SOFT LINEAR DISCRIMINANT ANALYSIS (SLDA) FOR PATTERN RECOGNITION WITH AMBIGUOUS REFERENCE LABELS: APPLICATION TO SOCIAL SIGNAL PROCESSING

Patrick Meyer, Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
38106 Braunschweig, Germany
{meyer,fingscheidt}@ifn.ing.tu-bs.de

ABSTRACT

While most pattern recognition approaches are designed and trained with clearly defined reference labels, there are a few new applications working with ambiguous ones. Since the linear discriminant analysis (LDA) is one of the most utilized methods in pattern recognition to reduce the dimensionality of feature vectors, typically increasing the robustness of the features, we propose in this work a modification of the LDA in order to be able to handle ambiguous reference labels in a soft-decision way. In the field of social signal processing (here: emotion recognition) we demonstrate that using a soft accuracy measure evaluating the classifier's confidence output by means of a soft-labeled emotional speech database really provides a degree of *similarity to (naturally ambiguous) human votes*. The adaptation of our classifier to such soft accuracy measure takes place by a retraining w.r.t. the human vote distribution. Applying this soft accuracy measure to emotion recognition with ambiguous reference labels both retraining the classifier and using the new soft LDA method leads to around 22% relative increase of accuracy.

Index Terms— linear discriminant analysis, ambiguous labels, ground truth, social signal processing, emotion recognition

1. INTRODUCTION

In pattern recognition two general approaches are distinguished: Supervised and unsupervised classification [1]. The first method classifies a pattern in accordance with a number of predefined classes, whereas the second approach has the objective to at first form appropriate classes itself and allocate the patterns to these so-far unknown classes. For both of them it is often the case that humans are able to label or classify the considered patterns unambiguously.

However, in recent years a new field of pattern recognition has received more and more attention: Social signal processing. In contrast to most other application fields of pattern recognition, this field deals with human interpretations of the behavior of people regarding their interactions (e.g., addressing, active listening), internal states (e.g., interest, emotions), personality (e.g., dominance) or social relations [2, 3]. This field of pattern recognition differs in one specific aspect from more classical recognition applications: It reveals considerable ambiguity in class assignment by humans. Since the perception and interpretation of human behavior varies from person to person, it is sometimes impossible to obtain an unambiguous ground truth or labels for this discipline.

Taking speech emotion recognition as an example, we find various studies about listeners' different perception of emotions from recorded speech, e.g., [4, 5]. Nevertheless, besides the consideration of emotions in a multidimensional continuous emotion space [6], the most familiar method of describing emotions is still to use a small number of discrete classes [6–8]. For this purpose, speech emotion databases are generally labeled by the *majority vote* of a number of human raters who evaluate the observed emotions by means of predefined classes. Accordingly, the classifiers are trained and evaluated with these henceforth called *hard decision* (HD) emotion labels, or *hard ground truth*.

Due to the fact that the interpretation of humans represents the only true reference labels for social signal processing, in previous work [9] we proposed a new evaluation methodology: By means of an emotional speech database with emotions simulated by actors, we were able to confirm with a perception test that listeners' perception of emotions from recorded speech varies like observed in [4]. Hence, we created a *soft decision* (SD) or *soft ground truth* consisting of the probability distribution of all votes of all raters and defined a new *soft accuracy measure* for such social signal processing approaches (see Section 2). This evaluates the recognition results not on the basis of a single majority vote, but effectively takes into account all votes of all raters in order to obtain a more human-like evaluation methodology. However, to obtain a smaller and more robust feature vector a conventional linear discriminant analysis (LDA) was employed, still reflecting the majority votes of the class labels.

The conventional LDA has already been used in a wide range of fields with different modifications and extensions. In automatic speech recognition, several approaches for class assignment exist, since this is not obvious in the case of employing an LDA together with continuous hidden Markov models and Gaussian mixture models [10–14]. Furthermore, in [15] an adaptive LDA is proposed, which considers individual class distributions, whereas other approaches focus on one specific kind of class distribution (e.g., on a heteroscedastic class distribution [16]). In the field of face recognition, a local LDA was proposed to handle multi-class nonlinear classification problems by applying a set of locally linear transforms. In this process, a *soft clustering* takes place whereby each data point belongs to each of the clusters with some posterior probability [17].

In order to break completely away from a hard ground truth, we pick up the idea of soft clustering [17] and propose in this paper a soft linear discriminant analysis (SLDA), which is part of a two-step approach towards processing ambiguous classes: First, we apply a neural network (NN), taking the probability distribution of our new soft ground truth as target signal in training. Moreover, all outputs of the NN are considered and adjusted as posterior probability distribution for the evaluation with the soft accuracy measure. Second, we derive and employ our SLDA, which analyzes the features not any longer on the basis of the hard ground truth, but on the soft one. This enables us to employ both training and test of the NN classifier in a soft-decision fashion, and also evaluation is performed by means of a soft accuracy measure. As a result we obtain a more human-like labeling, classification, and evaluation.

The rest of the paper is organized as follows: Accuracy definitions for both hard and soft decision are revisited in Section 2. This is followed by the mathematical model of the suggested soft linear discriminant analysis in Section 3. Section 4 describes an example application in the field of emotion recognition. An evaluation of the old and the new training method as well as the LDA and the SLDA with respect to the soft accuracy measure is examined in Section 5. Finally, conclusions are drawn in Section 6.

2. ACCURACY DEFINITIONS

In order to deal with ambiguous reference labels, we make use of a particular accuracy measure [9], which we will briefly revisit here.

Still in many applications of pattern recognition there exists a fixed ground truth, which consists of an unambiguous assignment to a class. For that reason one often targets a hard decision output of the classifier, which is then compared to the ground truth. It results in the *hard accuracy measure*

$$\mathrm{ACC}^{\mathrm{HD}} = \mathrm{ACC} = 1 - \frac{E}{R}\ , \tag{1}$$

with $E = S + I + D$ being the sum of substitutions, insertions, deletions, respectively, and R being the number of instances in the ground truth to be recognized. Note that in sentence-based emotion recognition we often have $E = S$, and $I = D = 0$.

However, as already mentioned in Section 1, there is an increasing number of applications that do not have a clearly defined ground truth with hard decisions. In accordance with the human reference it requires a soft evaluation method considering classifier confidence outputs and comparing them to the soft ground truth. This leads then to the *soft accuracy measure* [9]

$$\mathrm{ACC}^{\mathrm{SD}} = 1 - \frac{1}{2R}\sum_{t=1}^{R}\sum_{i=1}^{N}\left|\hat{\mathrm{P}}_{t,i} - \mathrm{P}_{t,i}\right|, \tag{2}$$

whereby the posterior probability $\hat{\mathrm{P}}_{t,i}$ characterizes the classifier's confidence output in terms of the recognized instance $t \in \{1, 2, \ldots, R\}$ and class $i \in \{1, 2, \ldots, N\}$ with N being the number of classes, while $\mathrm{P}_{t,i}$ describes the raters' distribution for instance t and class i, fulfilling $\sum_{i=1}^{N} \mathrm{P}_{t,i} = 1$.

3. (SOFT) LINEAR DISCRIMINANT ANALYSIS

In the following section, we describe the newly proposed SLDA, sketching first the classical LDA according to [18].

The LDA [18, 19] pursues the objective of reducing the dimensionality from a given feature vector in consideration of the best discrimination among N defined classes. This is realized by using a linear transformation from vector $\mathbf{x} \in \mathbb{R}^{d_x}$ to vector $\mathbf{y} \in \mathbb{R}^{d_y}$, whereby $d_y < d_x$. This mathematical operation can be expressed by

$$\mathbf{y} = \mathbf{A}^{\mathsf{T}}\mathbf{x}, \tag{3}$$

with $\{\}^{\mathsf{T}}$ being the transpose of the rectangular matrix $\mathbf{A} \in \mathbb{R}^{(d_x \times d_y)}$, which column vectors are linearly independent. In order to formulate a criterion of class separability, some statistical scatter matrices (within-class, between-class, or mixture scatter matrix) are used, which consist of two basic statistical definitions: The mean value and the covariance. The global mean value is given by

$$\boldsymbol{\mu} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t, \tag{4}$$

with the feature vector $\mathbf{x}_t, t \in \{1, 2, \ldots, T\}$, and T being the number of instances in some training data, and the class-specific mean value

$$\boldsymbol{\mu}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{x}_{t,i}, \tag{5}$$

with T_i being the number of instances of class i. By means of the class-assigned feature vector $\mathbf{x}_{t,i}$ the class-specific covariance matrix can be written as

$$\boldsymbol{\Sigma}_i = \frac{1}{T_i - 1}\sum_{t=1}^{T_i}(\mathbf{x}_{t,i} - \boldsymbol{\mu}_i)\cdot(\mathbf{x}_{t,i} - \boldsymbol{\mu}_i)^{\mathsf{T}}. \tag{6}$$

Note, that the normalization term $T_i - 1$ provides an unbiased estimate [18]. Furthermore, we need the probability of class i, simply expressed by

$$\mathrm{P}_i = \frac{T_i}{T}. \tag{7}$$

With the aid of these components the three desired scatter matrices can be defined: The within-class scatter matrix

$$\mathbf{S}_{\mathrm{w}} = \sum_{i=1}^{N}\mathrm{P}_i \cdot \boldsymbol{\Sigma}_i \tag{8}$$

describes the distribution of the samples around their class-specific mean value, whereby the number of classes is characterized by N. In contrast, the distribution of the class-specific mean values around the global mean value is represented by the between-class scatter matrix, which is expressed by

$$\mathbf{S}_{\mathrm{b}} = \sum_{i=1}^{N}\mathrm{P}_i \cdot (\boldsymbol{\mu}_i - \boldsymbol{\mu})\cdot(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathsf{T}}. \tag{9}$$

Finally, the addition of these two scatter matrices results in the mixture scatter matrix

$$\mathbf{S}_{\mathrm{m}} = \mathbf{S}_{\mathrm{w}} + \mathbf{S}_{\mathrm{b}} = \frac{1}{T - 1}\sum_{t=1}^{T}(\mathbf{x}_t - \boldsymbol{\mu})\cdot(\mathbf{x}_t - \boldsymbol{\mu})^{\mathsf{T}}, \tag{10}$$

which is the covariance matrix of all samples regardless of the class assignments.

A criterion for class separability based only on the first two scatter matrices is given by

$$J_1 = \mathrm{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b) \rightarrow \max, \tag{11}$$

with tr() being the trace of a quadratic $(n{\times}n)$-matrix. In order to obtain well-separated classes, the goal is to maximize the between-class distances while minimizing the within-class distances. This problem of optimization can be solved by means of the eigenvalue problem by which we obtain $N - 1$ non-zero eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$ [18]. It follows that we can map any $\mathbf{x_t}$ onto an $(N{-}1)$-dimensional subspace spanned by the $N - 1$ eigenvectors corresponding to the eigenvalues that are non-zero. So far the state of the art.

Now we assume to have available soft reference labels. As a result, a kind of class combination depending on the distribution of the labels over the classes is possible and the more the distribution over the classes is equal, the more is the separability between the classes reduced.

For the derivation from the classical LDA to the new soft LDA (SLDA) we expand the class-specific terms by the probability of the class assignment. In the special case of unambiguous reference labels, the SLDA shall correspond to the classical LDA. Now, the determination of the soft-class specific mean value considers all feature vectors weighted with their probability belonging to class i. This can be expressed by (compare to (5))

$$\boldsymbol{\mu}_i^{\mathrm{SD}} = \frac{\sum\limits_{t=1}^{T} \mathbf{x}_t \cdot \mathrm{P}_{t,i}}{\sum\limits_{t=1}^{T} \mathrm{P}_{t,i}}, \tag{12}$$

with the soft ground truth $\mathrm{P}_{t,i}$. The number of class instances T_i is replaced by the sum over all $\mathrm{P}_{t,i}$. Consequently, the soft probability of class i changes to (compare to (7))

$$\mathrm{P}_i^{\mathrm{SD}} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{P}_{t,i}. \tag{13}$$

The last modification affects the class-specific covariance matrix (6). Its soft equivalent results in

$$\boldsymbol{\Sigma}_i^{\mathrm{SD}} = \frac{\sum\limits_{t=1}^{T} \mathrm{P}_{t,i} \cdot (\mathbf{x}_t - \boldsymbol{\mu}_i^{SD}) \cdot (\mathbf{x}_t - \boldsymbol{\mu}_i^{SD})^{\mathsf{T}}}{-1 + \sum\limits_{t=1}^{T} \mathrm{P}_{t,i}}. \tag{14}$$

Since the global mean value does not depend on the distribution of the classes, it corresponds to the soft global mean value:

$$\boldsymbol{\mu}^{\mathrm{SD}} = \sum_{i=1}^{N} \mathrm{P}_i^{\mathrm{SD}} \cdot \boldsymbol{\mu}_i^{\mathrm{SD}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t = \boldsymbol{\mu} \tag{15}$$

Now all necessary definitions are given and we obtain the modified *soft scatter matrices*, whereby here again the mixture scatter matrix does not change, since it does not depend on the distribution of the classes:

$$\mathbf{S}_{\mathrm{w}}^{\mathrm{SD}} = \sum_{i=1}^{N} \mathrm{P}_i^{\mathrm{SD}} \cdot \boldsymbol{\Sigma}_i^{\mathrm{SD}} \tag{16}$$

$$\mathbf{S}_{\mathrm{b}}^{\mathrm{SD}} = \sum_{i=1}^{N} \mathrm{P}_i^{\mathrm{SD}} \cdot (\boldsymbol{\mu}_i^{\mathrm{SD}} - \boldsymbol{\mu}) \cdot (\boldsymbol{\mu}_i^{\mathrm{SD}} - \boldsymbol{\mu})^{\mathsf{T}} \tag{17}$$

$$\mathbf{S}_{\mathrm{m}}^{\mathrm{SD}} = \mathbf{S}_{\mathrm{w}}^{\mathrm{SD}} + \mathbf{S}_{\mathrm{b}}^{\mathrm{SD}} = \frac{1}{T} \sum_{t=1}^{T} (\mathbf{x}_t - \boldsymbol{\mu}) \cdot (\mathbf{x}_t - \boldsymbol{\mu})^{\mathsf{T}} = \mathbf{S}_{\mathrm{m}}. \tag{18}$$

4. APPLICATION: EMOTION RECOGNITION

One example field of application for the proposed SLDA is automatic speech emotion recognition. We briefly introduce the idea of soft reference labels in this field as well as a particular emotion recognition approach, which we will use for subsequent evaluation.

An emotion is a spontaneous reaction to an external stimulus that causes a change in physiological parameters and becomes observable, e.g., when these physiological parameters involve articulatory and phonatory processes. On the receiver side it is a matter of interpretation: Listeners' recognition and interpretation of emotions from recorded speech vary substantially [4]. The latter motivated us to rethink the standard annotation of emotional databases: In general the majority vote of the raters is reflected in the transcriptions of emotional speech. In order to target a human-like recognition of emotions, we believe that the ground truth as provided with an emotional speech database should not contain majority votes, but instead provide the underlying distribution of the labels in every case.

		Human votes						
		fea	**dis**	**hap**	**bor**	**neu**	**sad**	**ang**
Intended emotion	**fea**	88.4	2.2	4.3	0.0	2.2	0.2	2.7
	dis	3.3	81.1	0.4	1.8	1.8	11.6	0.0
	hap	3.1	1.4	84.0	0.0	7.0	1.2	3.3
	bor	0.0	0.0	0.0	95.7	3.3	1.0	0.0
	neu	0.0	0.0	1.3	5.3	92.6	0.4	0.4
	sad	4.6	0.0	0.0	16.9	9.7	68.8	0.0
	ang	0.0	0.1	0.1	0.1	0.0	0.1	99.6

Table 1. Average agreement results (in %) for the seven emotions (fear, disgust, happiness, boredom, neutral, sadness and anger) from the Berlin Database [20]. The ordinate denotes the intended emotion, whereas the abscissa represents the votes of the subjects in our perception test.

In our example application approach, we employ the feature extraction algorithm according to the ETSI Extended Advanced Front-End Recommendation [21] with a frame shift of 10 ms to obtain 13 mel frequency cepstral coefficients, one log-energy parameter, a pitch value, and the 1st- and 2nd-order derivatives of these. In addition, a voice activity detection flag is computed [22] resulting in a total of 46 frame-based features. Altogether we obtained 105 utterance-based statistical features, summarized in a feature vector $\mathbf{x}$ (further details to the feature composition can be found in [9]). We normalize these features and transform them with the SLDA regarding the $N = 7$ emotion classes [20] into a more robust vector $\mathbf{y}$ of length six, which is the input to the NN. After applying the NN, the output has to be adjusted to become a posterior probability confidence $\hat{\mathrm{P}}_{t,i}$. Concerning this we set all negative values to zero followed by a normalization to fulfill the stochastic constraint; a softmax operation could have been used equally.

Since in this work we focus on the (mathematical formulation of the) SLDA, we employ a very simple feedforward neural network (easy to resimulate for the reader, not intended for setting new benchmarks) consisting of one hidden layer with ten hidden layer neurons, six inputs and seven outputs (one for each emotion state). The network is trained with a quasi-Newton backpropagation algorithm, whereby the optimization of the feedforward network is performed with a mean absolute error function, in line with the definition of (2).

5. EMPIRICAL VALIDATION

We now compare LDA and SLDA with respect to the soft accuracy measure (2), employing the soft reference labels $\mathrm{P}_{t,i}$. For better comparison, we also investigate the hard accuracy measure (1).

5.1. Experimental Setup

For the experiments we used the Berlin Database of Emotional Speech [20] with speech recordings at 16 kHz sampling rate. It was recorded by ten actors (5 female and 5 male) expressing ten German utterances in seven different emotional states (fear, disgust, happiness, boredom, neutral, sadness and anger). With the aid of a perception test, the creators assessed the correctness of the acted emotions by the actors and labeled the formerly accepted 535 speech files. Since the Berlin Database does not include the detailed votes of the perception test, we repeated labeling with six subjects to

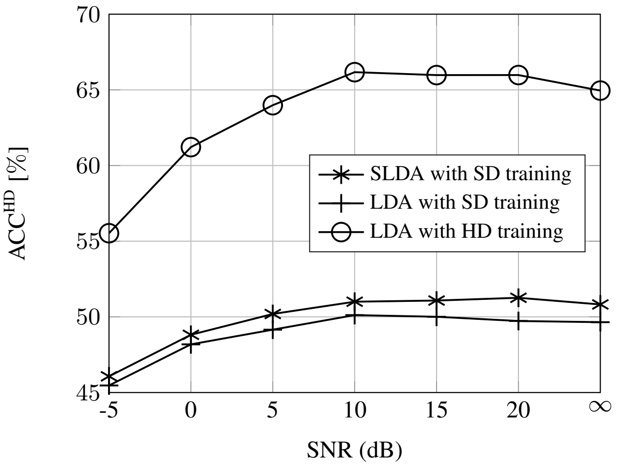

Fig. 1. Recognition results in terms of the state-of-the-art accuracy (% ACC$^{\text{HD}}$) vs. SNR (dB). Asterisk markers represent the new SLDA with SD training, while plus and circle markers denote the classical LDA with SD and HD training, respectively.

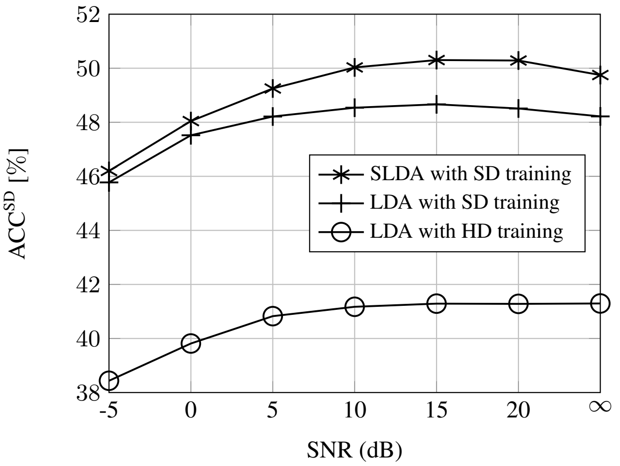

Fig. 2. Recognition results in terms of the (new) soft accuracy measure (% ACC$^{\text{SD}}$) vs. SNR (dB). Asterisk markers represent the new SLDA with SD training, while plus and circle markers denote the classical LDA with SD and HD training, respectively.

obtain a distribution over all emotions for each speech file, which is the soft ground truth. An overview of the overall result of our perception test is illustrated in percent as a confusion matrix in Table 1, whereby the x-axis represents the votes of the human raters and the y-axis depicts the intended emotion. It is evident that the perception test results in a distribution of the human votes over the predefined classes. For further investigations we have left out 13 speech files, in which the emotion class of the majority vote of the human raters did not agree with the intended emotion of the Berlin database. Hence, we took the remaining $T = 522$ speech files.

For a more realistic application case, the speech files were subject to additive car noise from the NTT Ambient Noise Database [23] at fixed SNR conditions (0 dB up to 20 dB active speech level in 5 dB steps) according to ITU-T Recommendation P.56 [24]. The considered SNR conditions are complemented with a clean speech condition denoted by an SNR of ∞ dB. Both emotional speech files and noise files were downsampled to 8 kHz sampling rate. The audio material was randomly split up into a training and a test set (70% / 30%) with disjoint speakers and then randomly assigned to the noise files. This process was repeated 40 times and the recognition results were averaged afterwards to obtain a reliable result. Furthermore, the NN was trained in multi-condition style on noisy data at an SNR of -5 dB and 10 dB as well as on clean data.

We analyzed three different cases, which were all evaluated with both the ACC$^{\text{HD}}$ (1) and the ACC$^{\text{SD}}$ (2) accuracy measures. For the first we used the classical LDA and trained our NN with the ground truth obtained from the majority votes of our perception test ("HD training"). Accordingly, only one output node of the NN was predefined as correct for each file in the training step. For the second and the third approach all output nodes of the NN were given our soft ground truth $\text{P}_{t,i}$ with $i \in \{1, \ldots, N\}$ as target signal ("SD training"). The difference consists in using either the classical LDA or the new SLDA.

5.2. Results and Discussion

Based on the two accuracy measures ACC$^{\text{HD}}$ and ACC$^{\text{SD}}$, Fig. 1 and Fig. 2 illustrate the recognition results of the three depicted cases. In both the circle markers represent the LDA with HD training result, whereas the plus and the asterisk markers characterize SD training with LDA and SLDA, respectively. Please note that a *comparison* of the respective accuracies in Fig. 1 and Fig. 2 is not meaningful, since their definition essentially differs.

With regard to the ACC$^{\text{HD}}$ the HD training approach with the classical LDA surpassed the two SD training approaches in Fig. 1, as expected, distinctly by up to 15% absolute. It is notable, that the SLDA approach achieved some better result than the classical LDA with SD training of around 1% absolute. In contrast, by applying the soft accuracy measure ACC$^{\text{SD}}$ it is clearly evident in Fig. 2 that the SD training approaches lead to a much better result than the HD training. Moreover, the SLDA method surpassed the classical LDA procedure by up to 1.8% absolute. It has to be borne in mind that the improvement of the SLDA is based on a posterior distribution with a very small variance since on average only 9.3% of all human votes differ from the desired emotion of the actors (Table 1). It can be expected that the improvement of the SLDA vs. the LDA rises even more in the case of a more uniform distribution. Employing SD training and the SLDA outperforms the conventional HD training approach by 22% relative in ACC$^{\text{SD}}$.

6. CONCLUSIONS

In this work, we proposed a two-step approach to classification tasks in social signal processing, consisting of a soft retraining of the classifier and a soft modification of the linear discriminant analysis (SLDA) in order to enable the possibility to exploit ambiguous class reference labels. We investigated the performance of the proposed SLDA method exemplarily for speech emotion recognition and used a soft accuracy measure for the evaluation being applied to the recognizer's confidence output. Using a soft ground truth (i.e., ambiguous reference class labels) the results show that both retraining and SLDA compared to the conventional training and LDA improves the performance by around 22% relative. In this paper we focused on the mathematics of the new soft linear discriminant analysis.

In future work we will expand the evaluation by applying some more sophisticated NNs (e.g., deep neural networks) and adding further databases with ambiguous labels, especially with natural emotions.

7. REFERENCES

[1] S. Watanabe, *Pattern Recognition: Human and Mechanical*, John Wiley & Sons, Inc., 1985.

[2] A. Vinciarell, M. Pantic, D. Heylen, D. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, Jan-Mar 2012.

[3] D. Gatica-Perez, "Automatic Nonverbal Analysis of Social Interactions in Small Groups: A review," *Image and Computing*, vol. 27, no. 12, pp. 1775–1787, Nov. 2009.

[4] M. Pakosz, "Attidudinal Judgements in Intonation: Some Evidence for a Theory," *Journal of Psycholinguistic Research*, vol. 12, no. 3, pp. 311–326, May 1983.

[5] L. Leinonen, T. Hiltunen, I. Linnankoski, and M.-L. Laakso, "Expression of Emotional-Motivational Connotations with a One-Word Utterance," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1853–1863, Sept. 1997.

[6] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-Based Evaluation and Estimation of Emotion in Speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, Oct.-Nov. 2007.

[7] R. Cowie and R.R. Cornelius, "Describing the Emotional States That are Expressed in Speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, Apr. 2003.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[9] P. Meyer and T. Fingscheidt, "A New Evaluation Methodology for Speech Emotion Recognition With Confidence Output," in *Proc. of 11th ITG Conf. on Speech Communication*, Erlangen, Germany, September 2014, pp. 81–84.

[10] R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," in *Proc. of 17th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, San Francisco, CA, USA, Mar. 1992, pp. 13–16.

[11] R. Haeb-Umbach, D. Geller, and H. Ney, "Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities," in *Proc. of 18th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Minneapolis, MN, USA, Apr. 1993, pp. 239–242.

[12] X. Aubert, R. Haeb-Umbach, and H. Ney, "Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context-Dependent Acoustic Models," in *Proc. of 18th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Minneapolis, MN, USA, Apr. 1993, pp. 648–651.

[13] R. Roth, J. Baker, J. Baker, L. Gillick, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, and F. Scattone, "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data," in *Proc. of 18th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Minneapolis, MN, USA, Apr. 1993, pp. 640–643.

[14] O. Siohan, "On the Robustness of Linear Discriminant Analysis as a Preprocessing Step for Noisy Speech Recognition," in *Proc. of 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, MI, USA, May 1995, pp. 125–128.

[15] C.M. Ayer, M.J. Hunt, and D.M. Brookes, "A discriminatively Derived Linear Transform for Improved Speech Recognition," in *Proc. of 3rd European Conf. on Speech Communication and Technology*, Berlin, Germany, Sept. 1993, vol. 1, pp. 583–586.

[16] N. Kumar and A.G. Andreou, "Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition," *Speech Communication*, vol. 26, no. 4, pp. 283–297, Dec. 1998.

[17] T.-K. Kim and J. Kittler, "Locally Linear Discriminant Analysis for Multimodally Distributed Classes for Face Recognition with a Single Model Image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 318–327, Mar. 2005.

[18] K. Fukunaga, *Introduction to Statistical Pattern Recognition (2nd ed.)*, Academic Press, School of Electrical Engineering, Purdue University, West Lafayette, Indiana, 1990.

[19] R.A. Fisher, "The Use of Multiple Measurement in Taxonomic Problems," *Annals of Human Genetics*, vol. 7, no. 2, pp. 179–188, Sept. 1936.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. of Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 1517–1520.

[21] ETSI, *ETSI ES 202 212 V1.1.2 Extended Advanced Front-End Feature Extraction Algorithm*, European Telecommunication Standards Institute, Nov. 2005.

[22] B. Fodor and T. Fingscheidt, "Reference-free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise," in *Proc. of 10th ITG Conf. on Speech Communication*, Braunschweig, Germany, Sept. 2012, pp. 199–202.

[23] NTT, *Ambient Noise Database for Telephonometry*, NTT Advanced Technology Corporation, 1996.

[24] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.