mathcal{L} = 0$.

C. Soft-label (conventional method)

Few speech samples achieve uniform likability assessments from all annotators. Assuming that the speech has also information for estimation of class, soft-labels are proposed for emotion recognition as they allow the use of ambiguous annotations [10], [11]. The target is calculated from annotated labels as follows:

$$t_{i,c} = \frac{\sum_{j \in J_i} t_{i,c}^{(j)}}{|J_i|} \tag{4}$$

where $t_{i,c}^{(j)}$ is the c-th element of the one-hot vector of annotated label $y_i^{(j)}$ and $|J_i|$ is the number of annotators in set J_i. The label equals the ratio of annotated labels for each call in this task. For example, when annotators $J_i = \{1, 2, 3\}$ exist for assessment of the i-th sample and they evaluate $y_i^{(1)} = 1$, $y_i^{(2)} = 1$ and $y_i^{(2)} = 0$, the target is $(t_{i,0}, t_{i,1}) = (1/3, 2/3)$. This method of determining labels allows us to use all annotated labels.

D. Problem of baseline and conventional methods

The performance of the baseline and conventional methods are often poor since they do not adequately account for annotator variability, which is the difference between each annotator's reliability. When some annotators have limited experience in assessing agent likability, the conventional methods often fail to output high-quality labels. This is most obvious when only a few annotations per call are available from an actual call-center.

The previous task is likability estimation of short speech samples, such as commands. Since likability labeling cost is not large, we can acquire a lot of labels for short utterances. Fig. 1 (a) shows the target labels when processing all annotators' labels for all calls. Although the fourth annotator in Fig. 1 (a) gives some labels that are different from other

Call i / Annotator j	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	0	1	-	-
$j=2$	0	-	-	1
$j=3$	-	1	0	-
$j=4$	-	-	1	0
Baseline $t_{i,1}$	**0**	**1**	**-**	**-**
Soft-label $t_{i,1}$	**0.00**	**1.00**	**0.50**	**0.50**
Full-annotated soft-label $t_{i,1}$	**0.00**	**1.00**	**0.25**	**0.75**

(b) Actual

Fig. 1. Ideal and actual annotations and target "likable" labels $t_{i,1}$ (Label "0" and "1" represent *non-likable* and *likable*, respectively. Decimal values of 0 to 1 indicate closeness to the *non-likable* and *likable* classes). Ideal labeling method annotates all samples by all annotators. In actual operation, we cannot obtain complete coverage of all calls and all annotators. By dividing the calls among the annotators, we can get only a few annotations per call.

annotators, the soft-labels reduce the impact of the difference in labels. It takes longer time to annotate long speech calls than to annotate short speech calls. In addition, annotators should be supervisors who have expert skill. Therefore, likability labeling costs are so high that it is not feasible to acquire many labels per call. In practice, only one or two supervisors label randomly assigned calls as shown in Fig. 1 (b). Since we can obtain only a few annotations per call, the training labels may include erroneous likability labels. In Fig. 1 (b), the soft-labels allow larger impact of the fourth annotator than the full-annotated soft-labels which is the same with soft-label of Fig. 1 (a). Therefore, it is necessary to obtain reasonable labels or to train a likability estimation model from just a few actual labels.

III. Proposed methods

This section describes the proposed methods; they train likability estimation models from just a few actual labels.

A. Approach

Our approach suppresses the effect of annotator variability by taking into account the individual annotator's reliability, which differs with regard to correctly assessing likability. Soft-label method assumes the all annotators exhibit the same variability in terms of likability annotation. However, actual annotators do not have the same variability, i.e., some annotators can evaluate likability correctly in accordance with call-center criteria while other annotators cannot. The conventional methods fail to adequate consider the difference in annotator variability. We propose two approaches. The first approach estimates target $t_{i,c}$ independent of annotator variability . Our method trains a likability estimation model by using the expectation of annotator-independent likability as an objective function. Our second approach propagates the suppressed loss to the likability estimation model $p(y_i = c|\mathbf{X}_i, \mathbf{\Theta})$ from annotator variability. The approach wraps the likability estimation model $p(y_i = c|\mathbf{X}_i, \mathbf{\Theta})$ under the layer that simulates annotator variability. When propagating, model $p(y_i = c|\mathbf{X}_i, \mathbf{\Theta})$ is expected to be trained independent of the layer simulating annotator variability. The rest of this section describes the methods.

B. Proposed1: Estimation of annotator-independent likability

To calculate annotator-independent target $t_{i,c}$, we assume the graphical model shown in Fig. 2. The model generates annotated likability $y_i^{(j)}$ from annotator-independent likability y_i, which is a latent variable, and annotator reliability parameters. Our approach estimates annotator-independent likability y_i from existing labels $y_i^{(j)}$. We assume that labels $y_i^{(j)}$ and y_i are generated in accordance with a Bernoulli distribution. The graphical model is formulated as follows:

$$y_i^{(j)} \sim p(y_i^{(j)}|y_i, \alpha_j, \beta_j, q) \quad (5)$$

$$= p(y_i^{(j)}|y_i, \alpha_j, \beta_j)p(y_i|q) \quad (6)$$

$$p(y_i|q) = \texttt{Bernoulli}(y_i|q) \quad (7)$$

$$p(y_i^{(j)}|\alpha_j, \beta_j, y_i = 1) = \texttt{Bernoulli}(y_i^{(j)}|\alpha_j) \quad (8)$$

$$p(y_i^{(j)}|\alpha_j, \beta_j, y_i = 0) = \texttt{Bernoulli}(y_i^{(j)}|\beta_j) \quad (9)$$

where Bernoulli$(x|\theta)$ is Bernoulli distribution formulated as $\theta^x(1-\theta)^{1-x}$, y_i is likability of the i-th call, α_j and β_j are the parameters of the j-th annotator reliability, which is the probability that annotators correctly answer *likable* or *non-likable*; q is a class distribution parameter that generates y_i.

From annotated likability $y_i^{(j)}$, we calculate parameters q, α_j and β_j based on the EM-algorithm [14], [15], [16]. After estimating the parameters, we estimate the expectation of latent variables $E[y_i = c]$. $E[y_i = c]$ which is independent from annotator j are used for target $t_{i,c}$. The target $t_{i,c}$ and expectation of likability $E[y_i = c]$ is calculated as follows:

$$t_{i,0} = E[y_i = 0] \propto (1-q)\prod_{j\in J_i} p(y_i^{(j)}|\alpha, \beta, y_i = 0) \quad (10)$$

$$t_{i,1} = E[y_i = 1] \propto q\prod_{j\in J_i} p(y_i^{(j)}|\alpha, \beta, y_i = 1) \quad (11)$$

The expectations $E[y_i = c]$ are obtained under the constraint of $\sum_{c\in C} E[y_i = c] = 1$.

C. Proposal2: Fine-tuning via marginalizing layer

Our second approach structures a neural network so as to include a layer that stochastically simulates the variability of annotators. When optimizing the neural network, annotator-independent loss is propagated to the likability estimation model via the layer by using annotated likability. The proposed loss function can be formulated as follows:

$$\mathcal{L} = -\sum_{c\in C} t_{i,c}^{(j)} \log p(y_i^{(j)} = c|\mathbf{X}_i, \mathbf{\Theta}, j) \quad (12)$$

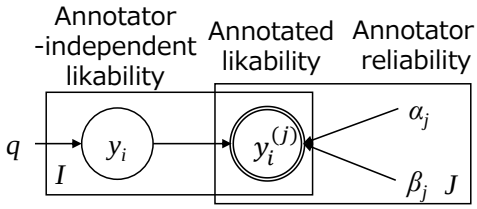

Fig. 2. Proposed generation model

where $\mathcal{L}$ is the loss of annotated likability. The objectiv function of the proposal is annotated likability.

Our proposed network is shown in Fig. 3. The netwo estimates annotated likability via the "Marginalizing layer whose input is the probability of call likability and t probabilities of the annotator reliability that judge, correctly or incorrectly, likability when call likability is *likable* or *non-likable*. The layer calculates a marginal probability from the joint probability of call likability and annotator reliability. The marginalizing layer is formulated as follows:

$$p(y_i^{(j)}|\mathbf{X}_i, \mathbf{\Theta}, j) = \sum_{y_i \in C} p\left(y_i^{(j)}|y_i, j\right) p(y_i|\mathbf{X}_i, \mathbf{\Theta}) \quad (13)$$

where $p(y_i|\mathbf{X}_i, \mathbf{\Theta})$ is the probability of the i-th annotator-independent likability estimated by the likability estimation model and $p\left(y_i^{(j)}|y_i, j\right)$ is the probability of the j-th annotator's reliability.

Converting annotator number j into the probability of annotator reliability is realized by the following layers:

$$\begin{aligned} p(y_i^{(j)}|y_i = 1, j) &= (\alpha_j, 1-\alpha_j) && (14)\\ &= \texttt{SOFTMAX}(\texttt{EMBED}(j; \mathbf{A})) && (15)\\ p(y_i^{(j)}|y_i = 0, j) &= (1-\beta_j, \beta_j) && (16)\\ &= \texttt{SOFTMAX}(\texttt{EMBED}(j; \mathbf{B})) && (17) \end{aligned}$$

where EMBED() is a linear transformational function to embed a symbol into a continuous vector, SOFTMAX() is a linear transformational function in () with softmax activation, $\mathbf{A}$ and $\mathbf{B}$ are the trainable parameters of EMBED and to give annotator reliability α_j, β_j, respectively. When estimating call likability, we use just the likability estimation model; the Marginalizing layer is used only for network fine-tuning.

D. Comparison of methods

The differences between the baseline and soft-label (conventional) and proposed methods are shown in Table I. Baseline, which uses unanimous annotations, does not consider variability in training. Baseline uses the dominant label for training the likability estimation model. The objective function $t_{i,c}$ yields binary value, i.e., $t_{i,c} \in \{0, 1\}$ according to Eq.(2) and Eq. (3). Both soft-label and the first proposal suppress label variability in training and calculate expectations, but our method also considers annotation variability. Soft-label method is nothing more than taking the mean of the annotations of each call as the expectation as indicated in Eq. (4). To suppress annotator variability, our proposal calculates the expectation of

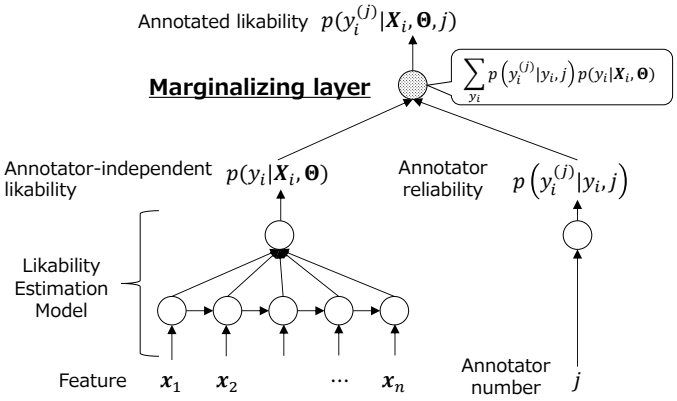

Fig. 3. The proposed network

TABLE I
DIFFERENCES OF METHODS

Methods	Label variability	Annotator variability	Supression of variability
Baseline	-	-	(Unanimous)
Conventional	✓	-	Expectation
Proposed1	✓	✓	Expectation
Proposed2	✓	✓	Marginalization

annotator-independent likability by also using parameters of annotator reliability α_j, β_j.

Our second proposal suppresses annotator variability by the marginalizing layer as indicated in Eq. (13). The layer takes as inputs the probability of annotator-independent likability and annotator reliability, and suppresses annotator variability when propagating estimated losses.

IV. EXPERIMENTS

To evaluate the proposed methods, we conducted likability estimations using actual call-center data.

A. Dataset

We used a dataset gathered from a telephone skill test, which was developed to evaluate the skill of actual call-center agents. The test setting was taken to be a corporate call-center. The recorded calls were evaluated according to a list of evaluation items, such as likability by 269 annotators skilled in call-center supervision. The evaluation items were scored using a 5-point scale (1: poor - 5:excellent). All calls were evaluated by two annotators. With regard to the evaluation item of likability, about 90% of the labels (5-point scale) were assigned scores of three or four. In this experiment, we binarized the scores into two bins of *likable* or *non-likable*: three or less, four or more, respectively. According to evaluation criterion of the telephone skill test, the level of *likable* means a call with better likability than usual telephone responses; that of *non-likable* means a call with usual or lower likability.

The telephone skill test contained 4,765 calls with total duration of about 230 hours. Most recorded calls were monaural. We manually separated the stereo calls to place agent and customer in different channels. Parts indicative of utterance overlap were excluded.

We obtained 1,417 stereo calls that were given the same binary label by two annotators. In order to allow comparison with baseline and conventional methods, one or two other annotators relabeled some calls. Finally, we obtained two datasets for use in the experiment as shown in Table II. 340 calls with unanimously assigned values from four annotators were used as the test set in our experiment. The sampling rate was 8 kHz.

B. Setup

44 dimensional acoustic, dialogue and linguistic features based on [17] were extracted as turn-level features. LSTMs with attention mechanism were used in the likability estimation model. The model takes turn-level features as input and outputs the posterior probability of agent likability. The model was LSTM with an attention layer and fully-connected layers with softmax function. The number of hidden units was 4 in the LSTM layer. Optimizer was Adam with the learning rate of 0.001; the dropout rate was 50%.

First, we calculated the expectation of annotator-independent likability for the proposed method from all annotations of 4,765 calls, i.e. the complete telephone skill test. Next, we measured performance in terms of accuracy by subjecting the stereo calls to 10-fold cross validation. We used 1/10 of the samples with four unanimous annotations as the test set. The training set for the baseline used only unanimously labels assigned by two, three or four annotators. The training set for the conventional and proposed methods also used labels assigned by two, three or four or less annotators. We trained all models 5 times in each condition and compared accuracy using the highest performance.

C. Results and discussion

Table III shows the accuracy of each training set. In the case of two agreed labels, the proposed methods achieved the highest performance in our experiment. Our methods yielded higher accuracy than the baseline and the conventional methods regardless of the number of annotations. Applying just the first proposed method, which suppress annotator variability based on the graphical model, yielded better accuracy than the baseline and conventional methods. This shows that the first proposed method is effective in suppressing annotator variability. Since applying the second proposal, fine-tuning the model via a marginalizing layer, yields better performance than applying only the first method, it is shown that the second proposal is also effective.

The conventional method is not sufficiently better than the baseline method in the same case. This may indicate that the soft-labels were impacted from erroneous likability annotations. The conventional method is effective for the task that allows multiple classes, such as emotion recognition [10], [11]. Since classes of *likable* and *non-likable* are conflicted in our task, soft-labels may not work well due to annotation errors.

TABLE II
DATASET ("ANNOT." MEANS "ANNOTATORS")

# of Annot.	# of Speech	# of Agreed speech	Dataset
≥ 2	1,417	1,369	Train
≥ 3	604	528	
4	454	340	Train & Test

TABLE III
ACCURACY OF LIKABILITY ESTIMATION.

# of Annot.	2	3	4
Baseline	.762	.756	.776
Conventional	.765	.756	.762
Proposed1	.782	.782	.779
Proposed1 + 2	.808	.791	.785

V. CONCLUSIONS

This paper proposed a new likability estimation technique for call-center agents. Most likability estimation models need numerous annotated speech samples to obtain high-quality training labels since the likability annotations often vary due to annotator disagreement. The performance of conventional likability estimation models is often poor since they do not adequately account for annotator variability; this is most obvious when there are few annotators per call as is true in actual call centers. Our approach suppresses annotator variability by assuming the factor of annotator reliability, which is the probability of correctly judging likability. To estimate target annotator-independent likability, we introduced a graphical model with annotator reliability and optimized the model by using the EM-algorithm. We also proposed a new neural network architecture to improve the model's performance. The architecture has a layer that takes as input the probability of speech likability and the probability of annotator reliability. To propagate losses of likability estimation model independent of annotator reliability, our proposal processes annotated labels via the proposed layer. Given just two annotations per call, our proposal yields better accuracy than the baseline and conventional methods.

REFERENCES

[1] G. Zweig, O. Siohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu, and B. Kingsbury, "Automated quality monitoring for call centers using speech and NLP technologies," in *Proc. NAACL HLT*, 2006, pp. 292–295.

[2] S. Roy, R. Mariappan, S. Dandapat, S. Sricastave, S. Galhotra, and B. Peddamuthu, "QART: A system for real-time hoilistic quality assurance for contact center dialogues," in *Proc. AAAI*, 2016, pp. 3768–3775.

[3] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. v. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The interspeech 2012 speaker trait challenge," in *Proc. Interspeech*, 2012, pp. 254–257.

[4] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, R. v. S. F. Burkhardt, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A survey on perceived speaker traits: Personality and likability and pathology and the first challenge," in *Computer Speech and Language*, vol. 29, 2015, pp. 100–131.

[5] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, ""would you buy a car from me?" – on the likability of telephone voices," in *Proc. Interspeech*, 2011, pp. 1557–1560.

[6] J. Pohjalainen, S. Kadioglu, and O. Räsänen, "Feature selection for speaker traits," in *Proc. of Interspeech*, 2012, pp. 270–273.

[7] D. Wu, "Genetic algorithm based feature selection for speaker trait classification," in *Proc. of Interspeech*, 2012, pp. 294–297.

[8] H. Buisman and E. Postma, "The log-gabor method: Speech classification using spectrogram image analysis," in *Proc. of Interspeech*, 2012, pp. 518–521.

[9] C. Montacié and M. Carat, "Pitch and intonation contribution to speakers' traits classification," in *Proc. of Interspeech*, 2012, pp. 526–529.

[10] H. M. Fayek, M. Lech, , and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *Proc. IJCNN*, 2016, pp. 566–570.

[11] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of all things the measure is man: Automatic classification of emotions and inter-labeler consistency," in *Proc. ICASSP*, 2005, pp. 317–320.

[12] X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, , and M. Zhang, "CDAS: A crowdsourcing data analytics system," in *Proc. PVLDB*, vol. 5, 2012, pp. 1040–1051.

[13] S. Hantke, E. Marchi, and B. Schuller, "Introducing the weighted trustability evaluator for crowdsoucing exemplified by speaker likability classification," in *Proc. LREC*, 2016, pp. 2156–2161.

[14] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," in *Journal of the Royal Statstical Society. Series C*, vol. 28, 1979, pp. 20–28.

[15] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: optimal integration of labels from labelers of unknown expertise," in *Proc. NIPS*, 2009, pp. 2035–2043.

[16] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima, "Accurate integration of crowdsouced labels using wokers' self-reported confidenxe scores," in *Proc. IJCAI*, 2013, pp. 2554–2560.

[17] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Hierarchical LSTMs with joint learning for estimating customer satisfaction from contact center calls," in *Proc. Interspeech*, 2017, pp. 1716–1720.